

Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, Dan Jurafsky
Stanford University

{ysmiura, zyh, ebtsai, langlotz, jurafsky}@stanford.edu

Abstract

Neural image-to-text radiology report generation systems offer the potential to improve radiology reporting by reducing the repetitive process of report drafting and identifying possible medical errors. However, existing report generation systems, despite achieving high performances on natural language generation metrics such as CIDEr or BLEU, still suffer from incomplete and inconsistent generations. Here we introduce two new simple rewards to encourage the generation of factually complete and consistent radiology reports: one that encourages the system to generate radiology domain **entities** consistent with the reference, and one that uses natural language inference to encourage these entities to be described in **inferentially consistent** ways. We combine these with the novel use of an existing semantic equivalence metric (BERTScore). We further propose a report generation system that optimizes these rewards via reinforcement learning. On two open radiology report datasets, our system substantially improved the F₁ score of a clinical information extraction performance by +22.1 ($\Delta + 63.9\%$). We further show via a human evaluation and a qualitative analysis that our system leads to generations that are more factually complete and consistent compared to the baselines.

1 Introduction

An important new application of natural language generation (NLG) is to build assistive systems that take X-ray images of a patient and generate a textual report describing clinical observations in the images (Jing et al., 2018; Li et al., 2018; Liu et al., 2019; Boag et al., 2020; Chen et al., 2020). Figure 1 shows an example of a radiology report generated by such a system. This is a clinically important task, offering the potential to reduce radiologists’ repetitive work and generally improve clinical communication (Kahn et al., 2009).

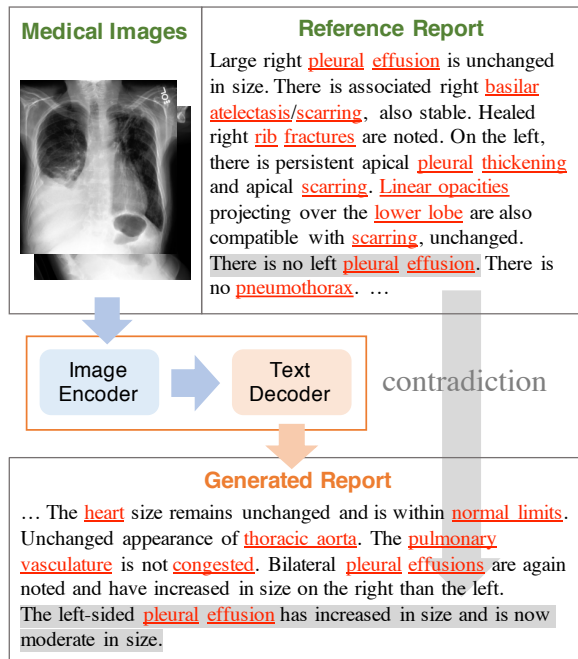


Figure 1: A (partial) example of a report generated from our system (with “...” representing abbreviated text). The system encodes images and generates text from that encoded representation. Underlined words are disease and anatomy entities. The shaded sentences are an example of a contradictory pair.

Automatic radiology report generation systems have achieved promising performance as measured by widely used NLG metrics such as CIDEr (Vedantam et al., 2015) and BLEU (Papineni et al., 2002) on several datasets (Li et al., 2018; Jing et al., 2019; Chen et al., 2020). However, reports that achieve high performance on these NLG metrics are not always factually complete or consistent. In addition to the use of inadequate metrics, the factual incompleteness and inconsistency issue in generated reports is further exacerbated by the inadequate training of these systems. Specifically, the standard teacher-forcing training algorithm (Williams and Zipser, 1989) used by most existing work can lead to a discrepancy between what the model sees during training and test time (Ran-

zato et al., 2016), resulting in degenerate outputs with factual hallucinations (Maynez et al., 2020). Liu et al. (2019) and Boag et al. (2020) have shown that reports generated by state-of-the-art systems still have poor quality when evaluated by their clinical metrics as measured with an information extraction system designed for radiology reports. For example, the generated report in Figure 1 is incomplete since it neglects an observation of *atelectasis* that can be found in the images. It is also inconsistent since it mentions *left-sided pleural effusion* which is not present in the images. Indeed, we show that existing systems are inadequate in factual completeness and consistency, and that an image-to-text radiology report generation system can be substantially improved by replacing widely used NLG metrics with simple alternatives.

We propose two new simple rewards that can encourage the factual completeness and consistency of the generated reports. First, we propose the Exact Entity Match Reward (fact_{ENT}) which captures the completeness of a generated report by measuring its coverage of entities in the radiology domain, compared with a reference report. The goal of the reward is to better capture disease and anatomical knowledge that are encoded in the entities. Second, we propose the Entailing Entity Match Reward ($\text{fact}_{\text{ENTNLI}}$), which extends fact_{ENT} with a natural language inference (NLI) model that further considers how inferentially consistent the generated entities are with their descriptions in the reference. We add NLI to control the overestimation of disease when optimizing towards fact_{ENT} . We use these two metrics along with an existing semantic equivalence metric, BERTScore (Zhang et al., 2020a), to potentially capture synonyms (e.g., “left and right” effusions are synonymous with “bilateral” effusions) and distant dependencies between diseases (e.g., a negation like “... but underlying consolidation or other pulmonary lesion not excluded”) that are present in radiology reports.

Although recent work in summarization, dialogue, and data-to-text generation has tried to address this problem of factual incompleteness and inconsistency by using natural language inference (NLI) (Falke et al., 2019; Welleck et al., 2019), question answering (QA) (Wang et al., 2020a), or content matching constraint (Wang et al., 2020b) approaches, they either show negative results or are not directly applicable to the generation of radiology reports due to a substantial task and do-

main difference. To construct the NLI model for $\text{fact}_{\text{ENTNLI}}$, we present a weakly supervised approach that adapts an existing NLI model to the radiology domain. We further present a report generation model which directly optimizes a Transformer-based architecture with these rewards using reinforcement learning (RL).

We evaluate our proposed report generation model on two publicly available radiology report generation datasets. We find that optimizing the proposed rewards along with BERTScore by RL leads to generated reports that achieve substantially improved performance in the important clinical metrics (Liu et al., 2019; Boag et al., 2020; Chen et al., 2020), demonstrating the higher clinical value of our approach. We make all our code and the expert-labeled test set for evaluating the radiology NLI model publicly available to encourage future research¹. To summarize, our contributions in this paper are:

1. We propose two simple rewards for image-to-text radiology report generation, which focus on capturing the factual completeness and consistency of generated reports, and a weak supervision-based approach for training a radiology-domain NLI model to realize the second reward.
2. We present a new radiology report generation model that directly optimizes these new rewards with RL, showing that previous approaches that optimize traditional NLG metrics are inadequate, and that the proposed approach substantially improves performance on clinical metrics (as much as $\Delta + 64.2\%$) on two publicly available datasets.

2 Related Work

2.1 Image-to-Text Radiology Report Generation

Wang et al. (2018) and Jing et al. (2018) first proposed multi-task learning models that jointly generate a report and classify disease labels from a chest X-ray image. Their models were extended to use multiple images (Yuan et al., 2019), to adopt a hybrid retrieval-generation model (Li et al., 2018), or to consider structure information (Jing et al., 2019). More recent work has focused on generating reports that are clinically consistent and accurate. Liu et al. (2019) presented a system that generates accurate reports by fine-tuning it with their Clinically

¹<https://github.com/yismiura/ifcc>

Coherent Reward. Boag et al. (2020) evaluated several baseline generation systems with clinical metrics and found that standard NLG metrics are ill-equipped for this task. Very recently, Chen et al. (2020) proposed an approach to generate radiology reports with a memory-driven Transformer. Our work is most related to Liu et al. (2019); their system, however, is dependent on a rule-based information extraction system specifically created for chest X-ray reports and has limited robustness and generalizability to different domains within radiology. By contrast, we aim to develop methods that improve the factual completeness and consistency of generated reports by harnessing more robust statistical models and are easily generalizable.

2.2 Consistency and Faithfulness in Natural Language Generation

A variety of recent work has focused on consistency and faithfulness in generation. Our work is inspired by Falke et al. (2019), Welleck et al. (2019), and Matsumaru et al. (2020) in using NLI to rerank or filter generations in text summarization, dialogue, and headline generations systems, respectively. Other attempts in this direction include evaluating consistency in generations using QA models (Durmus et al., 2020; Wang et al., 2020a; Maynez et al., 2020), with distantly supervised classifiers (Kryściński et al., 2020), and with task-specific content matching constraints (Wang et al., 2020b). Liu et al. (2019) and Zhang et al. (2020b) studied improving the factual correctness in generating radiology reports with rule-based information extraction systems. Our work mainly differs from theirs in the direct optimization of factual completeness with an entity-based reward and of factual consistency with a statistical NLI-based reward.

2.3 Image Captioning with Transformer

The problem of generating text from image data has been widely studied in the image captioning setting. While early work focused on combining convolutional neural network (CNN) and recurrent neural network (RNN) architectures (Vinyals et al., 2015), more recent work has discovered the effectiveness of using the Transformer architecture (Vaswani et al., 2017). Li et al. (2019) and Pan et al. (2020) introduced an attention process to exploit semantic and visual information into this architecture. Herdade et al. (2019), Cornia et al. (2020), and Guo et al. (2020) extended this architecture to learn geometrical and other relationships between input

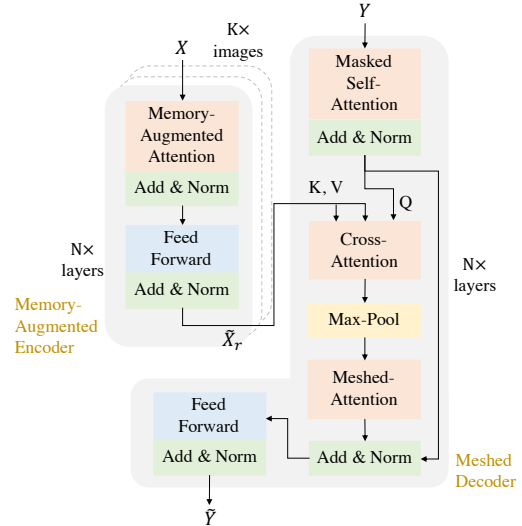


Figure 2: An overview of Meshed-Memory Transformer extended to multiple images.

regions. We find Meshed-Memory Transformer (Cornia et al., 2020) (\mathcal{M}^2 Trans) to be more effective in our radiology report generation task than the traditional RNN-based models and Transformer models (an empirical result will be shown in §4), and therefore use it as our base architecture.

3 Methods

3.1 Image-to-Text Radiology Report Generation with \mathcal{M}^2 Trans

Formally, given K individual images $x_{1\dots K}$ of a patient, our task involves generating a sequence of words to form a textual report \hat{y} , which describes the clinical observations in the images. This task resembles image captioning, except with multiple images as input and longer text sequences as output. We therefore extend a state-of-the-art image captioning model, \mathcal{M}^2 Trans (Cornia et al., 2020), with multi-image input as our base architecture. We first briefly introduce this model and refer interested readers to Cornia et al. (2020).

Figure 2 illustrates an overview of the \mathcal{M}^2 Trans model. Given an image x_k , image regions are first extracted with a CNN as $\mathbf{X} = \text{CNN}(x_k)$. \mathbf{X} is then encoded with a memory-augmented attention process $\mathcal{M}_{\text{mem}}(\mathbf{X})$ as

$$\mathcal{M}_{\text{mem}}(\mathbf{X}) = \text{Att}(W_q \mathbf{X}, \mathbf{K}, \mathbf{V}) \quad (1)$$

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{-d}\right) \mathbf{V} \quad (2)$$

$$\mathbf{K} = [W_k \mathbf{X}; \mathbf{M}_k] \quad (3)$$

$$\mathbf{V} = [W_v \mathbf{X}; \mathbf{M}_v] \quad (4)$$

where W_q, W_k, W_v are weights, $\mathbf{M}_k, \mathbf{M}_v$ are memory matrices, d is a scaling factor, and $[\ast; \ast]$

is the concatenation operation. $\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ is an attention process derived from the Transformer architecture (Vaswani et al., 2017) and extended to include memory matrices that can encode a priori knowledge between image regions. In the encoder, this attention process is a self-attention process since all of the query \mathbf{Q} , the key \mathbf{K} , and the value \mathbf{V} depend on \mathbf{X} . $\mathcal{M}_{\text{mem}}(\mathbf{X})$ is further processed with a feed forward layer, a residual connection, and a layer normalization to output $\tilde{\mathbf{X}}$. This encoding process can be stacked N times and is applied to K images, and n -th layer output of K image will be $\tilde{\mathbf{X}}_{n,K}$.

The meshed decoder first processes an encoded text \mathbf{Y} with a masked self-attention and further processes it with a feed forward layer, a residual connection, and a layer normalization to output $\ddot{\mathbf{Y}}$. $\ddot{\mathbf{Y}}$ is then passed to a cross attention $\mathcal{C}(\tilde{\mathbf{X}}_{n,K}, \ddot{\mathbf{Y}})$ and a meshed attention $\mathcal{M}_{\text{mesh}}(\tilde{\mathbf{X}}_{N,K}, \ddot{\mathbf{Y}})$ as

$$\mathcal{M}_{\text{mesh}}(\tilde{\mathbf{X}}_{N,K}, \ddot{\mathbf{Y}}) = \sum_n \alpha_n \mathcal{C}(\tilde{\mathbf{X}}_{n,K}, \ddot{\mathbf{Y}}) \quad (5)$$

$$\mathcal{C}(\tilde{\mathbf{X}}_{n,K}, \ddot{\mathbf{Y}}) = \max_K (\text{Att}(W_q \ddot{\mathbf{Y}}, W_k \tilde{\mathbf{X}}_{n,K}, W_v \tilde{\mathbf{X}}_{n,K})) \quad (6)$$

$$\alpha_n = \sigma (W_n [\mathbf{Y}; \mathcal{C}(\tilde{\mathbf{X}}_{n,K}, \ddot{\mathbf{Y}})] + b_n) \quad (7)$$

where \odot is element-wise multiplication, \max_K is max-pooling over K images, σ is sigmoid function, W_n is a weight, and b_n is a bias. The weighted summation in $\mathcal{M}_{\text{mesh}}(\tilde{\mathbf{X}}_{N,K}, \ddot{\mathbf{Y}})$ exploits both low-level and high-level information from the N stacked encoder. Differing from the self-attention process in the encoder, the cross attention uses a query that depends on \mathbf{Y} and a key and a value that depend on \mathbf{X} . $\mathcal{M}_{\text{mesh}}(\tilde{\mathbf{X}}_{N,K}, \ddot{\mathbf{Y}})$ is further processed with a feed forward layer, a residual connection, and a layer normalization to output $\tilde{\mathbf{Y}}$. As like in the encoder, the decoder can be stacked N times to output $\tilde{\mathbf{Y}}_N$. $\tilde{\mathbf{Y}}_N$ is further passed to a feed forward layer to output report \hat{y} .

3.2 Optimization with Factual Completeness and Consistency

3.2.1 Exact Entity Match Reward (fact_{ENT})

We designed an F-score entity match reward to capture factual completeness. This reward assumes that entities encode disease and anatomical knowledge that relates to factual completeness. A named entity recognizer is applied to \hat{y} and the corresponding reference report y . Given entities E_{gen} and E_{ref} recognized from y_{gen} and y_{ref} respectively,

precision (pr) and recall (rc) of entity match are calculated as

$$\text{pr}_{\text{ENT}} = \frac{\sum_{e \in E_{\text{gen}}} \delta(e, E_{\text{ref}})}{|E_{\text{gen}}|} \quad (8)$$

$$\text{rc}_{\text{ENT}} = \frac{\sum_{e \in E_{\text{ref}}} \delta(e, E_{\text{gen}})}{|E_{\text{ref}}|} \quad (9)$$

$$\delta(e, E) = \begin{cases} 1, & \text{for } e \in E \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

The harmonic mean of precision and recall is taken as fact_{ENT} to reward a balanced match of entities. We used Stanza (Qi et al., 2020) and its clinical models (Zhang et al., 2020c) as a named entity recognizer for radiology reports. For example in the case of Figure 1, the common entities among the reference report and the generated report are *pleural* and *effusion*, resulting to $\text{fact}_{\text{ENT}} = 33.3$.

3.2.2 Entailing Entity Match Reward ($\text{fact}_{\text{ENTNLI}}$)

We additionally designed an F-score style reward that expands fact_{ENT} with NLI to capture factual consistency. NLI is used to control the overestimation of disease when optimizing towards fact_{ENT} . In $\text{fact}_{\text{ENTNLI}}$, δ in Eq. 10 is expanded to

$$\phi(e, E) = \begin{cases} 1, & \text{for } e \in E \wedge \text{NLI}_e(\mathbf{P}, h) \notin \text{contradiction} \\ 1, & \text{for } \text{NLI}_e(\mathbf{P}, h) = \text{entailment} \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

$$\text{NLI}_e(\mathbf{P}, h) = \text{nli}(\hat{p}, h) \text{ where } \hat{p} = \arg \max_{p \in \mathbf{P}} \text{sim}(h, p) \quad (12)$$

where h is a sentence that includes e , \mathbf{P} is all sentences in a counter part text (if h is a sentence in a generated report, \mathbf{P} is all sentences in the corresponding reference report), $\text{nli}(*, *)$ is an NLI function that returns an NLI label which is one of {entailment, neutral, contradiction}, and $\text{sim}(*, *)$ is a text similarity function. We used BERTScore (Zhang et al., 2020a) as $\text{sim}(*, *)$ in the experiments (the detail of BERTScore can be found in Appendix A). The harmonic mean of precision and recall is taken as $\text{fact}_{\text{ENTNLI}}$ to encourage a balanced factual consistency between a generated text and the corresponding reference text. For example in the case of Figure 1, the sentence ‘‘The left-sided pleural effusion has increased in size and is now moderate in size.’’ will be contradictory to ‘‘There is no left pleural effusion.’’ resulting in *pleural* and *effusion* being rejected in y_{gen} .

3.2.3 Joint Loss for Optimizing Factual Completeness and Consistency

We integrate the proposed factual rewards into self-critical sequence training (Rennie et al., 2017). An RL loss \mathcal{L}_{RL} is minimized as the negative expectation of the reward r . The gradient of the loss is estimated with a single Monte Carlo sample as

$$r_{\theta} \mathcal{L}_{RL}(\theta) = -r_{\theta} \log P_{\theta}(\hat{y}_{sp}/x_{1...K}) (r(\hat{y}_{sp}) - r(\hat{y}_{gd})) \quad (13)$$

where \hat{y}_{sp} is a sampled text and \hat{y}_{gd} is a greedy decoded text. Paulus et al. (2018) and Zhang et al. (2020b) have shown that a generation can be improved by combining multiple losses. We combine a factual metric loss with a language model loss and an NLG loss as

$$L = \lambda_1 L_{NLL} + \lambda_2 L_{RL_NLG} + \lambda_3 L_{RL_FACT} \quad (14)$$

where \mathcal{L}_{NLL} is a language model loss, \mathcal{L}_{RL_NLG} is the RL loss using an NLG metric (e.g., CIDEr or BERTScore), \mathcal{L}_{RL_FACT} is the RL loss using a factual reward (e.g., fact_{ENT} or fact_{ENTNLI}), and λ are scaling factors to balance the multiple losses.

3.3 A Weakly-Supervised Approach for Radiology NLI

We propose a weakly-supervised approach to construct an NLI model for radiology reports. (There already exists an NLI system for the medical domain, MedNLI (Romanov and Shivade, 2018), but we found that a model trained on MedNLI does not work well on radiology reports.) Given a large scale dataset of radiology reports, a sentence pair is sampled and filtered with weakly-supervised rules. The rules are prepared to extract a randomly sampled sentence pair (s_1 and s_2) that are in an entailment, neutral, or contradiction relation. We designed the following 6 rules for weak-supervision.

Entailment 1 (E1) (1) s_1 and s_2 are semantically similar and (2) NE of s_2 is a subset or equal to NE of s_1 .

Neutral 1 (N1) (1) s_1 and s_2 are semantically similar and (2) NE of s_1 is a subset of NE of s_2 .

Neutral 2 (N2) (1) NE of s_1 are equal to NE of s_2 and (2) s_1 include an antonym of a word in s_2 .

Neutral 3 (N3) (1) NE types of s_1 are equal to NE types of s_2 and (2) NE of s_1 is different from NE of s_2 . NE types are used in this rule to

Training Data	#samples	Test Accuracy	
		RadNLI	MedNLI
MedNLI	13k	53.3	80.9
MedNLI + RadNLI	19k	77.8	79.8

Table 1: The accuracies of the NLI model trained with the weakly-supervised approach. RadNLI is the proposed NLI for radiology reports. The values are the average of 5 runs and the bold values are the best results of each test set.

introduce a certain level of similarity between s_1 and s_2 .

Neutral 4 (N4) (1) NE of s_1 are equal to NE of s_2 and (2) s_1 and s_2 include observation keywords.

Contradiction 1 (C1) (1) NE of s_1 is equal or a subset to NE of s_2 and (2) s_1 is a negation of s_2 .

The rules rely on a semantic similarity measure and the overlap of entities to determine the relationship between s_1 and s_2 . In the neutral rules and the contradiction rule, we included similarity measures to avoid extracting easy to distinguish sentence pairs.

We evaluated this NLI by preparing training data, validation data, and test data. For the training data, the training set of MIMIC-CXR (Johnson et al., 2019) is used as the source of sentence pairs. 2k pairs are extracted for E1 and C1, 0.5k pairs are extracted for N1, N2, N3, and N4, resulting in a total of 6k pairs. The training set of MedNLI is also used as additional data. For the validation data and the test data, we sampled 480 sentence pairs from the validation section of MIMIC-CXR and had them annotated by two experts: one medical expert and one NLP expert. Each pair is annotated twice swapping its premise and hypothesis resulting in 960 pairs and are split in half resulting in 480 pair for a validation set and 480 pairs for a test set. The test set of MedNLI is also used as alternative test data.

We used BERT (Devlin et al., 2019) as an NLI model since it performed as a strong baseline in the existing MedNLI system (Ben Abacha et al., 2019), and used Stanza (Qi et al., 2020) and its clinical models (Zhang et al., 2020c) as a named entity recognizer. Table 1 shows the result of the model trained with and without the weakly-supervised data. The accuracy of NLI on radiology data increased substantially by +24.5% with the addition

of the radiology NLI training set. (See Appendix A for the detail of the rules, the datasets, and the model configuration.)

4 Experiments

4.1 Data

We used the training and validation sets of MIMIC-CXR (Johnson et al., 2019) to train and validate models. MIMIC-CXR is a large publicly available database of chest radiographs. We extracted the *findings* sections from the reports with a text extraction tool for MIMIC-CXR², and used them as our reference reports as in previous work (Liu et al., 2019; Boag et al., 2020). *Findings* section is a natural language description of the important aspects in a radiology image. The reports with empty *findings* sections were discarded, resulting in 152173 and 1196 reports for the training and validation set, respectively. We used the test set of MIMIC-CXR and the entire Open-i Chest X-ray dataset (Demner-Fushman et al., 2012) as two individual test sets. Open-i is another publicly available database of chest radiographs which has been widely used in past studies. We again extracted the *findings* sections, resulting in 2347 reports for MIMIC-CXR and 3335 reports for Open-i. Open-i is used only for testing since the number of reports is too small to train and test a neural report generation model.

4.2 Evaluation Metrics

BLEU4, CIDEr-D & BERTScore: We first use general NLG metrics to evaluate the generation quality. These metrics include the 4-gram BLEU score (Papineni et al., 2002, BLEU4), CIDEr score (Vedantam et al., 2015) with gaming penalties (CIDEr-D), and the F₁ score of the BERTScore (Zhang et al., 2020a).

Clinical Metrics: However, NLG metrics such as BLEU and CIDEr are known to be inadequate for evaluating factual completeness and consistency. We therefore followed previous work (Liu et al., 2019; Boag et al., 2020; Chen et al., 2020) by additionally evaluating the clinical accuracy of the generated reports using a clinical information extraction system. We use CheXbert (Smit et al., 2020), an information extraction system for chest reports, to extract the presence status of a series of observations (i.e., whether a disease is present or not), and score a generation by comparing the values of these observations to those obtained from

²<https://github.com/MIT-LCP/mimic-cxr/tree/master/txt>

the reference³. The micro average of accuracy, precision, recall, and F₁ scores are calculated over 5 observations (following previous work (Irvin et al., 2019)) for: *atelectasis*, *cardiomegaly*, *consolidation*, *edema*, and *pleural effusion*⁴.

fact_{ENT} & fact_{ENTNLI}: We additionally include our proposed rewards fact_{ENT} and fact_{ENTNLI} as metrics to compare their values for different models.

4.3 Model Variations

We used \mathcal{M}^2 Trans as our report generation model and used DenseNet-121 (Huang et al., 2017) as our image encoder. We trained \mathcal{M}^2 Trans with the following variety of joint losses.

NLL \mathcal{M}^2 Trans simply optimized with NLL loss as a baseline loss.

NLL+CDr CIDEr-D and NLL loss is jointly optimized with $\lambda_1 = 0.01$ and $\lambda_2 = 0.99$ for the scaling factors.

NLL+BS The F₁ score of BERTScore and NLL loss is jointly optimized with $\lambda_1 = 0.01$ and $\lambda_2 = 0.99$.

NLL+BS+fc_E fact_{ENT} is added to NLL+BS with $\lambda_1 = 0.01$, $\lambda_2 = 0.495$, and $\lambda_3 = 0.495$.

NLL+BS+fc_{EN} fact_{ENTNLI} is added to NLL+BS with $\lambda_1 = 0.01$, $\lambda_2 = 0.495$, and $\lambda_3 = 0.495$.

We additionally prepared three previous models that have been tested on MIMIC-CXR.

TieNet We reimplemented the model of Wang et al. (2018) consisting of a CNN encoder and an RNN decoder optimized with a multi-task setting of language generation and image classification.

CNN-RNN² We reimplemented the model of Liu et al. (2019) consisting of a CNN encoder and a hierarchical RNN decoder optimized with CIDEr and Clinically Coherent Reward

³We used CheXbert instead of CheXpert (Irvin et al., 2019) since CheXbert was evaluated to be approximately 5.5% more accurate than CheXpert. The evaluation using CheXpert can be found in Appendix C.

⁴These 5 observations are evaluated to be most represented in real-world radiology reports and therefore using these 5 observations (and excluding others) leads to less variance and more statistical strength in the results. We include the detailed results of the clinical metrics in Appendix C for completeness.

Dataset	Model	NLG Metrics			Clinical Metrics (micro-avg)				Factual Rewards	
		BL4	CDr	BS	P	R	F ₁	acc.	f _{CE}	f _{CE_N}
MIMIC-CXR	<i>Previous models</i>									
	TieNet (Wang et al., 2018)	8.1	37.2	49.2	38.6	20.9	27.1	74.0		
	CNN-RNN ² (Liu et al., 2019)	7.6	44.7	41.2	66.4	18.7	29.2	79.0		
	R2Gen (Chen et al., 2020)	8.6	40.6	50.8	41.2	29.8	34.6	73.9		
	<i>Proposed approach without proposed optimization</i>									
	\mathcal{M}^2 Trans w/ NLL	10.5	44.5	51.2	48.9	41.1	44.7	76.5	27.3	24.4
	\mathcal{M}^2 Trans w/ NLL+CDr	13.3	67.0	55.9	50.0	51.3	50.6	76.9	35.2	32.9
	<i>Proposed approach</i>									
	\mathcal{M}^2 Trans w/ NLL+BS	12.2	58.4	58.4	46.3	67.5	54.9	74.4	35.9	33.0
	\mathcal{M}^2 Trans w/ NLL+BS+f _{CE}	11.1	49.2	57.2	46.3	73.2	56.7	74.2	39.5	34.8
\mathcal{M}^2 Trans w/ NLL+BS+f _{CE_N}	11.4	50.9	56.9	50.3	65.1	56.7	77.1	38.5	37.9	
Open-i	<i>Previous models</i>									
	TieNet (Wang et al., 2018)	9.0	65.7	56.1	46.9	15.9	23.7	96.0		
	CNN-RNN ² (Liu et al., 2019)	12.1	87.2	57.1	55.1	7.5	13.2	96.1		
	R2Gen (Chen et al., 2020)	6.7	61.4	53.8	27.0	17.3	21.1	94.9		
	<i>Proposed approach without proposed optimization</i>									
	\mathcal{M}^2 Trans w/ NLL	8.2	64.4	53.1	44.7	32.7	37.8	95.8	31.1	34.1
	\mathcal{M}^2 Trans w/ NLL+CDr	13.4	97.2	59.9	48.2	24.2	32.2	96.0	40.6	42.9
	<i>Proposed approach</i>									
\mathcal{M}^2 Trans w/ NLL+BS	12.3	87.3	62.4	47.7	46.6	47.2	95.9	41.5	44.1	
\mathcal{M}^2 Trans w/ NLL+BS+f _{CE}	12.0	99.6	62.6	44.0	53.5	48.3	95.5	44.4	46.8	
\mathcal{M}^2 Trans w/ NLL+BS+f _{CE_N}	13.1	103.4	61.0	48.7	46.9	47.8	96.0	43.6	47.1	

Table 2: Results of the baselines and our \mathcal{M}^2 Trans model trained with different joint losses. For the metrics, BL4, CDr, and BS represent BLEU4, CIDEr-D, and the F₁ score of BERTScore; P, R, F₁ and acc. represent the precision, recall, F₁, and accuracy scores output by the clinical CheXbert labeler, respectively. For the rewards, f_{CE} and f_{CE_N} represent fact_{ENT} and fact_{ENTNLI}, respectively.

which is a reward based on the clinical metrics.

R2Gen The model of Chen et al. (2020) with a CNN encoder and a memory-driven Transformer optimized with NLL loss. We used the publicly available official code and its checkpoint as its implementation.

For reproducibility, we include model configurations and training details in Appendix B.

5 Results and Discussions

5.1 Evaluation with NLG Metrics and Clinical Metrics

Table 2 shows the results of the baselines⁵ and \mathcal{M}^2 Trans optimized with the five different joint losses. We find that the best result for a metric or a reward is achieved when that metric or reward is used directly in the optimization objective. Notably, for the proposed factual rewards, the increases of +3.6 fact_{ENT} and +4.9 fact_{ENTNLI} are observed

⁵These MIMIC-CXR scores have some gaps from the previously reported values with some possible reasons. First, TieNet and CNN-RNN² in Liu et al. (2019) are evaluated on a pre-release version of MIMIC-CXR. Second, we used report-level evaluation for all models, but Chen et al. (2020) tested R2Gen using image-level evaluation.

on MIMIC-CXR with \mathcal{M}^2 Trans when compared against \mathcal{M}^2 Trans w/ BS. For the clinical metrics, the best recalls and F₁ scores are obtained with \mathcal{M}^2 Trans using fact_{ENT} as a reward, achieving a substantial +22.1 increase ($\Delta+63.9\%$) in F₁ score against the best baseline R2Gen. We further find that using fact_{ENTNLI} as a reward leads to higher precision and accuracy compared to fact_{ENT} with decreases in the recalls. The best precisions and accuracies were obtained in the baseline CNN-RNN². This is not surprising since this model directly optimizes the clinical metrics with its Clinically Coherent Reward. However, this model is strongly optimized against precision resulting in the low recalls and F₁ scores.

The results of \mathcal{M}^2 Trans without the proposed rewards and BERTScore reveal the strength of \mathcal{M}^2 Trans and the inadequacy of NLL loss and CIDEr for factual completeness and consistency. \mathcal{M}^2 Trans w/ NLL shows strong improvements in the clinical metrics against R2Gen. These improvements are a little surprising since both models are Transformer-based models and are optimized with NLL loss. We assume that these improvements are due to architecture differences such as memory matrices in the encoder of \mathcal{M}^2 Trans. The differ-

\mathcal{M}^2 Trans w/ BS <i>Proposed (simple)</i>	R2Gen (Chen et al., 2020)	No difference
36.5%	12.0%	51.5%

Table 3: The human evaluation result for randomly sampled 100 reports from the test set of MIMIC-CXR by two board-certified radiologists.

ence between NLL and NLL+CDr on \mathcal{M}^2 Trans indicates that NLL and CIDEr are unreliable for factual completeness and consistency.

5.2 Human Evaluation

We performed a human evaluation to further confirm whether the generated radiology reports are factually complete and consistent. Following prior studies of radiology report summarization (Zhang et al., 2020b) and image captioning evaluation (Vedantam et al., 2015), we designed a simple human evaluation task. Given a reference report (R) and two candidate model generated reports (C1, C2), two board-certified radiologists decided whether C1 or C2 is more factually similar to R. To consider cases when C1 and C2 are difficult to differentiate, we also prepared “No difference” as an answer. We sampled 100 reports randomly from the test set of MIMIC-CXR for this evaluation. Since this evaluation is (financially) expensive and there has been no human evaluation between the baseline models, we selected R2Gen as the best previous model and \mathcal{M}^2 Trans w/ BS as the most simple proposed model, in order to be able to weakly infer that all of our proposed models are better than all of the baselines. Table 3 shows the result of the evaluation. The majority of the reports were labeled “No difference” but the proposed approach received three times as much preference as the baseline.

There are two main reasons why “No difference” was frequent in human evaluation. First, we found that a substantial portion of the examples were normal studies (no abnormal observations), which leads to generated reports of similar quality from both models. Second, in some reports with multiple abnormal observations, both models made mistakes on a subset of these observations, making it difficult to decide which model output was better.

5.3 Estimating Clinical Accuracy with Factual Rewards

The integrations of fact_{ENT} and $\text{fact}_{\text{ENTNLI}}$ showed improvements in the clinical metrics. We further examined whether these rewards can be

Metric	ρ
BLEU4	0.092
CIDEr-D	0.034
BERTScore	0.155
fact_{ENT}	0.196
$\text{fact}_{\text{ENTNLI}}$	0.255

Table 4: The Spearman correlations ρ of NLG metrics and factual metrics against clinical accuracy. The strongest correlation among all metrics is shown in bold.

used to estimate the performance of the clinical metrics to see whether the proposed rewards can be used in an evaluation where a strong clinical information extraction system like CheXbert is not available. Table 4 shows Spearman correlations calculated on the generated reports of NLL+BS. $\text{fact}_{\text{ENTNLI}}$ shows the strongest correlation with the clinical accuracy which aligns with the optimization where the best accuracy is obtained with NLL+BS+ $\text{fact}_{\text{ENTNLI}}$. This correlation value is slightly lower than a Spearman correlation which Maynez et al. (2020) observed with NLI for the factual data (0.264). The result suggests the effectiveness of using the factual rewards to estimate the factual completeness and consistency of radiology reports, although the correlations are still limited, with some room for improvement.

5.4 Qualitative Analysis of Improved Clinical Completeness and Consistency

The evaluation with the clinically findings metrics showed improved generation performance by integrating BERTScore, fact_{ENT} , and $\text{fact}_{\text{ENTNLI}}$. As a qualitative analysis, we examined some of the generated reports to see the improvements. Example 1 in Figure 3 shows the improved factual completeness and consistency with BERTScore. The *atelectasis* is correctly generated and *left pleural effusion* is correctly suppressed with NLL+BS. Example 2 in Figure 4 shows the improved factual completeness with $\text{fact}_{\text{ENTNLI}}$. The *edema* is correctly generated and *atelectasis* is correctly suppressed with NLL+BS+ fact_{ENT} . These examples reveal the strength of integrating the three metrics to generate factually complete and consistent reports.

Despite observing large improvements with our model in the clinical finding metrics evaluation, the model is still not complete and some typical factual errors can be found in their generated reports. For example, Example 3 in Figure 4 includes a comparison of an observation against a previous study as


	Images	Reference	R2Gen	\mathcal{M}^2 Trans w/ NLL+BS
Example 1		Large right pleural effusion is unchanged in size. There is associated right basilar atelectasis/scarring, also stable. Healed right rib fractures are noted. On the left, there is persistent apical pleural thickening and apical scarring. Linear opacities projecting over the lower lobe are also compatible with scarring, unchanged. There is no left pleural effusion. There is no pneumothorax. Hilar and cardiomeastinal contours are difficult to assess, but appear unchanged. Vascular stent is seen in the left axillary/subclavian region.	PA and lateral chest views were obtained with patient in upright position. Analysis is performed in direct comparison with the next preceding similar study of DATE. The heart size remains unchanged and is within normal limits. Unchanged appearance of thoracic aorta. The pulmonary vasculature is not congested. Bilateral pleural effusions are again noted and have increased in size on the right than the left. The left-sided pleural effusion has increased in size and is now moderate in size.	As compared to prior chest radiograph from DATE, there has been interval improvement of the right pleural effusion. There is a persistent opacity at the right lung base. There is persistent atelectasis at the right lung base. There is no left pleural effusion. There is no pneumothorax. The cardiomeastinal and hilar contours are unchanged.

Figure 3: An example of radiology reports generated by R2Gen and by the proposed model with the optimization integrating BERTScore. Repeated sentences are removed from the example to improve readability.


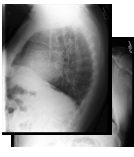
	Images	Reference	\mathcal{M}^2 Trans w/ NLL+BS	\mathcal{M}^2 Trans w/ NLL+BS+ f_{CEN}
Example 2		Frontal and lateral radiographs of the chest were acquired. There is new mild interstitial pulmonary edema. A small right pleural effusion may be minimally increased. There is also likely a trace left pleural effusion. There is no focal consolidation. The heart size is not significantly changed. There is no pneumothorax. Midline sternotomy wires are noted.	Assessment is limited by patient rotation. The patient is status post median sternotomy and CABG. Heart size is moderately enlarged. The aorta is tortuous and diffusely calcified. There is mild pulmonary vascular congestion. Small bilateral pleural effusions are present. Patchy opacities in the lung bases likely reflect atelectasis. No pneumothorax is identified. There are no acute osseous abnormalities.	The cardiomeastinal and hilar contours are stable. The aorta is tortuous. The patient is status post median sternotomy. The heart is mildly enlarged. The aorta is tortuous. The lung volumes are lower compared to the prior chest radiograph. Mild pulmonary edema is present. Small bilateral pleural effusions are present. There is no focal consolidation. No pneumothorax is seen. Median sternotomy wires and mediastinal clips are noted.
Example 3		A right-sided hemodialysis catheter terminates at the right atrium. Again seen are reticular interstitial opacities distributed evenly across both lungs, stable over multiple prior radiographs, previously attributed to chronic hypersensitivity pneumonitis on the chest CT from DATE. The cardiac and mediastinal silhouettes are unchanged. The central pulmonary vessels appear more prominent since the DATE study. Superimposed mild edema cannot be excluded. There is no focal consolidation, pleural effusion, or pneumothorax.	Right-sided dual lumen central venous catheter tip terminates in the lower SVC. Heart size remains mildly enlarged. The mediastinal and hilar contours are unchanged. There is no pulmonary edema. Minimal atelectasis is noted in the lung bases without focal consolidation. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities.	The cardiomeastinal and hilar contours are normal. The lung volumes are low. The lung volumes are present. There is mild pulmonary edema. There is no focal consolidation. No pleural effusion or pneumothorax is seen. A right-sided central venous catheter is seen with tip in the right atrium.

Figure 4: Examples of radiology reports generated by the proposed model with the optimization integrating BERTScore and f_{CEN} . Repeated sentences are removed from the examples to improve readability.

“... appear more prominent since ...” in the reference but our model (or any previous models) can not capture this kind of comparison since the model is not designed to take account the past reports of a patient as input. Additionally, in this example, *edema* is mentioned with uncertainty as “cannot be excluded” in the reference but the generated report with f_{CEN} simply indicates it as “There is mild pulmonary edema”.

6 Conclusion

We proposed two new simple rewards and combined them with a semantic equivalence metric to improve image-to-text radiology report generation systems. The two new rewards make use of radiology domain entities extracted with a named entity recognizer and a weakly-supervised NLI to capture the factual completeness and consistency of the generated reports. We further presented a Transformer-based report generation system that

directly optimizes these rewards with self-critical reinforcement learning. On two open datasets, we showed that our system generates reports that are more factually complete and consistent than the baselines and leads to reports with substantially higher scores in clinical metrics. The integration of entities and NLI to improve the factual completeness and consistency of generation is not restricted to the domain of radiology reports, and we predict that a similar approach might similarly improve other data-to-text tasks.

Acknowledgements

We would like to thank the anonymous reviewers and the members of the Stanford NLP Group for their very helpful comments that substantially improved this paper.

References

- Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. [Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.
- William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. 2020. [Baselines for Chest X-Ray Report Generation](#). In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116, pages 126–140.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. [Generating radiology reports via memory-driven transformer](#). In *Proceedings of The 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449.
- 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*, pages 10575–10584.
- Dina Demner-Fushman, Samee Antani, Simpson Matthew, and George R. Thoma. 2012. [Design and development of a multimodal biomedical information retrieval system](#). *Journal of Computing Science and Engineering*, 6(2):168–177.
- 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- 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*, pages 5055–5070.
- 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*, pages 2214–2220.
- Christiane Fellbaum, editor. 1998. *WordNet: A Lexical Database for English*. MIT Press.
- Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. [Normalized and geometry-aware self-attention network for image captioning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10324–10333.
- Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. [Image captioning: Transforming objects into words](#). In *Advances in Neural Information Processing*, volume 32, pages 11137–11147.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. [Densely connected convolutional networks](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. [CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.
- Baoyu Jing, Zeya Wang, and Eric Xing. 2019. [Show, describe and conclude: On exploiting the structure information of chest x-ray reports](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6570–6580.
- Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. [On the automatic generation of medical imaging reports](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586.
- Alistair E. W. 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(317).
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(16035).
- Charles E. Kahn, 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.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *International Conference for Learning Representations*.
- Wojciech Kryściński, 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*, pages 9332–9346.

- Christy Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2018. [Hybrid retrieval-generation reinforced agent for medical image report generation](#). In *Advances in Neural Information Processing*, volume 31, pages 1530–1540.
- Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. [Entangled transformer for image captioning](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8927–8936.
- Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. [Clinically accurate chest x-ray report generation](#). In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106, pages 249–269.
- Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. [Improving truthfulness of headline generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346.
- 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*, pages 1906–1919.
- Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. [X-linear attention networks for image captioning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2018. [A deep reinforced model for abstractive summarization](#). In *International Conference on Learning Representations*.
- Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. [Negbio: a high-performance tool for negation and uncertainty detection in radiology reports](#). In *AMIA 2018 Informatics Summit*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. [Sequence level training with recurrent neural networks](#). In *International Conference on Learning Representations*.
- 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*, pages 1179–1195.
- Alexey Romanov and Chaitanya Shivade. 2018. [Lessons from natural language inference in the clinical domain](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. [Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1519.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in neural information processing systems*, volume 30, pages 5998–6008.
- 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*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. [Show and tell: A neural image caption generator](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. [Asking and answering questions to evaluate the factual consistency of summaries](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.
- 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*, pages 9049–9058.
- Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020b. [Towards faithful neural table-to-text generation with content-matching constraints](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. [Dialogue natural language](#)

- [inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.
- Ronald J Williams and David Zipser. 1989. [A learning algorithm for continually running fully recurrent neural networks](#). *Neural computation*, 1(2):270–280.
- Zhaofeng Wu, Yan Song, Sicong Huang, Yuanhe Tian, and Fei Xia. 2019. [WTMED at MEDIQA 2019: A hybrid approach to biomedical natural language inference](#). In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 415–426.
- Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. [Automatic radiology report generation based on multi-view image fusion and medical concept enrichment](#). In *Medical Image Computing and Computer Assisted Intervention*, pages 721–729.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. [BERTScore: Evaluating text generation with BERT](#). In *International Conference on Learning Representations*.
- Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. [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*, pages 5108–5120.
- Yuhao Zhang, Yuhui Zhang, Peng Qi, Christopher D. Manning, and Curtis P. Langlotz. 2020c. [Biomedical and clinical English model packages in the Stanza Python NLP library](#). *arXiv preprint arXiv:2007.14640*.

A Detail of Radiology NLI

A.1 Rules & Examples of Weakly-Supervised Radiology NLI

We prepared the 6 rules (E1, N1–N4, and C1) to train the weakly-supervised radiology NLI. The rules are applied against sentence pairs consisting from premises (s_1) and hypotheses (s_2) to extract pairs that are in *entailment*, *neutral*, or *contradiction* relation.

Entailment Rule: E1

1. s_1 and s_2 are semantically similar.
2. The named entities (NE) of s_2 is a subset or equal to the named entities of s_1 as $NE(s_2) \subseteq NE(s_1)$.

We used BERTScore (Zhang et al., 2020a) as a similarity metric and set the threshold to $sim(s_1, s_2) \geq 0.7^6$. The clinical model of Stanza (Zhang et al., 2020c) is used to extract *anatomy* entities and *observation* entities. s_1 and s_2 are conditioned to be both negated or both non-negated. The negation is determined with a negation identifier or the existence of *uncertain* entity, using NegBio (Peng et al., 2018) as the negation identifier and the clinical model of Stanza is used to extract *uncertain* entities. s_2 is further restricted to include at least 2 entities as $|NE(s_2)| \geq 2$. These similarity metric, named entity recognition model, and entity number restriction are used in the latter neutral and contradiction rules. The negation restriction is used in the neutral rules but is not used in the contradiction rule. The following is an example of a sentence pair that matches E1 with entities in bold:

s_1 The **heart** is mildly **enlarged**.

s_2 The **heart** appears again mild-to-moderately **enlarged**.

Neutral Rule 1: N1

1. s_1 and s_2 are semantically similar.
2. The named entities of s_1 is a subset of the named entities of s_2 as $NE(s_1) \subsetneq NE(s_2)$.

Since s_1 is a premise, this condition denotes that the counterpart hypothesis has entities that are not

⁶*distilbert-base-uncased* with the baseline score is used as the model of BERTScore for a fast comparison and a smooth score scale. We swept the threshold value from $f0.6, 0.7, 0.8, 0.9g$ and set it to 0.7 as a relaxed boundary to balance between accuracy and diversity.

included in the premise. The following is an example of a sentence pair that matches N1 with entities in bold:

s_1 There is no **pulmonary edema** or definite **consolidation**.

s_2 There is no focal **consolidation**, **pleural effusion**, or **pulmonary edema**.

Neutral Rule 2: N2

1. The named entities of s_1 are equal to the named entities of s_2 as $NE(s_1) = NE(s_2)$.
2. The anatomy modifiers (NE_{mod}) of s_1 include an antonym (ANT) of the anatomy modifier of s_2 as $NE_{mod}(s_1) \cap ANT(NE_{mod}(s_2)) \neq \emptyset$.

Anatomy modifiers are extracted with the clinical model of Stanza and antonyms are decided using WordNet (Fellbaum, 1998). Antonyms in anatomy modifiers are considered in this rule to differentiate expressions like *left vs right* and *upper vs lower*. The following is an example of a sentence pair that matches N2 with antonyms in bold:

s_1 Moreover, a small **left** pleural effusion has newly occurred.

s_2 Small **right** pleural effusion has worsened.

Neutral Rule 3: N3

1. The named entity types (NE_{type}) of s_1 are equal to the named entity types of s_2 as $NE_{type}(s_1) = NE_{type}(s_2)$.
2. The named entities of s_1 is different from the named entities of s_2 as $NE(s_1) \cap NE(s_2) = \emptyset$.

Specific entity types that we used are *anatomy* and *observation*. This rule ensures that s_1 and s_2 have related but different entities in same types. The following is an example of a sentence pair that matches N3 with entities in bold:

s_1 There is minimal bilateral **lower lobe atelectasis**.

s_2 The **cardiac silhouette** is moderately **enlarged**.

Neutral Rule 4: N4

1. The named entities of s_1 are equal to the named entities of s_2 as $NE(s_1) = NE(s_2)$.

- s_1 and s_2 include observation keywords (KEY) that belong to different groups as $\text{KEY}(s_1) \neq \text{KEY}(s_2)$.

The groups of observation keywords are setup following the observation keywords of CheXpert labeler (Irvin et al., 2019). Specifically, $G1 = \{normal, unremarkable\}$, $G2 = \{stable, unchanged\}$, and $G3 = \{clear\}$ are used to determine words included in different groups as *neutral* relation. The following is an example of a sentence pair that matches N4 with keywords in bold:

s_1 **Normal** cardiomedial silhouette.

s_2 Cardiomedial silhouette is **unchanged**.

Contradiction Rule: C1

- The named entities of s_1 is a subset or equal to the named entities of s_2 as $\text{NE}(s_2) \subseteq \text{NE}(s_1)$.
- s_1 or s_2 is a negated sentence.

Negation is determined with the same approach as E1. The following is an example of a sentence pair that matches C1 with entities in bold:

s_1 There are also small bilateral **pleural effusions**.

s_2 No **pleural effusions**.

A.2 Validation and Test Datasets of Radiology NLI

We sampled 480 sentence pairs that satisfy the following conditions from the validation section of MIMIC-CXR:

- Two sentences (s_1 and s_2) have $\text{BERTScore}(s_1, s_2) \geq 0.5$.
- MedNLI labels are equally distributed over three labels: *entailment*, *neutral*, and *contradiction*⁷.

These conditions are introduced to reduce *neutral* pairs since most pairs will be *neutral* with random sampling. The sampled pairs are annotated twice swapping its premise and hypothesis by two experts: one medical expert and one NLP expert. For pairs that the two annotators disagreed, its labels are decided by a discussion with one additional NLP expert. The resulting 960 bidirectional pairs are splitted in half resulting in 480 pairs for a validation set and 480 pairs for a test set.

⁷We used the baseline BERT model of Wu et al. (2019) to assign MedNLI labels to the pairs.

A.3 Configuration of Radiology NLI Model

We used *bert-base-uncased* as a pre-trained BERT model and further fine-tuned it on MIMIC-III (Johnson et al., 2016) radiology reports with a masked language modeling loss for 8 epochs. The model is further optimized on the training data with a classification negative log likelihood loss. We used Adam (Kingma and Ba, 2015) as an optimization method with $\beta_1 = 0.9$, $\beta_2 = 0.999$, batch size of 16, and the gradient clipping norm of 5.0. The learning rate is set to $lr = 1e^{-5}$ by running a preliminary experiment with $lr = \{1e^{-5}, 2e^{-5}\}$. The model is optimized for the maximum of 20 epochs and a validation accuracy is used to decide a model checkpoint that is used to evaluate the test set. We trained the model with a single Nvidia Titan XP taking approximately 2 hours to complete 20 epochs.

B Configurations of Radiology Report Generation Models

B.1 \mathcal{M}^2 Trans

We used DenseNet-121 (Huang et al., 2017) as a CNN image feature extractor and pre-trained it on CheXpert dataset with the 14-class classification setting. We used GloVe (Pennington et al., 2014) to pre-train text embeddings and the pre-trainings were done on a training set with the embedding size of 512. The parameters of the model is set up to the dimensionality of 512, the number of heads to 8, and the number of memory vector to 40. We set the number of Transformer layer to $n_{layer} = 1$ by running a preliminary experiment with $n_{layer} = \{1, 2, 3\}$. The model is first trained against NLL loss using the learning rate scheduler of Transformer (Devlin et al., 2019) with the warm-up steps of 20000 and is further optimized with a joint loss with the fixed learning rate of $5e^{-6}$. Adam is used as an optimization method with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 48 for NLL loss and 24 for the joint losses. For λ , we first swept the optimal value of λ_1 from $\{0.03, 0.02, 0.01, 0.001\}$ using the development set. We have restricted λ_2 and λ_3 to have equal values in our experiments and constrained that all λ values sum up to 1.0. The model is trained with NLL loss for 32 epochs and further trained for 32 epochs with a joint loss. Beam search with the beam size of 4 is used to decode texts when evaluating the model against a validation set or a test set. We trained the model with a single Nvidia

Titan XP taking approximately 10 days to complete its optimization.

B.2 TieNet

We used ResNet-50 as a CNN image feature extractor with default ImageNet pre-trained weights. We used GloVe to pre-train text embeddings with the same configuration as \mathcal{M}^2 Trans. The parameters of the model is set up to the LSTM dimension of 256 and the number of global attentions to 5. The combination of NLL loss and the multi-label classification loss is used as its joint loss with the balance parameter $\alpha = 0.85$. The model is trained against the joint loss using a linear rate scheduler with the initial learning rate of $1e^{-4}$ and the multiplication of 0.5 per 8 epochs. The batch size is set to 32 and the model is trained with the joint loss for 32 epochs. Adam is used as an optimization method with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Beam search with the beam size of 4 is used to decode texts. We trained the model with a single Nvidia Titan XP taking approximately 2 days to complete its optimization.

B.3 CNN-RNN²

We used DenseNet-121 as a CNN image feature extractor with default ImageNet pre-trained weights. We used GloVe to pre-train text embeddings with the same configuration as \mathcal{M}^2 Trans. The parameters of the model is set up to the LSTM dimension of 256. We modified an information extraction system from CheXpert to CheXbert to improve the training speed of this model. The combination of CIDEr and Clinically Coherent Reward is used as its joint loss with the balance parameter $\lambda = 10.0$. The model is first trained against NLL loss using a linear rate scheduler with the initial learning rate of $1e^{-4}$ and the multiplication of 0.5 per 8 epochs. The model is further optimized with the joint loss with the fixed learning rate of $5e^{-6}$. Adam is used as an optimization method with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 32 for the NLL loss and 24 for the joint losses. The model is trained with NLL loss for 32 epochs and further trained for 32 epochs with the joint loss. Beam search with the beam size of 4 is used to decode texts. We trained the model with a single Nvidia Titan XP taking approximately 11 days to complete its optimization.

C Detailed Result of Clinical Metrics

Table 5 shows the detailed results of the clinical metrics for R2Gen, \mathcal{M}^2 Trans w/ BS, \mathcal{M}^2 Trans w/ BS+ f_{CE} , and \mathcal{M}^2 Trans w/ BS+ f_{CEN} . In most cases, the best F_1 scores are observed in the cases when $fact_{ENT}$ or $fact_{ENTNLI}$ is included in the joint losses. *Consolidation* is one exception where the best precisions, recalls, and F_1 scores vary among the joint losses. We assume this is due to the infrequent appearance of *consolidation* in both MIMIC-CXR and Open-i. For comparison against some past studies, we show the detailed results when CheXpert is used instead of CheXbert in Table 6. Since CheXbert is more or equally accurate for most observations than CheXpert, the scores in Table 6 follow similar trends against ones in Table 5. Table 7 shows the detailed results for the 9 remaining observations that are defined in CheXpert. Note that many of these observations are infrequent and have relatively weaker and unstable extraction performances compared to the 5 observations in Table 5.

Observation (# MIMIC-CXR / # Open-i)		MIMIC-CXR				Open-i			
		R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}	R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}
Micro Average (2713 / 654)	P	41.2	46.3	46.3	50.3	27.0	47.7	44.0	48.7
	R	29.8	67.5	73.2	65.1	17.3	46.6	53.5	46.9
	F ₁	34.6	54.9	56.7	56.7	21.1	47.2	48.3	47.8
	acc.	73.9	74.4	74.2	77.1	94.9	95.9	95.5	96.0
Atelectasis (604 / 216)	P	35.4	39.9	37.9	40.6	35.6	44.0	35.8	39.4
	R	27.8	67.4	80.5	76.2	7.4	35.6	47.7	45.4
	F ₁	31.1	50.2	51.6	53.0	12.3	39.4	40.9	42.2
	acc.	68.3	65.5	61.1	65.2	93.1	92.9	91.1	91.9
Cardiomegaly (535 / 225)	P	32.4	35.8	34.3	37.5	24.5	56.6	57.3	60.0
	R	53.5	73.3	81.3	61.3	33.8	55.6	55.6	46.7
	F ₁	40.4	48.1	48.2	46.6	28.4	56.1	56.4	52.5
	acc.	64.0	63.9	60.2	67.9	88.5	94.1	94.2	94.3
Consolidation (157 / 19)	P	14.3	10.5	19.6	19.2	0.0	10.9	15.2	14.3
	R	7.0	18.5	5.7	3.2	0.0	26.3	26.3	5.3
	F ₁	9.4	13.4	8.9	5.5	0.0	15.4	19.2	7.7
	acc.	91.0	84.0	92.1	92.6	99.0	98.4	98.7	99.3
Edema (645 / 75)	P	55.3	59.7	56.0	65.6	10.0	39.0	30.9	41.4
	R	24.3	59.2	69.9	52.7	4.0	30.7	50.7	32.0
	F ₁	33.8	59.5	62.2	58.5	5.7	34.3	38.4	36.1
	acc.	73.8	77.8	76.6	79.4	97.0	97.4	96.3	97.5
Pleural Effusion (772 / 119)	P	76.2	67.2	68.2	65.9	85.7	54.3	59.4	56.0
	R	24.1	80.6	78.5	82.0	15.1	63.0	66.4	66.4
	F ₁	36.6	73.3	73.0	73.1	25.7	58.4	62.7	60.8
	acc.	72.6	80.7	80.9	80.1	96.9	96.8	97.2	96.9

Table 5: The detailed results of R2Gen, \mathcal{M}^2 Trans w/ BS, \mathcal{M}^2 Trans w/ BS+f_{CE}, and \mathcal{M}^2 Trans w/ BS+f_{CEN} for the 5 observations. P is precision, R is recall, and acc. is accuracy. #MIMIC-CXR and #Open-i are the numbers of times that a corresponding observation has appeared as positive in the test set of MIMIC-CXR and Open-i, respectively.

Observation (# MIMIC-CXR / # Open-i)		MIMIC-CXR				Open-i			
		R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}	R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}
Micro Average (2713 / 654)	P	37.6	46.0	46.0	49.9	16.7	46.3	42.7	47.8
	R	29.1	67.2	72.9	64.6	17.1	45.8	52.5	46.3
	F ₁	32.8	54.6	56.4	56.3	16.9	46.1	47.1	47.0
	acc.	74.6	74.2	74.0	76.8	93.4	95.8	95.4	95.9
Atelectasis (604 / 216)	P	35.6	39.9	37.8	40.6	33.3	42.9	34.7	38.2
	R	23.9	67.6	80.6	76.4	7.1	35.4	47.2	44.8
	F ₁	28.6	50.2	51.5	53.0	11.7	38.8	40.0	41.2
	acc.	72.1	65.6	61.1	65.3	93.2	92.9	91.0	91.9
Cardiomegaly (535 / 225)	P	28.6	36.1	34.6	37.6	20.4	55.2	55.5	60.0
	R	49.2	72.8	81.1	60.5	31.3	53.5	53.0	45.7
	F ₁	36.2	48.2	48.5	46.4	24.7	54.3	54.2	51.9
	acc.	63.0	63.8	60.0	67.6	86.8	93.8	93.8	94.2
Consolidation (157 / 19)	P	10.7	10.5	19.6	19.2	1.1	10.9	14.7	14.3
	R	9.4	17.8	5.5	3.1	10.5	26.3	26.3	5.3
	F ₁	10.0	13.2	8.6	5.3	2.0	15.4	18.9	7.7
	acc.	88.4	83.7	91.9	92.4	94.0	98.4	98.7	99.3
Edema (645 / 75)	P	49.0	58.7	54.9	64.3	5.8	35.0	30.1	39.7
	R	29.1	59.0	69.5	52.3	4.1	28.8	50.7	31.5
	F ₁	36.5	58.8	61.4	57.7	4.8	31.6	37.8	35.1
	acc.	74.5	77.6	76.2	79.2	96.4	97.3	96.3	97.5
Pleural Effusion (772 / 119)	P	75.6	66.5	67.6	65.2	63.3	53.2	57.5	54.6
	R	23.4	80.3	78.2	81.6	16.4	63.8	66.4	66.4
	F ₁	35.8	72.7	72.5	72.5	26.0	58.0	61.6	59.9
	acc.	75.1	80.3	80.6	79.8	96.8	96.8	97.1	96.9

Table 6: The detailed results of R2Gen, \mathcal{M}^2 Trans w/ BS, \mathcal{M}^2 Trans w/ BS+f_{CE}, and \mathcal{M}^2 Trans w/ BS+f_{CEN} for the 5 observations evaluated with CheXpert instead of CheXbert.

Observation (# MIMIC-CXR / # Open-i)		MIMIC-CXR				Open-i			
		R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}	R2Gen	NLL + BS	NLL + BS + f _{CE}	NLL + BS + f _{CEN}
Enlarged Cardiome-diastinum (111 / 24)	P	4.4	5.1	4.8	4.6	3.8	0.7	2.0	4.0
	R	19.8	50.5	19.8	47.7	16.7	4.2	4.2	20.8
	F ₁	7.1	9.3	7.7	8.4	6.3	1.2	2.7	6.7
	acc.	75.6	53.4	77.5	50.6	96.4	95.1	97.8	95.8
Fracture (56 / 43)	P	0.0	40.0	10.7	26.1	0.0	0.0	0.0	3.1
	R	0.0	3.6	5.4	10.7	0.0	0.0	0.0	2.3
	F ₁	0.0	6.6	7.1	15.2	0.0	0.0	0.0	2.7
	acc.	97.6	97.6	96.7	97.1	98.7	98.6	98.1	97.8
Lung Lesion (97 / 89)	P	37.5	33.3	22.2	44.4	25.0	0.0	0.0	66.7
	R	3.1	1.0	2.1	4.1	1.1	0.0	0.0	4.5
	F ₁	5.7	2.0	3.8	7.5	2.2	0.0	0.0	8.4
	acc.	95.8	95.8	95.7	95.8	97.3	97.3	97.3	97.4
Lung Opacity (798 / 344)	P	44.5	48.8	53.5	54.9	43.1	50.5	57.8	41.1
	R	29.9	41.6	10.4	26.6	8.1	29.1	7.6	22.1
	F ₁	35.8	44.9	17.4	35.8	13.7	36.9	13.4	28.7
	acc.	63.5	65.3	66.5	67.6	89.4	89.7	89.9	88.7
No Finding (396 / 2319)	P	31.4	44.4	49.8	48.8	78.1	80.8	82.1	81.7
	R	43.9	35.9	41.7	39.9	84.1	93.4	91.5	88.4
	F ₁	36.6	39.7	45.4	43.9	81.0	86.6	86.5	84.9
	acc.	74.4	81.6	83.1	82.8	72.6	80.0	80.2	78.2
Pleural Other (39 / 29)	P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	F ₁	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	acc.	98.1	98.3	98.3	98.3	99.1	99.1	99.1	99.1
Pneumonia (424 / 109)	P	42.1	62.7	62.1	0.0	27.6	31.3	38.6	0.0
	R	15.8	16.3	17.0	0.0	7.3	19.3	24.8	0.0
	F ₁	23.0	25.8	26.7	0.0	11.6	23.9	30.2	0.0
	acc.	80.9	83.1	83.1	81.9	96.3	96.0	96.3	96.7
Pneumothorax (78 / 15)	P	60.0	28.7	37.0	50.0	100.0	40.0	0.0	100.0
	R	3.8	34.6	12.8	10.3	6.7	13.3	0.0	13.3
	F ₁	7.2	31.4	19.0	17.0	12.5	20.0	0.0	23.5
	acc.	96.7	95.0	96.4	96.7	99.6	99.5	99.6	99.6
Support Devices (624 / 41)	P	52.2	50.8	53.2	49.0	10.0	16.2	19.7	13.1
	R	68.9	83.5	78.7	89.7	12.2	43.9	36.6	56.1
	F ₁	59.4	63.2	63.5	63.3	11.0	23.7	25.6	21.3
	acc.	75.0	74.1	75.9	72.4	97.6	96.5	97.4	94.9

Table 7: The detailed results of R2Gen, \mathcal{M}^2 Trans w/ BS, \mathcal{M}^2 Trans w/ BS+f_{CE}, and \mathcal{M}^2 Trans w/ BS+f_{CEN} for the remaining 9 observations.