DEEP UNDERSTANDING AND GENERATION OF MEDICAL TEXT AND BEYOND

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF BIOMEDICAL
INFORMATICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Yuhao Zhang
March 2021

© 2021 by Yuhao Zhang. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/xg033vd7236

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Christopher Manning, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Russ Altman

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Curtis Langlotz

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Christopher Potts

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

Human language text plays a pivotal role in medicine. We use text to represent and store our biomedical knowledge, to communicate clinical findings, and to document various forms of medical data as well as healthcare outcomes. While deep language understanding techniques based on neural representation learning have fundamentally advanced our ability to process human language, can we leverage this advancement to transform our ability to understand, generate and utilize medical text? If so, how can we achieve this goal?

This dissertation aims to provide answers to these questions from three distinct perspectives. We first focus on a common form of medical text, biomedical scientific text, and study the long-standing challenge of extracting structured relational knowledge from this text. To handle the long textual context where biomedical relations are commonly found, we introduce a novel linguistically-motivated neural architecture that learns to represent a relation by exploiting the syntactic structure of a sentence. We show that this model not only demonstrates robust performance for biomedical relation extraction, but also achieves a new state of the art on relation extraction over general-domain text.

In the second part of this work, we focus on a different form of medical text, clinical report text, and more specifically, the radiology report text commonly used to describe medical imaging studies. We study the challenging problem of compressing long, detailed radiology reports into more succinct summary text. We demonstrate how a neural sequence-to-sequence model that is tailored to the structure of radiology reports can learn to generate fluent summaries with substantial clinical validity. We further present a reinforcement learning-based method that optimizes this system for correctness, a crucial metric in medicine. Our system has the potential of saving doctors from repetitive labor and improving clinical communications.

Finally, we connect text and image modalities in medicine, by addressing the challenge of transferring the knowledge that we learn from text understanding to understanding medical images. We present a novel method for improving medical image understanding by jointly modeling text and images in an unsupervised, contrastive manner. By leveraging the knowledge encoded in text, our method reduces the amount of labeled data needed for medical image understanding by an order of magnitude. Altogether, our studies demonstrate the great potential that deep language understanding and generation has in transforming medicine.

Acknowledgments

I realize that it is with great luck that I am writing down these words. The last few years of Ph.D. study at Stanford have been a truly unforgettable memory for me. While this dissertation is a summary of my Ph.D. research, I realize that it is also the result of what I have learned and received from all the great people around me, which I hope to sincerely acknowledge here.

I would like to start by giving my greatest thanks to my Ph.D. advisors, Curtis Langlotz and Christopher Manning. This dissertation is not possible without the support and advice that they have offered me over the last few years. I only got to know Curt at the end of the second year of my Ph.D. study. We met in a research collaboration meeting, but I ended up spending the first ten minutes on a fun conversation with him and listening to him talk about how proud he is of his son. It was that casual moment that made me realize how kind and caring Curt really is. The next few years of my Ph.D. life have proved me right — the amount of support and encouragement that I have received from Curt is beyond my expectation. Curt has his greatest vision about how artificial intelligence is able to transform healthcare, but at the same time, he keeps an unusually open mind to any new ideas that might lead to a difference. He encouraged me to pursue any directions that I was passionate about, but he also regularized me so that I was always on the right track to finishing my Ph.D. Curt is also extremely patient and willing to share. I still remember the day when Curt took me to his radiology examination room and taught me from scratch how radiologists work and interact with their tools. I also remember it when Curt showed me the photo of his old Ph.D. dissertation and told the story of how his first draft was rejected by his advisor. I believe that the way he thinks and the way he treats everyone else have had a profound impact on how I interact with other people around me.

I first knew about Chris when I worked on my undergraduate thesis project on information retrieval. His classic textbook *Introduction to Information Retrieval* has been a great inspiration to me. I have never thought back then that I would know Chris in person one day. Never have I imagined that Chris would ultimately become my Ph.D. advisor, and I would end up spending countless hours with him talking about my random research ideas and my life. Chris is always visionary about research but at the same time extremely down-to-earth. I still remember how amazed I was when I first started working with him and discovered every git commit that he made at 7 a.m. to the Stanford CoreNLP codebase. Chris always insists on the highest standards, but also provides the deepest care to his students. In private, he always pointed out every mistake or unfounded argument that I made; yet in public, he has never hesitated to help promote or defend my ideas. I can also recall the countless edits that Chris has made on my paper drafts, many of which were as specific as correcting a citation format or fixing a latex macro. Because of this extraordinary standard, I was in fact a bit scared of Chris when I started working with him; but as time goes by, he has become such a good friend. I am forever grateful for everything that I have received and learned from both Curt and Chris.

I would also like to thank my Ph.D. committee members, Russ Altman and Christopher Potts. I knew Russ before I even started my Ph.D., and have spent several quarters working with him. Russ has broad and natural curiosity in many things, and more importantly, he is always ready to learn. This personality of him has motivated and impacted me throughout my Ph.D. study. I only got to know Chris P after he returned from his sabbatical, and I was always amazed by how knowledgeable he is about linguistics (and pragmatics). Although we never collaborated before, when I reached out and asked Chris to serve on my reading committee, he immediately agreed and has since then provided me with great flexibility, for which I am very grateful.

I would like to extend my gratitude to all the mentors and professors who have helped me grow in the last few years. I worked with Mark Musen as a research assistant in my master's study. Mark introduced me to the Biomedical Informatics program at Stanford, and has been extremely generous and kind to me. In Mark's group, I also got to know Tania Tudorache and Matthew Horridge, who then became both great mentors as well as friends of mine. We had a great time working together and co-authored my first research paper in graduate school. I want to thank Dan Jurafsky and Percy Liang, who I have learned so much from through our collaborations or during the NLP group lunch events. Dan always gives positive energy to everyone around him, and has been incredibly supportive to every student in the NLP group. Percy is one of the sharpest people I have known, and is always able to give the most insightful comments on our research work. I want to thank Kai Zheng, who was my mentor when I did my research internship at the University of Michigan. Kai introduced me to the world of biomedical informatics, and has never hesitated to provide his help when I needed it. I want to thank Weizhu Chen, who was my mentor during my internship at Microsoft Research and has also been a good friend since then. I also want to thank Wei Deng, a visionary investor with whom I had the fortune to meet and became a friend during my undergraduate study. Wei persuaded me to choose Stanford for my graduate study, a decision that I have never regretted.

My Ph.D. journey would not have been so enjoyable without all the great friends that I made in the last few years. I want to start with my office mates, Peng Qi and Urvashi Khandelwal, with whom I shared Gates 232. I cannot remember how much time I have spent chatting with Peng. We talked about virtually anything, whether it was a random research idea, a frustration, or a life anecdote. Moreover, I was always amazed by how fast Peng can execute ideas, and have learned so much from collaborating with him. Urvashi is smart and extremely humble. She is also very kind-hearted and easy to talk to. She always gave her laughs at whatever bad joke that I told, and gave her encouragement in my difficult times. I will miss the good old days that I have shared with Peng and Urvashi in Gates 232. I am also very grateful to the KBP team. KBP is the first research project that I have worked on in the Stanford NLP group. The knowledge that I learned from this experience has ultimately led to a chapter in this dissertation. During my KBP days, I have spent numerous hours working with Danqi Chen, who is now an assistant professor at Princeton. Danqi always asks the right questions and look at things at the finest level of detail. She is also a good friend and always listens to others' needs. What I learned from collaborating with Danqi has benefited me throughout my Ph.D. study. I have also learned a lot from Arun Chaganty, another KBP teammate. Arun is generous and always willing to help: whenever I had a technical problem, I knew that I can ask Arun for answers. Other KBP teammates that I have truly enjoyed working with include Ashwin Paranjape, Gabor

Angeli, Jason Bolton, Victor Zhong, Kevin Clark, and Abi See. There are other members and alumni of the Stanford NLP group, the Stanford AIMI center and the BMI program that I have enjoyed working with and am grateful to, including but not limited to Matt Lamm, Tim Dozat, Will Monroe, Robin Jia, Jiwei Li, Sebastian Schuster, Ziang Xie, Albee Ling, Alice Yu, Steven Bagley, etc.

I want to thank all the people that have collaborated with me during my Ph.D., who have helped me grow and made the Ph.D. journey not alone. For the Stanza NLP toolkit that we built during my Ph.D., I am grateful for having Yuhui Zhang, Peng Qi, John Bauer, and Jason Bolton working with me. I recruited Yuhui to work with us when he just finished his undergraduate study at Tsinghua. He has become such a great help to us since then. John Bauer has led the work on maintaining Stanza since its release and has always stayed on top of things. Other close collaborators include Daisy Ding, Hang Jiang, Derek Merck, Yasuhide Miura, Ian Pan, Bethany Percha, Tianpei Qian, Jinfeng Rao, and Devendra Sachan. The great work from them has led to this dissertation as well as other research papers that I have published during my Ph.D.

Last, but certainly not least, I want to thank my parents, Mantang and Caichun, and my wife, Xiaochen, for their unconditional love and support. My parents are ordinary people — they have lived through all the hardship and struggles that their generations had to suffer, but they never hesitated to provide the best they can to me at every stage of my life. This is something I am forever grateful for. I met Xiaochen during my undergraduate years. She traveled half of the earth with me to pursue our dreams. Xiaochen has been greatly supportive since day one of my Ph.D. study, and has accompanied me through all the ups and downs. I will not take their love for granted.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Text is ubiquitous in medicine. Scientists use human language text as the central medium for publishing and communicating their scientific discoveries about biomedicine. Patients use text to express their healthcare needs in conversations or online forums. Doctors use text to communicate clinical procedures and findings, and furthermore, to document various medical conditions as well as healthcare outcomes. For these reasons, text represents one of the most crucial data sources in medicine. It is thus a long-standing goal for the biomedical informatics and natural language processing (NLP) communities to build automated systems that can understand and generate medical text in various contexts (Cohen and Demner-Fushman, 2014).

The last few years have witnessed the success of deep learning and the changes it has brought to NLP research. Deep language understanding techniques based on neural networks have fundamentally advanced our abilities to analyze unstructured text (Chen and Manning, 2014), answer questions (Seo et al., 2017), automate conversations (Li et al., 2016), and translate between human languages (Wu et al., 2016).

Situated in this need to understand medical text and these changes in our ability to process language, this dissertation focuses on answering the following key questions:

- What is the value of understanding and generating medical text?

- How can we leverage deep language understanding techniques to better understand and generate medical text?
- Can understanding and generating medical text inspire us to improve the robustness and efficiency of our techniques?

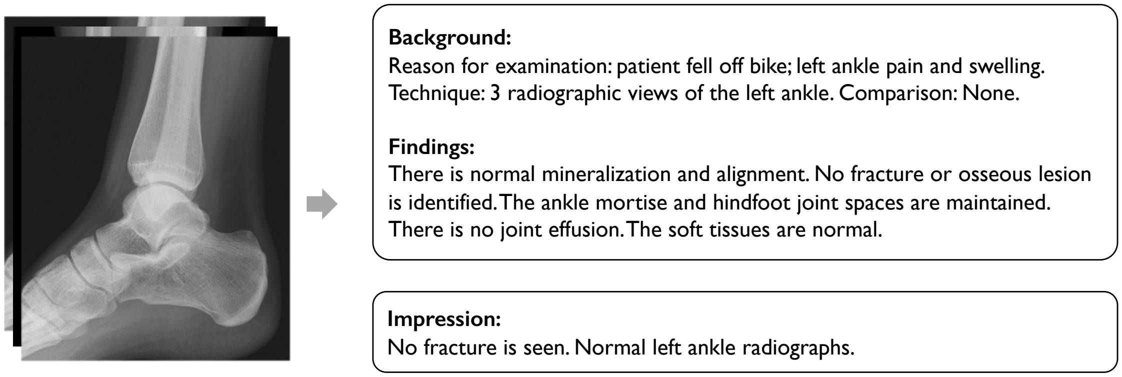
Before we move on and discuss how we approach these questions in this dissertation, let us first consider the following hypothetical story of a hospital visit by Mr. Lee.

Mr. Lee had an accident two days ago when he was biking on a mountain road. He almost collided with a car at a sharp turn and fell off his bike. As a result, he scratched his left ankle skin. His left ankle had been swelling and hurting since then. He decided to visit the clinic to see if his ankle was fractured.

During the office visit, the attending physician asked about Mr. Lee's condition, and wrote down the following medical notes:

> [...]
> *Chief complaint: Left ankle swelling and persistent pain.*
> *History of present illness: Mr. Lee is a 35-year-old male. He fell off his bike during biking two days before the visit. He has had a swollen left ankle since then. He can move his left ankle but has experienced slight persistent pain in the area. The condition has not worsened since the accident.*
> [...]

After visually examining his ankle, the physician ordered radiographs to further determine where a fracture was present. The radiograph study produced a series of X-ray images of his left ankle, as shown in Figure 1.1, which were sent to a radiologist for interpretation. The radiologist read Mr. Lee's images as well as the medical notes written by the physician, and documented the clinical interpretation of the images in a detailed radiology report (see Figure 1.1). Next, the radiologist wrote a summary of the report, called an "Impression" statement, which summarizes and concisely highlights the most important findings. The textual radiology report was then sent back to the referring physician, who read the summary and communicated the findings to the patient.



Background:
Reason for examination: patient fell off bike; left ankle pain and swelling.
Technique: 3 radiographic views of the left ankle. Comparison: None.

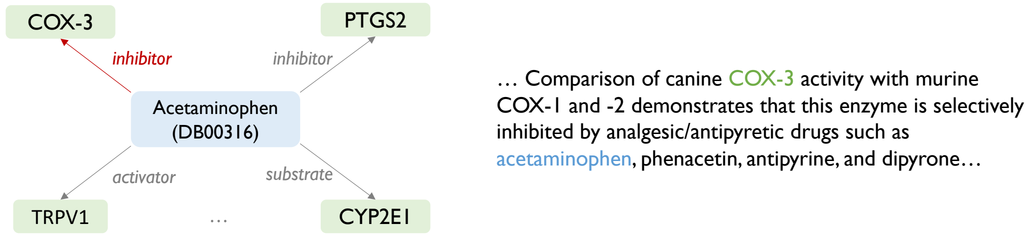
Findings:
There is normal mineralization and alignment. No fracture or osseous lesion is identified. The ankle mortise and hindfoot joint spaces are maintained. There is no joint effusion. The soft tissues are normal.

Impression:
No fracture is seen. Normal left ankle radiographs.

Figure 1.1: An example series of radiographic images along with the paired clinical radiology report. The report is drafted by a radiologist, and consists of a detailed description of the patient's conditions and clinical findings (top right), as well as a more concise summary of the study (bottom right).

Let us now pause this story and review what we have seen. In this typical clinic visit by Mr. Lee, both the initial medical notes and the radiology report represent a first category of text that is commonly produced and used in medicine: **clinical text**. Although its context may vary, clinical text is a pivotal component in healthcare and serves as a central medium for **documentation and communication** in medicine. For example, in our story, the medical notes were used to communicate the condition of the patient to the radiologists, and the radiology report as well as the impression were used to communicate the results of the imaging study to the referring physician and the patient. While it seems natural for medical practitioners to produce accurate clinical notes like these, in reality, writing clinical text is often an error-prone process (Wagner and Hogan, 1996; Gershanik et al., 2011). Moreover, report drafting is a repetitive task that may take away time from doctors that could otherwise be spent on the important work of diagnosing and treating disease (Ammenwerth and Spötl, 2009). These errors can be considerably avoided and much time can be saved if we can develop systems that understand existing medical notes or generate future notes. In other words, **understanding and generating medical text can help improve the quality and efficiency of communications in medicine**.

Furthermore, in the radiology exam shown in Figure 1.1, the textual radiology report encodes the clinically important information from the medical images in a natural language format. In other words, the radiology report provides natural language explanations of the



… Comparison of canine COX-3 activity with murine COX-1 and -2 demonstrates that this enzyme is selectively inhibited by analgesic/antipyretic drugs such as acetaminophen, phenacetin, antipyrine, and dipyrone…

Figure 1.2: A subset of the drug-protein relations for the drug *acetaminophen* stored in the DrugBank knowledge base. On the right we also show how one of the relations was described with natural language in a scientific publication (Chandrasekharan et al., 2002).

images that we can utilize to guide future care or to educate new learners. If we can build machines that understand the significance of various findings in the report text and link them with the visual features in the corresponding image regions, this can naturally help us build systems that better understand medical images, and that may ultimately automate key aspects of the imaging diagnosis process. Therefore, **understanding medical text offers opportunities to enhance the utility of other forms of medical data, such as images**.

Let us return to Mr. Lee's office visit. Because the radiology exam found no severe clinical abnormality in Mr. Lee's ankle, the physician decided to order some medications to help ease the pain and swelling. As a result, the physician ordered him *acetaminophen*, a widely used pain relief medication. Now Mr. Lee can return home and rest assured that his ankle will recover from the accident.

While today adults in the United States can easily get access to acetaminophen as a pain or fever relief medication, the development and testing of a drug like acetaminophen is not easy. In fact, it usually requires us to fundamentally understand how a chemical like acetaminophen interacts with various proteins in the human body such that it can take effect, be metabolized and get safely excreted from the body. To aim our understanding, numerous biomedical knowledge bases were constructed to store and represent structured knowledge about biomedicine. For example, Figure 1.2 presents the known relations between acetaminophen and different proteins stored in the DrugBank knowledge base (Wishart et al., 2008). Knowledge bases such as the DrugBank are therefore invaluable for the development of new treatments and the testing of new drugs (Zhu et al., 2019).

Despite their importance, the development of biomedical knowledge bases such as the DrugBank has been a long-standing challenge. This is mainly because our scientific discoveries about biomedicine are mostly published in text format, as is shown in Figure 1.2. The paragraph on the right represents a second category of text that we see in medicine: **biomedical scientific text**. While text like this is easy to read and understand for scientists, it is extremely difficult for machines to process it. As a result, traditionally, the development of large biomedical knowledge bases has heavily relied on manual work by scientific curators, whose job is to read the relevant scientific papers and manually organize the structured knowledge into the knowledge bases (Hewett et al., 2002; Wishart et al., 2008). This has made the construction of biomedical knowledge bases extremely expensive and hard to sustain. Therefore, developing automated systems that understand the scientific text and extract structured knowledge from it can save us from the manual work and provide more comprehensive knowledge for the development of new treatments. Or in other words, **understanding medical text can help us obtain actionable biomedical knowledge**.

This dissertation is written around these aforementioned perspectives of what understanding and generating medical text means to us. Namely, a central theme of this dissertation is to showcase via several distinct studies that understanding and generating medical text can have a positive impact on medicine by helping us obtain actionable knowledge, improve communications, and understand other forms of medical data.

The importance of medical text analysis leads to the following important question: **why does understanding and generating medical text present substantial computational challenges**? To answer this question, let us go back to the previous examples. For an automated system to process the biomedical scientific text in Figure 1.2, the system must be capable of navigating through the long context of words in the paragraph and discerning biomedical relations (e.g., *inhibitor* vs. *activator*) encoded in drastically different forms. Similarly, to understand the clinical note in Figure 1.1, a system must recognize each term and know how to interpret various clinical findings in the context of the patient's conditions; to further complete such a note, it must know what content to generate and how to compose it in a fluent, efficient and grammatical manner. Moreover, for a system to read the medical images and the text in Figure 1.1 together, it has to obtain a joint understanding

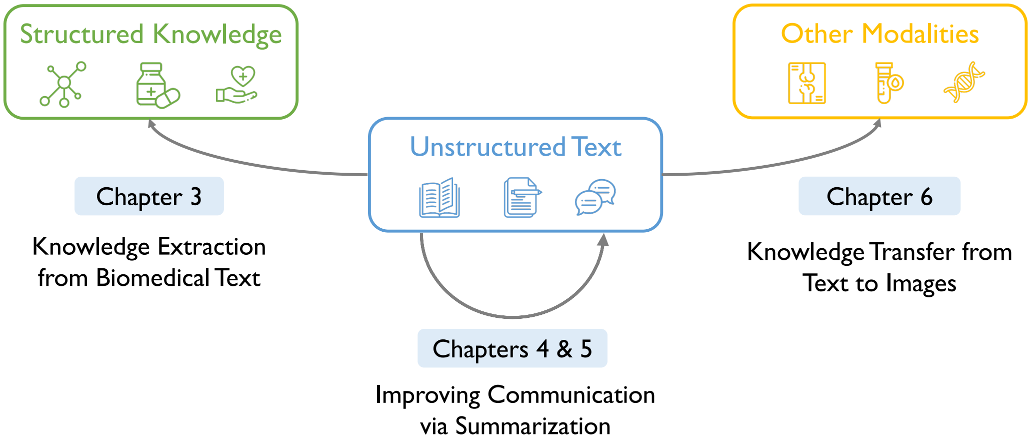

Structured Knowledge
Other Modalities
Unstructured Text
Chapter 3
Knowledge Extraction from Biomedical Text
Chapter 6
Knowledge Transfer from Text to Images
Chapters 4 & 5
Improving Communication via Summarization

Figure 1.3: An overview of the main chapters in this dissertation.

of both and be capable of grounding the nuances of meanings expressed in the text (e.g., *fracture* vs. *mineralization*) to the relevant visual features in the image. In fact, none of these tasks is computationally trivial. While numerous research efforts have been made in the past to understand biomedical and clinical text (Hunter and Cohen, 2006; Cohen and Demner-Fushman, 2014; Pons et al., 2016), historical work has relied mostly on expert rules and statistical models built with sparse, hand-engineered features. A second theme of this dissertation is to demonstrate with empirical evidence that deep learning language understanding techniques based on dense vector representations of text substantially outperform traditional rule or feature-based systems for all the aforementioned tasks.

1.2 Dissertation Outline

This dissertation is organized according to the central themes that we have introduced above. In Chapter 2, we first provide a detailed overview of previous work closely relevant to this dissertation. The subsequent chapters will focus on three distinct applications of understanding and generating medical text, as shown in Figure 1.3.

In Chapter 3, we first focus on the understanding of biomedical scientific text, and study the long-standing challenge of extracting structured relational knowledge from this text. To handle the long textual context where biomedical relations are commonly found, we introduce a novel linguistically-motivated neural architecture that learns to represent a relation encoded in a sentence by exploiting the syntactic structure of the sentence. On several widely used benchmark datasets, we show that our model not only demonstrates robust performance for biomedical relation extraction, but also achieves a new state of the art on relation extraction over general-domain text. This chapter is based on work first published as Zhang et al. (2018b).

In Chapter 4, we shift our focus to the clinical report, and more specifically, the radiology report commonly used to describe medical imaging studies. We study the challenging problem of automated summarization of long, detailed radiology reports into the more succinct impression text shown in Figure 1.1. We demonstrate how a neural sequence-to-sequence model tailored to the structure of radiology reports can learn to generate fluent summaries that overlap substantially with human-written ones. We also show on real-world radiology report datasets that our model outperforms traditional extractive summarization models based on sparse modeling of the report text. In a human evaluation, a radiologist has indicated that our model output is as least as good as the human-written summary in 67% of the examples, suggesting substantial clinical validity. This chapter is based on work first published as Zhang et al. (2018a).

Next, in Chapter 5, we extend our study in Chapter 4, by addressing a critical issue of our neural model, that the generated summaries tend to be factually incomplete and incorrect. We do this by first identifying the imperfect objectives that we use to train and evaluate our model, and then proposing a new information extraction-based framework that evaluates a text generation based on its factual content. We further present a reinforcement learning-based method that optimizes this new metric, and demonstrate via both automatic and human evaluation that this new method leads to radiology summaries that are more correct and have higher clinical validity. This chapter is based on work first published as Zhang et al. (2020c).

In Chapter 6, we connect text and image modalities in medicine by addressing the challenge of transferring knowledge learned from text understanding to understanding medical images. We present a novel method for improving medical image understanding by jointly modeling text and images in an unsupervised, contrastive manner. We show via experiments on multiple medical image classification and retrieval tasks that the proposed method improves the accuracy of the learned image encoders, and substantially outperforms existing methods based on ImageNet pretraining or image-only contrastive learning. A preprint of this work is available on arXiv as Zhang et al. (2020b).

Lastly, we conclude this dissertation and discuss future directions in Chapter 7.

1.3 Contributions

In summary, this dissertation makes the following detailed contributions to the field of understanding and generating medical text:

- We propose a novel neural architecture that improves upon existing models for the task of extracting biomedical relations from scientific text.
- We develop the first neural system that completes a clinical radiology report by summarizing the radiology findings into more concise impression statements. We develop a novel framework to optimize this system for the crucial correctness metric. Our system has the potential to save healthcare providers from repetitive labor and improving clinical communications.
- We develop a joint neural architecture for improving representations of medical images by understanding their paired clinical reports. Our method reduces the amount of labeled data needed for medical image understanding by an order of magnitude, and breaks new ground for effectively utilizing existing medical text data for understanding data of other modalities.

Further, this dissertation makes the following contributions to general language understanding and generation research:

- We demonstrate that our linguistically-motivated architecture for relation extraction generalizes to relation extraction from news articles and web text, and achieves superior performance on standard benchmarks.
- We pioneer in the direction of optimizing the factual correctness of a neural text summarization model. Our study represents the first success in improving the factual correctness of a summarization model via reinforcement learning.

- We pioneer the improvement of visual representations via cross-modality contrastive pretraining with human-written descriptive text. Our study has recently been successfully applied at much larger scale by Radford et al. (2021) and led to state-of-the-art general visual recognition capabilities.

Chapter 2

Related Work

In this chapter, we provide an overview of previous work relevant to this dissertation, and discuss its connections to our studies. For each individual area, we focus on both the development of general methodologies and existing applications in biomedicine.

In Section 2.1, we start by giving a general overview of research on the automatic population of knowledge bases from text, and discuss the applications of these systems to extracting biomedical knowledge.

In Section 2.2, we delve into relation extraction, a core component of knowledge base population systems. We provide an overview of general relation extraction research as well as its applications in biomedicine, and discuss how our study in Chapter 3 is connected to existing work.

In Section 2.3, we provide an overview of existing work in text summarization, with a focus on neural text summarization methods and existing applications of summarization in biomedicine. We also discuss how our studies in Chapter 4 and Chapter 5 are related to existing work in these areas.

Finally, in Section 2.4, we review existing work in medical image understanding and the joint modeling of medical image and text data. We also discuss previous work in contrastive visual representation learning and text-image pretraining, which inspired our study in Chapter 6.

2.1 Knowledge Base Population

Large-scale knowledge bases that store structured facts are widely used in many domains and power numerous downstream applications. For example, DBPedia[1], Freebase[2] or Wikidata[3] are widely used general-domain knowledge bases. In the biomedical domain, DrugBank (Wishart et al., 2008) and PharmGKB (Hewett et al., 2002) are among the most widely used, storing structured facts about drugs and proteins. Depending on the domain and scale, these knowledge bases are often constructed via crowdsourcing or expert curation, sometimes with the help of automated knowledge base population systems, to which our study in Chapter 3 is tightly related.

Knowledge base population[4] (KBP) is a task that aims at taking a large collection of unstructured text, and using it to populate a structured knowledge base. This collection of text can be Wikipedia or web articles, as is often the case for general-domain KBP systems; or it can be a collection of scientific documents or abstracts, as is often the case for biomedical KBP systems. The structured facts (i.e., the output of KBP systems) are often represented in the form of (s, r, o) triples, where s is a subject entity, o an object entity and r a relation type, often drawn from a fixed schema.

Systematic research in KBP systems has been facilitated by a number of community shared tasks or challenges in this area, among which the yearly TAC KBP challenge (McNamee and Dang, 2009) is a representative one. These challenges provide opportunities to evaluate the end-to-end performance of KBP systems at extracting entities and relation triples, given a shared collection of documents such as newswire articles. A typical KBP system submission to these tasks consists of several individual components in a pipeline, as shown in Figure 2.1. These components include:

- Syntactic annotation components, which split the input text into individual sentences and tokens, and annotate the text with syntactic information.

[1] https://wiki.dbpedia.org/

[2] https://en.wikipedia.org/wiki/Freebase_(database)

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

[4] While in medicine, the word “population” is often used to refer to a specific group of people, here the word “population” refers to the action of populating a knowledge base with structured facts.

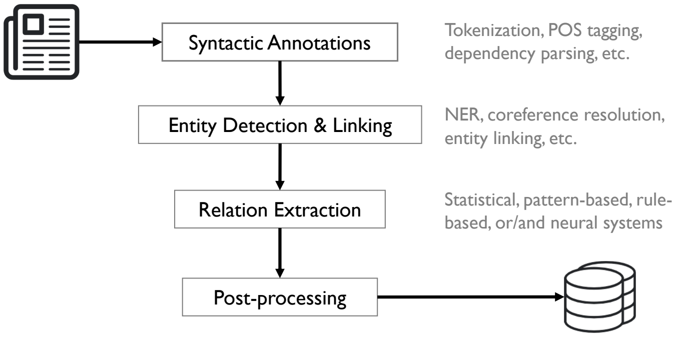

Syntactic Annotations
Tokenization, POS tagging, dependency parsing, etc.
Entity Detection & Linking
NER, coreference resolution, entity linking, etc.
Relation Extraction
Statistical, pattern-based, rule-based, or/and neural systems
Post-processing

Figure 2.1: Overview of a typical pipeline-based knowledge base population system. This figure is adapted from Zhang et al. (2016).

- Entity detection and linking, which aims at recognizing the spans of entity mentions, and linking the detected entities to an existing taxonomy. For general-domain text these entities often include *person* or *organization* entities, and are linked to their unique Wikipedia page. For biomedical domain, these entities of interest often include *drug*, *protein* or *disease* mentions, and are often linked to an existing biomedical ontology.

- Relation extraction component, which aims at discerning whether a relation exists between a pair of extracted entity mentions, and if so, the type of the relation. Example relations for general-domain text include the *date_of_birth* relation between a person and a date mention; example relations for the biomedical domain include the *inhibitor* or the *activator* relation between a drug and a protein mention.

- Post-processing components, which are typically designed to guarantee the validity of the produced knowledge base (e.g., to remove duplicates or resolve conflicts among the extracted facts).

In addition to previous work on KBP, there are research efforts that focus on inferring new facts in a knowledge base that do not utilize any text at all (Socher et al., 2013; Lin et al., 2015). This task is often referred to as knowledge base completion. Moreover, the task of semantic parsing aims at converting a natural language query into a logical form, which can be used to query against an existing knowledge base (Zelle and Mooney, 1996;

Zettlemoyer and Collins, 2007; Berant and Liang, 2014). We skip an in-depth overview of these areas as it is not directly related to this dissertation.

Concurrent to the development of various KBP systems, numerous efforts have been made to apply these systems to **biomedical knowledge base population**. For example, Garten et al. (2010) discussed research efforts on constructing pharmacogenomics knowledge bases from text with automated systems. Thorn et al. (2013) further introduced how a pharmacogenomics knowledge base, the PharmGKB, was enhanced with a combination of KBP techniques and human curation. Literome (Poon et al., 2014) is a system that aims at extracting genomic knowledge from PubMed articles and making this knowledge available via a cloud service. Life-iNet (Ren et al., 2017) is a system that supports automated construction of a knowledge base from life science articles and querying against the extracted structured facts. KnowLife (Ernst, 2017) is an effort that aims at leveraging KBP techniques to construct a health knowledge base that covers a wide range of biomedical entities (e.g., gene, disease, or anatomy), from text covering a wide range of genres. Similar to general KBP research, biomedical KBP research has been facilitated by numerous shared tasks such as the BioNLP shared tasks (Pyysalo et al., 2012) or the BioCreative challenge series (Wei et al., 2015).

Our study in Chapter 3 is focused on a core component in pipeline-based KBP system (see Figure 2.1), the relation extraction component, and its application to extracting biomedical knowledge. We now describe related work in relation extraction in detail.

2.2 Relation Extraction

At the core of a typical KBP system as discussed above is a **relation extraction** model. The task of relation extraction involves discerning whether a relation exists between two entity mentions in a piece of text, such as a sentence. Here the two entity mentions are often referred to as a *subject* and an *object* mention, respectively. For example, given the following sentence:

> *This property may be able to explain the ability of [chloroquine]$_{subject}$ to inhibit [CYP2D6]$_{object}$-mediated metabolism in vitro and in vivo.*

where the drug *chloroquine* is a subject entity and the protein *CYP2D6* an object, a relation extraction model should be able to extract the following relation triple based on the context:

(*chloroquine*, *inhibitor*, *CYP2D6*)

where *inhibitor* is the relation type. A relation extraction model like this will be the focus of our study in Chapter 3.

Traditionally, the task of relation extraction has been studied with broadly three different approaches:

- Fully-supervised methods, where a relation classifier is trained on a supervised dataset of sentences with entity and relation annotations;
- Distant supervision (Mintz et al., 2009; Surdeanu et al., 2012), where a weakly-supervised dataset is created by mapping relation triples in a knowledge base to sentences in a large corpus, and the dataset is used in place of a supervised one;
- Open information extraction (Mausam et al., 2012; Angeli et al., 2015), where syntactic structures of sentences are exploited to extract open-domain relation triples, instead of using a pre-defined relation set.

Among these three approaches our study in Chapter 3 is most tightly related to the fully-supervised method for relation extraction.

At the core of fully-supervised and distantly-supervised approaches are statistical classifiers, which traditionally are built with sparse hand-engineered features. In particular, many of these traditional classifiers find syntactic information of the input sentences beneficial to the relation extraction task. For example, Mintz et al. (2009) explored adding syntactic features to a statistical classifier and found them to be particularly useful when sentences are long. In parallel to these feature-based classifiers are kernel-based methods for relation extraction, where relations are classified based on their similarity with existing examples in a kernel space. A number of these approaches also leverage syntactic information to construct the kernel space, finding that tree-based kernels (Zelenko et al., 2003) and dependency path-based kernels (Bunescu and Mooney, 2005b) are particularly effective for this task.

Recent years have seen the success of neural representation learning-based approaches for relation extraction. For example, Zeng et al. (2014) first applied a one-dimensional convolutional neural network (CNN) combined with hand-engineered features to encode relations and found it to outperform traditional methods on standard benchmarks. Vu et al. (2016) showed that combining a CNN architecture with a recurrent neural network (RNN) through a voting scheme can further improve performance. Zhou et al. (2016) and Wang et al. (2016) found that attention mechanisms over RNN and CNN architectures are useful for relation extraction. Despite the success of increasingly complex forms of neural architectures, Adel et al. (2016) and Zhang et al. (2017) have shown that relatively simple neural models (CNN and augmented LSTM, respectively) can achieve comparable or superior performance to more complex models when trained on larger datasets.

Apart from neural models over word sequences, incorporating syntactic structures such as dependency trees into neural models has also been shown to improve relation extraction performance by capturing long-distance relations. For example, Xu et al. (2015c) generalized the idea of dependency path kernels by applying a long short-term memory (LSTM) network, a special form of RNN, over the shortest dependency path between the entity mentions. In their experiments, this model outperformed a similar LSTM model applied over the original sentence sequence. Liu et al. (2015) first applied a recursive network over the subtrees rooted at the words on the dependency path and then applied a CNN over the path. Miwa and Bansal (2016) applied a Tree-LSTM (Tai et al., 2015), a generalized form of LSTM over dependency trees originally developed for encoding the semantic meaning of a sentence, in a joint entity and relation extraction setting. They found it to be most effective when applied to the subtree rooted at the lowest common ancestor of the two entities.

In parallel to this development, many efforts have been made on applying these techniques to **biomedical relation extraction**, which focuses on extracting relations between biomedical entities from scientific text. For example, Bunescu and Mooney (2005a) applied a subsequence-based kernel method to the task of extracting protein-protein interactions. Riedel and McCallum (2011) proposed a joint statistical model for biomedical entity and event extraction. Peng and Lu (2017) applied a multichannel convolutional network enhanced with lexical and syntactic features to extracting protein-protein relations. Lim and Kang (2018) applied the Tree-LSTM model to extracting chemical-gene relations from biomedical abstracts. Quirk and Poon (2017) extended biomedical relation extraction to a cross-sentence setting, and proposed a distant supervision-based method for this task. Peng et al. (2017) further improved cross-sentence relation extraction with a syntactically augmented neural sequence model, and showed that it improves the state of the art for extracting drug-mutation-gene relations. Apart from methods that rely on supervised or distantly supervised learning, there has been work that studied the unsupervised discovery and extraction of biomedical relationships from scientific literature text (Quan et al., 2014; Percha and Altman, 2015, 2018). Finally, in addition to these individual studies, research work in this area has relied heavily on resources released as part of the BioNLP shared tasks (Pyysalo et al., 2012; Nédellec et al., 2013) or the BioCreative challenges (Wei et al., 2015; Krallinger et al., 2017).

Our study in Chapter 3 is built on top of this existing work in supervised relation extraction, especially existing work based on modeling the syntactic structures with neural architectures. Our method closely connects to the studies of Liu et al. (2015) and Miwa and Bansal (2016) by extending their models with a new neural architecture. Furthermore, our study relies on the general-domain relation extraction resource released by Zhang et al. (2017), and biomedical relation extraction resources by Peng et al. (2017) and Krallinger et al. (2017).

2.3 Text Summarization and Its Applications in Medicine

In this section, we focus on reviewing previous work on text summarization, an area closely related to the summarization of medical reports in Chapter 4 and Chapter 5.

The task of **text summarization** aims at compressing a long document into a shorter text while preserving the key facts in the original document. Text summarization systems can be applied in many practical domains, among which news summarization is most commonly studied (Dang, 2005). In news summarization, a system takes a news article as the input document, and outputs a one-sentence or multi-sentence textual summary that preserves the gist of the news article. For example, for the following (truncated) news article in the DUC summarization dataset (Dang, 2005; Grusky et al., 2018):

MAPUTO, Mozambique (AP) – Just as aid agencies were making headway in

> *feeding hundreds of thousands displaced by flooding in southern and central Mozambique, new floods hit a remote northern region Monday. The Messalo River overflowed [...]*

A single-sentence summary written by a human expert is:

> *Floods hit north Mozambique as aid to flooded south continues.*

In practice, text summarization systems are often evaluated against this human-written summary as an oracle reference.

Early work on text summarization mainly focuses on **extractive summarization**, where the summaries are generated by scoring and selecting sentences from the input. The systems are trained either in an unsupervised fashion, where the text units are selected based on the document structure, or in a supervised fashion, where oracle sentences are used as supervision signals. Luhn (1958) proposed to represent the input by topic words and score each sentence by the amount of topic words it contains. Kupiec et al. (1995) studied statistical methods for text summarization and proposed to score sentences with a feature-based statistical classifier. Barzilay et al. (1999) studied multi-document summarization and proposed an information fusion model for it that combined sentence fragment selection from the documents and rule-based paraphrasing with rules derived from corpus analysis. Steinberger and Jezek (2004) applied latent semantic analysis to cluster the topics in a document and then select sentences that cover the most topics. Meanwhile, various graph-based methods, such as the LexRank (Mihalcea and Tarau, 2004) and the TextRank algorithm (Erkan and Radev, 2004), were applied to model sentence dependency by representing sentences as vertices and similarities as edges. Sentences are then scored and selected via modeling of the graph properties. In Chapter 4 and Chapter 5 we treat these early extractive summarization systems as baselines and compare our models with them.

The application of neural networks, especially neural sequence-to-sequence learning methods (Sutskever et al., 2014) has enabled **abstractive summarization** systems, where new words and phrases are generated to form the summaries. Rush et al. (2015) first applied an attention-based neural encoder and a neural language model decoder for neural abstractive summarization. Nallapati et al. (2016b) extended the previous method and used RNN models for both the encoder and the decoder. Nallapati et al. (2016a) further compared the RNN-based architecture for neural abstractive and extractive summarization.

Meanwhile, a series of **hybrid summarization** systems that combine the advantages of abstractive and extractive summarization were proposed and studied. For example, to address the limitation that neural models with a fixed vocabulary cannot handle out-of-vocabulary words, a pointer-generator model was proposed which uses an attention mechanism that copies elements directly from the input (Nallapati et al., 2016b; Merity et al., 2017; See et al., 2017). See et al. (2017) further proposed a coverage mechanism to address the repetition problem in the generated summaries. Chen and Bansal (2018) proposed a hybrid system that first selects sentences from the input document and then rewrites the selected sentences to form abstract summaries. Gehrmann et al. (2018) proposed a bottom-up approach where an abstractive summarization system is restricted to only consider the selected sentences in the document as input. Our study in Chapter 4 is closely related to existing work in neural abstractive and hybrid summarization; our model is directly inspired by the pointer-generator system as in (See et al., 2017).

While it is common practice to train neural summarization systems in an end-to-end supervised manner by maximizing the likelihood of the reference summaries, reinforcement learning (RL) has been explored as an alternative training strategy and shown useful in previous work (Paulus et al., 2018; Chen and Bansal, 2018; Dong et al., 2018). Specifically, Paulus et al. (2018) found that directly optimizing an abstractive summarization model on the ROUGE metric (Lin, 2004) via RL can improve the summary ROUGE scores. Chen and Bansal (2018) explored training their select-and-rewrite hybrid summarization system with RL. Dong et al. (2018) proposed to model extractive summarization as a contextual bandit problem, and designed an RNN-based neural architecture for this setting. They similarly optimized the proposed architecture with RL for end metrics. Our study in Chapter 5 is directly inspired by this line of work.

As neural summarization models achieve increasingly higher performance on benchmark datasets as measured by common metrics such as the ROUGE scores, a recent line of work has focused on the **factual correctness** or consistency of these systems. Kryściński et al. (2019a) critically evaluated a collection of state-of-the-art summarization systems and found that they tend to have poor factual consistency with the input document. To improve the correctness of these systems, Cao et al. (2017) proposed to extract fact triples from the input document with an open information extraction system, and then attend to these fact triples during the decoding process. Their study, despite being an early attempt, did not focus on an explicit measurement of factual consistency or correctness of the generated summaries. Goodrich et al. (2019) proposed to evaluate the factual accuracy of generated text with an information extraction system. They found that while existing information extraction systems are generally inadequate for this task, systems that are based on a fixed schema perform better than open information extraction systems. Falke et al. (2019) explored using natural language inference (NLI) systems to evaluate the correctness of generated summaries, with the intuition that summaries that are consistent with the original document should be evaluated as "entailment" by a well-trained NLI system. However, they arrived at a negative conclusion that current NLI models trained on existing datasets tend to be inadequate for this task, and that more advanced NLI models or better datasets need to be constructed to improve the robustness of NLI systems. Kryściński et al. (2019b) took a different approach, and proposed to evaluate factual consistencies in the generated summaries using a weakly-supervised fact verification model. Their proposed model was based on a strong pretrained transformer architecture (i.e., BERT), and was trained with a weakly-supervised dataset that contain both consistent and inconsistent document-summary pairs. Our study in Chapter 5 is closely related to this line of work, and our findings regarding the correctness of existing systems are in line with those by Kryściński et al. (2019a). Moreover, while none of this work has shown a notable success in directly optimizing a summarization system for factual correctness, our study in Chapter 5 represents the first success in this direction.

In addition to single-document summarization systems, there are systems that are designed for the settings of multi-document summarization (Lin and Hovy, 2002; Haghighi and Vanderwende, 2009), or multi-modality summarization (Gross et al., 2000). Furthermore, text summarization has also been applied to other domains such as the scientific domain (Teufel and Moens, 2002) or legal domain (Sharma et al., 2019). We will not discuss these areas in detail, except for the applications of summarization in the medical domain, which we review below.

Numerous efforts have been made to apply text summarization methods to the medical domain. Existing work on **medical applications of text summarization** can be broadly clustered into three areas (Mishra et al., 2014). The first area focuses on the summarization of online medical articles or literature for answering medical or clinical questions. For example, Demner-Fushman and Lin (2006) proposed a system that can answer medical questions by combining an extractive article summarization system with a document semantic clustering system. Chen and Verma (2006) described a system that answers a user's medical queries by summarizing articles retrieved from the database and selecting summaries that are similar to the user's queries. Cao et al. (2011) proposed the AskHERMES system, which answers medical questions by combining a question categorization system, an information retrieval system that retrieves relevant articles, and an extractive summarization system that extracts passages and sentences to answer the questions.

The second area focuses on summarizing biomedical scientific articles into short abstracts. Reeve et al. (2006) proposed an extractive approach for summarizing biomedical articles based on the frequency distribution of concepts in sentences. Plaza et al. (2008) described an ontology and graph-based extractive method that views sentences in an article as nodes in a graph and scores sentences to form the summaries. Sarkar (2009) combined domain-specific features with other commonly used features for sentence ranking and abstract generation. Plaza et al. (2012) further enhanced existing systems with knowledge-based word sense disambiguation methods to improve summarization quality.

The third area focuses on the summarization of clinical patient records (Pivovarov and Elhadad, 2015). Different from the aforementioned work, related work in this area often uses a combination of unstructured text and structured data as the input, and focuses on extracting and displaying key information rather than generating free-text summaries. Powsner and Tufte (1997) described a system that summarizes psychiatric patient records by extracting and visualizing key psychiatric variables. Liu and Friedman (2004) introduced a system that extracts a patient's key problems from the narrative clinical records, and displays them in a tree-structured view. Bui et al. (2007) described a system that summarizes the clinical reports of brain tumor patients by selecting and displaying the most crucial image and textual information from the records. Hirsch et al. (2015) introduced

HARVEST, a system that summarizes a collection of patient records by extracting important concepts and organizing key information in a temporal order. Our studies in Chapter 4 and Chapter 5 are most closely related to this third area of work in that they both focus on summarizing clinical reports. They however differ from this existing work in that they focus on generating abstractive, free-text summaries using neural models.

Finally, despite the numerous efforts in applying summarization to medical documents, work on **summarization of radiology reports** has been limited. Most early work that attempts to “summarize” radiology reports focused on classifying and extracting information from the report text (Friedman et al., 1995; Hripcsak et al., 1998; Elkins et al., 2000; Hripcsak et al., 2002). More recently, Hassanpour and Langlotz (2016) studied extracting various clinical named entities from multi-institutional radiology reports using traditional feature-based classifiers. Goff and Loehfelm (2018) built an NLP pipeline to identify asserted and negated disease entities in the “Impression” section of radiology reports as a step towards report summarization. Cornegruta et al. (2016) proposed to use a recurrent neural network architecture to model radiological language in solving the medical named entity recognition and negation detection tasks on radiology reports. Our work in Chapter 4 and Chapter 5 pioneer in the direction of creating free-text summaries of radiology reports.

To summarize, our studies in Chapter 4 and Chapter 5 are directly built on top of existing work in abstractive and hybrid neural text summarization, and are inspired by existing work on applying RL to summarization systems. Chapter 5 is concurrent to existing work that studies factual correctness of neural summarization systems, and represents the first success in explicitly optimizing a neural summarization system with a correctness objective. Furthermore, both studies are closely related to existing applications of summarization to the medical domain, but they differ from existing work and pioneer in the direction of generating abstractive, free-text summaries of clinical reports.

2.4 Joint Text and Image Understanding

Lastly, in this section we review related work on medical image understanding and the joint modeling of text and image data, to which our study in Chapter 6 is closely related.

Our study is most relevant to existing work on deep learning for **abnormality detection**

from medical images, especially work on classifying radiographic images. For example, Gulshan et al. (2016) and Abràmoff et al. (2016) studied the automatic detection of diabetic retinopathy from retinal fundus photographs and showed the success of CNN-based architectures. Esteva et al. (2017) studied skin cancer detection from clinical skin images using deep neural networks. De Fauw et al. (2018) showed the success of deep learning on the detection of sight-threatening retinal diseases from three-dimensional optical tomography scans. Wang et al. (2017) introduced the first public hospital-scale chest X-ray image dataset covering many disease categories, and evaluated the performance of CNN-based models on this dataset. Rajpurkar et al. (2018b) provided in-depth comparisons of different CNN architectures for chest X-ray image classification. Raghu et al. (2019) studied the effect of transfer learning from ImageNet pretraining on medical image classification tasks. Wang and Wong (2020) studied the applicability of deep learning for the detection of COVID-19 pneumonia based on chest radiographic images. In addition, there are studies that focus on applying deep learning to understanding medical images of other modalities, such as MRI (Mazurowski et al., 2019) or ultrasound (Liu et al., 2019b). We will not conduct an in-depth review of the work in other imaging modalities.

Our study in Chapter 6 is also closely related to work on **joint medical image and text modeling**. Much work in this area focuses on improving medical image representation by understanding or mining the related medical text. Shin et al. (2015) introduced the first system that utilizes a radiology report for improving a CNN-based medical image encoder. Their approach focuses on a topical clustering of the reports rather than a fine-grained understanding of them. Wang et al. (2017) described a method to create a large-scale chest X-ray image dataset by mining the chest radiology reports with word patterns and syntactic rules. Irvin et al. (2019) improved the accuracy of their patterns and further extended their method to handle uncertainty in radiology reports. The goal of our study in Chapter 6 is aligned with this line of work. We however differ from existing work substantially in the development of a domain-agnostic, joint statistical model of the image and text, instead of hand-crafted patterns.

Previous work on joint medical image and text modeling has also focuses on generating textual radiology reports from medical images (Wang et al., 2018; Jing et al., 2018; Liu et al., 2019a). We will not review these studies in detail as they are not directly related to our work.

In addition to the aforementioned areas, our study is directly inspired by the recent line of work on image view-based **contrastive visual representation learning**, which aims to improve image representations by contrasting cropped areas from natural images (Hénaff et al., 2020; Chen et al., 2020a; He et al., 2020; Grill et al., 2020). Our study is a generalization of these methods to a multi-modality setting, where the contrasted pairs are sampled from image and text modality, respectively. To some extent, our method is conceptually related to the multi-view contrastive coding framework by Tian et al. (2020).

Another line of work related to our study is **visual-linguistic representation learning**. A number of recent studies have explored the use of transformer models for joint modeling of image and text data (Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2020). These models are typically trained with a combination of the masked language model objective, an image object prediction objective and a binary image-text pairing objective. Among existing studies, Ilharco et al. (2020) and Gupta et al. (2020) used a cross-modality contrastive objective related to our study in Chapter 6, but for the purpose of probing visual-linguistic models and learning phrase grounding, respectively. Our study differs from this line of work in several crucial ways: 1) while existing work in visual-linguistic learning focuses on learning visual representations from paired text via a binary contrastive prediction task, we show the superior performance of a new cross-modality objective based on noise contrastive estimation; 2) existing work has primarily used object representations extracted from image segmentation models in their preprocessing steps, making them less applicable to medical image understanding tasks where anatomical segmentations are extremely hard to obtain; 3) while existing work has evaluated primarily on visual-linguistic tasks such as visual question answering, we instead focus on evaluation with classification and retrieval tasks which are at the center of medical image understanding research.

To summarize, our study in Chapter 6 is directly related to work on medical image understanding, and uses resources created by the studies of Irvin et al. (2019) and Wang and Wong (2020). Our study extends and improves upon existing work that relies on handcrafted patterns to mine radiology reports for improving medical image representations. Furthermore, our study is directly inspired by recent work on contrastive visual representation learning; it is related to but differs from recent work on visual-linguistic learning in the training objectives and evaluation strategies.

Chapter 3

Understanding Relations in Medical Text and Beyond

Text is the major data format that we use to store and communicate our biomedical knowledge. The biomedical science community reports new scientific discoveries by encoding them into free-text research papers and making them available through publication platforms such as the scientific journals. This textual content is further indexed by scientific databases such as the PubMed platform[1] and becomes the foundation of future medical discoveries and practices.

While this textual knowledge is easy for human beings to read and understand, it becomes cumbersome when we need to query or represent this knowledge with computers, or combine relevant knowledge to make new discoveries, a common need in the development of health information technologies or new drugs (Himmelstein et al., 2017; Zhu et al., 2019). To this end, numerous biomedical knowledge bases, such as DrugBank (Wishart et al., 2008) or PharmGKB (Hewett et al., 2002), have been developed and heavily used by the scientific community. These knowledge bases store structured or semi-structured information about core biomedical entities such as drugs, proteins/genes or diseases, and more importantly, the known relations between these entities. For example, Table 3.1 shows several proteins that are known to have a relation with the drug *chloroquine* in the DrugBank knowledge base. Relational representations like these have made the query, display and

[1] https://pubmed.ncbi.nlm.nih.gov/

Text	Drug	Protein	Relation
Our results indicate that **chloroquine**-mediated inhibition of **TNF**-alpha, IL-1beta and IL-6 synthesis occurs through different modes in lipopolysaccharide-stimulated human monocytes/macrophages.	chloroquine	TNF	inhibitor
...neutrophil stimulation was not prevented by immobilization of bacterial DNA or by wortmannin or **chloroquine**, two agents that inhibit **TLR9** signaling.	chloroquine	TLR9	inhibitor
In addition to the well-known functions of **chloroquine** such as elevations of endosomal pH, the drug appears to interfere with terminal glycosylation of the cellular receptor, angiotensin-converting enzyme 2 (**ACE2**).	chloroquine	ACE2	modulator
This property may explain the ability of **chloroquine** to inhibit **CYP2D6**-mediated metabolism in vitro and in vivo.	chloroquine	CYP2D6	inhibitor
...			

Table 3.1: Drug-protein relational knowledge stored in the DrugBank database, along with the text that indicates the relations in the linked literature. We only show a subset of the relational knowledge related to the drug *chloroquine*. The mention spans of the drugs and proteins are highlighted in bold.

processing of biomedical knowledge much easier. Furthermore, they offer opportunities to conduct complex reasoning within the space and directly help the discovery of biological pathways, a key component in the development of new treatments for diseases (Apic et al., 2005).

Despite their cruciality, the curation of knowledge bases such as DrugBank or PharmGKB has been a long-standing challenge for the scientific community (Klein et al., 2001). Traditionally, this is done manually by human curators, who need to retrieve and read the relevant scientific papers (with text similar to that shown in Table 3.1), identify specific entities and relations from these papers, and add them into the knowledge base. However, unlike annotating general-domain text which can be distributed easily via crowdsourcing, curating biomedical knowledge requires domain experts who have gone through substantial training in the relevant fields. As a result, the curation of these knowledge bases has been an extremely slow and expensive process.

Even after a biomedical knowledge base is successfully constructed, maintaining the

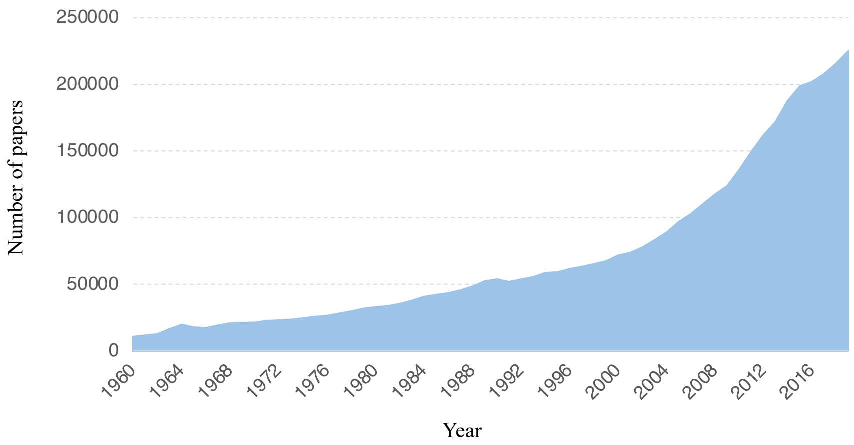


Figure 3.1: Yearly number of papers indexed by PubMed that are relevant to the keyword *cancer*.

knowledge base and keeping its knowledge update-to-date is yet another significant challenge. Imagine that we are building a *cancer* knowledge base that stores knowledge about drugs and proteins that are relevant to the cause and treatment of cancer, it is natural to frequently update the entities and relations in this knowledge base as new biomedical discoveries in this domain are made available. However, this task is far from trivial. As shown in Figure 3.1, the yearly number of papers that are indexed by PubMed and relevant to cancer is growing at an exponential speed. Reading these papers and distilling knowledge from them is beyond the capability of any individual scientific group.

For these reasons, it is imperative to develop systems that can read the free-text scientific literature and distill knowledge from it in an automated fashion. In fact, this is exactly what **relation extraction**, an area of research within natural language processing, aims at solving. Formally, given a piece of text such as a sentence, relation extraction involves discerning whether a relation exists between two entities in the sentence (often termed *subject* and *object*, respectively). Successful relation extraction will not only contribute to the construction of biomedical knowledge bases (Quirk and Poon, 2017), but also serve as the cornerstone of general applications requiring relational understanding of unstructured text on a large scale, such as question answering (Yu et al., 2017).

While some simple relations encoded in the input text can be recognized via the writing of patterns as done in pattern-based relation extraction systems (Auger and Barrière, 2008), thus making relation extraction a simple computational problem, relations in real-world text can be encoded in very diverse forms and therefore difficult to be captured by any specific group of patterns. This is especially true in the case of biomedical scientific text. For example, as shown in Table 3.1, the same *inhibitor* relation between a drug and a protein mention can be expressed in drastically different forms. This has motivated us to develop statistical relation extraction models, especially models based on neural networks (Zelenko et al., 2003; Zeng et al., 2014; Miwa and Bansal, 2016; Zhang et al., 2017), that are able to learn the patterns of a specific relation from a collection of human-annotated examples.

In addition, it is common for two entity mentions in a biomedical relation to span over a long context. As shown by the third example in Table 3.1, the drug mention *chloroquine* and the protein mention *ACE2* are far apart in the sentence where they co-occur, forming a *long-range relation*. Long-range relations like this are extremely common in biomedical scientific text, and pose significant challenge to relation extraction models that directly work on the surface word sequence. This has motivated us to study linguistically-motivated models that instead work on dependency trees, a specific form of syntactic structure of the input sentences (Nivre et al., 2016, 2020). As we will show later, this helps substantially reduce the context that a model has to work with, and thus helps the relation extraction model achieve more robust performance, especially in recognizing long-range relations commonly seen in biomedical text.

In this chapter, we focus on the problem of identifying relations from biomedical text and propose a novel neural network-based architecture for this task. We then demonstrate via experiments that the proposed model not only outperforms existing model types for biomedical relation extraction, but also generalizes to recognizing relations in general-domain text and advances the state of the art on relevant benchmark datasets. Our detailed contributions include:

1) We propose a neural model for relation extraction based on graph convolutional networks, which allows it to efficiently pool information over the syntactic structures of the input sentences;

2) We further present a new path-centric pruning technique to help dependency-based models maximally remove irrelevant information without damaging crucial content

to improve their robustness;

3) We conduct experiments on two widely used biomedical relation extraction datasets that involve recognizing drug-protein and drug-mutation relations, as well as two general-domain relation extraction datasets that involve newswire and web text, and show that the proposed model advances the state of the art for relation extraction;

4) We present detailed analysis on the model and the pruning technique, and show that our model is more robust to long-range relations and have complementary strengths with sequence models.

This chapter is organized as follows. In Section 3.1, we first discuss the limitations of existing dependency-based relation extraction models and motivate the development of our model. Then in Section 3.2 and Section 3.3, we introduce our proposed model architecture and the proposed tree pruning technique, respectively. Next, in Section 3.4 and Section 3.5, we present our experiments on two biomedical relation extraction benchmarks and two general-domain benchmarks, respectively. Lastly, we present some in-depth analysis in Section 3.6 to understand our model and the contributions of its individual components.

3.1 Relation Extraction with Dependency Trees

One of the most commonly used form of syntactic structure is a *dependency parse tree*. For instance, a dependency parse tree representation of the sentence "*Chloroquine does not inhibit infection in the human lung cell Calu-3 with SARS-COV-2*" is shown in Figure 3.2. Each edge in this tree representation marks a typed dependency relationship between a *head* word and a *dependent* word. These head-dependent relations provide an approximation to the semantic relationship between predicates and their arguments that makes them directly useful for many downstream applications (Jurafsky and Martin, 2020). As a result, models making use of dependency parse trees of the input sentences, or *dependency-based models*, have proven to be very effective for relation extraction, because they capture long-range syntactic relations that are obscure from the surface form alone (e.g., when long clauses or complex scoping are present). For instance, in Figure 3.2, the distance between the entity

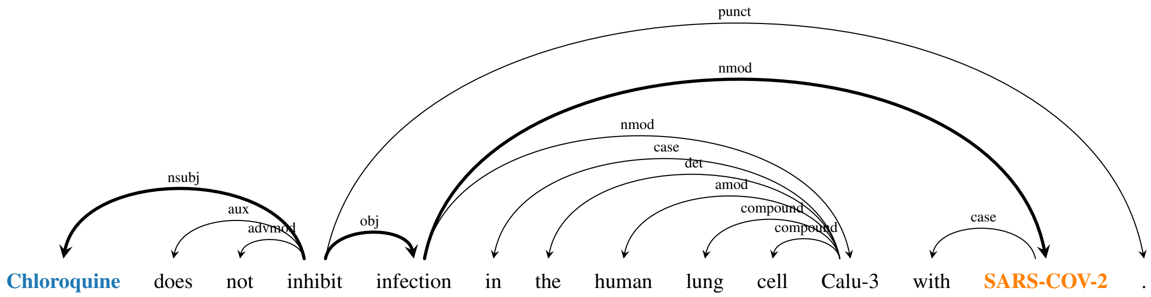


Figure 3.2: Dependency parse tree representation of an example sentence. The dependency parse tree is produced according to the Universal Dependencies formalism (Nivre et al., 2016). The tree is rooted at the word *inhibit*. Each edge in the tree goes from a *head* word to a *dependent* word, representing a grammatical dependency relationship. Moreover, each edge is typed with a particular *dependency label*, such as the *nsubj* label. Note that the distance between *chloroquine* and *SARS-COV-2* is sharply reduced to 3 in the dependency representation.

words *chloroquine* and *SARS-COV-2* is sharply reduced from 11 in the original sentence surface form to 3 in the dependency representation, making the relation extraction task easier.

Traditional feature-based models for relation extraction are able to represent dependency information by featurizing dependency trees as overlapping paths along the trees (Culotta and Sorensen, 2004; Kambhatla, 2004). However, these models face the challenge of sparse feature spaces and are brittle to lexical variations. More recent neural models address this problem with distributed representations built from their computation graphs formed along parse trees. One common approach to leverage dependency information is to perform bottom-up or top-down computation along the parse tree or the subtree below the lowest common ancestor (LCA) of the entities (Miwa and Bansal, 2016). Another popular approach, inspired by Bunescu and Mooney (2005b), is to reduce the parse tree to the *shortest dependency path* between the entities (Xu et al., 2015b,c).

However, these models suffer from several drawbacks. Neural models operating directly on parse trees are usually difficult to parallelize and thus computationally inefficient, because aligning trees for efficient batch training is usually non-trivial. Models based on

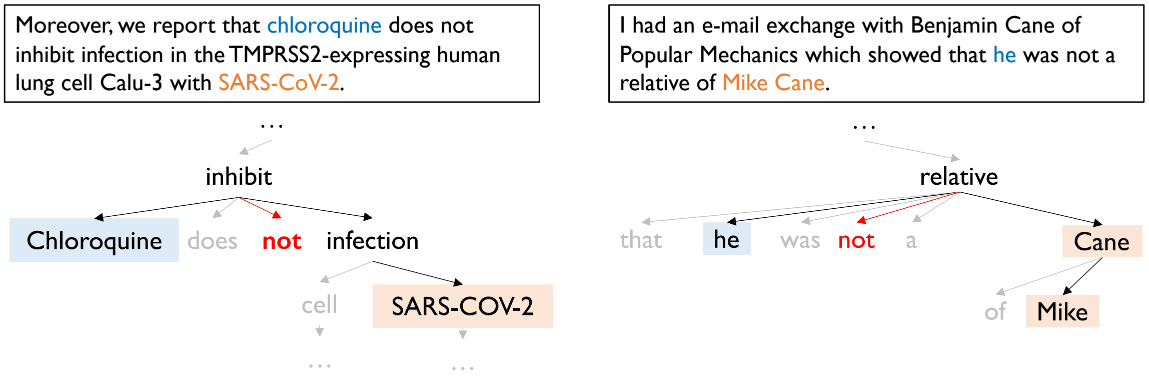

Moreover, we report that chloroquine does not inhibit infection in the TMPRSS2-expressing human lung cell Calu-3 with SARS-CoV-2.
...
inhibit
Chloroquine
does
not
infection
cell
SARS-COV-2
...
...
I had an e-mail exchange with Benjamin Cane of Popular Mechanics which showed that he was not a relative of Mike Cane.
...
relative
that
he
was
not
a
Cane
of
Mike

Figure 3.3: Example sentences along with their dependency trees that encode relations between two entity mentions. Left: an example drawn from a scientific paper (Hoffmann et al., 2020); right: an example modified from the TAC KBP challenge corpus (Getman et al., 2017). For both examples, the subtrees of the original dependency tree between the subject mention (highlighted in blue) and the object mention (highlighted in orange) are also shown, where the shortest dependency paths between the mentions are highlighted in bold. Dependency labels on the tree edges are not shown. Note that in both examples, the negation ("not") is off the dependency path.

the shortest dependency path between the subject and object are computationally more efficient, but this simplifying assumption has major limitations as well. Figure 3.3 shows real-world examples of both biomedical and general-domain text where crucial information (i.e., negation) would be excluded when the model is restricted to only considering the dependency path between the entity mentions.

Motivated by these observations, we propose a novel extension of the graph convolutional network (Kipf and Welling, 2017; Marcheggiani and Titov, 2017) that is tailored for encoding and understanding relations expressed in free text. Our model encodes the dependency structure over the input sentence with efficient graph convolution operations, then extracts entity-centric representations to make robust relation predictions. We also apply a novel *path-centric pruning* technique to remove irrelevant information from the tree while maximally keeping relevant content, which further improves the performance of several dependency-based models including ours.

3.2 Models

In this section, we first describe graph convolutional networks (GCNs) over dependency tree structures, and then we introduce an architecture that uses GCNs at its core for relation extraction.

3.2.1 Graph Convolutional Networks over Dependency Trees

The graph convolutional network (Kipf and Welling, 2017) is an adaptation of the convolutional neural network (LeCun et al., 1998) for encoding graphs. Given a graph with n nodes, we can represent the graph structure with an $n \times n$ adjacency matrix $\mathbf{A}$ where $A_{ij} = 1$ if there is an edge going from node i to node j. In an L-layer GCN, if we denote by $h_i^{(l-1)}$ the input vector and $h_i^{(l)}$ the output vector of node i at the l-th layer, a graph convolution operation can be written as

$$h_i^{(l)} = \sigma\big(\sum_{j=1}^{n} A_{ij} W^{(l)} h_j^{(l-1)} + b^{(l)}\big), \tag{3.1}$$

where $W^{(l)}$ is a linear transformation, $b^{(l)}$ a bias term, and σ a nonlinear function (e.g., ReLU (Nair and Hinton, 2010)). Intuitively, during each graph convolution, each node gathers and summarizes information from its neighboring nodes in the graph.

We adapt the graph convolution operation to model dependency trees by converting each tree into its corresponding adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if there is a dependency edge between tokens i and j. However, naively applying the graph convolution operation in Equation (3.1) could lead to node representations with drastically different magnitudes, since the degree of a token varies a lot. This could bias our sentence representation towards favoring high-degree nodes regardless of the information carried in the node (see details in Section 3.2.2). Furthermore, the information in $h_i^{(l-1)}$ is never carried over to $h_i^{(l)}$, since nodes never connect to themselves in a dependency tree.

We resolve these issues by normalizing the activations in the graph convolution before

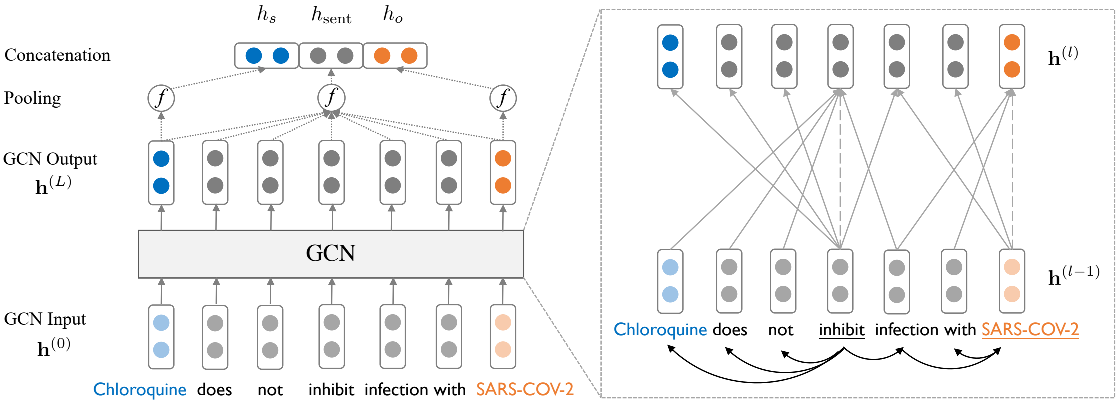


Figure 3.4: Overview of a graph convolutional network for relation extraction. The left side shows the overall model architecture. On the right side, we only show the detailed graph convolution computation for the words "inhibit" and "SARS-COV-2" for clarity. A full unlabeled dependency parse of the sentence is also provided at the bottom for reference.

feeding it through the nonlinearity, and adding self-loops to each node in the graph:

$$h_i^{(l)} =\sigma\big(\sum_{j=1}^{n} \tilde{A}_{ij}W^{(l)}h_j^{(l-1)}/d_i + b^{(l)}\big), \tag{3.2}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ with $\mathbf{I}$ being the $n \times n$ identity matrix, and $d_i = \sum_{j=1}^{n} \tilde{A}_{ij}$ is the degree of token i in the resulting graph.

Stacking this operation over L layers gives us a deep GCN network, where we set $h_1^{(0)}, \ldots, h_n^{(0)}$ to be input word vectors, and use $h_1^{(L)}, \ldots, h_n^{(L)}$ as output word representations. All operations in this network can be efficiently implemented with matrix multiplications, making it ideal for batching computation over examples and running on GPUs. For example, one layer of information propagation in Equation (3.2) can be implemented as:

$$\mathbf{h}^{(l)} = \sigma\big(W^{(l)}\mathbf{h}^{(l-1)}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-1} + b^{(l)} \otimes 1_n\big), \tag{3.3}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{D}$ is the diagonal matrix where $D_{ii} = d_i + 1, \forall i$, and 1_n represents the n-dimensional vector with all ones. Moreover, the propagation of information between tokens occurs in parallel, and the runtime does not depend on the depth of the dependency tree.

Note that the GCN model presented above uses the same parameters for all edges in the dependency graph. We also experimented with: (1) using different transformation matrices W for top-down, bottom-up, and self-loop edges; and (2) adding dependency relation-specific parameters for edge-wise gating, similar to (Marcheggiani and Titov, 2017). We found that modeling directions does not lead to improvement,[2] and adding edge-wise gating further hurts performance. We hypothesize that this is because the presented GCN model is usually already able to capture dependency edge patterns that are informative for classifying relations, and modeling edge directions and types does not offer additional discriminative power to the network before it leads to overfitting. For example, the relations entailed by "*A*'s son, *B*" and "*B*'s son, *A*" can be readily distinguished with "'s" attached to different entities, even when edge directionality is not considered.

3.2.2 Encoding Relations with GCN

We now formally define the task of relation extraction. Let $\mathcal{X} = [x_1, ..., x_n]$ denote a sentence, where x_i is the i^{th} token. A subject entity and an object entity are identified and correspond to two spans in the sentence: $\mathcal{X}_s = [x_{s_1}, \ldots, x_{s_2}]$ and $\mathcal{X}_o = [x_{o_1}, \ldots, x_{o_2}]$. Given $\mathcal{X}$, $\mathcal{X}_s$, and $\mathcal{X}_o$, the goal of relation extraction is to predict a relation $r \in \mathcal{R}$ (a predefined relation set) that holds between the entities or "no relation" otherwise.

After applying an L-layer GCN over word vectors, we obtain hidden representations of each token that are directly influenced by its neighbors no more than L edges apart in the dependency tree. To make use of these word representations for relation extraction, we first obtain a sentence representation as follows (see also Figure 3.4 left):

$$h_{\text{sent}} = f\big(\mathbf{h}^{(L)}\big) = f\big(\text{GCN}(\mathbf{h}^{(0)})\big), \tag{3.4}$$

where $\mathbf{h}^{(l)}$ denotes the collective hidden representations at layer l of the GCN, and $f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^d$ is a max pooling function that maps from n output vectors to the sentence vector by keeping the maximum value for each dimension in the output vectors.

We also observe that information close to entity tokens in the dependency tree is often central to relation classification. Therefore, we also obtain a subject representation h_s from

[2] We therefore treat the dependency graph as undirected, i.e. $\forall i, j, A_{ij} = A_{ji}$.

$\mathbf{h}^{(L)}$ as follows

$$h_s = f\big(\mathbf{h}^{(L)}_{s_1:s_2}\big), \tag{3.5}$$

as well as an object representation h_o similarly.

Inspired by recent work on relational learning between entities (Santoro et al., 2017; Lee et al., 2017), we obtain the final representation used for classification by concatenating the sentence and the entity representations, and feeding them through a feed-forward neural network (FFNN):

$$h_{\text{final}} = \text{FFNN}\big([h_{\text{sent}}; h_s; h_o]\big). \tag{3.6}$$

This h_{final} representation is then fed into a linear layer followed by a softmax operation to obtain a probability distribution over relations.

The max pooling function in Equation (3.4) collects representations from all tree nodes as features for the classifier. In our experiments we found that the output vector h_{sent} tends to have large magnitude, and therefore adding the following regularization term to the cross entropy loss of each example improves the results:

$$\ell_{\text{reg}} = \lambda \cdot \|h_{\text{sent}}\|^2. \tag{3.7}$$

Here, ℓ_{reg} functions as an l_2 regularization on the learned sentence representations. λ controls the regularization strength and we set $\lambda = 0.003$. We empirically found this to be more effective than applying l_2 regularization on the convolutional weights.

3.2.3 Contextualized GCN

The network architecture introduced so far learns effective representations for relation extraction, but it also leaves a few issues inadequately addressed. First, the input word vectors do not contain contextual information about word order or disambiguation. Second, the GCN highly depends on a correct parse tree to extract crucial information from the sentence (especially when pruning is performed), while existing parsing algorithms produce imperfect trees in many cases.

To resolve these issues, we further apply a Contextualized GCN (C-GCN) model, where the input word vectors are first fed into a bi-directional long short-term memory (LSTM) network to generate contextualized representations, which are then used as $\mathbf{h}^{(0)}$ in the original model. This BiLSTM contextualization layer is trained jointly with the rest of the network. We show empirically in Section 3.5 that this augmentation substantially improves the performance over the original model.

We note that this relation extraction model is conceptually similar to graph kernel-based models (Zelenko et al., 2003), in that it aims to utilize local dependency tree patterns to inform relation classification. Our model also incorporates crucial off-path information, which greatly improves its robustness compared to shortest dependency path-based approaches. Compared to tree-structured models (e.g., Tree-LSTM (Tai et al., 2015)), it not only is able to capture more global information through the use of pooling functions, but also achieves substantial speedup by not requiring recursive operations that are difficult to parallelize. For example, we observe that on a Titan Xp GPU, training a Tree-LSTM model over a minibatch of 50 examples takes 6.54 seconds on average, while training the original GCN model takes only 0.07 seconds, and the C-GCN model 0.08 seconds.

3.3 Incorporating Off-path Information with Path-centric Pruning

Dependency trees provide rich structures that one can exploit in relation extraction, but most of the information pertinent to relations is usually contained within the subtree rooted at the lowest common ancestor (LCA) of the two entities. Previous studies (Xu et al., 2015c; Miwa and Bansal, 2016) have shown that removing tokens outside this scope helps relation extraction by eliminating irrelevant information from the sentence. It is therefore desirable to combine our GCN models with tree pruning strategies to further improve performance. However, pruning too aggressively (e.g., keeping only the dependency path) could lead to loss of crucial information and conversely hurt robustness. For instance, the

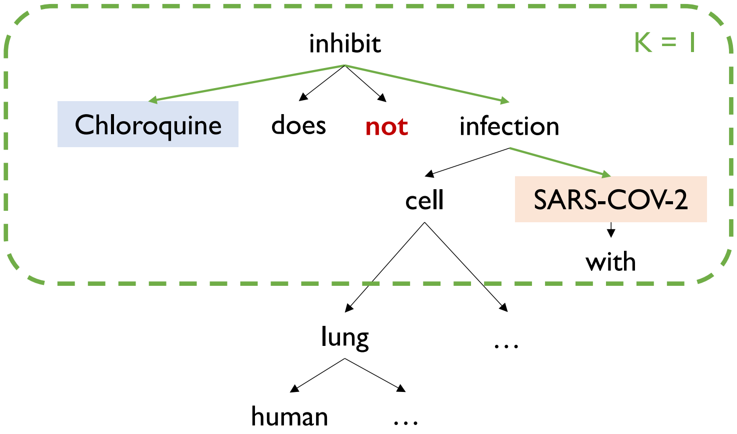


Figure 3.5: An example dependency tree pruned with the proposed path-centric pruning technique. The dependency structure shown was derived from the sentence "*Moreover, we report that chloroquine does not inhibit infection in the TMPRSS2-expressing human lung cell Calu-3 with SARS-CoV-2*". The dependency path between the two mentions is highlighted with green edges, and the tree structure kept after pruning with $K = 1$ is shown in the green box. Note that the critical negation word "*not*" is kept after the pruning.

negation in Figure 3.3 is neglected when a model is restricted to only looking at the dependency path between the entities. Similarly, in the sentence "*She was diagnosed with cancer last year, and succumbed this June*", the dependency path *She*←*diagnosed*→*cancer* is not sufficient to establish that *cancer* is the cause of death for the subject unless the conjunction dependency to *succumbed* is also present.

Motivated by these observations, we propose *path-centric pruning*, a novel technique to incorporate information off the dependency path. This is achieved by including tokens that are up to distance K away from the dependency path in the LCA subtree. $K = 0$, corresponds to pruning the tree down to the path, $K = 1$ keeps all nodes that are directly attached to the path (see Figure 3.5), and $K = \infty$ retains the entire LCA subtree. We combine this pruning strategy with our GCN model, by directly feeding the pruned trees into the graph convolutional layers.[3] We show that pruning with $K = 1$ achieves the best balance between including relevant information (e.g., negation and conjunction) and keeping irrelevant content out of the resulting pruned tree as much as possible. We also empirically show that $K = 1$ works uniformly better than other pruning strategies in Section 3.6.1.

[3] For our C-GCN model, the LSTM layer still operates on the full sentence regardless of the pruning.

We note that a technique similar to path-centric pruning has been applied to reduce the space of possible arguments in semantic role labeling (He et al., 2018). The authors showed pruning words too far away from the path between the predicate and the root to be beneficial, but reported the best pruning distance to be 10, which almost always retains the entire tree. Our method differs in that it is applied to the shortest dependency path between entities, and we show that in our technique the best pruning distance is 1 for several dependency-based relation extraction models.

3.4 Experiments: Understanding Biomedical Relations

In this section, we first evaluate the performance of our proposed model on the task of understanding biomedical relations found in biomedical literature text. We then study the generalizability of our proposed model architecture on recognizing general-domain relations found in web and newswire text in the next section.

3.4.1 Experimental Setup

To evaluate the effectiveness of our proposed model on the task of understanding biomedical relations, we conduct separate experiments on two biomedical relation extraction datasets:

- **ChemProt** (Krallinger et al., 2017): First introduced in the BioCreative VI shared tasks, the ChemProt dataset aims at evaluating automatic systems that are able to automatically detect relations between chemical compounds/drug and genes/proteins from free text such as PubMed abstracts. It includes 10,060 sentences sampled from PubMed abstracts, each with manual annotations of chemical compound mentions, gene/protein mentions, and one of 13 types of chemical-protein relations. Examples of such relations include the *inhibitor* relation, which indicates that a particular chemical compound inhibits the expression of a particular protein/gene, or the *indirect_upregulator* relation, which indicates that a particular compound indirectly up-regulates the expression of a protein/gene. This dataset has been widely used as a testbed of biomedical relation extraction systems, and we report the micro-averaged F_1 scores as is conventional.

- **Drug-Mutation** (Peng et al., 2017): The Drug-Mutation dataset was originally created for evaluating automatic systems for the task of extracting binary and ternary relations held between drug, gene and mutation mentions. A drug-gene-mutation interaction is broadly defined as an association between the drug efficacy and the mutation in the given gene, and this relational knowledge is often of clinical importance to molecular tumor boards for cancer treatment. The dataset was constructed by sampling sentences from biomedical literature in the PubMed Central database[4] and annotating them with distant supervision. For our experiments, we focus on the binary task of recognizing drug-mutation relations from text, which includes 3,192 positive examples and an equal number of negative examples. Due to the relatively small dataset size, we follow Peng et al. (2017) and report the average test accuracy over five-fold cross validation.

For experiments on both datasets, we follow Zhang et al. (2017) and employ an "entity mask" strategy where we replace each subject entity with a special *SUBJ-<NER>* token, and each object entity with a special *OBJ-<NER>* token. For instance, the example sentence "*[Cyanopindolol]$_{Chemical}$, an antagonist of the [serotonin terminal autoreceptor]$_{Protein}$, also prolonged the clearance of 5-HT from the CA3 region*" is converted to "*SUBJ-Chemical, an antagonist of the OBJ-Protein OBJ-Protein OBJ-Protein, also prolonged the clearance...*" in preprocessing. This not only provides our model with the entity position information, but also prevents the model from overfitting to the actual entity mentions during training.

3.4.2 Baseline Models

We compare our models with several competitive dependency-based models and neural sequence models.

Dependency-based models. In our biomedical relation extraction experiments, we compare with three types of dependency-based models widely adopted in previous work. (1) Shortest Dependency Path LSTM (SDP-LSTM) (Xu et al., 2015c), which applies a neural

[4] http://www.ncbi.nlm.nih.gov/pmc/

sequence model on the shortest path between the subject and object entities in the dependency tree. (2) Tree-LSTM (Tai et al., 2015), which is a recursive model that generalizes the LSTM to arbitrary tree structures. We investigate the child-sum variant of Tree-LSTM, and apply it to the dependency tree (or part of it). In practice, we find that modifying this model by concatenating dependency label embeddings to the input of forget gates improves its performance on relation extraction, and therefore use this variant in our experiments. (3) Graph-LSTM (Peng et al., 2017), which modifies the original LSTM reccurence by using additional nodes connected by dependency edges as input. This model was first proposed for the task of n-ary relation extraction task. We compare to this model in our experiments on the Drug-Mutation dataset.

Neural sequence model. We presented in previous work (Zhang et al., 2017) a competitive sequence model that processes the input sentence with an LSTM, and applies an attention mechanism on top of it to select hidden states more relevant to the prediction. The attention mechanism is position-aware, in that it encodes the relative position of each token in the sentence with respect to the subject and the object with position embeddings. We showed that for the task of relation extraction it outperforms several CNN and dependency-based models by a substantial margin. We refer to this model as position-aware LSTM (PA-LSTM), and compare our proposed model with this strong baseline.

3.4.3 Results

We present our main results on the ChemProt dataset in Table 3.2. We first observe that that shorted dependency path-based model (SDP-LSTM) and the Tree-LSTM model achieve comparable scores on this dataset, with the Tree-LSTM model slightly more effective than the SDP-LSTM model. On the other hand, the strong sequence-based PA-LSTM model outperforms both dependency-based baselines notably, by a margin of up to 6.7 F_1, despite no linguistic information being used by this model. Overall, we find that our proposed C-GCN model achieves the best performance, outperforming the best dependency-based baseline by notable 7.9 F_1 points and the PA-LSTM model by 3.6 F_1 points. This suggests

System	Test F_1
SDP-LSTM (Xu et al., 2015c)	67.3
Tree-LSTM (Tai et al., 2015)	69.7
PA-LSTM (Zhang et al., 2017)	74.0*
GCN (Ours)	**73.1***
C-GCN (Ours)	**77.6***

Table 3.2: Micro-averaged test F_1 scores on the ChemProt dataset. All results are obtained under the same setup by using the model's open implementation. $*$ marks statistically significant improvements over the SDP-LSTM model with $p < .05$ under a bootstrap test.

System	Test Accuracy
SDP-LSTM‡ (Xu et al., 2015c)	70.2
Tree-LSTM‡ (Tai et al., 2015)	75.9
Graph-LSTM† (Peng et al., 2017)	75.6
C-GCN (Ours)	**84.2**

Table 3.3: Average test accuracy scores over five-fold cross validation on the Drug-Mutation dataset. † marks results reported in the original paper, and ‡ marks results obtained by using the open implementations.

not only that our proposed GCN architecture improves the model's ability of utilizing linguistic structure, but also that the additional use of linguistic information has improved the model's ability at discerning relations expressed in the input text.

We further present the results on the Drug-Mutation dataset in Table 3.3. We confirm the effectiveness of our proposed C-GCN model, and find that it achieves much higher performance than other baseline models under this cross validation setup, outperforming the best baseline model (Tree-LSTM) by 8.3 F_1.

3.5 Experiments: Understanding General-domain Relations

In this section, we generalize our evaluation in Section 3.4, and study the effectiveness of our proposed models on recognizing general-domain relations found in web and newswire text.

3.5.1 Experimental Setup

We conduct experiments on two general-domain relation extraction datasets:

- **TACRED** (Zhang et al., 2017): TACRED aims at evaluating the performance of automated systems at understanding relational facts present in newswire and web data. It was created by sampling over 106k mention pairs drawn from the newswire and discussion forum text offered by the yearly TAC KBP[5] challenge. Mentions in TACRED are divided into subject and object mentions, with subject mentions categorized into either person or organization types, and object mentions categorized into 16 fine-grained types (e.g., date, location). The mention pairs (as well as the sentences containing them) are then crowd-annotated with one of 41 relation types and a special *no_relation* class when the mention pair does not have a relation between them within these categories. Example relation types in TACRED include the *per:title* relation, which connects a *person* mention with its corresponding job *title* mention, or the *org:founded_by* relation, which connects an *organization* mention with its founding *person* mention. For this dataset, we report micro-averaged precision, recall and F_1 scores as is conventional.

- **SemEval 2010 Task 8** (Hendrickx et al., 2009): The SemEval dataset is widely used in previous work for evaluating systems' performance at understanding relations in web text, but is significantly smaller with 8,000 examples for training and 2,717 for testing. Sentences in the SemEval dataset are collected via pattern-based Web search and annotated by a group of expert annotators. Unlike TACRED, it contains

[5] https://tac.nist.gov/2017/KBP/index.html

19 relation classes over untyped mention pairs: 9 directed relations and a special *Other* class. Examples of the directed relations include the *Cause-Effect* relation, which suggests that an event or object leads to an effect indicated by another object, or the *Component-Whole* relation, indicating that an object is a component of a larger whole. For SemEval, we follow the convention and report the official macro-averaged F_1 scores.

For fair comparisons on the TACRED dataset, we follow the same evaluation protocol used in (Zhang et al., 2017) by selecting the model with the median dev F_1 from 5 independent runs and reporting its test F_1. We also use the same "entity mask" strategy as we did in Section 3.4. For all models, we also adopt the "multi-channel" strategy as in (Zhang et al., 2017) by concatenating the input word embeddings with POS and NER embeddings.

Traditionally, evaluation on SemEval is conducted without entity mentions masked. However, as we will discuss in Section 3.6.5, we found this method to encourage models to overfit to these mentions and therefore fail to test their actual ability to generalize. We therefore report results on the SemEval dataset with two separate evaluation protocols: (1) *with-mention*, where mentions are kept for comparison with previous work; and (2) *mask-mention*, where they are masked to test the generalization of our model in a more realistic setting.

3.5.2 Baseline Models

We again compare our models with several competitive dependency-based and neural sequence models.

Dependency-based models. Similarly, we use the SDP-LSTM (Xu et al., 2015c) and Tree-LSTM (Tai et al., 2015) as our baseline dependency-based models. Additionally, we compare against a logistic regression (LR) classifier which combines dependency-based features with hand-tuned features, including dependency path-based features, lemmatized n-gram features and NER/POS tag features, as this model was shown to achieve competitive performance on the TACRED dataset (Zhang et al., 2017), despite its simplicity in model architecture.

System	P	R	F_1
LR† (Zhang+2017)	**73.5**	49.9	59.4
SDP-LSTM† (Xu+2015c)	66.3	52.7	58.7
Tree-LSTM‡ (Tai+2015)	66.0	59.2	62.4
PA-LSTM† (Zhang+2017)	65.7	64.5	65.1
GCN	69.8	59.0	64.0
C-GCN	69.9	63.3	66.4*
C-GCN + PA-LSTM	71.3	**65.4**	**68.2***

Table 3.4: Micro-averaged test precision, recall and F_1 scores on the TACRED dataset. Underscore marks highest number among single models; bold marks highest among all. † marks results reported in (Zhang et al., 2017); ‡ marks results produced with our implementation. $*$ marks statistically significant improvements over PA-LSTM with $p < .01$ under a bootstrap test.

Neural sequence model. As in Section 3.4, we again compare against the PA-LSTM model as it was shown to achieve state-of-the-art performance on the TACRED dataset (Zhang et al., 2017).

3.5.3 Results

We present our main results on the TACRED test set in Table 3.4. We first observe that, similar to the biomedical relation extraction experiments in Section 3.4, our GCN model again outperforms all dependency-based models by at least 1.6 F_1. By using contextualized word representations, the C-GCN model further outperforms the strong PA-LSTM model by 1.3 F_1. In addition, we find our model improves upon other dependency-based models in both precision and recall. Comparing the C-GCN model with the GCN model, we find that the gain mainly comes from improved recall. We hypothesize that this is because the C-GCN is more robust to parse errors by capturing local word patterns (see also Section 3.6.3).

As we will show in Section 3.6.3, we find that our GCN models have complementary strengths for recognizing relations when compared to the PA-LSTM. To leverage this result, we further experiment with a simple interpolation strategy to combine these models. Given the output probabilities $P_G(r|x)$ from a GCN model and $P_S(r|x)$ from the sequence model

System	*with-m*	*mask-m*
SVM† (Rink+2010)	82.2	–
SDP-LSTM† (Xu+2015c)	83.7	–
SPTree† (Miwa+2016)	84.4	–
PA-LSTM‡ (Zhang+2017)	82.7	75.3
Our Model (C-GCN)	**84.8***	**76.5***

Table 3.5: Macro-averaged F$_1$ scores on the SemEval dataset. † marks results reported in the original papers; ‡ marks results produced by using the open implementation. The last two columns show results from *with-mention* evaluation and *mask-mention* evaluation, respectively. $*$ marks statistically significant improvements over PA-LSTM with $p < .05$ under a bootstrap test.

for any relation r, we calculate the interpolated probability as

$$P(r|x) = \alpha \cdot P_G(r|x) + (1 - \alpha) \cdot P_S(r|x)$$

where $\alpha \in [0, 1]$ is chosen on the dev set and set to 0.6. This simple interpolation between a C-GCN and a PA-LSTM model achieves an F$_1$ score of 68.2, outperforming each model alone by at least 1.8 F$_1$.

We present additional results on the SemEval test set in Table 3.5. We find that under the conventional *with-entity* evaluation, our C-GCN model again outperforms all existing dependency-based neural models on this separate dataset, confirming its effectiveness on understanding general-domain relations. Notably, by properly incorporating off-path information, our model outperforms the previous shortest dependency path-based model (SDP-LSTM). Under the *mask-entity* evaluation, our C-GCN model also outperforms PA-LSTM by a substantial margin, suggesting its generalizability even when entities are not seen.

3.6 Analysis & Discussion

In this section, we analyze our proposed model architecture by running additional experiments on both the biomedical and general-domain datasets. Our experiments aim at answering three main questions. First, how does individual component in the proposed model

Pruning Strategy	ChemProt F_1	TACRED F_1
$K = 0$	76.2	67.0
$K = 1$	**77.6**	**67.6**
$K = 2$	76.1	67.3
$K = \infty$	75.9	67.2
Full Tree	74.1	66.8

Table 3.6: Performance of the C-GCN model under different pruning strategies. We report F_1 scores on the dev set of the ChemProt dataset and the TACRED dataset. K represents the pruning distance variable; $K = \infty$ indicates a pruning strategy where we keep the entire subtree rooted at the LCA of the two entities; "Full tree" indicates that no pruning is used.

contribute to its effectiveness? Second, does the linguistically-motivated architecture provide our model with complementary strength over sequence-based models? And lastly, how can we interpret predictions from our proposed model?

3.6.1 Effect of Path-centric Pruning

To understand the effectiveness of the introduced path-centric pruning technique, we compare the performance of the C-GCN model on the ChemProt biomedical relation extraction dataset and the TACRED general-domain relation extraction dataset, when the pruning distance K is varied. We experiment with $K \in \{0, 1, 2, \infty\}$ and a setting where the full tree is used as input, and present the results in Table 3.6. For both tasks, we find that the performance of the C-GCN model peaks when $K = 1$, outperforming their respective dependency path-based counterpart ($K = 0$). This confirms our hypothesis in Section 3.3 that incorporating off-path information is crucial to relation extraction. We further find that for both tasks the C-GCN model becomes less effective when the entire dependency tree is present, indicating that including extra information can hurt the performance.

To study how these findings generalize to other dependency-based model architectures, we run additional experiments with varying pruning distance K for the Tree-LSTM model on the TACRED dataset. We find that the Tree-LSTM model achieves dev F_1scores of 64.1, 64.7, 64.5 and 64.1 with $K = 0, 1, 2, \infty$, respectively. Miwa and Bansal (2016) reported that a Tree-LSTM achieves similar performance when the dependency path and the LCA

Model	Dev F_1
Best C-GCN	67.4
– h_s, h_o, and Feedforward (FF)	66.4
– LSTM Layer	65.5
– Dependency tree structure	64.2
– FF, LSTM, and Tree	57.1
– FF, LSTM, Tree, and Pruning	47.4

Table 3.7: An ablation study of the best C-GCN model. Scores are reported on the dev set of the TACRED dataset and are median results from 5 independently trained models.

subtree are used. Our experiments confirm their finding, and further show that the result can be improved by path-centric pruning with $K = 1$.

3.6.2 Ablation Study

To study the contribution of each component in the C-GCN model, we ran an ablation study on the TACRED dev set, and present the results in Table 3.7. We find that: (1) The entity representations and feedforward layers contribute 1.0 F_1. (2) When we remove the dependency structure from the model (i.e., setting $\tilde{\mathbf{A}}$ to $\mathbf{I}$), the score drops by a notable 3.2 F_1. (3) F_1 drops markedly by 10.3 when we remove the feedforward layers, the LSTM component and the dependency structure altogether. (4) Removing the pruning (i.e., using full trees as input) further hurts the result by another 9.7 F_1.

3.6.3 Complementary Strengths of GCNs and PA-LSTMs

To understand what the GCN models are capturing and how they differ from a sequence model such as the PA-LSTM, we compared their performance over examples in the TACRED dev set. Specifically, for each model, we trained it for 5 independent runs with different seeds, and for each example we evaluated the model's accuracy over these 5 runs. For instance, if a model correctly classifies an example for 3 out of 5 times, it achieves an accuracy of 60% on this example. We observe that on 847 (3.7%) dev examples, our C-GCN model achieves an accuracy at least 60% higher than that of the PA-LSTM, while on 629

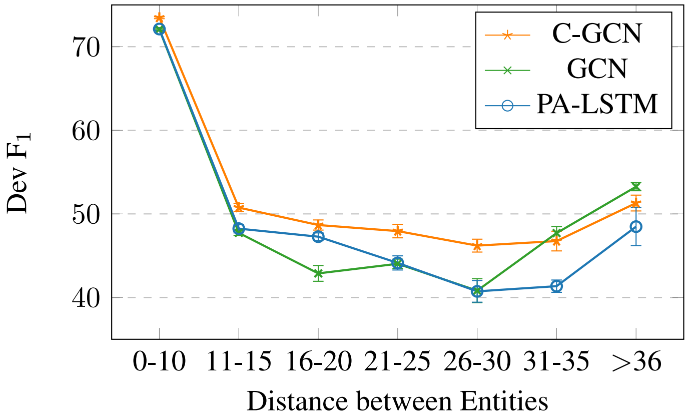


Figure 3.6: Relation extraction performance with regard to distance between the entities in the sentence for C-GCN, GCN and PA-LSTM. We report results on the dev set of the TACRED dataset. Error bars indicate standard deviation of the mean estimate over results from 5 independently trained models.

(2.8%) examples the PA-LSTM achieves 60% higher. This complementary performance explains the gain we see in Table 3.4 when the two models are combined.

We further show that this difference is due to each model's competitive advantage (see Figure 3.6): dependency-based models are better at handling sentences with entities farther apart, while sequence models can better leverage local word patterns regardless of parsing quality.

3.6.4 Understanding Model Behavior

To gain more insights into the C-GCN model's behavior, we visualized the partial dependency tree it is processing and how much each token's final representation contributed to h_{sent} (Figure 3.7). We find that the model often focuses on the dependency path, but sometimes also incorporates off-path information to help reinforce its prediction. The model also learns to ignore determiners (e.g., "the") as they rarely affect relation prediction.

To further understand what dependency edges contribute most to the classification of different relations, we scored each dependency edge by summing up the number of dimensions each of its connected nodes contributed to h_{sent}. We present the top scoring edges in Table 3.8. As can be seen in the table, most of these edges are associated with indicative

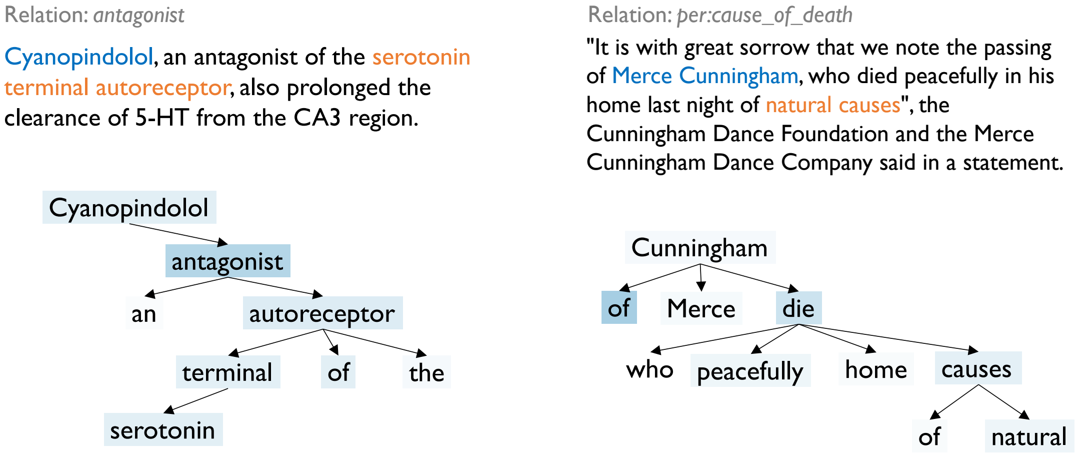

Relation: antagonist
Cyanopindolol, an antagonist of the serotonin terminal autoreceptor, also prolonged the clearance of 5-HT from the CA3 region.
Cyanopindolol
antagonist
an
autoreceptor
terminal
of
the
serotonin
Relation: per:cause_of_death
"It is with great sorrow that we note the passing of Merce Cunningham, who died peacefully in his home last night of natural causes", the Cunningham Dance Foundation and the Merce Cunningham Dance Company said in a statement.
Cunningham
of
Merce
die
who
peacefully
home
causes
of
natural

Figure 3.7: Examples and the pruned dependency trees where the C-GCN predicted correctly. The left example is drawn from the ChemProt dataset, while the right example is drawn from the TACRED dataset. In both examples, words are shaded by the number of dimensions they contributed to h_{sent} in the pooling operation, with punctuation omitted. In the left example, the word "antagonist" contributed the most to the prediction, while in the right example, "die" and "of" are the top contributors to the prediction.

nouns or verbs of each relation.[6]

3.6.5 Entity Bias in the SemEval Dataset

In our study, we observed a high correlation between the entity mentions in a sentence and its relation label in the SemEval general-domain relation extraction dataset. To further understand this phenomenon, we ran further experiments with the PA-LSTM model (Zhang et al., 2017) on this dataset.[7] We started by simplifying every sentence in the SemEval training and dev sets to "*subject* and *object*", where *subject* and *object* are the actual entities in the sentence. Surprisingly, a trained PA-LSTM model on this data is able to achieve 65.1 F_1 on the dev set if GloVe is used to initialize word vectors, and 47.9 dev F_1 even without GloVe initialization. This is significantly higher than the F_1 score of 4.9 generated by random guess, suggesting that substantial biases encoded in the entities are utilized by the

[6] We do notice the effect of dataset bias as well: the name "Buffett" is too often associated with contexts where shareholder relations hold, and therefore ranks top in that relation.

[7] We choose the PA-LSTM model because it is more amenable to our experiments with simplified examples.

Relation	Dependency Tree Edges		
per:children	S-PER ← son	son → O-PER	S-PER ← survived
per:other_family	S-PER ← stepson	niece → O-PER	O-PER ← stepdaughter
per:employee_of	a ← member	S-PER ← worked	S-PER ← played
per:schools_attended	S-PER ← graduated	S-PER ← earned	S-PER ← attended
org:founded	founded → O-DATE	established → O-DATE	was ← founded
org:number_of_employees	S-ORG ← has	S-ORG → employs	O-NUMBER ← employees
org:subsidiaries	S-ORG ← O-ORG	S-ORG → 's	O-ORG → division
org:shareholders	buffett ← O-PER	shareholder → S-ORG	largest ← shareholder

Table 3.8: The three dependency edges that contribute the most to the classification of different relations in the TACRED dev set. Edge contribution is the sum of pooling dimensions contributed by tokens on either end of the edge, calculated over the entire dev set. For clarity, we removed edges which 1) connect to common punctuation (i.e., commas, periods, and quotation marks), 2) connect to common prepositions (i.e., of, to, by), and 3) connect between tokens within the same entity. We use PER, ORG for entity types of PERSON, ORGANIZATION. We use S- and O- to denote subject and object entities, respectively.

models. To further evaluate the model in a more realistic setting, we trained one model with the original SemEval training set (unmasked) and one with mentions masked in the training set, following our standard preprocessing procedures with other datasets (masked). While the unmasked model achieves a 83.6 F_1 on the original SemEval dev set, F_1 drops drastically to 62.4 if we replace dev set entity mentions with a special <*UNK*> token to simulate the presence of unseen entities. In contrast, the masked model is unaffected by unseen entity mentions and achieves a stable dev F_1 of 74.7. This suggests that models trained without entities masked generalize poorly to new examples with unseen entities. Our findings call for more careful evaluation that takes dataset biases into account in future relation extraction studies.

3.7 Summary & Future Directions

In this chapter, we have presented a novel neural architecture for relation extraction based on a graph convolutional network that can efficiently pool information from the dependency tree of the input sentence. We also presented a path-centric pruning technique that improves the robustness of dependency-based models by removing irrelevant content without ignoring crucial information.

We have shown via experiments on a widely used chemical-protein and a drug-mutation relation extraction dataset, that our proposed model substantially outperforms existing model types, including sequence-based models and dependency-based models, for the task of biomedical relation extraction. We further demonstrated via experiments on two general-domain relation extraction dataset that our model is effective for the general relation extraction task, and advances the state of the art on these datasets. We also showed through detailed analysis that our model has complementary strengths to sequence models, and that the proposed pruning technique can be effectively applied to other dependency-based models.

There are several promising directions that our method can be further extended:

- **The integration of large pretrained language models**. Since the original publication of our study, neural language models (LMs) pretrained on large unsupervised corpora have advanced the state of the art of various NLP benchmark tasks, and become a common building block for NLP models. Early examples of these LMs include ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Since then, various studies have demonstrated the benefits of using LMs for relation extraction (Alt et al., 2019; Peters et al., 2019; Joshi et al., 2020). Furthermore, Lee et al. (2019); Peng et al. (2019); Gu et al. (2020) have applied large LMs to the task of biomedical relation extraction and shown state-of-the-art results on several benchmark datasets. Despite these advances, it remains an open research question how to best leverage pretrained LMs for the task of relation extraction.

- **Unsupervised methods for learning relation encoders**. While our method has demonstrated progress over existing methods on the aggregated relation extraction results over many relation types, we notice that some common relation types (such as the *antagonist* relation or the *per:title* relation) have seen better performance than other relations where supervised data tends to be sparse and harder to collect. More generally speaking, the difficulty of collecting data for relation types in the long tail has substantially limited the performance of supervised models (Zhang et al., 2017) and has been a long-lasting challenge. It is therefore important to study methods that combine unsupervised learning with state-of-the-art neural architectures for the relation extraction task. Baldini Soares et al. (2019) have pioneered this direction by exploring contrastive methods for learning relation encoders based on the transformer architecture.

- **Combining pretraining with linguistic structures**. We have shown in our study that by exploiting the syntactic structure of a sentence, our model can achieve better performance on longer range relations. While existing studies with pretrained LMs have largely relied on an end-to-end transformer architecture with self-attention heads, it remains an interesting direction to study whether these pretrained models can still benefit from the integration of syntactic structures. Some studies have pioneered this direction since the publication of our work: Sachan et al. (2020) have explored the integration of syntactic attention heads in pretrained transformers for relation extraction; Xu et al. (2020) have shown that more benefits can be obtained from syntactic structures if the structures are used in the pretraining process.

- **Exploiting and combining with state-of-the-art question answering systems**. End-to-end neural question answering (QA) systems (also referred to as reading comprehension systems) have demonstrated outstanding results in recent years and provide alternative means for obtaining knowledge from unstructured text (Chen et al., 2017; Devlin et al., 2019). It remains an open question how we can better combine the power of these QA systems with traditional relation extraction and KB systems to improve knowledge acquisition from biomedical text. On the one hand, it is possible to convert the relation extraction problem into a QA problem (Levy et al., 2017) and exploit the power of existing end-to-end QA systems for extracting relational knowledge. On the other hand, can we combine the flexibility of a QA system with the reasoning capability and interpretability of relation extraction and KB systems for more robust information acquisition?

While understanding biomedical scientific text is a highly important task in medicine that helps advance our knowledge, scientific text is not the only form of text in medicine. In the next two chapters, we will shift our focus to the understanding of a different text genre in medicine, namely clinical report text, which is commonly used by healthcare providers to document the diagnosis and treatment of diseases, as well as healthcare outcomes. We will also identify a unique opportunity which natural language understanding offers in helping with the generation of this important form of text in medicine.

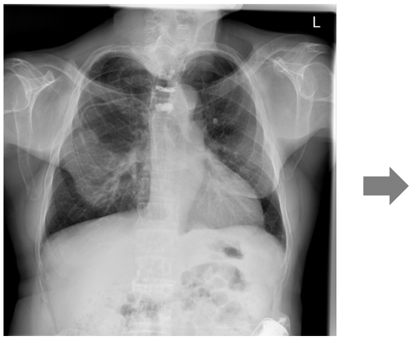
Chapter 4

Summarizing Medical Reports as Text Generation

In the previous chapter, we focused on the understanding of biomedical scientific text, and presented a neural network-based model that identifies the crucial relations between core entities such as proteins and drugs. While biomedical scientific text is a crucial medium that stores and communicates medical knowledge, an equally important type of text in medicine is the *clinical report*, one element of the textual content in the *medical record* (Spooner and Pesaturo, 2013), which medical practitioners use to document or communicate the conditions of a patient or the findings of a clinical exam.

Clinical reports come in various forms depending on their purpose or the type of exam that they aim to document. In addition, reports can have varying degrees of structure. For example, some clinical reports, such as a blood test report, or a pathology report, may have embedded tables filled with numerical values or clinical status, making them structured or semi-structured. Meanwhile, a radiology report, a common type of clinical report used to document and communicate crucial findings in a medical imaging study (such as an X-ray or CT exam), is largely unstructured and encodes information in a free-text format (Kahn Jr et al., 2009).

Figure 4.1 presents an example chest radiographic image along with its corresponding radiology report. As shown in the figure, a standard radiology report usually consists of several sections that are written with free text: 1) a *Background* section, which describes



Background:
Comparison: None. Indication: Preprocedure evaluation prior to bone marrow transplant.

Findings:
The lungs appear clear.The heart and pulmonary XXXX appear normal.There is severe kyphotic deformity of the chest involving prior fractures of thoracic vertebral bodies and the sternum.There are multiple XXXX fractures identified involving upper thoracic vertebral bodies and a single upper lumbar vertebral body.The patient is status post vertebroplasty at multiple levels.The pleural spaces appear clear. There is right-sided chest XXXX, the distal tip in the upper right atrium. Mediastinal contours appear normal.

Impression:
No evidence of acute cardiopulmonary disease. Changes of acute kyphotic deformity and of the thorax as described above.

Figure 4.1: An example chest radiographic image and the corresponding radiology report. Both the image and the report is drawn from the Indiana University Chest X-ray Dataset (Demner-Fushman et al., 2016). *XXXX* represents words that are de-identified and removed from the original clinical report.

the exam and patient information, such as the purpose of the exam or the patient's illness history, etc.; 2) a *Findings* section, which describes the imaging features and clinical observations of various anatomies or body parts in detail; and 3) an *Impression* section, which is a summary of the overall clinical findings and conclusions, along with possible recommendations (Kahn Jr et al., 2009). In a typical radiology workflow, a radiologist first drafts the detailed findings into the report, and then summarizes the salient findings into the more concise Impression section based also on the condition of the patient. The radiology report is then delivered to the referring physician or the patient to communicate the results.

The Impression is the most significant part of a radiology report. This is not only because that it summarizes the radiology findings with the most concise language (see Figure 4.1), but also because previous studies have shown that over 50% of referring physicians read only the impression statements in a report (Lafortune et al., 1988; Bosmans et al., 2011). This has made it especially important to compose the Impression section in a clear, concise and accurate fashion.

Despite its importance, the generation of the impression statement is often error-prone in clinical practice. For example, crucial findings may be forgotten by the radiologist and

Background: history: swelling; pain. technique: 3 views of the left ankle were acquired. comparison: no prior study available.

Findings: there is normal mineralization and alignment. no fracture or osseous lesion is identified. the ankle mortise and hindfoot joint spaces are maintained. there is no joint effusion. the soft tissues are normal.

Human Impression:
normal left ankle radiographs.

Extractive Baseline:
there is no joint effusion.

Pointer-Generator:
normal right ankle.

Our model:
normal radiographs of the left ankle.

Figure 4.2: An example radiology report with impression statements from a human and different systems. The report body contains study background information in a Background Section, and radiology findings in a Findings Section. The human-written summary (or impression) and predicted summaries from different models are also shown. The extractive summarization baseline system does not summarize well, the baseline pointer-generator model generates a spurious sequence, while our model gives a correct summary by incorporating the background information.

therefore be omitted from the impression statements, which would cause significant miscommunications (Gershanik et al., 2011) and medical errors. Additionally, the process of writing the impression statements is time-consuming and highly repetitive with the dictation of the Findings section. Saving the radiologists from this repetitive work means that they can focus their attention on the interpretation of the images and the diagnosis of diseases, the most critical part of their work. These suggest a crucial need to develop systems that automate the radiology impression generation process.

For these reasons, in this chapter, we study the automated generation of radiology impressions from textual findings descriptions with natural language generation techniques.

In particular, we argue that this task could be viewed as a *text summarization* problem, where the source text sequence is the radiology findings and the target text sequence the impression statements. We collect a dataset of radiology reports from actual hospital radiographic studies, and find that this task involves both *extractive* summarization where some descriptions of radiology observations can be taken directly from the findings, and *abstractive* summarization where new words and phrases, such as conclusions of the study, need to be generated from scratch.

We empirically evaluate existing popular summarization systems on this task, including both extractive and abstractive systems (Gambhir and Gupta, 2017), especially systems that are based on neural sequence-to-sequence learning (Sutskever et al., 2014). We find that, while existing neural models such as the pointer-generator network (See et al., 2017) can generate plausible summaries, they sometimes fail to model the study background information and thus generate spurious results, as shown by the example in Figure 4.2. To address this problem, we propose a summarization model that is tailored for the structure of a radiology report, can properly encode the study background information, and use the encoded information to guide the decoding process.

We show that our model outperforms existing non-neural and neural baselines on our dataset measured by the standard ROUGE metrics (Lin, 2004) designed for text summarization. Moreover, in a blind experiment, a board-certified radiologist indicated that 67% of sampled system summaries are at least as good as the reference summaries written by well-trained radiologists, suggesting significant clinical validity of the resulting system. We further show through detailed analysis that our model could be transferred to radiology reports from another organization, and that the model can sometimes summarize radiology studies for body parts unseen during training.

To summarize, this chapter makes the following contributions:

1) We propose to automate the generation of impression statements in clinical radiology reports via summarizing the free-text radiology findings with neural sequence-to-sequence learning;

2) We propose a new customized summarization model for this task that improves over existing methods by better leveraging study background information;

3) We show on a real-world radiology report dataset collected from a hospital that our proposed model outperforms existing models on standard summarization metrics;

4) We further show via a radiologist evaluation that the summaries generated by our model have significant clinical validity;

5) We show via analysis that our model presents cross-institutional and cross-body part transferability as measured by standard text generation metrics, and identify common mistakes output by the model.

This chapter is organized as follows. In Section 4.1 we start by providing a formal definition of the task of summarizing radiology reports. Next, in Section 4.2 we describe our proposed neural model for this task. We then describe in Section 4.3 how we collect a real-world radiology report dataset, and compare our system against baseline summarization models on this dataset. We show our detailed experimental results in Section 4.4, and provide further analysis on the model output and the transferability of our model in Section 4.5. Lastly, we provide a summary of our study in Section 4.6 and highlight several important directions for future work.

4.1 Task Definition

We define the task of summarizing radiology findings as follows. Given a passage of findings represented as a sequence of tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, with N being the length of the findings, our goal is to find a sequence of tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$ that best summarizes the salient and clinically significant findings in $\mathbf{x}$, with L being an arbitrary length of the summary.[1] Note that the mapping between $\mathbf{x}$ and $\mathbf{y}$ can either be modeled in an unsupervised way (as done in unsupervised summarization systems), or be learned from a dataset of findings-summary pairs.

[1] While the name “impression” is often used in clinical settings, we use “summary” and “impression” interchangeably.

4.2 Models

In this section we introduce our model for the task of summarizing radiology findings. As our model builds on top of existing work on neural sequence-to-sequence learning and the pointer-generator model, we start by introducing them.

4.2.1 Neural Sequence-to-Sequence Model

At a high-level, our model implements the summarization task with an encoder-decoder architecture, where the encoder learns hidden state representations of the input, and the decoder decodes the input representations into an output sequence.

For the encoder, we use a Bi-directional Long Short-Term Memory (Bi-LSTM) network. Given the findings sequence $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, we encode $\mathbf{x}$ into hidden state vectors with:

$$\mathbf{h} = \text{Bi-LSTM}(\mathbf{x}), \tag{4.1}$$

where $\mathbf{h} = \{h_1, h_2, \ldots, h_N\}$. Here h_N combines the last hidden states from both directions in the encoder.

After the entire input sequence is encoded, we generate the output sequence step by step with a separate LSTM decoder. Formally, at the t-th step, given the previously generated token y_{t-1} and the previous decoder state s_{t-1}, the decoder calculates the current state s_t with:

$$s_t = \text{LSTM}(s_{t-1}, y_{t-1}). \tag{4.2}$$

For the initial decoder state we set $s_0 = h_N$. We then use s_t to predict the output word by applying a linear layer to s_t followed by a softmax layer over the output vocabulary.

The vanilla sequence-to-sequence model that uses only s_t to predict the output word has a major limitation: it generates the entire output sequence based solely on a vector representation of the input (i.e., h_N), which may result in significant information loss. For better decoding we therefore employ the attention mechanism (Bahdanau et al., 2015; Luong et al., 2015), which uses a weighted sum of all input states at every decoding step.

Given the decoder state s_t and an input hidden state h_i, we calculate an input distribution

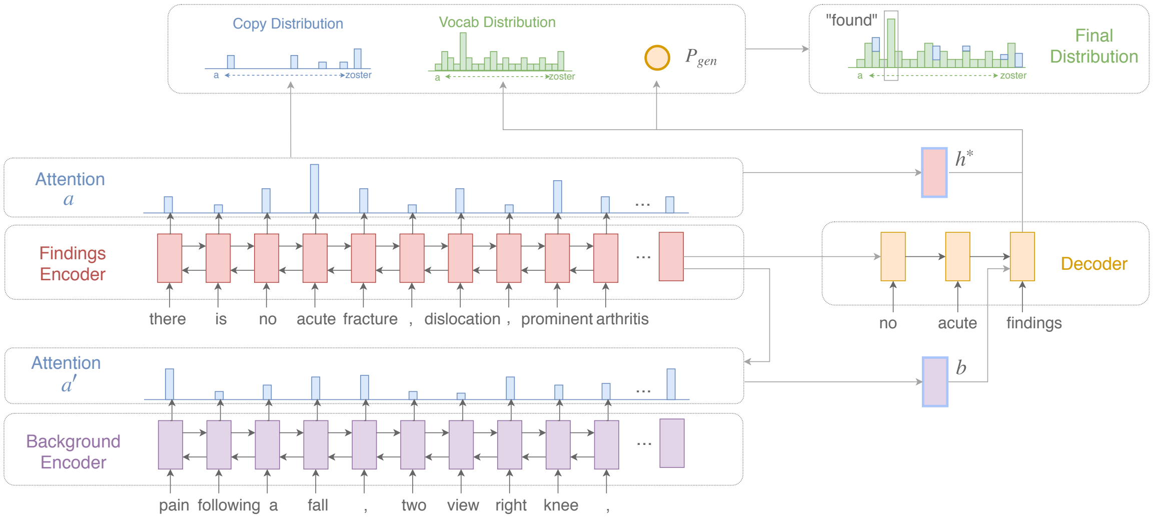


Figure 4.3: Overall architecture of our summarization model.

a^t as:

$$e_i^t = v^\top \tanh(W_h h_i + W_s s_t), \tag{4.3}$$

$$a^t = \mathrm{softmax}(e^t), \tag{4.4}$$

where W_h, W_s and v are learnable parameters.[2] We then calculate a weighted input vector as:

$$h_t^* = \sum_i a_i^t h_i. \tag{4.5}$$

h_t^* encodes the salient input information that is useful at decoding step t. Lastly, we obtain the output vocabulary distribution at step t as:

$$P(y_t|\mathbf{x}, y_{<t}) = \mathrm{softmax}(V' \tanh(V[s_t; h_t^*])), \tag{4.6}$$

where V' and V are learnable parameters.

[2] For clarity we leave out the bias terms in all linear layers.

4.2.2 Pointer-Generator Network

While the encoder-decoder framework described above can generate impressions from a fixed vocabulary, the model can clearly benefit from being able to "copy" salient observations directly from the input findings. To add such "copying" capacity into the model, we use a pointer-generator network similar to the one described in See et al. (2017).

The main idea is that at each decoding step t, we allow the model to either generate a word from the vocabulary with a generation probability p_{gen}, or copy a word directly from the input sequence with probability $1 - p_{\text{gen}}$. We model p_{gen} as:

$$p_{\text{gen}} = \sigma(w_{h^*}^{\top} h_t^* + w_s^{\top} s_t + w_y y_{t-1}), \tag{4.7}$$

where y_{t-1} denotes the previous decoder output, w_{h^*}, w_s and w_y learnable parameters and σ a sigmoid function. For the copy distribution, we reuse the attention distribution a^t calculated in (4.4). Therefore, the overall output distribution in the pointer-generator network is:

$$P(y_t|\mathbf{x}, y_{<t}) = p_{\text{gen}} P_{\text{vocab}}(y_t) + (1 - p_{\text{gen}}) \sum_{i:x_i=y_t} a_i^t, \tag{4.8}$$

where $P_{\text{vocab}}(y_t)$ is the same as the output distribution in (4.6).

4.2.3 Incorporating Study Background Information

The background part of a radiology report is also important, since crucial information such as the purpose of the study, the body part involved and the condition of the patient are often mentioned only in the Background section. A straightforward way of incorporating the background information is to prepend all the background text to the findings, and treat the entire sequence as input to the pointer-generator network. However, as we will show in Section 4.4, this naive method in fact hurts the summarization quality, presumably because the model cannot sufficiently distinguish between the findings and the background information, which as a result leads to insufficient modeling of both the findings and the background. To solve this, we propose to encode the background text with a separate attentional encoder, and use the resulting background representation to guide the decoding process in the summarization model (Figure 4.3).

To differentiate the background part of a report from the actual findings section as shown in Figure 4.2, we now use $\mathbf{x}^b$ to denote the background token sequence, and $\mathbf{x}$ to denote the actual findings section. Our goal is then to find $\mathbf{y}$ that maximizes $P(\mathbf{y}|\mathbf{x}, \mathbf{x}^b)$. To do this, we again obtain the hidden state vectors $\mathbf{h}$ of the findings section as in (4.1). Similarly, we obtain the hidden state vectors of the background text with $\mathbf{x}^b$ as input using a separate Bi-LSTM encoder:

$$\mathbf{h}^b = \text{Bi-LSTM}^b(\mathbf{x}^b). \tag{4.9}$$

Next, we calculate a distribution over $\mathbf{h}^b$ as:

$$e'_i = v'^{\top} \tanh(W_b h_i^b + W_h h_N), \tag{4.10}$$

$$a' = \text{softmax}(e'), \tag{4.11}$$

where W_b and W_h are learnable parameters and h_N the last hidden state of the findings encoder. The distribution a' models the importance of tokens in the background section. We then obtain a weighted representation of the background text as:

$$b = \sum_i a'_i h_i^b, \tag{4.12}$$

where vector b has the same size as h^b, and encodes the salient background information.

Lastly, we use the background vector b to guide the decoding process, by modifying the recurrent kernel of the decoder LSTM in (4.2) to be:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ u_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot \begin{bmatrix} s_{t-1} \\ y_{t-1} \\ b \end{bmatrix}, \tag{4.13}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t, \tag{4.14}$$

$$s_t = o_t \odot \tanh(c_t), \tag{4.15}$$

where i_t, f_t, o_t denotes the input, forget, and output gates, W the weight matrix and c_t the internal cell of the LSTM respectively, and $\odot$ represents an element-wise multiplication. Again for clarity we leave out the bias terms in (4.13). As a result, every state in the decoding process is directly influenced by the information encoded by the background vector b. The rest of the model, including the calculation of the vocabulary distribution and the copy distribution, remains the same.

4.3 Experiments

To test the effectiveness of our summarization model, we collected reports of radiographic studies from the picture archiving and communication system (PACS) at the Stanford Health Care. In this section we describe our data collection process, baseline models and experimental setup, and present the results and discussions in Section 4.4.

4.3.1 Data Collection

We collected the reports of all radiographic studies from 2000 to 2014 from the Stanford Health Care.[3] For preprocessing, we first tokenized all reports with Stanford CoreNLP (Manning et al., 2014), and filtered the dataset by excluding reports where (1) no findings or impression section can be found; (2) multiple findings or impression sections can be found but cannot be aligned; or (3) the findings have fewer than 10 words or the impression has fewer than 2 words.

We removed body parts where only a small number of cases are available, and included reports of the top 12 body parts in the PACS system to maintain generalizability. For common body parts with more than 10k reports (e.g., chest), we subsampled 10,000 reports from them.

This results in a dataset with a total of 87,127 reports. We further randomly split the dataset into a 70% training, a 10% development and a 20% test set. We show the dataset statistics in Table 4.1 and its detailed statistics split by body part in Figure 4.4.

[3] Our retrospective study has been approved by the corresponding institutional review boards with waiver of consent.

Data Split	Number of Reports	Percentage
Train	60,990	70%
Development	8,712	10%
Test	17,425	20%
Total	87,127	–

Table 4.1: Overall statistics of the collected Stanford Health Care dataset.

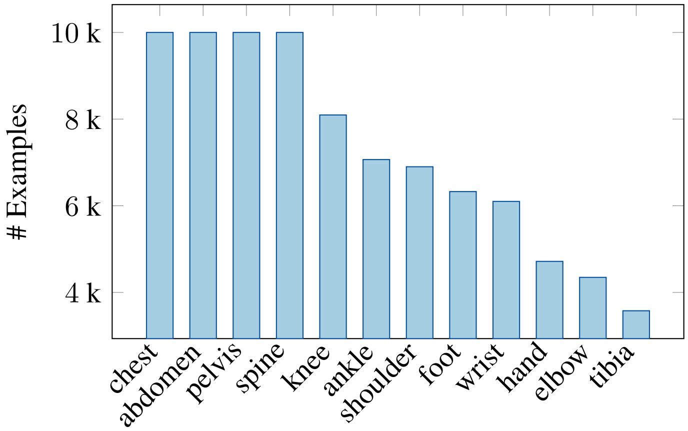


Figure 4.4: Number of examples split by body part in the collected Stanford Health Care dataset.

4.3.2 Baseline Models

For our main experiments, we compare our model against several competitive non-neural and neural systems on the collected dataset. Unless otherwise stated, the baseline models take only the findings section as input.[4] We now describe each baseline model.

S&J-LSA. This is an extractive approach described by Steinberger and Jezek (2004), which applies Latent Semantic Analysis (LSA) to text summarization. This method first takes the input radiology findings paragraph, and applies singular value decomposition (SVD) to the term-by-sentence co-occurrence matrix derived from the passage. The results

[4] We find that when the background section is prepended to the input, the extractive baseline models may select sentences in the background part as the summary, resulting in deteriorated performance.

of this procedure are “concept” (i.e., word) clusters as scored by the SVD process, with more important words assigned with higher scores. Finally, the system selects the top N sentences with the highest scored concepts to form the summary.

LexRank. LexRank is another popular extractive model introduced by Erkan and Radev (2004). In LexRank, an input findings paragraph is first represented as a graph of sentences, and a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph. Sentences are then scored by the eigenvector centrality in the graph, and the top N highest scored sentences in the input paragraph are then kept as the summary.

Pointer-Generator. We also run the baseline pointer-generator model as introduced by See et al. (2017). We find the “coverage” mechanism described in their paper did not improve summary quality in our task and therefore did not use it for simplicity. We compare our model with two versions of the pointer-generator model: one with only the findings section as input and another one with the background sections prepended to the findings section as input.

4.3.3 Experimental Setup

Evaluation Metrics. In our main experiments we evaluate the models with the widely-used ROUGE metric (Lin, 2004), as is done in previous text summarization work. We report the F_1 scores for ROUGE-1, ROUGE-2 and ROUGE-L, which measure the word-level unigram-overlap, bigram-overlap and the longest common sequence between the reference summary and the system predicted summary respectively.

Word Vectors. To enable knowledge transfer from a larger corpus, we applied the GloVe algorithm (Pennington et al., 2014) to a corpus of 4.5 million radiology reports of all modalities (e.g., X-ray, CT) and body parts. We used the resulting 100-dimensional word vectors to initialize all word embedding layers in our neural models, and empirically found this to improve the performance of our neural models by about 1 ROUGE-L score.

Implementations & Model Details. For the two non-neural extractive baselines, we use their open implementations offered by the Sumy library.[5] For both of them, we select the top N scored sentences to form the summary and treat N as a hyperparameter. We use $N = 3$ in our experiments as it yields best scores on the dev set of our collected dataset. We implemented all neural models with PyTorch.[6] To train the neural models we append a special `<EOS>` token to the end of every reference summary. We then employ the standard teacher-forcing with the reference summaries and optimize the negative log-likelihood loss using the Adam optimizer (Kingma and Ba, 2015). We tune all hyperparameters on the dev set. We use 2-layer Bi-LSTM for all encoders, and set the hidden size to be 100 for each direction; 1-layer LSTM for the decoder and set the hidden size to be 200. During inference, we employ the standard beam search with a beam size of 5. We stop decoding whenever a `<EOS>` token is predicted, and otherwise use a maximum output sequence length of 100.

4.4 Results

In this section we present our experimental results on the collected Stanford radiology reports dataset. We first present our main results with automatic evaluation metrics, as long as some qualitative examples from our system. We then present results from a radiologist evaluation where we evaluate the generated summaries based on their clinical correctness and validity, and our findings from this experiment.

4.4.1 Automatic Evaluation Results

We present results of our main experiments in Table 4.2. We find that the two non-neural extractive models perform comparably, and both are able to obtain non-trivial subsequence overlap with the reference summaries as measured by ROUGE scores. However, a baseline neural pointer-generator that combines the sequence generation and the copy mechanism beats the non-neural baselines substantially on all metrics. We confirm that naively incorporating the study background information by prepending the background section directly

[5] https://github.com/miso-belica/sumy

[6] https://pytorch.org/

System	ROUGE-1	ROUGE-2	ROUGE-L
Extractive Baseline: S&J-LSA	29.39	16.27	27.38
Extractive Baseline: LexRank	30.48	17.09	28.49
Pointer-Generator	46.51	33.39	45.07
Pointer-Generator ($\oplus$ Background)	45.39	32.60	44.05
Our model	**48.56**	**35.25**	**47.06**

Table 4.2: Automatic evaluation results on the test set of the Stanford reports. "$\oplus$ Background" represents prepending the background section to the findings section to form the input to the model. All the ROUGE scores have a 95% confidence interval of at most ± 0.50 as calculated by the official ROUGE script.

to the input findings in the pointer-generator model in fact hurts the performance (noted by $\oplus$ Background). In comparison, our model benefits from using the separately encoded background vector to guide the decoding process, and achieves best scores on all ROUGE metrics.

We also present sampled test examples and system output in Figure 4.5. We find that compared to the non-neural extractive baselines, the neural models are not limited by sentences in the findings section and therefore generate summaries of better quality. For example, the neural models learn to compose the summary by combining observation phrases from different sentences, or by generating new conclusive phrases such as "negative study". Compared to the pointer-generator model, our model learns to correctly utilize relevant background information (e.g., previous study or exam information) to improve the summary.

4.4.2 Understanding Clinical Validity with Radiologist Evaluation

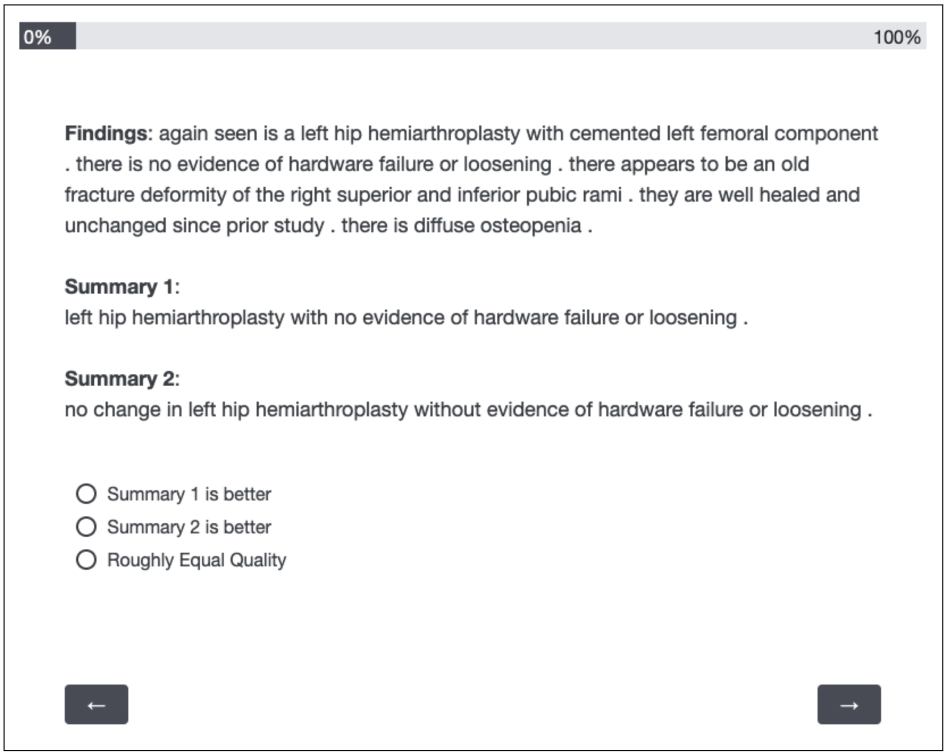
A limitation of using the ROUGE metrics to evaluate the systems is that they only measure the similarity between the predicted summary and the reference summary, but do not sufficiently reflect the overall grammaticality or utility of the predictions. Therefore, to understand the clinical validity of our system generated summaries, we also conducted evaluations with a board-certified radiologist.

In this evaluation, we randomly sampled 100 examples from our test set. We ran our

Background: radiographic examination of the abdomen. clinical history: xx years of age, male, please obtain upright and lateral decub. comparison: abdominal x-ray <date>. procedure comments: two views of the abdomen. **Findings:** median sternotomy wires are seen in the anterior chest wall in addition to several mediastinal clips and an aicd. trace bilateral pleural effusions are noted. interval increase in small bowel dilatation compared to previous study with multiple air-fluid levels, consistent with small bowel obstruction. there is a paucity of colonic gas. no pneumoperitoneum.	**Background:** three views of the right shoulder and three views of the left shoulder: <date>. clinical history: an xx-year-old female with bilateral shoulder pain. **Findings:** three views of the right shoulder consisting of external rotation, axillary, and scapular views demonstrate no evidence of fracture or dislocation. the joint spaces are well-maintained without evidence of degenerative change. there is normal mineralization throughout. three views of the left shoulder . . . are well-maintained without evidence of degenerative change. mineralization is normal throughout.	**Background:** three views of the abdomen: <date>. comparison: <date>. clinical history: a xx-year-old male status post hirschsprung's disease repair. **Findings:** the supine, left-sided decubitus and erect two views of the abdomen show increased dilatation of the small bowel since the prior exam on <date>. there are multiple air-fluid levels, suggesting bowel obstruction. no free intraperitoneal gas is present.
Human: small bowel dilatation with multiple air-fluid levels and colonic decompression consistent with small bowel obstruction.	**Human:** unremarkable radiographs of bilateral shoulders.	**Human:** increased dilatation of the small bowel with multiple air-fluid levels, suggesting bowel obstruction. no free intraperitoneal gas.
Extractive Baseline: median sternotomy wires are seen in the anterior chest wall in addition to several mediastinal clips and an aicd.	**Extractive Baseline:** three views of the right shoulder consisting of external rotation, axillary, and scapular views demonstrate no evidence of fracture or dislocation.	**Extractive Baseline:** the supine, left sided decubitus and erect two views of the abdomen show increased dilatation of the small bowel since the prior exam on <data>.
Pointer-Generator: interval increase in bowel dilatation, consistent with bowel obstruction.	**Pointer-Generator:** no evidence of fracture or dislocation of the right shoulder.	**Pointer-Generator:** increased dilatation of small bowel, suggesting small bowel obstruction.
Our model: interval increase in small bowel dilatation compared to abdominal x-ray dated <date> with multiple air-fluid levels, consistent with small bowel obstruction.	**Our model:** unremarkable bilateral shoulders.	**Our model:** increased dilatation of small bowel, suggesting bowel obstruction. no free intraperitoneal gas.

Figure 4.5: Sampled test examples and system predictions from the Stanford Health Care dataset. First example: our model learns to relate the summary with a previous study mentioned only in the background section. Second: our model correctly summarizes the body part involved in the study. Third: our model correctly includes more crucial information as found in the human summary.

best model over these 100 examples, and presented each example along with the corresponding system predicted summary and reference human-written summary to the radiologist. We randomly ordered the predicted and reference summary such that the correspondence cannot be guessed from the order. The radiologist was then asked to select which of the two summaries was better, or that they have roughly equal quality. Figure 4.6 shows an example annotation interface presented to the radiologist.



Findings: again seen is a left hip hemiarthroplasty with cemented left femoral component . there is no evidence of hardware failure or loosening . there appears to be an old fracture deformity of the right superior and inferior pubic rami . they are well healed and unchanged since prior study . there is diffuse osteopenia .

Summary 1:
left hip hemiarthroplasty with no evidence of hardware failure or loosening .

Summary 2:
no change in left hip hemiarthroplasty without evidence of hardware failure or loosening .

○ Summary 1 is better
○ Summary 2 is better
○ Roughly Equal Quality

Figure 4.6: An example annotation interface shown to the radiologist.

Category	Percentage
Human Summary Wins	33
System Prediction Wins	16
Roughly Equal Quality	51

Table 4.3: Radiologist evaluation result on 100 sampled test examples from the Stanford Health Care dataset. For a total of 67 examples, the radiologist indicated that the system summary is at least as good as the human-written summary.

Table 4.3 presents the result of radiologist evaluation. For 51 examples, the radiologist indicated that the human-written and system-generated summaries are equivalent. For 16 examples, the radiologist preferred the system summary, and for the remaining 33 examples, the radiologist preferred the human-written summary. Overall, for a total of 67% test examples the radiologist indicated that the system summary is at least as good as the human-written summary. Note that under our setting, a randomly generated sequence

System	ROUGE-1	ROUGE-2	ROUGE-L
LexRank	15.42	5.65	14.60
Our model	35.02	20.79	34.56

Table 4.4: Cross-organization evaluation results on the Indiana University chest x-ray dataset. All the ROUGE scores have a 95% confidence interval of at most ± 1.10 as calculated by the official ROUGE script.

would have almost zero chance to be indicated as good as the human-written summary. We therefore believe the result suggests substantial clinical validity of our system.

4.5 Analysis

In this section, we analyze the performance of our model under different settings, with the focus on answering three important questions: 1) How well can the model generalize to reports from another organization? 2) How well can the model generalize to body parts unseen during training? 3) What types of common mistakes does the model make?

4.5.1 Generalizability to Different Organizations

Deploying a clinical NLP system at an organization different from the one where the training data comes from is a common need. However, this is challenging in that medical practitioners including radiologists from different organizations tend to go through different training and follow different templates or styles when writing medical text. Here we aim to understand the cross-organization generalizability of our summarization model.

For this evaluation setting, we use the publicly available Indiana University Chest X-ray Dataset (Demner-Fushman et al., 2016), which consists of chest X-ray images paired with the corresponding radiology reports. We filtered the reports with the same set of rules and arrived at a collection of 2,691 unique reports. We used this dataset as the test set, and ran our best model trained on our own dataset directly on it. The results are shown in Table 4.4 and sampled examples are shown in the first two columns of Figure 4.7. We find that our model again outperforms the baseline extractive model substantially in this transfer setting,

Cross-organization	Cross-organization	Cross-body part: Knee
Background: indication: xxxx year old male with end-stage renal disease on hemodialysis **Findings:** the heart size is mildly enlarged. there is tortuosity of the thoracic aorta. no focal airspace consolidation, pleural effusions or pneumothorax. no acute bony abnormalities.	**Background:** indication: xxxx year old female, hypoxia. comparison: pa lateral views of the chest dated xxxx. **Findings:** bilateral emphysematous again noted and lower lobe fibrotic changes. postsurgical changes of the chest including cabg procedure, stable. stable valve artifact. there are no focal areas of consolidation. no large pleural effusions. no evidence of pneumothorax. . . . contour abnormality of the posterior aspect of the right 7th rib again noted, stable.	**Background:** radiographic examination of the knee: <date> <time>. clinical history: xx-year-old man with right knee pain. comparison: none. procedure comments: 2 views of the right knee were performed. **Findings:** there is no visible fracture or malalignment. likely small joint effusion. mild fullness in the popliteal region of the right knee may represent a baker ’s cyst. mild soft tissue swelling along the medial aspect of the knee is present.
Human: cardiomegaly without acute pulmonary findings.	**Human:** no acute cardiopulmonary abnormality. stable bilateral emphysematous and lower lobe fibrotic changes.	**Human:** no acute bony abnormality. likely joint effusion and soft tissue swelling along the medial aspect of the knee.
Our model: mild cardiomegaly. no radiographic evidence of acute cardiopulmonary process.	**Our model:** stable postsurgical changes of the chest as described above. no evidence of pneumothorax.	**Our model:** mild soft tissue swelling along the medial aspect of the knee. no fracture or malalignment.

Figure 4.7: Examples that demonstrate generalization to notes from different organizations and to unseen body parts. First two columns: sampled examples from the Indiana University dataset and system output in the cross-organization evaluation. Last column: sampled test example of a “knee” study in our cross-body part evaluation.

and the generated summaries are both grammatical and clinically meaningful.

4.5.2 Generalizability to Unseen Body Parts

Radiology studies conducted on different body parts often include vastly different observations and diagnosis. For example, while “lung base opacity” is a common observation in chest radiographic studies, it does not exist in musculoskeletal studies. In practice, an organization may not have adequate report data that covers some rare body parts. It is therefore interesting to test to what extent our summarization model can generalize to reports for

Body Part	ROUGE-1	ROUGE-2	ROUGE-L
Chest	31.24	17.99	30.38
Abdomen	28.90	17.23	27.83
Knee	48.78	35.07	47.49

Table 4.5: Cross-body part evaluation results of our neural model on the Stanford Health Care dataset. All the ROUGE scores have a 95% confidence interval of at most ± 0.75 as calculated by the official ROUGE script.

body parts unseen during training.

We study this by simulating the condition where a specific body part is not present in the training data. Given the entire dataset $\mathcal{D}$, and a subset of the dataset $\mathcal{D}_B$ that corresponds to a body part B, we reserved the entire subset $\mathcal{D}_B$ as test data, and used $\mathcal{D} - \mathcal{D}_B$ for training (90%) and validation (10%). Table 4.5 presents the evaluation results for body part "chest", "abdomen" and "knee". We find that for "chest" and "abdomen", the system summaries degrade substantially when the corresponding data were not seen during training. However, the predicted summaries degrade less for "knee" when reports of it were not seen during training, presumably because the model can learn to summarize reasonably well from reports of other close musculoskeletal studies such as "ankle" or "elbow" studies. We confirm this by examining the model predictions: in the example shown in the last column of Figure 4.7, the model learns to compose the summary with salient observations such as "tissue swelling" and "fracture", while being able to copy the anatomy "knee" (unseen during training) from the findings section.

4.5.3 Error Analysis

Lastly, to understand the common types of errors that our model makes, we run a detailed error analysis on 100 sampled dev examples. We focus on four types of errors in our analysis: (1) missing critical information, if the predicted summary fails to include some clinically important information; (2) inaccurate/spurious information, if the predicted summary contains observations or conclusions that are inaccurate, or that do not exist in the findings;

Category	Percentage
Good Summary	63
Missing Critical Information	24
Inaccurate/Spurious Information	8
Redundant	4
Ungrammatical	6

Table 4.6: Error analysis on 100 sampled dev examples from the Stanford Health Care dataset.

(3) redundant summary, if the predicted summary is repetitive or over-verbose; and (4) ungrammatical summary, if the predicted summary contains significant grammatical errors. For each example, we examine whether it contains any of the errors by comparing it with the reference summary; otherwise we classify it as a good summary. Note that under our analysis setting, an example can be assigned to more than one error category.

We present the result of our error analysis in Table 4.6, and include examples of different error types in Figure 4.8. We find that 63% of examples are qualitatively close to the reference summary, which aligns well with the radiologist evaluation result. Among the four error categories, missing critical information is the most common error with 24% of examples, suggesting that the summaries may be improved with explicit modeling of the importance of different radiology findings. We also find through qualitative analysis that the model tends to miss on followup procedures recommended by the human radiologist, since these procedures are often not included in the findings section and generating them needs significant understanding of the study and domain knowledge. This highlights the importance of incorporating domain knowledge for the further improvement of these summarization systems.

4.6 Summary & Future Directions

In this chapter, we studied the automated generation of radiology summaries (i.e. impression statements) from the textual radiology findings written by radiologists. We proposed to model this process via neural sequence-to-sequence learning, and designed a customized

Error type: missing critical information	Error type: redundant summary	Error type: ungrammatical summary
Background: radiographic examination of the lumbar spine: <time>. clinical history: <age>, lower back pain. comparison: none. procedure comments: 4 views of the lumbar spine. **Findings:** five non-rib bearing lumbar type vertebral bodies are present. there is trace retrolisthesis of l5 on s1. there is no evidence of instability on flexion and extension views. the spinal alignment is otherwise normal. the disc spaces and vertebral body heights are preserved. there is no visible fracture. no visible facet joint arthropathy or pars defects.	**Background:** radiographic examination of the shoulder: <time>. clinical history: <age> years of age, pain in joint involving shoulder region. comparison: outside study dated <date>. procedure comments: single axillary view of the left shoulder. **Findings:** single axillary view of the shoulder again demonstrates a highly comminuted fracture of the humeral head and likely fracture of the scapular body. the humeral head appears located on the glenoid.	**Background:** radiographic examination of the shoulder: <time>. clinical history: <age> years of age, xray exam of lower spine 2 or 3 views. x-ray exam of right shoulder complete. comparison: none. procedure comments: three views of the right shoulder. **Findings:** a calcification of the rotator cuff is seen above the greater tuberosity. there is no fracture or malalignment. the soft tissues and visualized lung are unremarkable.
Human: trace retrolisthesis of l5 on s1 with no evidence of instability with motion. otherwise normal lumbar spine.	**Human:** redemonstration of a highly comminuted fracture of the humeral head and likely fracture of the scapular body . the humeral head appears to be located on the glenoid .	**Human:** no acute bony or joint abnormality, but there is calcification of the rotator cuff that may be due to calcific tendinitis.
Our model: no acute bony or articular abnormality.	**Our model:** highly comminuted fracture of the scapular body and likely fracture of the scapular body.	**Our model:** calcification acute bony or joint abnormality.

Figure 4.8: Examples of different types of errors that our system makes on the Standord Hospital dataset. Words that are missing from or are erroneously included in the model predictions are highlighted.

neural model for this task which is based on the pointer-generator network and additionally uses encoded background information to guide the decoding process. We collected a dataset from actual hospital studies and showed that our model not only outperforms non-neural and neural baselines, but also generates summaries with substantial clinical validity and cross-organization generalizability. We further found via error analysis that 63% of the generated summaries are of high quality, and that missing critical information and including inaccurate or spurious information are the most common types of errors in the generated summaries.

Our findings shed light on the following important directions for further improving the practical usability of radiology report summarization systems:

1) **Improving the factual correctness and consistency of the generated summaries**. While factual errors or inconsistencies are more tolerable in general-domain summarization tasks (e.g., news summarization), they can result in significant consequences in the summarization of medical documents, such as diagnostic errors or miscommunications of results. For this reason, improving the factual correctness of neural summarization models is especially critical for the medical domain. Furthermore, our findings suggest that improvements need to be made on both the **metrics** and **model architectures**: summarization metrics need to go beyond measuring simple textual overlap and need to either rely on deeper semantic match between the text, or involve explicit comparisons of the factual content; summarization models need to be explicitly optimized for factual consistency in their outputs.

2) **Exploring model architectures that integrate domain knowledge into the summarization process**. As our results and examples have revealed, current summarization models often make mistakes that demonstrate a lack of necessary domain knowledge. For example, the model may generate contradictory claims in a single summary paragraph that are easily detectable or avoidable by modeling relevant medical knowledge. In other cases, the model may not be able to generate a "recommended follow-up" statement that is often present in radiologist-written summaries. To solve this, future work may explore customized neural architectures that either integrate relevant domain ontologies such as the UMLS (Bodenreider, 2004) or the RadLex ontology (Langlotz, 2006), or involve explicit representation and modeling of relational medical knowledge in the summarization process.

3) **Improving the generalizability and robustness of such systems on radiology reports of different styles and domains**. This is critical for the practical usability of such systems, as in practice these systems are often trained on data from a specific organization and deployed on others. In such a cross-organization setting, a

model needs to be robust enough to text of different forms, styles or even domains. We have shown in our experiments that existing models can sometimes experience a degradation of performance as measured by ROUGE scores when transferred to data from a different organization. This often means degradation of grammaticality and correctness in the output, and may translate to medical errors. Viable methods to improve the generalizability include training the summarization system on data that is aggregated from multiple organizations and is representative of different genres and domains, or utilizing encoder or decoder models that are pretrained unsupervisedly on a much larger scale of medical text.

In the next chapter, we will extend our work in the first direction described above, by studying the problem of improving factual correctness of our radiology summarization system. We will present a framework to measure the factual correctness of a generated summary against its reference, and a reinforcement learning-based strategy to optimize our summarization model for its correctness. We will leave integrating domain knowledge and improving the generalizability of existing systems as future work.

Chapter 5

Towards Factually Correct Summarization

In the previous chapter, we studied the problem of automated generation of clinical radiology impression (i.e., summary) statements, which summarize the most important clinical findings in a radiology study in a clear and concise manner. In particular, we have formalized the task as a text summarization problem, where the source text sequence is the free-text radiology findings written by radiologists, and the target text sequence the summary statements. We showed that a neural abstractive summarization model tailored for the structure of a radiology report is able to generate radiology summaries that are fluent and achieve high overlap with summaries written by radiologists.

While this neural model demonstrates promising performance and could save radiologists from the repetitive work of writing the radiology summaries, we also showed via our blind radiologist evaluation that about 33% of total summaries generated by this model have quality lower than the human-written ones. Our error analysis further revealed that incorrectness is the most frequent type of mistake: about 30% of the outputs from this model contain factual errors or inconsistencies when compared with the human-written summaries. This has made such a system unusable in practice, as factual correctness is critically important in this domain to prevent medical errors.

Why do systems make factual mistakes? We identify that a core issue behind the factually incorrect generations is that neural summarization models such as the one we have

Background: radiographic examination of the chest. clinical history: 80 years of age, male ...

Findings: frontal radiograph of the chest demonstrates repositioning of the right atrial lead possibly into the ivc. [...] a right apical pneumothorax can be seen from the image. moderate right and small left pleural effusions continue. no pulmonary edema is observed. heart size is upper limits of normal.

Human Summary: pneumothorax is seen. bilateral pleural effusions continue.

Summary A (ROUGE-L = 0.77):
no pneumothorax is observed. bilateral pleural effusions continue.

Summary B (ROUGE-L = 0.44):
pneumothorax is observed on radiograph. bilateral pleural effusions continue to be seen.

Figure 5.1: A (truncated) radiology report and summaries with their ROUGE-L scores. Compared to the human summary, Summary A has high textual overlap (i.e., ROUGE-L) but makes a factual error; Summary B has a lower ROUGE-L score but is factually correct.

presented have been trained and evaluated with imperfect objectives. At training time, the models are primarily optimized with the conditional language modeling objective (i.e., the likelihood of the reference, human-written summaries) (Rush et al., 2015; See et al., 2017). While this objective encourages generations that are fluent and mimic human references (Paulus et al., 2018), it does not guarantee factually correct summaries. On the other hand, at evaluation time, the models are primarily compared using textual overlap-based metrics such as the ROUGE scores (Lin, 2004). While these metrics may sometimes correlate positively with human ratings in terms of informativeness and overall quality (Novikova et al., 2017), comparing model outputs with them can be misleading and adversarially favor suboptimal generations that contain factual errors, as is demonstrated by the example in Figure 5.1. Here, while the second radiology report summary is an overall correct and higher-quality summary, it is undesirably penalized with a lower ROUGE-L score, due to its overall lower overlap with the human-written summary.

Despite the importance of factual correctness in generations, existing attempts at improving the correctness of abstractive summarization models have seen very limited success. For example, Cao et al. (2017) explored augmenting the attention mechanism of neural summarization models with factual relation triples extracted from the document with an open information extraction system. Falke et al. (2019) studied using natural language inference systems to rerank generated summaries based on their factual consistency with the input document. Kryściński et al. (2019b) proposed to verify factual consistency of generated summaries with a weakly-supervised model trained with synthetic positive and negative document-summary pairs. Despite these efforts, none of the existing work has focused explicitly on optimizing an abstractive summarization system with a correctness objective. As a result, even state-of-the-art systems trained with ample data still produce summaries with a substantial number of factual errors (Goodrich et al., 2019; Kryściński et al., 2019a; Maynez et al., 2020).

In this chapter, we extend our study on summarizing clinical radiology reports, with a focus on optimizing the factual correctness of our neural summarization systems. Moreover, we hope that our investigation can offer insights for improving the correctness of general summarization systems. In fact, the task of summarizing radiology reports has several key properties that make it ideal for studying factual correctness in summarization models. First, the clinical facts or observations present in radiology reports have less ambiguity compared to open-domain text, which as we will show allows objective comparison of facts in the generated text. Second, radiology reports involve a relatively limited space of facts, which makes automatic measurement of factual correctness in the generated text approachable. Lastly, as factual correctness is a crucial metric in this domain, improving factual correctness will directly lead to an ability to use the resulting system.

To this end, we design a framework where an external information extraction system is used to extract information in the generated summary and produce a factual accuracy score by comparing it against the human reference summary. We further develop a training strategy where we combine a factual correctness objective, a textual overlap objective and a language model objective, and jointly optimize them via reinforcement learning (RL).

Similar to the last chapter, we again evaluate our method on real-world radiology reports collected from hospitals. On two datasets of reports collected from different hospitals, we show that our training strategy substantially improves the factual correctness of the summaries generated by a competitive neural summarization system. Moreover, we observe for the first time that, even in the absence of a factual correctness objective, optimizing a textual overlap-based metric substantially improves the factual correctness of the resulting system compared to using only maximum likelihood training. We further show via human evaluation and analysis that our training strategy leads to summaries with higher overall quality and correctness and which are closer to the human-written ones.

To summarize, our main contributions in this chapter include:

1) We propose a general framework and a training strategy for improving the factual correctness of summarization models by optimizing a multi-part objective via reinforcement learning; to our knowledge, our study represents the first attempt in this direction;

2) We apply our proposed strategy to the summarization of radiology reports, and empirically show that it improves the factual correctness of the generated summaries on two real-world radiology report datasets;

3) We demonstrate via radiologist evaluation that our system is able to generate summaries with clinical validity close to human-written ones.

This chapter is organized as follows. In Section 5.1 we start by briefly reviewing the definition of our task and the baseline customized summarization model that we developed in Chapter 4. We then extend our study by describing a framework to fact-check the generated summary against its reference and a method for optimizing the correctness of our model in Section 5.2. In Section 5.3 we introduce how we collect two real-world chest radiographic report datasets, and evaluate our systems on these datasets. We then describe our experimental results in Section 5.4 and present detailed analysis of our model in Section 5.5. We close this chapter in Section 5.6 by summarizing our key takeaways and discussing the limitations of our study.

5.1 Task and Baseline Pointer-Generator Model

We briefly review the task of summarizing radiology findings, as well as the background-augmented pointer-generator summarization model, which we described in Chapter 4.

Formally, given a passage of radiology findings represented as a sequence of tokens $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, with N being the length of the findings, our summarization task involves finding a sequence of tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$ that best summarizes the salient and clinically significant findings in $\mathbf{x}$.

To model the summarization process, we use the previously described background-augmented pointer-generator network as the backbone of our method. This abstractive summarization model extends a pointer-generator (See et al., 2017) with a separate background section encoder. At a high level, this model first encodes the input sequence $\mathbf{x}$ into hidden states with a Bi-LSTM network, and then generates an output sequence $\mathbf{y}$ with a separate LSTM decoder with attention to the input sequence (Bahdanau et al., 2015).

To make this encoder-decoder network more suitable for summarizing radiology findings with multiple sections, we added two augmentations to its attentional encoder-decoder model. First, a copy mechanism (Vinyals et al., 2015; See et al., 2017) is added to copy words from the input. At each step of decoding, we calculate a generation probability, and this generation probability is used to blend the original output vocabulary distribution and a copy distribution to generate the next word. Second, we separately encode the background section (as shown in Figure 5.1), and inject the representation into the decoding process by concatenating it with the input.

5.2 Fact Checking in Summarization

Summarization models such as the one described in Section 5.1 are commonly trained with the teacher-forcing algorithm (Williams and Zipser, 1989) by maximizing the likelihood of the reference, human-written summaries. However, this training strategy results in a significant discrepancy between what the model sees during training and test time, a problem often referred to as the *exposure bias* issue (Ranzato et al., 2016). This issue has been shown to often lead to degenerate output at test time for generation tasks.

An alternative training strategy is to directly optimize standard metrics such as ROUGE scores (Lin, 2004) with RL and this was shown to improve summarization quality (Paulus et al., 2018). Nevertheless, this method still provides no guarantee that the generated summary is factually accurate and complete, since the ROUGE scores merely measure the superficial text overlap between two sequences and do not account for the factual alignment between them. To illustrate this, consider a reference sentence *pneumonia is seen* and a generated sentence *pneumonia is not seen*. These two sentences have substantial text overlap and thus the generated sentence would achieve a high ROUGE score against the reference sentence, despite that it conveys an entirely opposite fact. A similar case is also illustrated by the example in Figure 5.1. In this section we first introduce a method to verify the factual correctness of the generated summary against the reference summary, and then describe a training strategy to directly optimize a factual correctness objective to improve summary quality.

5.2.1 Evaluating Factual Correctness via Fact Extraction

A convenient way to explicitly measure the factual correctness of a generated summary against the reference is to first extract and represent the facts in a structured format. To this end, we define a *fact extractor* to be an information extraction module, denoted as f, which takes in a summary sequence y and returns a structured fact vector $\mathbf{v}$:

$$\mathbf{v} = f(y) = (v_1, ..., v_m). \tag{5.1}$$

Here we consider v_i as a categorical variable that we want to measure via fact checking and m the total number of such variables. For example, in the case of summarizing radiology reports, v_i can be a binary variable that describes whether an event or a disease such as *pneumonia* is present or not in a radiology study.

Given a fact vector $\mathbf{v}$ output by f from a reference summary and $\hat{\mathbf{v}}$ from a generated summary, we further define a *factual accuracy* score s to be the ratio of variables in $\hat{\mathbf{v}}$ which equal the corresponding variables in $\mathbf{v}$, namely:

$$s(\hat{\mathbf{v}}, \mathbf{v}) = \frac{\sum_{i=1}^{m} \mathbb{1}[v_i = \hat{v}_i]}{m} \tag{5.2}$$

where $s \in [0, 1]$. Note that our definition of factual accuracy here requires a summary to be both precise and complete in order to achieve a high s score: missing out a positive variable or falsely claiming a negative variable will be equally penalized by our score.

Our general definition of the fact extractor module f allows it to have different realizations for different domains. For our task of summarizing radiology findings, we make use of the open-source CheXpert radiology report labeler (Irvin et al., 2019).[1] At its core, the CheXpert labeler parses the input sentences into dependency structures and runs a series of surface and syntactic rules to extract the presence status of 14 clinical observations seen in chest radiology reports. It was evaluated to have over 95% overall F_1 when compared against oracle annotations from multiple radiologists on a large-scale radiology report dataset. For the purpose of this study, we used a subset of the variables extracted by the CheXpert labeler, and discuss the reasons and our inclusion criteria in Section 5.3.3.

5.2.2 Improving Factual Correctness via Policy Learning

The fact extractor module introduced above not only enables us to measure the factual accuracy of a generated summary, but also provides us with an opportunity to directly optimize the factual accuracy as an objective. This can be achieved by viewing our summarization model as an agent, the actions of which are to generate a sequence of words to form the summary $\hat{y}$, conditioned on the input x.[2] The agent then receives rewards $r(\hat{y})$ for its actions, where the rewards can be designed to measure the quality of the generated summary. Our goal here is to learn an optimal policy $P_\theta(y|x)$ for the summarization model, parameterized by the network parameters θ, which achieves the highest expected reward under the training data.

Formally, we achieve this goal by minimizing loss $\mathcal{L}$, the negative expectation of the reward $r(\hat{y})$ over our training data:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\hat{y}\sim P_\theta(y|x)}[r(\hat{y})]. \tag{5.3}$$

[1] https://github.com/stanfordmlgroup/chexpert-labeler

[2] For clarity, going forward we drop the bold symbol and use x and y (instead of $\mathbf{x}$ and $\mathbf{y}$) to represent the input and output sequences, respectively.

Note that here the reward $r(\hat{y})$ is a general function of $\hat{y}$ (and optionally also y), which allows for different realizations that we will explore later.

The calculation of the gradient of $\mathcal{L}$ with respect to our parameters θ can be derived with the REINFORCE algorithm (Williams, 1992) as:

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{\hat{y} \sim P_\theta(y|x)}[\nabla_\theta \log P_\theta(\hat{y}|x) r(\hat{y})]. \tag{5.4}$$

This gradient calculation is essentially an negative expectation of the gradient of the log probability of the generated sequence, weighted by its respective reward. However, the exact value of the gradient is still hard to compute, due to the required expectation over all possible generations $\hat{y}$ from the model, which is difficult to enumerate in most scenarios. In practice, we approximate this gradient over a training example with a single Monte Carlo sample. To reduce the variance of this estimation, we further deduct a *baseline reward* from the original reward calculated from each $\hat{y}$:

$$\nabla_\theta \mathcal{L}(\theta) \approx -\nabla_\theta \log P_\theta(\hat{y}_s|x)(r(\hat{y}_s) - \bar{r}), \tag{5.5}$$

where $\hat{y}_s$ is a sampled sequence from the model and $\bar{r}$ a baseline reward. Related work has used different strategies to obtain a good baseline reward, such as estimating the value of the baseline reward with a separately trained classifier. Here we adopt the *self-critical training* strategy (Rennie et al., 2017), where we obtain the baseline reward $\bar{r}$ by applying the same reward function r to a greedily decoded sequence $\hat{y}_g$:

$$\bar{r} = r(\hat{y}_g). \tag{5.6}$$

We empirically find that using this self-critical baseline reward helps stabilize the training of our summarization model.

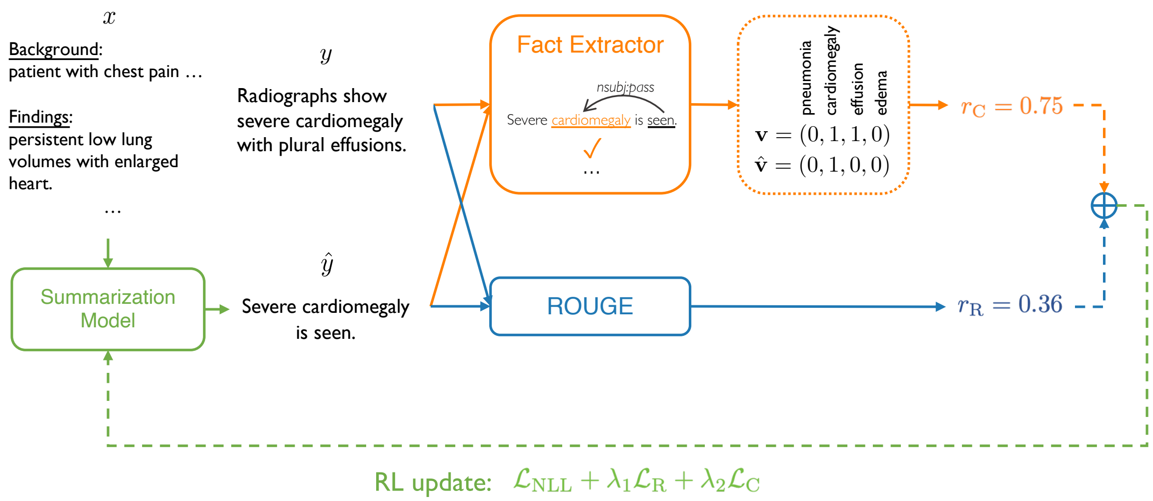

x
Background:
patient with chest pain …
Findings:
persistent low lung volumes with enlarged heart.
…
Summarization Model
y
Radiographs show severe cardiomegaly with plural effusions.
$\hat{y}$
Severe cardiomegaly is seen.
Fact Extractor
nsubj:pass
Severe cardiomegaly is seen.
✓
…
pneumonia
cardiomegaly
effusion
edema
$\mathbf{v} = (0, 1, 1, 0)$
$\hat{\mathbf{v}} = (0, 1, 0, 0)$
$r_{\mathrm{C}} = 0.75$
ROUGE
$r_{\mathrm{R}} = 0.36$
RL update: $\mathcal{L}_{\mathrm{NLL}} + \lambda_1 \mathcal{L}_{\mathrm{R}} + \lambda_2 \mathcal{L}_{\mathrm{C}}$

Figure 5.2: Our proposed reinforcement learning-based training strategy. Compared to existing work which relies only on a ROUGE reward r_{R}, we add a factual correctness reward r_{C} which is enabled by a fact extractor, which is realized by a separate information extraction model. The summarization model is updated via RL, using a combination of the NLL loss, a ROUGE-based loss and a factual correctness-based loss. For simplicity we only show a subset of the clinical variables in the fact vectors $\mathbf{v}$ and $\hat{\mathbf{v}}$.

5.2.3 Reward Function

The learning strategy in Equation (5.5) provides us with the flexibility to optimize arbitrary reward functions. Here we decompose our reward function into two parts:

$$r = \lambda_1 r_{\mathrm{R}} + \lambda_2 r_{\mathrm{C}}, \tag{5.7}$$

where $r_{\mathrm{R}} \in [0, 1]$ is a ROUGE reward, namely the ROUGE-L score (Lin, 2004) of the predicted sequence $\hat{y}$ against the reference y; $r_{\mathrm{C}} \in [0, 1]$ is a correctness reward, namely the factual accuracy s of the predicted sequence against the reference sequence, as in Equation (5.2); $\lambda_1, \lambda_2 \in [0, 1]$ are scalar weights that control the balance between the two. To measure the similarity between the reference and the generation, we also experimented with more recent metrics that rely on neural representations of text, such as the BERTScore (Zhang et al., 2020a). However, we found that these metrics, mostly trained on web and newswire data, generalize poorly to our domain of text. We leave the exploration of better reward functions as future work.

Paulus et al. (2018) found that directly optimizing a reward function without the original negative log-likelihood (NLL) objective as used in teacher-forcing can hurt the readability of the generated summaries, and proposed to alleviate this problem by combining the NLL objective with the RL loss. Here we adopt the same strategy, and our final loss during training is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{R}} + \lambda_2 \mathcal{L}_{\text{C}} + \lambda_3 \mathcal{L}_{\text{NLL}}, \tag{5.8}$$

where $\lambda_3 \in [0, 1]$ is an additional scalar that controls the weight of the NLL loss.

Our overall training strategy is illustrated in Figure 5.2. Our final loss jointly optimizes three aspects of the summaries: $\mathcal{L}_{\text{NLL}}$ serves as a conditional language model that optimizes the fluency and relevance of the generated summary, $\mathcal{L}_{\text{R}}$ controls the brevity of the summary and encourages summaries which have high overlap with human references, and $\mathcal{L}_{\text{C}}$ encourages summaries that are factually accurate when compared against human references.

5.3 Experiments

To evaluate our proposed training strategy against baseline methods, we collected two separate real-world radiology report datasets from hospitals. In this section, we describe the data collection process, the baseline models, our evaluation protocols and our model implementation and training details.

5.3.1 Data Collection

We collected anonymized chest radiographic reports within a certain period of time from two collaborating hospitals: the Stanford University Hospital and the Rhode Island Hospital (RIH).[3] Note that unlike in Chapter 4 where we used radiographic studies of all types, here we focus on studies of chest examinations, which constitute the commonest type of radiographic studies in a typical hospital.

[3] Our retrospective study has been approved by the corresponding institutional review boards with waiver of consent.

	Time Coverage	
Split	Stanford	RIH
Train	2009/01 – 2014/04	2017/11 – 2018/06
Dev	2014/05 – 2014/08	2018/07 – 2018/09
Test	2014/09 – 2014/12	2018/10 – 2018/12

Table 5.1: Time coverage of different splits in the Stanford and RIH datasets.

	Number of Examples	
Split	Stanford	RIH
Train	89,992 (68.8%)	84,194 (60.3%)
Dev	22,031 (16.8%)	25,966 (18.6%)
Test	18,827 (14.4%)	29,494 (21.1%)
Total	130,850	139,654

Table 5.2: Detailed statistics of the Stanford and RIH datasets.

For both datasets, we ran preprocessing following the same procedure described in Chapter 4, which involves tokenizing all reports with Stanford CoreNLP (Manning et al., 2014) and filtering out reports that are too short. For the purpose of evaluation, we additionally replaced all date and time mentions in the preprocessed reports with special tokens (e.g., <*DATE*>).

Instead of using random stratification, we stratified each dataset temporally into training, dev and test splits. We employed this stratification strategy to test whether our model generalizes to future data when trained on historical data. We include the stratification details of both datasets in Table 5.1 and the statistics of them in Table 5.2.

5.3.2 Models

As we use the augmented pointer-generator network described in Section 5.1 and Chapter 4 as the backbone of our method, we mainly compare against it as the baseline model (PG Baseline).

For the proposed RL-based training strategy, we compare three variants of it: training with only the ROUGE reward ($\mathrm{RL_R}$), with only the factual correctness reward ($\mathrm{RL_C}$), or with both ($\text{RL}_{\text{R+C}}$). All three variants have the NLL component in the training loss as in Equation (5.8). For all variants, we initialize the model with the best baseline model trained with standard teacher-forcing on the training data, and then finetune it on the same training data with the corresponding RL loss, until it reaches the best validation score.

To understand the difficulty of the task and evaluate the necessity of using abstractive summarization models on the two datasets, similar to the experiments in Chapter 4, we additionally evaluate two extractive summarization methods: (1) LexRank (Erkan and Radev, 2004), a widely-used non-neural extractive summarization algorithm; and (2) BanditSum (Dong et al., 2018), a state-of-the-art RL-based neural extractive summarization model. For both methods we use their open implementations. We include other model implementation and training details in Section 5.3.4.

5.3.3 Evaluation

Metrics. We use two sets of automatic metrics to evaluate model performance at the corpus level. First, we use the standard **ROUGE** scores (Lin, 2004), and report the F_1 scores for ROUGE-1, ROUGE-2 and ROUGE-L, which compare the word-level unigram, bigram and longest common sequence overlap with the reference summary, respectively.

For factual correctness evaluation, we use a **Factual $\mathbf{F}_1$** score. While the factual accuracy score s that we use in the reward function evaluates how factually accurate a specific summary is, comparing it at the corpus level can be misleading, for the same reason that accuracy is a misleading measure in information retrieval (Manning et al., 2008). To understand this, imagine the case where a clinical variable v has rare presence in the corpus. A model which always generates a negative summary for it (i.e., $v = 0$; the disease is not present) can have high accuracy, but is useless in practice. Instead, for each variable, we obtain a model's predictions over all test examples and calculate its F_1 score. We then macro-average the F_1 of all variables to obtain the overall factual F_1 score of the model.

Note that the CheXpert labeler that we use is specifically designed to run on radiology summaries, which usually have a different style than the radiology findings section of the reports (see further analysis in Section 5.5). As a result, we found the labeler to be less accurate when applied to the findings section. For this reason, we were not able to estimate the factual F_1 scores on the summaries generated by the two extractive summarization models.

Clinical Variables Inclusion Criteria. While the CheXpert labeler that we use is able to extract status for 14 clinical variables, we found that several variables are very rarely represented in our corpora and therefore using all of them makes the calculation of the factual F_1 score very unstable. For example, we found that training the same model using different random initializations would result in highly varying F_1 scores for these variables. For this reason, for both datasets we removed from the factual F_1 calculation all variables which have less than 3% positive occurrences on the validation set. We further removed the variables "Pleural Other" and "Support Devices" due to their ambiguity. This process results in a total of 9 variables for the Stanford dataset and 8 for the RIH dataset. Additionally, apart from the positive and negative status, the CheXpert labeler is also able to generate an *uncertain* status for a variable, capturing observations with uncertainty, such as in the sentence "*pneumonia is likely present*". While we can modify the factual accuracy score to take uncertainty into account, for simplicity in this work we do not make the distinction between a positive status and an uncertain status.

5.3.4 Model Implementation and Training Details

For the baseline background-augmented pointer-generator model, we reuse the same implementation as in Chapter 4. We use a 2-layer LSTM as the findings encoder, 1-layer LSTM as the background encoder, and a 1-layer LSTM as the decoder. For all LSTMs we use a hidden size of 200. For the embedding layer we use 100-dimensional GloVe vectors (Pennington et al., 2014) which we pretrained on about 4 million radiology reports. We apply dropout (Srivastava et al., 2014) with $p = 0.5$ to the embeddings. At decoding time, we use the standard beam search with a beam size of 5 and a maximum decoding length of 50.

For the training and finetuning of the models, we use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of $1e^{-3}$. We use a batch size of 64 and clip the gradient with a norm of 5. During training we evaluate the model on the dev set every 500 steps and decay the learning rate by 0.5 whenever the validation score does not increase after 2500 steps. Since we want the model outputs to have both high overlap with the human references and high factual correctness, for training we always use the average of the dev ROUGE score and the dev factual F_1 score as the stopping criteria. We tune the scalar weights in the loss function on the dev sets and use weights of $\lambda_1 = 0.97$, $\lambda_2 = 0.97$ and $\lambda_3 = 0.03$ for both datasets.

For extractive summarization model LexRank, we use its open implementation in the Sumy Python library.[4] For the BanditSum model, we use authors' original open-source implementation.[5] We use default values for all hyperparameters as in Dong et al. (2018). For both models we again select the top $N = 3$ scored sentences to form the summary, which yields the highest ROUGE-L scores on the dev sets.

For ROUGE evaluation, we use the Python ROUGE implementation released by Google Research.[6] We empirically find it to provide very close results to the original Perl ROUGE implementation by Lin (2004), but is substantially faster to execute.

5.4 Results

In this section, we first present our automatic evaluation results on the two collected datasets. We then verify our findings, by presenting a human evaluation with board-certified radiologists where we compare the summaries generated by humans, the baseline and our proposed model.

5.4.1 Automatic Evaluation

Our main results on both datasets are shown in Table 5.3. We first notice that while the neural extractive model, BanditSum, outperforms the non-neural extractive method on ROUGE scores, our PG baseline model substantially outperforms both of them, suggesting that on both datasets abstractive summarization is necessary to generate summaries comparable to human-written ones. We further show that this difference is likely due to the different

[4] https://github.com/miso-belica/sumy

[5] https://github.com/yuedongP/BanditSum

[6] https://github.com/google-research/google-research/tree/master/rouge

	Stanford				**RIH**			
System	R-1	R-2	R-L	Factual F_1	R-1	R-2	R-L	Factual F_1
LexRank (Erkan and Radev, 2004)	26.8	16.3	23.6	—	20.6	10.7	18.3	—
BanditSum (Dong et al., 2018)	32.7	20.9	29.0	—	26.1	14.0	23.3	—
PG Baseline	48.3	38.8	46.6	55.9	54.1	44.7	52.2	69.3
PG + RL_R	**52.0**	**41.1**	**49.5**	63.2	**58.0**	**47.2**	**55.7**	73.3
PG + RL_C	50.7	39.7	48.0	**65.9**	55.2	45.4	52.9	**75.4**
PG + $\text{RL}_\text{R+C}$	**52.0**	41.0	49.3	64.5	57.0	46.6	54.7	74.8

Table 5.3: Main results of different summarization models on the Stanford and RIH datasets. R-1, R-2, R-L represent the ROUGE scores. PG Baseline represents our baseline augmented pointer-generator; RL_R, RL_C and $\text{RL}_\text{R+C}$ represent RL training with the ROUGE reward alone, with the factual correctness reward alone and with both. All the ROUGE scores have a 95% confidence interval of at most ±0.6. F_1 scores for extractive models were not evaluated for the reason discussed in Section 5.3.3.

styles of language (see Section 5.5): while radiologists tend to use more compressed language when writing the summaries, extractive methods produce more verbose summaries that fail to capture this difference.

On the Stanford dataset, training the pointer-generator model with ROUGE reward alone (RL_R) leads to improvements on all ROUGE scores, with a gain of 2.9 ROUGE-L scores. Training with the factual correctness reward alone (RL_C) leads to the best overall factual F_1 with a substantial gain of 10% absolute, however with consistent decline in the ROUGE scores compared to RL_R training. Combining the ROUGE and the factual correctness rewards ($\text{RL}_\text{R+C}$) achieves a balance between the two, leading to an overall improvement of 2.7 on ROUGE-L and 8.6% on factual F_1 compared to the baseline. This indicates that $\text{RL}_\text{R+C}$ training leads to both higher overlap with references and improved factual correctness.

Most surprisingly, while ROUGE has been criticized for its poor correlation with human judgment of quality and insufficiency for evaluating correctness of the generated text (Chaganty et al., 2018), we find that optimizing ROUGE reward jointly with NLL leads to substantially more factually correct summaries than the baseline, shown by the notable gain of 7.3% factual F_1 from the RL_R training.

All of our findings are consistent on the RIH dataset, with $\text{RL}_\text{R+C}$ achieving an overall

	Stanford			**RIH**		
Variable	PG Baseline	$\text{RL}_{\text{R+C}}$	Δ	PG Baseline	$\text{RL}_{\text{R+C}}$	Δ
No Finding	77.3	81.5	+4.2*	91.0	92.0	+1.0*
Cardiomegaly	29.5	40.4	+10.9*	21.1	33.8	+12.7*
Airspace Opacity	64.6	74.9	+10.3*	80.4	83.5	+3.1*
Edema	58.4	70.9	+12.5*	73.4	80.2	+6.8*
Consolidation	46.3	53.2	+6.9*	–	–	–
Pneumonia	46.7	46.8	+0.2	63.5	69.2	+5.7*
Atelectasis	48.8	56.3	+7.5*	60.5	66.5	+6.0*
Pneumothorax	69.5	82.9	+13.4*	89.7	93.2	+3.5*
Pleural Effusion	62.0	73.4	+11.4*	74.3	79.9	+5.6*
Macro Avg.	55.9	64.5	+8.6*	69.3	74.8	+5.5*

Table 5.4: Test set factual F_1 scores for all variables on the Stanford and RIH datasets. $*$ marks statistically significant improvements with $p < .01$ under a bootstrap test.

improvement of 2.5 ROUGE-L and 5.5% factual F_1 scores.

Fine-grained Correctness. To understand how improvements in individual variables contribute to the overall improvement, we show the fine-grained factual F_1 scores for all variables on the Stanford and RIH datasets in Table 5.4. We find that on both datasets, improvements in $\text{RL}_{\text{R+C}}$ can be observed on all variables tested. We further find that, as we change the initialization across different training runs, while the overall improvement on factual F_1 stays approximately unchanged, the distribution of the improvement on different variables can vary substantially. Developing a training strategy for fine-grained control over different variables is an interesting direction for future work.

Qualitative Results. In Figure 5.3 we present two example reports along with the human references, the PG baseline outputs and $\text{RL}_{\text{R+C}}$ outputs. In the first example, while baseline output seems generic and does not include any meaningful observation, the summary from the $\text{RL}_{\text{R+C}}$ model aligns well with the reference, and therefore achieves a higher factual accuracy score. In the second example, the baseline model wrongly copied an observation from the findings although the actual context is *no longer evident*, while the $\text{RL}_{\text{R+C}}$ model correctly recognizes this and produces a better summary. We present more example reports

Stanford Dataset

Background: radiographic examination of the chest ...

Findings: continuous rhythm monitoring device again seen projecting over the left heart. persistent low lung volumes with unchanged cardiomegaly. again seen is a diffuse reticular pattern with interstitial prominence demonstrated represent underlying emphysematous changes with superimposed increasing moderate pulmonary edema. small bilateral pleural effusions. persistent bibasilar opacities left greater than right which may represent infection versus atelectasis.

Human: increased moderate pulmonary edema with small bilateral pleural effusions. left greater than right basilar opacities which may represent infection versus atelectasis.

PG Baseline ($s = 0.33$): no significant interval change.

RL$_{\textbf{R+C}}$ ($s = 1.00$): increasing moderate pulmonary edema. small bilateral pleural effusions. persistent bibasilar opacities left greater than right which may represent infection versus atelectasis.

RIH Dataset

Background: history: lobar pneumonia, unspecified organism ...

Findings: lines/tubes: none. lungs: right middle lobe airspace disease seen on prior radiographs from <date> and <date> is no longer evident. bilateral lungs appear clear. pleura: there is no pleural effusion or pneumothorax. heart and mediastinum: no cardiomegaly. thoracic aorta appears calcified and mildly tortuous. bones: ...

Human: no acute cardiopulmonary abnormality.

PG Baseline ($s = 0.75$): right middle lobe airspace disease could represent atelectasis, aspiration or pneumonia.

RL$_{\textbf{R+C}}$ ($s = 1.00$): no acute cardiopulmonary abnormality.

Figure 5.3: Truncated examples from the test sets along with human, PG baseline and RL$_{\text{R+C}}$ outputs. Factual accuracy scores (s) are also shown for the model outputs. For the Stanford example, clinical observations in the summaries are marked for clarity; for RIH, a wrongly copied observation is marked.

in Figure 5.4.

5.4.2 Human Evaluation

To study whether the improvements in the factual correctness scores lead to improvement in summarization quality under expert judgment, we run a comparative human evaluation following previous work (Chen and Bansal, 2018; Dong et al., 2018) as well as our experiments in Chapter 4. We sampled 50 test examples from the Stanford dataset, and for each example we presented to two board-certified radiologists the full radiology findings along

Stanford Dataset

Background: radiographic examination of the chest: <date> <time>. clinical history: <age> years of age, with concern for pulmonary edema. procedure comments: 3 single views of the chest...

Findings: in the first chest radiograph from <date> at <time> there is interval intubation. left arm-picc line remains in place. grossly unchanged persistent cardiomegaly, bilateral pleural effusion, and mild pulmonary edema. severe djd of the left gh joint is noted. in the second chest radiograph there is interval placement of a trialysis catheter in the left ij. no other significant changes are noted. in the third chest radiograph from <date> at <time> there is an increased left basilar opacity likely reflecting basilar consolidation, atelectasis or aspiration.

Human: in the final chest radiograph there is increased left basilar opacity likely reflecting basilar consolidation, atelectasis or aspiration.

PG Baseline: interval intubation with placement of a trialysis catheter in the left ij. grossly unchanged cardiomegaly, bilateral pleural effusion, and mild pulmonary edema.

$\text{RL}_{\text{R+C}}$: interval placement of a trialysis catheter in the left ij. an increased left basilar opacity likely reflecting basilar consolidation, atelectasis or aspiration or aspiration.

RIH Dataset

Background: post op cardiac surgery - check lines and tubes. technique: single view of the chest obtained at <time> <date>...

Findings: lines/tubes: right ij sheath with central venous catheter tip overlying the svc. on initial radiograph, endotracheal tube between the clavicular heads, and enteric tube with side port at the ge junction and tip below the diaphragm off the field-of-view; these are removed on subsequent film. mediastinal drains and left thoracostomy tube are unchanged. lungs: low lung volumes. retrocardiac airspace disease, slightly increased on most recent film. pleura: small left pleural effusion. no pneumothorax. heart and mediastinum: postsurgical widening of the cardiomediastinal silhouette. aortic arch calcification. bones: intact median sternotomy wires.

Human: left basilar airspace disease and small left pleural effusion. lines and tubes positioned as above.

PG Baseline: lines and tubes as above. retrocardiac airspace disease, which may represent atelectasis, aspiration, or pneumonia.

$\text{RL}_{\text{R+C}}$: lines and tubes as described above. retrocardiac airspace disease, slightly increased on most recent film. small left pleural effusion.

Figure 5.4: More examples from the test splits of both datasets along with human, PG baseline and $\text{RL}_{\text{R+C}}$ summaries. In the first example, the baseline output successfully copied content from the context, but missed important observations. In the second example, the baseline output included some spurious facts that were not mentioned, and again neglected some important observations. In neither examples the $\text{RL}_{\text{R+C}}$ outputs make perfect summaries, but they represent better summaries than the baseline outputs.

with blinded summaries from 1) the human reference, 2) the PG baseline and 3) our $\text{RL}_{\text{R+C}}$ model. We shuffled the three summaries such that the correspondence cannot be guessed

Metric	Win	Tie	Lose
Our Model vs. PG Baseline			
Fluency	7%	60%	33%
Factual Correctness	31%	55%	14%
Overall Quality	48%	24%	28%
Our Model vs. Human Reference			
Fluency	17%	54%	29%
Factual Correctness	23%	49%	28%
Overall Quality	44%	17%	39%

Table 5.5: Results of the radiologist evaluation. The top three rows present results when comparing our $\mathrm{RL}_{\mathrm{R+C}}$ model output versus the baseline model output; the bottom three rows present results when comparing our model output versus the human-written summaries.

from the ordering of them. We then asked the radiologists to carefully read and compare the three summaries based on the following three metrics:

1) **Fluency**: whether the presented summary uses fluent language and is easily understood by domain experts. Examples of reports with bad fluency include those with significant grammatical errors, or with unnatural or repetitive language.

2) **Factual correctness and completeness**: whether the summary contains correct and complete information. Examples of factually incorrect summaries include those which misses important observations, or include spurious observations that are not consistent with the findings.

3) **Overall quality**. Examples of summaries with low overall quality include those that are not fluent, factually incorrect or incomplete, highly verbose, or completely irrelevant to the findings, etc.

For each metric we asked the radiologists to rank the three summaries, with ties allowed. After the evaluation, we converted each ranking into two binary comparisons: 1) our model versus the baseline model, and 2) our model versus human reference.

The results of this radiologist evaluation experiment are shown in Table 5.5. Comparing our model against the baseline model, we find that: 1) in terms of fluency our model

System	Stanford perplexity	RIH perplexity
Human	6.7	5.5
LexRank	10.8	36.9
BanditSum	9.9	40.9
PG Baseline	4.8	3.8
PG + $\text{RL}_{\text{R+C}}$	6.5	4.8

Table 5.6: Perplexity scores obtained from the test set human references and model predictions. All perplexity scores shown are evaluated by a neural language model trained on a radiology impression dataset.

is less preferred, although a majority of the results (60%) are ties; and 2) our model wins substantially more on factual correctness and overall quality. Comparing our model against human references, we find that: 1) human wins more on fluency; 2) factual correctness results are close, with 72% of our model outputs being at least as good as human; and 3) surprisingly, in terms of overall quality our model was slightly preferred by the radiologists compared to human references. This may be because factual correctness is a much more important characteristic of radiology reports than fluency. Lastly, when comparing the baseline model against human references, we find that outputs from the baseline model are much less correct and lower-quality than human summaries.

5.5 Analysis

In this section we run an analysis on the summaries produced by our proposed summarization model, with the goal of understanding how and why the generated summaries differ from those produced by the baseline models.

Fluency and Style of Summaries. Our human evaluation results in Section 5.4.2 suggest that in terms of fluency our model output is less preferred than human reference and baseline output. To further understand the fluency and style of summaries from different models at a larger scale, we trained a neural language model (LM) for radiology summaries following previous work (Liu et al., 2018). Intuitively, radiology summaries which are more

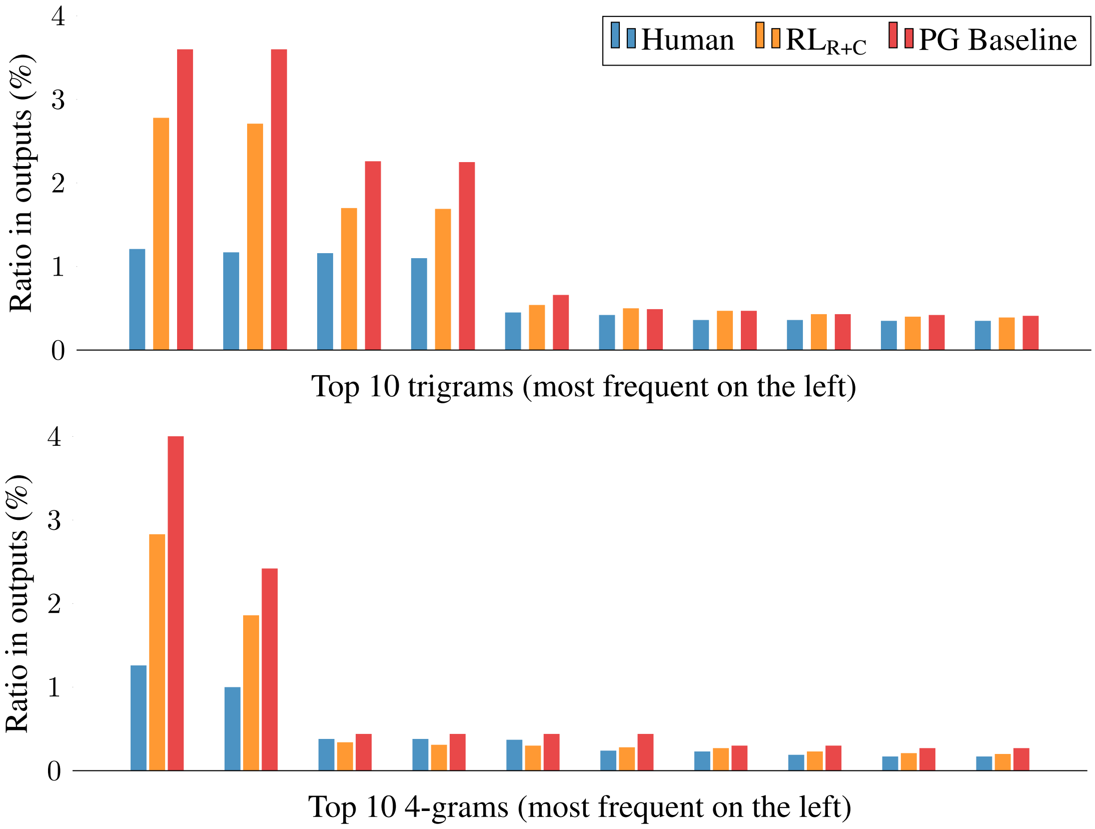


Figure 5.5: Distributions of the most frequent n-grams from model outputs. The upper figure shows the distributions of the top 10 most frequent trigrams, while the lower figure shows that of the top 10 most frequent 4-grams. Results shown were generated from the Stanford test set. In both figures the RL$_{R+C}$model presents more diverse use of n-grams that are closer to human-written summaries.

fluent and consistent with humans in style should be able to achieve a lower perplexity under this in-domain LM, and vice versa. To this end, we collected all human-written summaries from the training and dev split of both datasets, which in total gives us about 222,000 summaries. We then trained a strong Mixture of Softmaxes LM (Yang et al., 2018) on this corpus, and evaluated the perplexity of test set outputs for all models.

The results are shown in Table 5.6. We find that while extractive models can achieve non-trivial overlap with references, their perplexity scores tend to be much higher than humans. We conjecture that this is because radiologists are trained to write the summaries with more compressed language than when they are writing the findings, therefore sentences directly extracted from the findings tend to be more verbose than needed.

We further observe that the baseline model achieves even lower perplexity than humans, and our proposed method leads to a perplexity score much closer to human references. We hypothesize that this is because models trained with teacher-forcing are prone to generic generations which are fluent and relevant but may not be factually correct. Training with the proposed rewards alleviates this issue, leading to summaries more consistent with humans in style. For example, we find that *no significant interval change* is a very frequent generation from the baseline, regardless of the actual input. This sentence occurs in 34% of the baseline outputs on the Stanford dev set, while the number for $\text{RL}_{\text{R+C}}$ and human are only 24% and 17%.

Our hypothesis is further confirmed when we plot the distribution of the top 10 most frequent trigrams and 4-grams from different models in Figure 5.5: while the baseline heavily reuses the few most frequent trigrams, our model $\text{RL}_{\text{R+C}}$ tends to have more diverse summaries which are closer to human references. The same trend is observed for 5-grams.

5.6 Summary and Limitations

In this chapter we extended the neural summarization model that we developed in Chapter 4, and presented a general framework and a training strategy to improve its factual correctness. Our method relies on an information extraction system to fact-check the generated summary against its reference, and then on a reinforcement learning-based training technique for optimization. We applied our approach to the summarization of radiology reports on datasets collected from two separate hospitals, and showed its success via both automatic and radiologist evaluation. We further showed via examples and analysis that our model leads to improved correctness for all clinical variables tested, and is able to produce summaries that are more diverse and have styles more consistent with human-written summaries.

Our study also yields some general takeaways for developing neural text summarization systems:

- In a domain with a limited space of facts such as radiology reports, a carefully implemented IE system can be used to improve the factual correctness of neural summarization models via RL;
- Even in the absence of a reliable IE system, optimizing the ROUGE metrics via RL can substantially improve the factual correctness of the generated summaries.

Limitations While we showed the success of our approach, we also recognize several important limitations of our study.

- Our proposed training strategy crucially depends on the availability of an external IE module. While this IE module is relatively easy to implement for a domain with a limited space of facts, how to generalize this method to open-domain summarization remains unsolved.
- Our study was based on a rule-based IE system, which often suffers from limited robustness and generalizability. The use of a more robust statistical IE model can potentially improve the results.
- We mainly focus our study on key factual errors which result in a flip of the binary outcome of an event (e.g., presence of disease), whereas factual errors in generated summaries can occur in other forms such as wrong adjectives or coreference errors (Kryściński et al., 2019a). These errors tend to be more subtle and the detection of them often requires deeper understanding of the text, or with the help of other text analysis systems such as a dependency parser or a coreference resolution system.

Since the original publication of our study, some alternative means for verifying the factual consistencies of neural summarization models have been proposed and studied, with some addressing the limitations mentioned above. In particular, Kryscinski et al. (2020) have proposed a transformer-based consistency checking model and a weakly supervised approach to train this model for detecting inconsistencies in newswire summarization. Maynez et al. (2020) conducted a systematic study of the factuality and faithfulness of existing neural summarization models, and found that large pretrained models tend to be more consistent than models trained from scratch with a particular dataset. Wang et al.

(2020) and Durmus et al. (2020) studied QA-based methods for detecting inconsistencies in the generated summaries for newswire summarization. Despite these efforts, how to detect factual inconsistencies and how to further improve the consistency of generated text in different domains remain open research questions.

So far in this dissertation, we have focused on the understanding of biomedical scientific text and clinical text such as the radiology report. Although these forms of text are major media for documentation and communication in medicine, it is important to realize that text is not the only modality of data in medicine. Among all other data modalities, image data is a particularly important one. Medical images provide valuable information of the interior of a human body for clinical analysis and medical intervention, and are thus produced routinely in medical practice. For this reason, the automated understanding and processing of medical images has been a long-standing mission of the artificial intelligence community, but the complexity of these images has posed a significant challenge.

Fortunately, these images often co-occur or are used in conjunction with clinical text data for communication purposes. In the next chapter, we will discuss why this provides unique opportunities as well as challenges to transfer our understanding of the textual knowledge to the understanding of these images, and present a novel framework based on unsupervised learning to accomplish this goal.

Chapter 6

Joint Medical Text and Image Understanding

In the previous chapters, we have focused on the understanding and generation of different genres of medical text, such as biomedical scientific text or clinical report text. While our studies have revealed important means to access actionable biomedical knowledge and to improve communication in healthcare, we have restricted our scope to this single data modality. It is important to realize that text often co-occurs, or is used in conjunction with other important data modalities in medicine (Raghupathi and Raghupathi, 2014; Belle et al., 2015).

Among all these modalities, images are a particularly crucial one. Medical images, such as X-ray or pathology images, provide interior views of a human body and visual representations of a tissue's function, and are an important means for clinical diagnosis and medical intervention (Branstetter, 2009). Moreover, medical images are abundant in healthcare. For example, it was estimated that approximately 1 billion radiologic imaging examinations are performed worldwide annually (Bruno et al., 2015). This statistic does not include other common imaging types such as nuclear or ultrasound imaging, and the trend is still growing. This represents a huge proportion of the data produced and used in the healthcare systems worldwide.

The interpretation of these medical images is typically done manually by relevant medical experts such as radiologists. However, this task can be highly challenging for even the most skilled experts, and as a result, the estimated prevalence of radiologic error can go from 4% to a striking ratio of 30% depending on the exam types and patient population (Bruno et al., 2015). This motivates the need for the development of systems that automate the understanding and interpretation of medical images.

The recent surge of deep neural architectures for visual recognition has driven rapid progress in automated medical image understanding (Gulshan et al., 2016; Esteva et al., 2017; De Fauw et al., 2018; Rajpurkar et al., 2018b). However, with expert-level performance achieved only in some specialties and under specific circumstances, medical image understanding remains a difficult task for the majority of medical specialties, mainly due to its challenging nature and the extreme scarcity of annotated data.

Existing work on medical image understanding has followed two general approaches to obtain annotations for medical imaging tasks. The first approach has been using high-quality annotations created by medical experts (Abràmoff et al., 2016; Gulshan et al., 2016; Shih et al., 2019; Wang and Wong, 2020). However, obtaining annotations from medical experts is usually much more expensive than crowdsourcing from non-experts. As a result, datasets created in this way are often orders of magnitude smaller than natural image datasets such as ImageNet (Russakovsky et al., 2015). To remedy this, existing work on medical imaging has relied heavily on transferring model weights from ImageNet pretraining (Wang et al., 2017; Esteva et al., 2017; Irvin et al., 2019). This approach is suboptimal because, as shown in Figure 6.1, medical image understanding often requires representations of very fine-grained visual features that are drastically different from those required for identifying objects in natural scenes. For example, understanding that the left image in Figure 6.1 represents "cardiomegaly" requires a model to recognize the enlarged cardiac contour as represented by the gray shadow in the white box; similarly, recognizing the "pleural effusion" in the right image requires a model to identify the subtle blunted angle near the lateral lower lung, due to the presence of excess fluid in the area. As a result of these disparate image characteristics between medical and natural images, Raghu et al. (2019) found that ImageNet pretraining often provides little to no benefit compared to simple random initialization.

A second popular approach is motivated by the observation that a medical image is often produced in conjunction with its clinical textual descriptions in a typical imaging

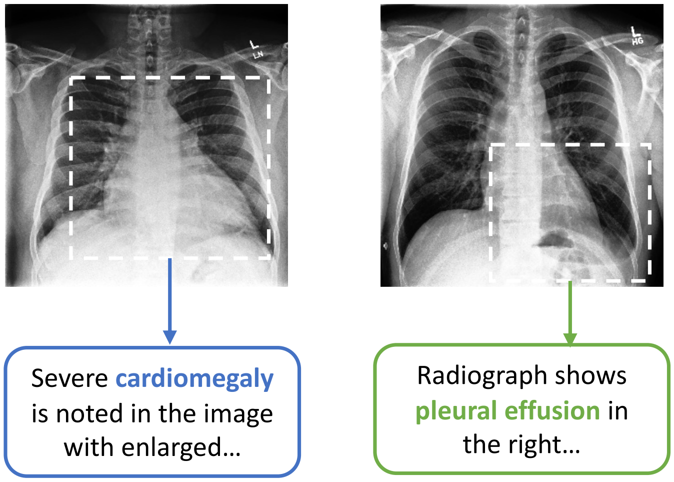

Severe cardiomegaly is noted in the image with enlarged...
Radiograph shows pleural effusion in the right...

Figure 6.1: Two example chest radiograph images along with sentences from their paired textual reports. The two images correspond to different abnormality categories. For both images we also show example views (in dashed box) indicative of their characteristics.

workflow. As we have already shown in previous chapters (see Figure 4.1), the free-text radiology report is a common way for radiologists to document and communicate their image interpretations. Figure 6.1 also demonstrates the expert-written textual descriptions of two example medical images. Therefore, this second approach attempts to address annotation scarcity by using expert-crafted rules or patterns to extract labels from the textual reports accompanying the medical images. This approach has led to datasets of larger scale, since the text data paired with medical images are often produced naturally by medical experts and are abundant in a typical hospital's IT systems. Nevertheless, this rule-based label extraction approach has three important limitations. First, the rules or patterns used in the systems are often inaccurate, leading to misleading extracted labels. Second, the rules are often limited to a few major abnormality categories (Wang et al., 2017), resulting in very inefficient use of the textual report data. And third, these rules are often domain-specific and sensitive to the style of the text, making cross-domain and cross-institution generalization difficult (Irvin et al., 2019). An example of a rule-based label extraction approach, but also evidence of its domain-specificity and sensitivity to textual style was seen in the CheXpert information extraction system used in Chapter 5.

In efforts to make more efficient use of unlabeled image data, several recent studies have shown promising results on classifying natural images via the use of unsupervised contrastive representation learning methods (Chen et al., 2020a; He et al., 2020; Grill et al., 2020). However, as we will show in this chapter, applying these image view-based contrastive methods to medical images provides only marginal benefits compared to ImageNet pretraining, a result mostly due to the high inter-class similarity of the medical images as in Figure 6.1.

For these reasons, in this chapter, we extend our work on clinical text understanding, but focus on improving visual representations of medical images by harnessing the power of deep understanding of the abundant textual data accompanying the images. Meanwhile, we achieve this goal via unsupervised statistical learning, eliminating the inaccuracy and inefficiency of traditional rule-based approaches. To this end, we present ***Con****trastive* ***VI****sual* ***R****epresentation Learning from* ***T****ext (ConVIRT)*, a framework for learning visual representations by exploiting the naturally occurring pairing of images and textual data. ConVIRT improves visual representations by maximizing the agreement between true image-text pairs versus random pairs via a bidirectional contrastive objective between the image and text modalities. We apply ConVIRT to the pretraining of medical image encoders, and show that it leads to higher-quality in-domain image representations that capture the subtlety of visual features required for medical image understanding tasks.

Compared to existing methods, ConVIRT has the advantages of utilizing the paired text data in a way agnostic to the medical specialty and requiring no additional expert input. This allows us to evaluate ConVIRT by transferring its pretrained weights to 4 different medical image classification tasks covering 2 different medical specialties. We find that the resulting models outperform all baseline initialization approaches, including the standard ImageNet pretraining and several strong baselines that also utilize the paired text data. ConVIRT further improves upon popular image-only unsupervised learning methods such as SimCLR (Chen et al., 2020a) and MoCo v2 (Chen et al., 2020b). Most notably, we show that in all 4 classification tasks, ConVIRT requires only 10% as much labeled training data as an ImageNet initialized counterpart to achieve better or comparable performance. We further evaluate ConVIRT on two new zero-shot retrieval tasks, an image-image and a text-image retrieval task, and also find it superior to all baselines, corroborating the high quality of the learned representations.

To summarize, this chapter makes the following key contributions:

1) We propose ConVIRT, a general unsupervised framework for learning visual representations from the paired textual data;

2) We apply ConVIRT to pretrain medical image encoders on two unsupervised medical image-text datasets of different specialties;

3) We evaluate the pretrained encoders on 4 benchmark medical image classification tasks and show that they outperform both ImageNet pretraining and baseline image view-based contrastive learning;

4) We further collect two new zero-shot medical image retrieval datasets, and show that encoders pretrained with ConVIRT substantially outperform baselines on these tasks.

This chapter is organized as follows. In Section 6.1, we give a formal definition of our representation learning framework, and describe ConVIRT and its implementation in detail. In Section 6.2, we describe our pretraining and evaluation datasets, baseline models as well as our experimental settings in detail; we also introduce how we collect the new zero-shot medical image retrieval datasets. We present our experimental results in Section 6.3, and provide more in-depth analysis of ConVIRT in Section 6.4. Lastly, we summarize our key findings in Section 6.5, and highlight some directions for future work.

6.1 Methods

6.1.1 Learning Framework Definition

To formally define our representation learning setting, we assume paired input $(\mathbf{x}_v, \mathbf{x}_u)$ where $\mathbf{x}_v$ represents one or a group of images, and $\mathbf{x}_u$ represents a text sequence which describes the imaging information in $\mathbf{x}_v$. Our goal is to learn a parameterized image encoder function f_v, which maps an image to a fixed-dimensional vector. We are then interested in transferring the learned image encoder function f_v into downstream tasks, such as classification or image retrieval. In this work, we model the encoder function f_v as a convolutional neural network (CNN), as is commonly done in medical image understanding work.

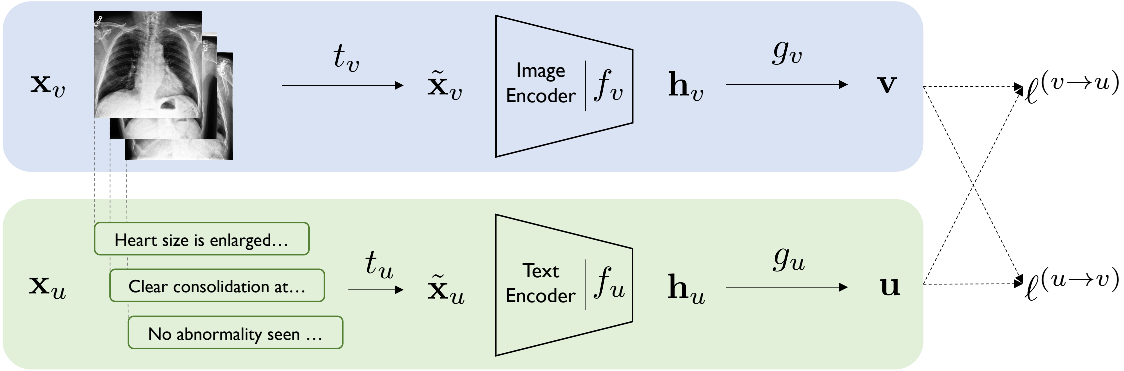


Figure 6.2: Overview of the proposed ConVIRT framework. The blue and green shades represent the image and text encoding pipelines, respectively. $\mathbf{x}_v$ and $\mathbf{x}_u$ represent input batches of images and text, respectively; t_v and t_u represent image and text transformation functions; f_v and f_u represent encoders; and g_v and g_u represent projection functions. Our method relies on maximizing the agreement between the true image-text representation pairs with bidirectional losses $\ell^{(v\to u)}$ and $\ell^{(u\to v)}$.

We note that while input image-text pairs $(\mathbf{x}_v, \mathbf{x}_u)$ are often non-trivial to obtain for natural images, they naturally exists for many medical domains and are readily usable. In particular, medical experts such as radiologists produce textual descriptions of images as part of their routine workflow, and as a result, paired medical image-text data is often stored in abundance in a typical hospital's IT system, some of which are also made available as public resources (Demner-Fushman et al., 2016; Johnson et al., 2019).

6.1.2 Contrastive Visual Representation Learning from Text

An overview of our method, ConVIRT, for learning the image encoder f_v is shown in Figure 6.2. At a high level, our method converts each input image $\mathbf{x}_v$ and text $\mathbf{x}_u$ into d-dimensional vector representations $\mathbf{v}$ and $\mathbf{u}$ respectively, following a similar processing pipeline. Our method then learns the image and text representation functions by maximizing the agreement between true image-text representation pairs, while minimizing the agreement between randomly sampled pairs.

For each input image $\mathbf{x}_v$, our method starts by drawing a random view $\tilde{\mathbf{x}}_v$ from $\mathbf{x}_v$ with a sampled transformation function $t_v \sim \mathcal{T}$, where $\mathcal{T}$ represents a family of stochastic image transformation functions described later. Next, the encoder function f_v transforms $\tilde{\mathbf{x}}_v$ into a fixed-dimensional vector $\mathbf{h}_v$, followed by a non-linear projection function g_v which further transforms $\mathbf{h}_v$ into vector $\mathbf{v}$:

$$\mathbf{v} = g_v(f_v(\tilde{\mathbf{x}}_v)), \tag{6.1}$$

where $\mathbf{v} \in \mathbb{R}^d$. Similarly, for each text input $\mathbf{x}_u$, we draw a span $\tilde{\mathbf{x}}_u$ from it following a sampling function t_u, and obtain a text representation $\mathbf{u}$ with a text encoder f_u and a projection function g_u as:

$$\mathbf{u} = g_u(f_u(\tilde{\mathbf{x}}_u)), \tag{6.2}$$

where $\mathbf{u} \in \mathbb{R}^d$. The projection functions g_v and g_u project representations for both modalities from their encoder space to the same d-dimensional space for contrastive learning.

At training time, we sample a minibatch of N input pairs ($\mathbf{x}_v$, $\mathbf{x}_u$) from training data, and calculate their representation pairs ($\mathbf{v}$, $\mathbf{u}$). We now use ($\mathbf{v}_i$, $\mathbf{u}_i$) to denote the i-th pair in the input batch. The training objective of ConVIRT involves two loss functions. The first loss function is an image-to-text contrastive loss for the i-th pair:

$$\ell_i^{(v \rightarrow u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}, \tag{6.3}$$

where $\langle \mathbf{v}_i, \mathbf{u}_i \rangle$ represents the cosine similarity, i.e., $\langle \mathbf{v}, \mathbf{u} \rangle = \mathbf{v}^\top \mathbf{u} / \|\mathbf{v}\| \|\mathbf{u}\|$; and $\tau \in \mathbb{R}^+$ represents a temperature parameter. This loss takes the same form as the InfoNCE loss (Oord et al., 2018), and it was shown that minimizing this loss leads to encoders that maximally preserve the mutual information between the true pairs under the representation functions. Intuitively, it is also the log loss of an N-way classifier that tries to predict ($\mathbf{v}_i$, $\mathbf{u}_i$) as the true pair. Note that unlike previous work which use a contrastive loss between inputs of the same modality (Chen et al., 2020a; He et al., 2020), our image-to-text contrastive loss is asymmetric for each input modality. We therefore define a similar text-to-image contrastive loss as:

$$\ell_i^{(u \rightarrow v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}. \tag{6.4}$$

Our final training loss is then computed as a weighted combination of the two losses averaged over all positive image-text pairs in each minibatch:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \Big(\lambda \ell_i^{(v \to u)} + (1 - \lambda) \ell_i^{(u \to v)} \Big), \tag{6.5}$$

where $\lambda \in [0, 1]$ is a scalar weight.

6.1.3 Realization

We note that our ConVIRT framework defined above is agnostic to the specific choice of image and text encoders, transformations and projection functions. In our particular implementation, following previous work (Chen et al., 2020a), we model g_v and g_u as separate learnable single-hidden-layer neural networks, i.e., $g_v(\cdot) = \mathbf{W}^{(2)} \sigma(\mathbf{W}^{(1)}(\cdot))$ where σ is a ReLU non-linearity, and similarly for g_u.

For the image encoder f_v, we use the ResNet50 architecture (He et al., 2016) for all experiments, as it is the architecture of choice for much medical imaging work and is shown to achieve competitive performance. For the text encoder f_u, we use a BERT encoder (Devlin et al., 2019) followed by a max-pooling layer over all output vectors. We also experimented with using a mean-pooling layer or using the special [CLS] token representation from BERT as our pooling strategy (Reimers and Gurevych, 2019), and found max-pooling achieved the best overall performance, and therefore used it consistently across all experiments.

For the image transformation family $\mathcal{T}$ where t_v is sampled from, we use sequential applications of five random transformations: *cropping*, *horizontal flipping*, *affine transformation*, *color jittering* and *Gaussian blur*. Different from previous work on contrastive visual representation learning (Chen et al., 2020a,b), we only apply brightness and contrast adjustments in *color jittering*, due to the monochrome nature of the medical images. For the text transformation function t_u, we apply a simple uniform sampling of a sentence from the input document $\mathbf{x}_u$ (i.e., $\tilde{\mathbf{x}}_u$ is a randomly sampled sentence from $\mathbf{x}_u$ for each minibatch). We did not use a more aggressive transformation mainly because sampling at the sentence level can preserve the semantic meaning of the sampled spans.

6.2 Data & Experiments

We now introduce the paired datasets that we used for contrastive pretraining and the downstream tasks and datasets that we used to evaluate the pretrained image encoders. We then introduce the baseline methods that we compare our contrastive pretraining method against in our experiments. We also include our implementation and model training details, as well as our dataset collection details at the end of this section.

6.2.1 Data for Pretraining

We test our ConVIRT framework by pretraining two separate image encoders covering different medical specialties using two separate paired image-text datasets:

- **Chest** image encoder: We use version 2 of the public **MIMIC-CXR** database (Johnson et al., 2019), which is a collection of chest radiograph images paired with their text reports, and since its release has become a standard resource for studying multi-modal modeling of medical images. After preprocessing, this dataset contains a total of about 217k image-text pairs, with each pair containing an average of 1.7 images and 6.0 sentences.

- **Bone** image encoder: We obtain a collection of musculoskeletal image-text pairs from the Rhode Island Hospital system. Following chest images, musculoskeletal images constitute the second most common type of radiograph images in a typical hospital. This dataset contains a total of 48k image-text pairs, with each pair containing an average of 2.5 images and 8.0 sentences.

We include model implementation and pretraining details in Section 6.2.4.

6.2.2 Evaluation Tasks & Data

We evaluate our pretrained image encoders on three downstream medical imaging tasks: image classification, image-image retrieval and text-image retrieval. We now describe each of the evaluation settings as well as the datasets used.

Image Classification. We evaluate our pretrained image representations on four representative medical image classification tasks:

- **RSNA Pneumonia Detection** (Wang et al., 2017; Shih et al., 2019): this task involves binary classification of a chest radiograph image into either a *pneumonia* or a *normal* category. We used the original version of this dataset available at its Kaggle page,[1] which contains 25184/1500/3000 annotated images in its training/validation/test sets, respectively.

- **CheXpert** image classification (Irvin et al., 2019): this task involves multi-label binary classification of a chest image for five individual labels, i.e., *atelectasis*, *cardiomegaly*, *consolidation*, *edema* and *pleural effusion*. We downloaded the original version of this dataset from its official website.[2] Since the original expert-labeled test set of this dataset is hidden and not included as part of the release, we instead followed Raghu et al. (2019) and used the original expert-labeled validation set as our test set, and randomly sampled 5000 images from the original training set for validation purpose. The resulting dataset contains 218414/5000/234 images in each split.

- **COVIDx** image classification (Wang and Wong, 2020): This task involves multi-class classification of a chest image into one of *COVID19*, *non-COVID pneumonia* or *normal* categories. We prepared this dataset following the scripts provided by its authors.[3] We used the version 4 of this dataset, the latest version at the time of this work. We additionally randomly sampled 300 images from the training set for validation, resulting in a dataset with 13598/300/300 images in each split.

- **MURA** bony abnormality detection (Rajpurkar et al., 2018a): This task involves binary classification of a musculoskeletal image into *abnormal* or *normal*. We downloaded the original version of this dataset from its website.[4] Similar to the CheXpert dataset, we again used the original validation set as our test set, and randomly sampled 10% images from the training set for validation, resulting in a dataset with 33078/3730/3197 images

[1] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

[2] https://stanfordmlgroup.github.io/competitions/chexpert/

[3] https://github.com/lindawangg/COVID-Net

[4] https://stanfordmlgroup.github.io/competitions/mura/

in each split. Different from the other 3 datasets, the MURA dataset uses patient-level evaluation, meaning that the prediction results from different images of the same patient needs to be aggregated to produce a final prediction for the patient, which is then scored against the gold patient label. We therefore followed Rajpurkar et al. (2018a) and at test time aggregated result for a patient by averaging the predicted probabilities from multiple images.

We report test accuracy for COVIDx given its balanced test set, and report the standard area under the receiver operating characteristic curve (AUC) metric for other tasks following previous work.

Following previous work on unsupervised visual representation learning (Hénaff et al., 2020; Chen et al., 2020a; He et al., 2020), for all classification tasks, we evaluate each pretrained image encoder under two individual settings: a **linear classification** setting, where the pretrained CNN weights are frozen and only a randomly initialized linear classification head is trained for the task; and a **fine-tuning** setting, where both the CNN weights and the linear head are fine-tuned together. The two settings complement each other for evaluation purposes: while the linear setting directly evaluates the quality of the extracted image features with the pretrained CNN, the fine-tuning setting more closely resembles how the pretrained CNN weights are used in practical applications.

To further compare the **data efficiency** of different pretraining methods, for each setting we evaluate the image encoders with **1%**, **10%** and **all** training data, respectively (except for the COVIDx dataset where we omit the 1% setting due to the scarcity of data for some categories). To control the variance in results, for all settings and models, we report average results aggregated over 5 independent training runs.

Zero-shot Image-image Retrieval. This evaluation is similar to the conventional content-based image retrieval setting in which we search for images of a particular category using a representative *query* image. For evaluation, a group of query images and a larger collection of *candidate* images, each with a categorical label, are given to a pretrained CNN encoder. We encode each query and candidate image with this encoder, and then for each query, rank all candidates by their cosine similarities to the query in descending order. Since a widely-used annotated benchmark for this setting is not available, we create our own dataset by re-using existing annotations in the CheXpert dataset (Irvin et al., 2019) and additional expert annotations from a board-certified radiologist. The resulting dataset covers 8 different chest abnormality categories, each with 10 expert-annotated query and 200 candidate images. We include the detailed collection and annotation procedure in Section 6.2.5, and refer to this dataset as **CheXpert 8×200 Retrieval Dataset**. We focus our evaluation on retrieval precision, and evaluate our models with Precision@k metrics where $k = 5, 10, 100$.

Zero-shot Text-image Retrieval. This setting is similar to the image-image retrieval setting, but instead of using query images, we retrieve images of a particular category with textual queries. For this purpose, we ask a radiologist to write 5 diverse and representative textual descriptions for each of the 8 abnormality categories for the same CheXpert 8x200 candidate images (see Section 6.2.5 for details). At test time, for each query we encode its text with the learned text encoder f_u and then retrieve from candidate images in a similar way. This evaluation not only evaluates the quality of the learned image representations, but also the alignment between the text representations and the image representations. We again use Precision@k metrics where $k = 5, 10, 100$.

6.2.3 Baseline Methods

We compare ConVIRT against the following standard or competitive initialization methods:

- **Random Init.**: For all tasks we initialize the ResNet50 image encoder with its default random initialization.

- **ImageNet Init.**: We initialize ResNet50 with weights pretrained on the standard ImageNet ILSVRC-2012 task (Russakovsky et al., 2015). We include this as a baseline since ImageNet pretraining remains a dominant approach for medical imaging work (Raghu et al., 2019).

- **Caption-LSTM**: We initialize the ResNet50 weights by first pretraining it with an image captioning task using the standard CNN-LSTM with attention architecture (Xu et al., 2015a). For the captioning task, we train the model to decode the paired textual report from the encoded image representations. Compared to the random or ImageNet initializations, this is an "in-domain" initialization baseline which uses the paired text data for representation learning.

- **Caption-Transformer**: In this initialization method we replace the CNN-LSTM model in Caption-LSTM with a CNN-Transformer-based captioning model in Cornia et al. (2020), which recently achieves state-of-the-art results on the COCO image captioning benchmark (Lin et al., 2014).

- **Contrastive-Binary**: This baseline differs from our method by contrasting the paired image and text representations with a binary classification head, as is widely done in visual-linguistic pretraining work (Tan and Bansal, 2019; Su et al., 2020). For each input pair, we first project encoder outputs $\mathbf{h}_v$ and $\mathbf{h}_u$ into the same dimension with linear layers, concatenate them, and use a MLP network to predict a binary probability of whether the input is a real or a "fake" pair, which we train with a standard binary cross-entropy loss. During training, for each $(\mathbf{x}_v, \mathbf{x}_u)$ pair in the training set, we construct a "fake" pair by replacing $\mathbf{x}_u$ with a randomly sampled one from the dataset. We expect that this binary classification task requires the encoder to learn reasonable representations of the input images, and therefore is a stronger in-domain initialization baseline.

For fair comparison, for all baselines that require paired image-text data, we use the same paired datasets as in our contrastive pretraining. For the captioning-based methods, we use the model checkpoints that achieve the best CIDEr score (Vedantam et al., 2015) on a held-out validation set.

6.2.4 Model Implementation and Training Details

Pretraining Dataset Preprocessing. For the MIMIC-CXR chest radiograph dataset, we use the publicly available JPG version of it.[5] For both the MIMIC-CXR chest dataset and the Rhode Island Hospital bone image datasets, we resize the image files to have a size of 256 on the larger side. For the textual radiology report data, we first tokenize all reports with the default English tokenizer in version 4.0.0 of the CoreNLP library (Manning et al.,

[5] https://physionet.org/content/mimic-cxr-jpg/2.0.0/

2014). Next, we keep only the *Findings* and *Impression* sections and remove all other sections. We remove all image-text pairings from the dataset where the text section is empty or has less than 3 tokens. This preprocessing procedure gives us about 217k total image-text pairs for pretraining our chest image encoder and 48k total pairs for pretraining our bone image encoder.

Image and Text Encoders. For the image encoder, we use the standard ResNet50 implementation provided by the torchvision library. For the text encoder, we use the BERT base encoder offered by the Transformers library (Wolf et al., 2020) and initialize it with the ClinicalBERT model (Alsentzer et al., 2019) pretrained on the MIMIC clinical notes. We also experimented with training a specialized BERT encoder on a large collection of radiology notes but found that it made no substantial difference in the pretraining results. At pretraining time we freeze the embeddings and the first 6 layers of this BERT encoder, and only fine-tune the last 6 layers for our contrastive task.

Other Hyperparameters of ConVIRT. For contrastive learning, we use projection layers with an output dimension $d = 512$, a temperature value $\tau = 0.1$, a loss weight $\lambda = 0.75$. These hyperparameter settings are obtained by comparing the linear evaluation validation scores on the RSNA image classification task with the pretrained ResNet50 weights. For the image transformation family $\mathcal{T}$, we adopt the implementations offered by the torchvision library.[6] We apply *random cropping* with a ratio sampled from $[0.6, 1.0]$; *horizontal flipping* with $p = 0.5$; *affine transformation* with a degree sampled from $[-20, 20]$, max horizontal and vertical translation fractions of 0.1, and a scaling factor sampled from $[0.95, 1.05]$; *color jittering* with brightness and contrast adjustment ratios sampled from $[0.6, 1.4]$; and *Gaussian blur* with $\sigma \in [0.1, 3.0]$. All images are resized to 224×224 after the transformation t_v is applied. Limited by computational resources, we arrive at these image transformation parameters via preliminary experiments rather than a systematic search.

ConVIRT Pretraining Details. At pretraining time, for each dataset, we randomly sample 5k image-text pairs to form a held-out validation set. We we use the Adam optimizer

[6] https://github.com/pytorch/vision

(Kingma and Ba, 2015) with an initial learning rate of 1e-4 and weight decay of 1e-6. We initialize the image encoder with ImageNet pretrained weights at the beginning of pretraining, and use a fixed batch size of 32. We calculate the validation loss every 5000 steps, and if the validation loss does not decrease after 5 straight evaluation runs, we anneal the learning rate by a factor of 0.5. We stop pretraining after 200 evaluation runs, and save the model checkpoint that achieves the lowest validation loss. For efficiency, we employ mixed-precision training, and for reference, the whole pretraining run on the MIMIC-CXR dataset took about 3 days on a single Titan RTX GPU card.

Classification Model Training Details. For all classification models that require ImageNet pretrained initialization, we use the pretrained weights from torchvision, which achieves an ImageNet top-5 error rate of 7.13%. For all datasets, we first zero-pad the input image to be square, and then resize it to be 224×224. For training, we use the Adam optimizer with an initial learning rate of 1e-3 for the COVIDx task and 1e-4 for the other three tasks. We additionally apply a weight decay of 1e-6 and a dropout before the last classification layer with $p = 0.2$ in all tasks. All classification models are trained with a batch size of 64. In the fine-tuning evaluation setting, we first "warmup" the classification head by freezing the CNN weights and only training the classification head with a learning rate of 1e-3 for 200 steps, after which we unfreeze the CNN weights and fine-tune the entire network together. Validation score is obtained after each epoch of training and we anneal the learning rate by a factor of 0.5 if the validation score is not improved after 3 epochs. The training is stopped after no validation improvement is observed for 10 straight epochs, at which point the model checkpoint with the highest validation score is evaluated on the test set.

6.2.5 Collection of Zero-Shot Retrieval Datasets

We now describe our collection procedures for the zero-shot image-image and text-image retrieval datasets used in our experiments.

Image-image Retrieval Dataset Collection

We create the CheXpert 8×200 Retrieval Dataset with 8 different abnormality categories commonly found in Chest radiograph images, including *atelectasis*, *cardiomegaly*, *edema*, *fracture*, *pleural effusion*, *pneumonia*, *pneumothorax* and a special *no finding* category indicating that no obvious abnormality is found in the image. We create the dataset by reusing existing rule-labeled annotations in the CheXpert dataset (Irvin et al., 2019) and additional expert annotations. To create the candidate images for a category label ℓ, we go through all images in the CheXpert training set, and keep an image as a candidate image if only its label for ℓ is positive and all other categories negative. We only include images with this "exclusive positivity" as candidate images, mainly to avoid confounding results between categories in retrieval evaluation.

To create the query images for a category ℓ, we again first pre-select 50 exclusively positive images for this category in the CheXpert training set (with all candidate images excluded). Next, we ask a board-certified radiologist to examine each of the 50 images, and exclude images that: 1) might indicate additional abnormalities other than ℓ, 2) have uncommon color or contrast distortions in the image, or 3) are not well posed during the capture of the image. This procedure is mainly to avoid including query images that have uncommon features and may therefore bias the retrieval evaluation results. At the end, we aggregate the annotation results from the radiologist and keep 10 query images for each abnormality category.

Text-image Retrieval Dataset Collection

For the text-image retrieval dataset, we first reuse all candidate images from the CheXpert 8×200 image-image retrieval dataset described above, with 200 images for each of 8 categories. To create the textual queries for each abnormality category, we ask a board-certified radiologist to write at least 5 different sentences that he will use to describe this abnormality in radiology reporting. We additionally set the following requirements: 1) the sentences must describe the category with no ambiguity and must not include other categories; 2) the sentences must be diverse from each other; and 3) the sentences should not include very

Image Category	Example Textual Query
Atelectasis	Platelike opacity likely represents atelectasis.
Cardiomegaly	The cardiac silhouette is enlarged.
Edema	The presence of hazy opacity suggests interstitial pulmonary edema.
Fracture	A cortical step off indicates the presence of a fracture.
Pleural Effusion	The pleural space is partially filled with fluid.
Pneumonia	A pulmonary opacity with ill defined borders likely represents pneumonia.
Pneumothorax	A medial pneumothorax is present adjacent to the heart.
No Finding	No clinically significant radiographic abnormalities.

Table 6.1: Example textual queries for each of the 8 categories in the text-image retrieval task. Only one example query is shown for each category.

specific anatomic locations or rare clinical observations. At the end, we aggregate the results and keep 5 textual queries for each abnormality category. For reference, we present example textual queries in Table 6.1.

6.3 Results

We now describe the experimental results of the classification and zero-shot retrieval tasks, and highlight our main findings.

6.3.1 Classification Tasks

Linear Classification. We present all linear classification results for the medical imaging tasks in Table 6.2a. We find that compared to random initialization, ImageNet initialization provides markedly better representations, despite pretrained on a very different domain of images; in-domain image initialization methods that use paired image-text data further improve over ImageNet initialization in almost all settings. Among the in-domain initialization methods, our proposed ConVIRT pretraining achieves the best overall results in all settings. Notably, we find that on three out of the four tasks, with only 1% training data ConVIRT is able to achieve classification results better than the default ImageNet initialization with 100% training data, highlighting the high quality of the learned representations from ConVIRT.

(a) Linear Classification.

	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
Method	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	55.0	67.3	72.3	58.2	63.7	66.2	69.2	73.5	50.9	56.8	62.0
ImageNet Init.	82.8	85.4	86.9	75.7	79.7	81.0	83.7	88.6	63.8	74.1	79.0
In-domain initialization methods											
Caption-Transformer	84.8	87.5	89.5	77.2	82.6	83.9	80.0	89.0	66.5	76.3	81.8
Caption-LSTM	89.8	90.8	91.3	85.2	85.3	86.2	84.5	**91.7**	75.2	81.5	84.1
Contrastive-Binary	88.9	90.5	90.8	84.5	85.6	85.8	80.5	90.8	76.8	81.7	85.3
ConVIRT (Ours)	**90.7**	**91.7**	**92.1**	**85.9**	**86.8**	**87.3**	**85.9**	**91.7**	**81.2**	**85.1**	**87.6**

(b) Fine-tuning.

	RSNA (AUC)			CheXpert (AUC)			COVIDx (Accu.)		MURA (AUC)		
Method	1%	10%	all	1%	10%	all	10%	all	1%	10%	all
General initialization methods											
Random Init.	71.9	82.2	88.5	70.4	81.1	85.8	75.4	87.7	56.8	61.6	79.1
ImageNet Init.	83.1	87.3	90.8	80.1	84.8	87.6	84.4	90.3	72.1	81.8	87.0
In-domain initialization methods											
Caption-Transformer	86.3	89.2	92.1	81.5	86.4	**88.2**	88.3	92.3	75.2	83.2	87.6
Caption-LSTM	87.2	88.0	91.0	83.5	85.8	87.8	83.8	90.8	78.7	83.3	87.8
Contrastive-Binary	87.7	89.9	91.2	86.2	86.1	87.7	89.5	90.5	80.6	84.0	88.4
ConVIRT (Ours)	**88.8**	**91.5**	**92.7**	**87.0**	**88.1**	88.1	**90.3**	**92.4**	**81.3**	**86.5**	**89.0**

Table 6.2: Results for the medical image classification tasks: (a) linear classification setting; (b) fine-tuning setting. All results are averaged over 5 independently trained models. Best results for each setting are shown in boldface. The 1% setting for the COVIDx dataset is omitted due to the scarcity of labels in COVIDx.

Fine-tuning. We show the fine-tuning evaluation results in Table 6.2b. Similar to the linear setting, we find that: 1) ImageNet initialization is again better than random initialization with smaller margins; 2) all in-domain initialization methods are better than the popular ImageNet initialization in most settings; and 3) our proposed ConVIRT pretraining again achieves the best overall results in 10 out of the 11 settings, with the exception of the CheXpert dataset with all training data used, where the result of ConVIRT is similar to that of the Caption-Transformer result. Most notably, on all datasets, with only 10%

	Image-Image Retrieval			Text-Image Retrieval		
Method	Prec@5	Prec@10	Prec@50	Prec@5	Prec@10	Prec@50
Random	12.5	12.5	12.5	12.5	12.5	12.5
ImageNet	14.8	14.4	15.0	–	–	–
In-domain initialization methods						
Caption-Transformer	29.8	28.0	23.0	–	–	–
Caption-LSTM	34.8	32.9	28.1	–	–	–
Contrastive-Binary	38.8	36.6	29.7	15.5	14.5	13.7
ConVIRT (Ours)	**45.0**	**42.9**	**35.7**	**60.0**	**57.5**	**48.8**
Fine-tuned						
ConVIRT + CheXpert Supervised	56.8	56.3	48.9	–	–	–

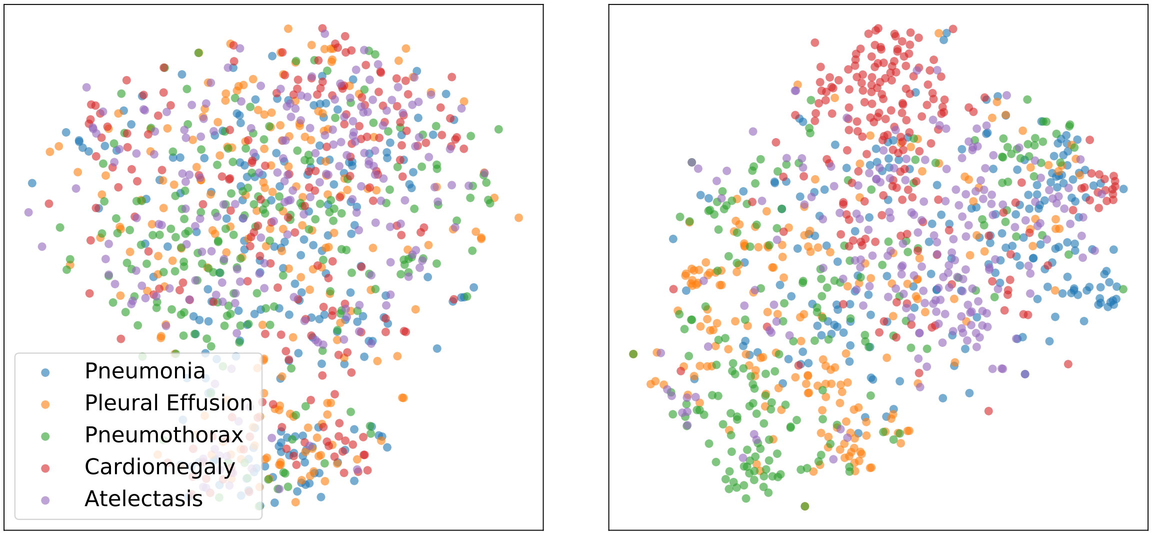
Table 6.3: Zero-shot image-image and text-image retrieval results on the CheXpert 8×200 datasets. *Random* shows results from a random guess; *ConVIRT + CheXpert Supervised* shows results from further fine-tuning the ConVIRT pretrained weights with supervised training data. Text-image retrieval results are not obtained for some methods due to the lack of text encoders.

labeled training data ConVIRT achieves classification results that are better or close to the ImageNet initialization with 100% training data results.

We also notice that our results for using ImageNet versus random initialization are different from Raghu et al. (2019): while they showed comparable results from the two strategies, we find that using ImageNet initialization is still superior than random initialization in most results, justifying its popularity. Upon closer examination, we conjecture that this is likely due to under-optimization of their models: while our ResNet50 with random initialization achieves an average AUC of 85.8 on the CheXpert dataset, their ResNet50 model only achieved 83.5 AUC on the same evaluation set.

6.3.2 Retrieval Tasks

We present the zero-shot image-image and text-image retrieval results in Table 6.3. For the image-image retrieval setting, we present additional results from fine-tuning our pretrained model on all CheXpert training data, and use them as “upper bounds” of the results obtained from the use of supervised labels. We find that: 1) using ImageNet pretrained CNN weights in a zero-shot image retrieval setting is only better than random guess by small margins;



(a) ImageNet Pretraining

(b) ConVIRT Pretraining

Figure 6.3: t-SNE visualizations of encoded image representations from ImageNet and ConVIRT pretraining.

2) all in-domain pretrained CNN weights achieve much better retrieval performance than ImageNet weights; and 3) our proposed ConVIRT pretraining achieves the best overall retrieval results on all metrics. We find that while Contrastive-Binary performs notably better than other baselines and approaches that of the ConVIRT results in the image-image retrieval setting, its text-image retrieval results are far from ConVIRT pretraining. We conjecture that the lack of an explicit similarity-based loss function in the Contrastive-Binary baseline model results in misaligned representations in the image and text space, leading to poor results in text-image retrieval.

To understand how well ConVIRT pretraining helps separate images from different abnormality categories in its encoding space, in Figure 6.3 we present t-SNE plots (Maaten and Hinton, 2008) of candidate images in the CheXpert 8x200 dataset for five selected categories, from the ImageNet pretrained CNN encoder and the ConVIRT pretrained encoder. It is worth noting that clustering images in our setting is much more challenging than that in the general object classification setting due to the high inter-class similarity of the medical images. Nevertheless we find that ConVIRT pretraining achieves a better clustering of the images in the t-SNE plots. On the other hand, the lack of clear separations between groups

Settings	RSNA Linear (1%, AUC)	Image-Image (Prec@10)	Text-Image (Prec@10)
ConVIRT (default)	90.7	42.9	57.5
$\tau = 0.01$	90.7	40.5	21.0
$\tau = 1$	89.6	25.0	31.0
bs = 16	90.3	40.0	55.8
bs =128	90.3	39.3	50.3
linear proj.	90.6	40.8	55.8

Table 6.4: Results with different hyperparameters for ConVIRT pretraining. Results are shown for the RSNA 1% data linear evaluation, image-image and text-image retrieval tasks. Our default model uses $\tau = 0.1$, bs $= 32$ and non-linear projections.

suggests room for further improvement.

6.4 Analysis and Discussion

We now present analysis and discussion about factors that influence the performance of ConVIRT pretraining and its comparisons to existing image-only unsupervised pretraining methods.

6.4.1 Hyperparameter Analysis

Similar to previous work on unsupervised image representation learning (Chen et al., 2020a; He et al., 2020), we first find that the effectiveness of ConVIRT pretraining method is most sensitive to the temperature value τ. As shown in Table 6.4, using a temperature much lower than the ideal value ($\tau = 0.01$) hurts the retrieval results, and a temperature much larger ($\tau = 1$) notably hurts the performance on all tasks. Unlike previous work, we find that using a smaller or larger batch size hurts the retrieval performance, but neither setup brings substantial impact to the classification results. Lastly, we find that replacing the non-linear projection heads in g_v and g_u with linear layers hurts the retrieval results moderately, suggesting worse representations. However, this is again not reflected notably in the RSNA classification results.

6.4.2 Comparisons to Image-only Contrastive Learning

ConVIRT shows superior results against baselines in evaluation, but an important question remains as to how it compares against existing image-only contrastive visual representation learning methods. We study this by pretraining image encoders with two popular such methods, SimCLR (Chen et al., 2020a) and MoCo v2 (Chen et al., 2020b). For a fair comparison, in both experiments we use the exact same set of images from the MIMIC-CXR dataset that we use in the pretraining of our method and the baselines. Our settings for each method are:

- **SimCLR**: We use the open PyTorch implementation available at https://github.com/sthalles/SimCLR. For image encoder we use ResNet50. We use cosine similarity in the loss function, set the temperature value to 0.1 and set the output dimension to 128. We use the default image augmentation functions in the paper except for the *color jittering* transformation where we set the saturation and hue adjustment to 0 due to the monochrome nature of our medical images. For training, we use the Adam optimizer with an initial learning rate of 3e-4 and weight decay of 1e-4. We set batch size to 128 and run training on a single GPU card for 100 epochs, as we find that increasing the batch size or number of epochs does not lead to improved results. We use the default settings for all other parameters.

- **MoCo v2**: We use the authors' original PyTorch implementation available at https://github.com/facebookresearch/moco. For image encoder we use ResNet50. We follow the default MoCo v2 setting and use a temperature value of 0.07 and an output dimension of 128. Similarly, we adopt the default image augmentation functions except for the *color jittering* transformation where we set the saturation and hue adjustment to 0. For training, we use the SGD optimizer with a learning rate of 0.0075 and weight decay of 1e-4. We use a batch size of 64 and a queue size of 4096, and run parallel training on two GPU cards for 100 epochs, as we find that further increasing the batch size or number of epochs does not lead to improved results. During training, we anneal the learning rate by a factor of 0.1 at the 60th and 80th epochs.

Method	RSNA Linear (1%, AUC)	CheXpert Linear (1%, AUC)	Image-Image (Prec@10)
ImageNet	82.8	75.7	14.4
SimCLR	86.3	77.4	17.6
MoCo v2	86.6	81.3	20.6
ConVIRT	90.7	85.9	42.9

Table 6.5: Comparisons of ConVIRT to image-only unsupervised image representation learning approaches. For RSNA and CheXpert we present AUC scores under linear classification with 1% training data.

Experimental Results. We again run classification and retrieval experiments with both SimCLR and MoCo v2 pretraining, and present the results in Table 6.5. We find that compared to ImageNet initialization, both contrastive methods lead to marginal to moderate improvements on the classification and retrieval tasks, suggesting some domain adaptation from the use of in-domain images. However, the relatively small amount of improvement on all tasks suggests that the pretraining procedure fails to make efficient use of the input medical images. In contrast, our ConVIRT pretraining method substantially outperforms both methods on all tasks. This difference can be explained by the different objectives used in previous methods and our method: the high inter-class similarity of medical images has resulted in very similar sampled views from images of different categories, leading to inefficient contrastive learning in SimCLR and MoCo; the contrast between correct and randomly sampled image-text pairs in ConVIRT, on the other hand, is not impacted by the inter-class similarity and makes efficient use of additional information in the textual data.

Visualization. To understand the representational difference that has led to this difference in performance, for all four initialization methods, we visualize in Figure 6.4 the saliency maps (Simonyan et al., 2014) corresponding to the correct class on sampled images from the CheXpert dataset. Models for all initialization methods are trained with 1% CheXpert training data under the linear classification setting (with pretrained CNN weights frozen). We find that ImageNet pretraining has led to models that focus on trivial visual features that are mostly irrelevant to the task, and that the model with ConVIRT pretrained weights has

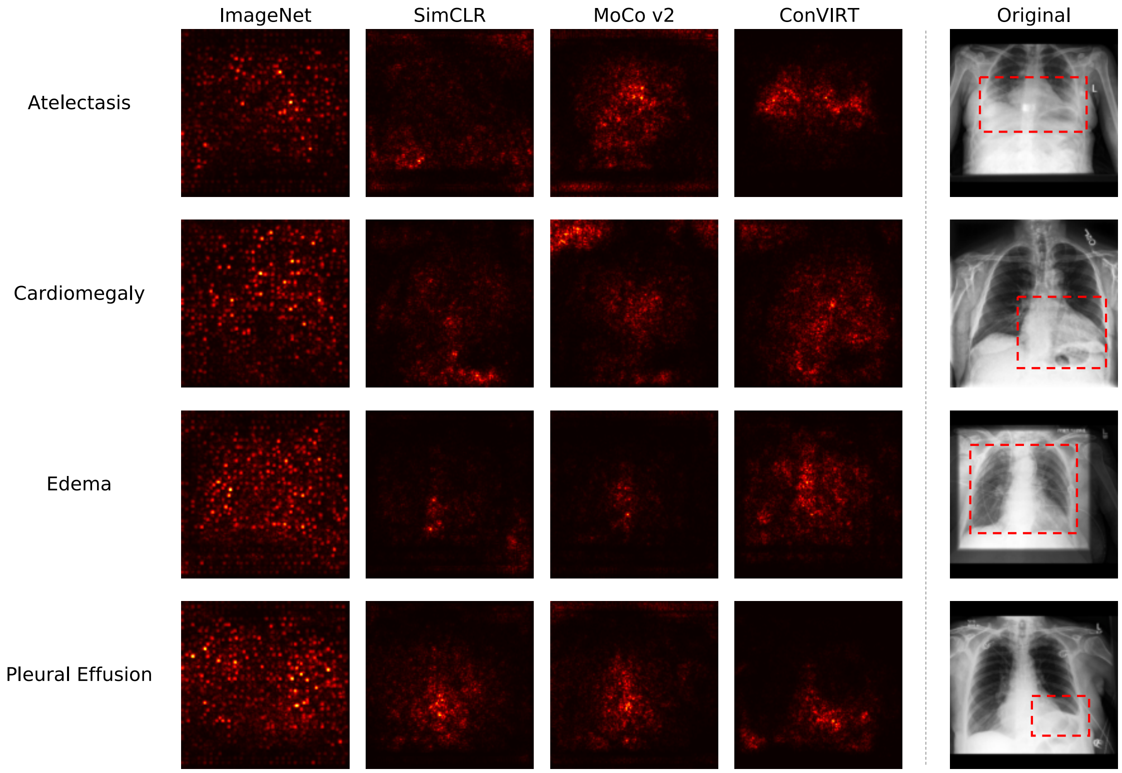
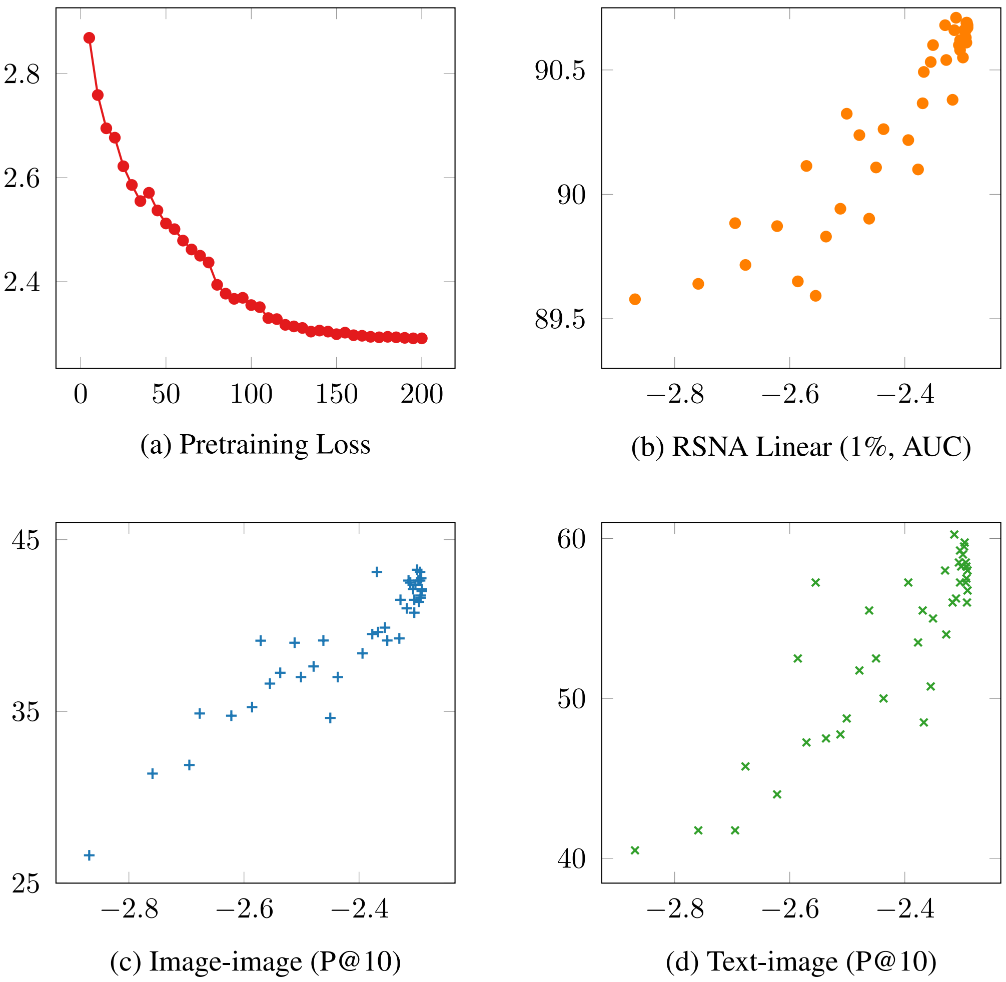


Figure 6.4: Saliency maps on sampled images for 4 abnormality categories in the CheXpert dataset. For each image we present maps for ImageNet, SimCLR, MoCo v2 and our ConVIRT initializations. Ground truth regions that are indicative of the abnormalities are shown as red boxes in the original images on the right, and are seen to most closely match the regions found by ConVIRT.

focused on much more relevant areas than those with SimCLR and MoCo v2 pretraining, suggesting more effective representation learning. For example, for *atelectasis*, while the ConVIRT model has correctly focused on the bottom of the lung regions, the SimCLR model has much more scattered focus and the MoCo model has incorrectly focused on the heart region.

6.4.3 Correlation between Contrastive Loss and End Task Results

To understand the relation between a model's performance on the ConVIRT pretraining task and its performance on the downstream tasks, we ran an analysis where for every 5 epochs during the pretraining, we transferred the pretrained checkpoint to the downstream tasks



(a) Pretraining Loss

(b) RSNA Linear (1%, AUC)

(c) Image-image (P@10)

(d) Text-image (P@10)

Figure 6.5: Correlation between ConVIRT loss and end task performance. (a) shows pretraining validation loss at different epochs; (b)-(d) shows correlation between the pretraining loss and the performance of three end tasks. For (a) the x-axis shows the training epoch number, and for (b)-(d) the x-axis shows the negative value of the pretraining loss (i.e., $-\mathcal{L}$) on a held-out validation set.

and evaluate its performance. The pretraining was run for a total of 200 epochs, and 40 points were obtained with varying validation loss and end task results. Figure 6.5 presents the models' validation loss on the pretraining task and their achieved performance on the RSNA 1% data linear evaluation and the two retrieval tasks. For all three tasks, we find a clear positive correlation between the pretraining performance and the end task performance. This corroborates that by learning with the ConVIRT objective, the image encoder learns gradually improved representations for the end tasks, and suggests that further improvement on the pretraining task may have positive impact on the end task performance.

6.5 Summary and Future Directions

In this chapter we presented ConVIRT, an unsupervised method for learning medical visual representations from naturally occurring pairing of images and text. ConVIRT learns high-quality representations of medical images, by contrasting the image representations with the paired text data via a bidirectional objective between the two modalities. Its success relies on the efficient contrastive objective and high-quality pretrained text representations.

We empirically showed that on 4 medical image classification tasks, ConVIRT outperformed other strong in-domain initialization methods that also use the text data, and led to representations with markedly higher quality. Most notably, compared to ImageNet pretraining, ConVIRT is able to achieve the same level of classification accuracy with an order of magnitude less labeled data. We further collected two zero-shot medical image retrieval datasets, an image-image and a text-image retrieval dataset, and showed on these datasets that ConVIRT has led to image encoders that are more effective at retrieval tasks too. We showed through in-depth analysis that ConVIRT substantially outperformed existing image view-based contrastive learning methods such as SimCLR and MoCo on medical imaging tasks, and that a model's performance on the ConVIRT pretraining task positively correlates with its performance on end tasks.

Our experiments and analysis also revealed the following directions to further extend our work:

- **Developing better negative sampling strategies for ConVIRT**. Similar to SimCLR (Chen et al., 2020a), ConVIRT uses in-batch negative examples which are constructed via uniform random sampling. This strategy might be suboptimal and may introduce an undesirably large ratio of false negative examples. For example, a common textual description like "*no abnormality is seen*" might occur multiple times in the same batch, polluting the contrastive learning process. This might help explain why further increasing the batch size of ConVIRT does not help learn better representations (see Section 6.4.1). While we show that ConVIRT can still benefit from the large scale of pretraining data and learn useful representations, developing better negative sampling strategies that can help reduce false negative pairs may further enhance ConVIRT's performance.

- **Searching for more effective learning components**. Another direction for improving ConVIRT is to search for better alternatives to its individual learning components. For example, ConVIRT uses a simple sentence-level uniform sampling function as its text transformation function. Better transformation functions that can change the surface text form without changing its semantic meanings might serve as additional data augmentations and may improve the learning performance.

- **Developing methods for learning representations of other types of medical images**. We have designed ConVIRT to take one-to-one image-text pairs as input. However, this setting might not be flexible enough for some other types of medical images. For example, computed tomography (CT), another common medical imaging technique, often produces hundreds of 2D images representing scanning results at different depth of a human body. This leads to a hundred-to-one mapping from the images to the textual report. How to generalize the framework of ConVIRT to this setting and learn meaningful visual representations from the textual knowledge remains an open question.

- **Scaling ConVIRT to more paired data covering more medical specialties**. We have pretrained separate image encoders with ConVIRT on two individual image-text paired datasets covering different domains (i.e., chest and bone). But can the pretrained encoders in ConVIRT benefit from even larger scale of paired data? Can we pretrain a single image encoder on aggregated data covering many different medical specialties and still achieve outstanding transfer learning performance on all downstream tasks? These are all interesting research questions for future studies.

In relevance to our ConVIRT framework, several papers concurrent to ours have studied the problem of learning visual representations from text data and proposed different strategies for this task (Sariyildiz et al., 2020; Desai and Johnson, 2021). Most notably, since the original release of our work, ConVIRT has been applied at much larger scales in several general visual recognition studies, including the CLIP model (Radford et al., 2021), which uses a simplified version of the ConVIRT approach, and the ALIGN model (Jia et al., 2021). These successful applications have confirmed that ConVIRT is a promising strategy for learning visual representations from human-written descriptive text, and that it has the potential to further advance the state of the art for general visual recognition tasks.

Chapter 7

Conclusions

In this dissertation, we focused on the transformative role that deep language understanding plays in helping us understand and generate medical text. We have shown this via several distinctive perspectives:

In Chapter 3, we focused on the understanding of biomedical scientific text, and studied the challenging problem of extracting structured relational knowledge from this text. We introduced a novel linguistically-motivated neural architecture that learns to represent a relation encoded in a sentence by exploiting the syntactic structure of the sentence. We showed that this model has the key advantages of being robust to the long context where biomedical relations are commonly found, and being more computationally efficient compared to recursive architectures. On several widely used benchmark datasets, we showed that our model not only demonstrates superior performance for biomedical relation extraction, but also achieves a new state of the art on relation extraction over general-domain text.

In Chapter 4, we focused on the clinical report text, and more specifically, the radiology report text used in medical imaging studies. We studied the problem of summarizing long, detailed radiology reports written by radiologists into more succinct summary statements. On real-world radiology report datasets collected from hospitals, we demonstrated how a neural abstractive summarization model that is tailored to the structure of radiology reports outperforms traditional extractive models based on sparse modeling of the report text, and generates fluent summaries that overlap notably with human-written summaries. We further showed via radiologist evaluation that the predictions from our model demonstrate substantial clinical potential.

In Chapter 5, we extended our study in Chapter 4, by identifying a crucial shortcoming of our neural summarization model, that the generated summaries tend to be factually incomplete and incorrect. We addressed this problem by proposing a new information extraction-based framework that evaluates a generated summary by comparing its factual content with the reference. We further presented a reinforcement learning-based method that optimizes this new metric, and demonstrated via both automatic and human evaluation that this new method has led to radiology summaries that are more correct and have higher clinical validity. Our study provides novel methods to optimize the factual correctness of a neural text summarization model, and the resulting system has the potential to save healthcare providers from repetitive labor and to improve clinical communications.

In Chapter 6, we connected the text and image modalities in medicine, by focusing on transferring the knowledge that we learn from text understanding to understanding medical images. We presented a novel unsupervised framework that improves medical image understanding by contrasting an image with both true and randomly paired report text. Compared to existing medical imaging work that relies on either ImageNet pretraining or extracting labels from the unstructured report text with patterns, our framework jointly models the image and text in an end-to-end manner, is fully unsupervised and is agnostic to the medical imaging domain. On multiple medical image classification and retrieval datasets covering two medical specialties, we showed that the proposed method improves the accuracy of the learned image encoders, and substantially outperforms existing methods based on ImageNet pretraining or image-only contrastive learning. Our study pioneers the general direction of cross-modality contrastive pretraining, and opens up new directions for effectively utilizing large-scale medical text data for understanding images and other clinical data.

Altogether, this dissertation has conveyed the following key insights: First, understanding and generating medical text has the potential to transform medicine by helping us obtain actionable biomedical knowledge, by improving communications in healthcare, and by improving the understanding of other data modalities in medicine. Second, deep language understanding techniques based on dense vector representations of text outperform traditional rule-based or sparse feature-based methods on all the tasks studied. Third, the complex, noisy nature of medical text and the low tolerance for errors in this application domain have provided us with unique opportunities to improve the robustness and efficiency of our techniques.

Lastly, we hope to briefly highlight the following important directions for future work:

- **Efficient learning methods for understanding medical data.** Structured annotations of medical data are often extremely expensive to obtain. Moreover, the low tolerance for error in this domain means that we cannot simply deploy an inaccurate system and wait to collect real-world feedback for it. Thus, it is especially critical to develop methods that can efficiently learn from the structure of the data, rather than learning from scratch from human annotations or engagements. The recent progress on self-supervised learning for language and image understanding (Devlin et al., 2019; Lee et al., 2019; Chen et al., 2020a) is in line with this direction of research.

- **Models that represent and ground predictions in real-world knowledge.** We have shown via our summarization studies that the current generation of neural models lacks necessary real-world knowledge. This is especially critical for medicine as medical decisions are often made with a complex body of knowledge, which for human beings typically takes years of training to acquire. While neural models are capable of representing some of this knowledge implicitly in their parameter space, it is imperative for us to redesign our model architectures, such that they can represent and ground their predictions in real-world knowledge in an explicit manner. This explicit grounding will enable us to deploy these models with confidence.

- **Methods for understanding noisy medical text.** Medical data is often noisy when produced in real clinical settings. Yet, the particularly low tolerance for error requires us to develop methods that are more robust to noise, such as incomplete text or spelling errors. Furthermore, it would be exciting to develop ambient intelligence technologies that can understand noisy medical conversations and engage with users when necessary, such that we can ultimately free healthcare providers from the repetitive, tedious documentation work.

- **Learning with medical data of heterogeneous modalities.** We have showcased the benefits of end-to-end joint modeling of medical text and image data. However, images are not the only modality other than text. It therefore would be exciting to develop methods that combine the understanding of other forms of medical data, such as semi-structured tabular data or genomic data, to ultimately improve the quality of healthcare.

Bibliography

Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative Ophthalmology & Visual Science*, 57(13):5200–5206.

Heike Adel, Benjamin Roth, and Hinrich Schütze. 2016. Comparing convolutional neural networks to traditional models for slot filling. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Elske Ammenwerth and H-P Spötl. 2009. The time needed for clinical documentation versus direct patient care. *Methods of Information in Medicine*, 48(01):84–91.

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Gordana Apic, Tijana Ignjatovic, Scott Boyer, and Robert B Russell. 2005. Illuminating drug discovery with biological pathways. *FEBS Letters*, 579(8):1872–1877.

Alain Auger and Caroline Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 14(1):1.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *The 2015 International Conference on Learning Representations*.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ashwin Belle, Raghuram Thiagarajan, SM Soroushmehr, Fatemeh Navidi, Daniel A Beard, and Kayvan Najarian. 2015. Big data analytics in healthcare. *BioMed Research International*, 2015.

Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Jan ML Bosmans, Joost J Weyler, Arthur M De Schepper, and Paul M Parizel. 2011. The radiology report as seen by radiologists and referring clinicians: Results of the COVER and ROVER surveys. *Radiology*, 259(1):184–195.

Barton F Branstetter. 2009. *Practical imaging informatics: foundations and applications for PACS professionals*. Springer.

Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676.

Alex AT Bui, Denise R Aberle, and Hooshang Kangarloo. 2007. TimeLine: visualizing integrated patient records. *IEEE Transactions on Information Technology in Biomedicine*, 11(4):462–473.

Razvan Bunescu and Raymond J Mooney. 2005a. Subsequence kernels for relation extraction. In *Advances in Neural Information Processing Systems*.

Razvan C Bunescu and Raymond J Mooney. 2005b. A shortest path dependency kernel for relation extraction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP 2005)*.

Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *Journal of Biomedical Informatics*, 44(2):277–288.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2017. Faithful to the original: Fact aware neural abstractive summarization. *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-2017)*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*.

NV Chandrasekharan, Hu Dai, K Lamar Turepu Roos, Nathan K Evanson, Joshua Tomsik, Terry S Elton, and Daniel L Simmons. 2002. Cox-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: cloning, structure, and expression. *Proceedings of the National Academy of Sciences*, 99(21):13926–13931.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ping Chen and Rakesh Verma. 2006. A query-based medical information summarization system using ontology knowledge. In *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pages 37–42. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 2018 Annual Meeting of the Association of Computational Linguistics (ACL 2018)*.

Kevin Bretonnel Cohen and Dina Demner-Fushman. 2014. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company.

Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. 2016. Modelling radiological language with bidirectional long short-term memory networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI)*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory Transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference (DUC)*.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan ODonoghue, Daniel

Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Karan Desai and Justin Johnson. 2021. VirTex: Learning visual representations from textual annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob S Elkins, Carol Friedman, Bernadette Boden-Albala, Ralph L Sacco, and George Hripcsak. 2000. Coding neuroradiology reports for the northern Manhattan stroke study: a comparison of natural language processing and manual review. *Computers and Biomedical Research*, 33(1):1–10.

Günes Erkan and Dragomir R Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Patrick Ernst. 2017. Biomedical knowledge base construction from text and its applications in knowledge-based systems. *PhD Dissertation, Saarland University and State Library*.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*.

Carol Friedman, George Hripcsak, William DuMouchel, Stephen B Johnson, and Paul D Clayton. 1995. Natural language processing in an operational clinical information system. *Natural Language Engineering*, 1(1):83–108.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1):1–66.

Yael Garten, Adrien Coulet, and Russ B Altman. 2010. Recent progress in automatically extracting information from the pharmacogenomic literature. *Pharmacogenomics*, 11(10):1467–1489.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*.

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie M Strassel. 2017. Overview of linguistic resources for the TAC-KBP 2017 evaluations: Methodologies and results. In *Proceedings of the Text Analysis Conference (TAC)*.

Daniel J Goff and Thomas W Loehfelm. 2018. Automated radiology report summarization using an open-source natural language processing pipeline. *Journal of Digital Imaging*, 31(2):185–192.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 19).*

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*.

Ralph Gross, Michael Bett, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Towards a multimodal meeting record. In *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME2000).*

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).*

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779.*

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. In *Proceedings of the 16th European Conference on Computer Vision (ECCV).*

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Saeed Hassanpour and Curtis P Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66:29–39.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shexia He, Zuchao Li, Hai Zhao, and Hongxiao Bai. 2018. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning (ICML)*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*.

Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. 2002. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Research*, 30(1):163–165.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

Jamie S Hirsch, Jessica S Tanenbaum, Sharon Lipsky Gorman, Connie Liu, Eric Schmitz, Dritan Hashorva, Artem Ervits, David Vawdrey, Marc Sturm, and Noémie Elhadad. 2015. HARVEST, a longitudinal patient record summarizer. *Journal of the American Medical Informatics Association*, 22(2):263–274.

Markus Hoffmann, Kirstin Mösbauer, Heike Hofmann-Winkler, Artur Kaul, Hannah Kleine-Weber, Nadine Krüger, Nils C Gassen, Marcel A Müller, Christian Drosten, and Stefan Pöhlmann. 2020. Chloroquine does not inhibit infection of human lung cells with SARS-CoV-2. *Nature*, 585(7826):588–590.

George Hripcsak, John HM Austin, Philip O Alderson, and Carol Friedman. 2002. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*, 224(1):157–163.

George Hripcsak, Gilad J Kuperman, and Carol Friedman. 1998. Extracting findings from narrative reports: software transferability and sources of physician disagreement. *Methods of Information in Medicine*, 37(01):01–07.

Lawrence Hunter and K Bretonnel Cohen. 2006. Biomedical language processing: What's beyond PubMed? *Molecular Cell*, 21(5):589–594.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2020. Probing text models for common ground with visual representations. *arXiv preprint arXiv:2005.00619*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019)*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Daniel Jurafsky and James H Martin. 2020. Speech and language processing, Chapter 14: Dependency parsing. *Third Edition Draft of December 30, 2020*. `https://web.stanford.edu/~jurafsky/slp3/14.pdf`.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 Interactive Poster and Demonstration Sessions*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Teri E Klein, Jeffrey T Chang, Mildred K Cho, Katrina L Easton, Ray Fergerson, Micheal Hewett, Zhen Lin, Y Liu, S Liu, DE Oliver, et al. 2001. Integrating genotype and phenotype information: An overview of the PharmGKB project. *The Pharmacogenomics Journal*, 1(3):167–170.

Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, et al. 2017. Overview of the BioCreative VI chemical-protein interaction track. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

M Lafortune, G Breton, and JL Baudouin. 1988. The radiological report: What is useful for the referring physician? *Canadian Association of Radiologists*, 39(2):140–143.

Curtis P Langlotz. 2006. RadLex: A new method for indexing online educational materials. *RadioGraphics*, 26(6):1595–1597.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sangrak Lim and Jaewoo Kang. 2018. Chemical–gene relation extraction using recursive neural network. *Database*, 2018.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: ACL Workshop*.

Chin-Yew Lin and Eduard Hovy. 2002. From single to multi-document summarization. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019a. Clinically accurate chest X-ray report generation. *arXiv preprint arXiv:1904.02633*.

Hongfang Liu and Carol Friedman. 2004. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. *Studies in Health Technology and Informatics*, 107(Pt 1):639–643.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *The 2018 International Conference for Learning Representations (ICLR)*.

Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, and Tianfu Wang. 2019b. Deep learning in medical ultrasound analysis: a review. *Engineering*, 5(2):261–275.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. 2015. A dependency-based neural network for relation classification. *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. 2019. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954.

Paul McNamee and Hoa Trang Dang. 2009. Overview of the TAC 2009 knowledge base population track. In *Proceedings of the Text Analysis Conference (TAC)*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *The 2017 International Conference on Learning Representations (ICLR)*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the*

47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2016a. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-2017)*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016b. Abstractive text summarization using sequence-to-sequence RNNs and beyond. *The SIGNLL Conference on Computational Natural Language Learning (CoNLL), 2016*.

Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. 2013. Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajic, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*.

Jekaterina Novikova, Ondej Duek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations (ICLR)*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence N-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*.

Yifan Peng and Zhiyong Lu. 2017. Deep learning for extracting protein-protein interactions from biomedical literature. In *Proceedings of the BioNLP 2017 Workshop*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *Proceedings of the 18th BioNLP Workshop and Shared Task*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bethany Percha and Russ B Altman. 2015. Learning the structure of biomedical relationships from unstructured text. *PLOS Computational Biology*, 11(7):e1004216.

Bethany Percha and Russ B Altman. 2018. A global network of biomedical relationships derived from text. *Bioinformatics*, 34(15):2614–2624.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Laura Plaza, Alberto Díaz, and Pablo Gervás. 2008. Concept-graph based biomedical automatic summarization using ontologies. In *Proceedings of the 3rd Textgraphs Workshop on Graph-based Algorithms for Natural Language Processing*.

Laura Plaza, Mark Stevenson, and Alberto Díaz. 2012. Resolving ambiguity in biomedical text to improve summarization. *Information Processing & Management*, 48(4):755–766.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

Hoifung Poon, Chris Quirk, Charlie DeZiel, and David Heckerman. 2014. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*.

Seth M Powsner and Edward R Tufte. 1997. Summarizing clinical psychiatric data. *Psychiatric Services*, 48(11):1458–1460.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Dan Sullivan, Chunhong Mao, Chunxia Wang, Bruno Sobral, Jun'ichi Tsujii, and Sophia Ananiadou. 2012. Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011. *BMC Bioinformatics*, 13:1–26.

Changqin Quan, Meng Wang, and Fuji Ren. 2014. An unsupervised text mining method for relation extraction from biomedical literature. *PloS One*, 9(7):e102039.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. 2019. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*.

Wullianallur Raghupathi and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):1–10.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. 2018a. MURA: Large dataset for abnormality detection in musculoskeletal radiographs. In *Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL)*.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. 2018b. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Medicine*, 15(11):e1002686.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *The 2016 International Conference on Learning Representations (ICLR)*.

Lawrence H Reeve, Hyoil Han, Saya V Nagori, Jonathan C Yang, Tamara A Schwimmer, and Ari D Brooks. 2006. Concept frequency distribution in biomedical text summarization. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Xiang Ren, Jiaming Shen, Meng Qu, Xuan Wang, Zeqiu Wu, Qi Zhu, Meng Jiang, Fangbo Tao, Saurabh Sinha, David Liem, et al. 2017. Life-iNet: A structured network-based knowledge exploration and analytics system for life sciences. In *Proceedings of ACL 2017, System Demonstrations*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sebastian Riedel and Andrew McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bryan Rink and Sanda Harabagiu. 2010. UTD: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Devendra Singh Sachan, Yuhao Zhang, Peng Qi, and William Hamilton. 2020. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*.

Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning visual representations with caption annotations. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*.

Kamal Sarkar. 2009. Using domain knowledge for text summarization in medical domain. *International Journal of Recent Trends in Engineering*, 1(1):200.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *The 2017 Annual Meeting of the Association of Computational Linguistics (ACL 2017)*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *The 2017 International Conference on Learning Representations (ICLR)*.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. 2019. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041.

Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M Summers. 2015. Interleaved text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*.

Linda M Spooner and Kimberly A Pesaturo. 2013. The medical record. *Fundamental Skills for Patient Care in Pharmacy Practice*, 37.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proceedings of the 2004 International Conference on Information System Implementation and Modeling*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Caroline F Thorn, Teri E Klein, and Russ B Altman. 2013. PharmGKB: The pharmacogenomics knowledge base. In *Pharmacogenomics*, pages 311–320. Springer.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*.

Ngoc Thang Vu, Heike Adel, Pankaj Gupta, and Hinrich Schütze. 2016. Combining recurrent and convolutional neural networks for relation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*.

Michael M Wagner and William R Hogan. 1996. The accuracy of medication data in an outpatient electronic medical record. *Journal of the American Medical Informatics Association*, 3(3):234–244.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Linda Wang and Alexander Wong. 2020. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *arXiv preprint arXiv:2003.09871*.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 2018. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the BioCreative V chemical disease relation (CDR) task. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280.

David S Wishart, Craig Knox, An Chi Guo, Dean Cheng, Savita Shrivastava, Dan Tzur, Bijaya Gautam, and Murtaza Hassanali. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(suppl_1):D901–D906.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rmi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015a. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015b. Semantic relation classification via convolutional neural networks with simple negative sampling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015c. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Nan Duan, and Daxin Jiang. 2020. Syntax-enhanced pre-trained model. *arXiv preprint arXiv:2012.14116*.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. 2018. Breaking the softmax bottleneck: A high-rank RNN language model. In *The 2018 International Conference for Learning Representations (ICLR)*.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2014)*.

Luke Zettlemoyer and Michael Collins. 2007. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *The 2020 International Conference for Learning Representations (ICLR)*.

Yuhao Zhang, Arun Tejasvi Chaganty, Ashwin Paranjape, Danqi Chen, Jason Bolton, Peng Qi, and Christopher D Manning. 2016. Stanford at TAC KBP 2016: Sealing pipeline leaks and understanding Chinese. In *Proceedings of the Text Analysis Conference (TAC)*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018a. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020b. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020c. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yuhao Zhang, Peng Qi, and Christopher D Manning. 2018b. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Yongjun Zhu, Olivier Elemento, Jyotishman Pathak, and Fei Wang. 2019. Drug knowledge bases and their applications in biomedical informatics research. *Briefings in Bioinformatics*, 20(4):1308–1321.