Efficient and Effective Models for Machine Reading Comprehension

Adams Wei Yu
Carnegie Mellon University

Collaborators

David Dohan  Thang Luong  Rui Zhao  Kai Chen  Mohammad Norouzi  Hongrae Lee  Quoc Le
Bio

- Ph.D Candidate @ MLD, CMU
  - Advisor: Jaime Carbonell, Alex Smola
  - Research: **Efficient** AI, ML, NLP.
    - Algorithm: Large scale optimization
    - Model: Machine reading comprehension and question answering

This Talk: Question Answering
Question Answering is important
Most NLP tasks can be formulated as QA

- Translation
- Summarization
- Natural language inference
- Sentiment analysis
- Semantic role labeling
- Relation extraction
- Goal oriented dialogue
- Semantic parsing
- Pronoun resolution

Natural Language Decathlon
[McCann, 2018]
How good are the current QA systems?

Concrete Answer

No clear answer
Early Success
Watson: complex multi-stage system

DeepQA: The Technology Behind Watson
An example of a new software paradigm

DeepQA generates and scores many hypotheses using an extensible collection of Natural Language Processing, Machine Learning and Reasoning Algorithms. These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.

http://www.aaai.org/Magazine/Watson/watson.php
We Design End-to-End (Deep Learning) System
Deep Learning: Sparking the New Wave of AI

AlphaGo

Smart Assistant

Machine Translation

Self-driving car
Bottleneck: Speed

ImageNet Classification

ResNet-152: 10 days to train
Bottleneck: Speed

Google Machine Translation

English $\rightarrow$ French
6 days to train
96 GPUs
Bottleneck: Speed

4 weeks to train, 1920 CPUs and 280 GPUs
Wow! I just finished watching this and it is amazing. I’m completely blown away by the stunning cinematography. The movie was almost entirely filmed in Maui and I didn’t want to take my eyes off the gorgeous colors on the screen. The beaches, trails, mountains and roads that Michael and Jessi run on throughout the movie, are absolutely breathtaking. There is even a bonus feature section that displays Michael’s photography in a zen-like production of photographs that he captured during his runs.

The movie is designed as easy to follow, clear, concise chapters. It takes the viewer from the first step of shedding your shoes and putting your bare feet to the earth to running like a child again with a light foot and a free spirit. I am not a runner, yet this movie encourages me to take the initiative of trying few of those shoes. I plan to watch it over and over again until I reach what I’d like to. This is an amazing resource guide put to life!

I think this movie is a terrific instructional video for experienced athletes too, who want to halt running shoes. The book also gives information on the development of strength and speed. Overall, an amazing addition to your video collection.

This movie is filled with information such as (i) discussing “how to” go barefoot, (ii) what kind of shoe to purchase, (iii) exercises to help your body heal and be in proper health and ready again, (iv) exercises to help your body heal with overall health. It documents a serious accident that Michael had, which was the catalyst in getting him into barefoot running in the first place. Michael’s story is incredibly inspirational and to see the strength of his muscles up close through the filming, after viewing the footage of what his body had been through prior to going barefoot, is fascinating.

Love it!
Why Should We Care?

1. High turnaround time for experimentation.
2. Hard to train on large data.
3. Hard for real time applications.
4. High monetary cost.

......
This Talk: Efficient and Effective Deep Learning Models for Reading Comprehension
Roadmap

1. Skipping Irrelevant Information [ACL’17]
2. Discarding Recurrence [ICLR’18]
3. Future Work
Recurrent Neural Networks

The weather is nice today

Hidden state:

\[ \theta_m \]

Word embedding:

\[
\begin{align*}
\mathbf{x}_1 & \to h_1, \\
\mathbf{x}_2 & \to h_2, \\
\mathbf{x}_3 & \to h_3, \\
\mathbf{x}_4 & \to h_4, \\
\mathbf{x}_5 & \to h_5
\end{align*}
\]

Discrepancy between groundtruth and prediction:

\[
f(\theta_m)
\]
First Challenge: Hard to Capture Long Dependency

Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write too much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcibly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book. Fans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.
Second Challenge: Hard to Compute in Parallel

Strictly sequential
Q: What is dropout?

- Graphical Models
- Statistics
- Deep Learning

Skim

Model
Algorithm
Regularization
Q: What is dropout?

**Dropout Regularization For Neural Networks**

Dropout is a regularization technique for neural network models proposed by Srivastava, et al. in their 2014 paper Dropout: A Simple Way to Prevent Neural Networks from Overfitting (download the PDF).

Dropout is a technique where randomly selected neurons are ignored during training. They are "dropped-out" randomly. This means that their contribution to the activation of downstream neurons is temporarily removed on the forward pass and any weight updates are not applied to the neuron on the backward pass.

As a neural network learns, neuron weights settle into their context within the network. Weights of neurons are tuned for specific features providing some specialization. Neighboring neurons become to rely on this specialization, which if taken too far can result in a fragile model too specialized to the training data. This reliance on context for a neuron during training is referred to as colex co-adaptations.
Roadmap

1. Skipping Irrelevant Information [ACL’17]

2. Discarding Recurrence [ICLR’18]

3. Future Work
Learning to Skim Text

Yu et al.
ACL 2017
Texts are usually redundant ... Humans don’t always read word by word.

This review is from: Barefoot Running - The Movie: How to Run Light and Free by Getting in Touch with the Earth (NTSC/US Version) (DVD)

Wow! I just finished watching this and it is amazing. I’m completely blown away by the stunning cinematography. The movie was almost entirely filmed in Maui and I didn’t want to take my eyes off the gorgeous colors on the screen. The beaches, trails, mountains and roads that Michael and Jessie run on throughout the movie, are absolutely breathtaking. There is even a bonus feature section that displays Michael’s photography in a zen-like production of photographs that he captured during his runs.

The movie is designed as easy to follow, clear, concise chapters. It takes the viewer from the first step of shedding your shoes and putting your bare feet to the earth to running like a child again with a light foot and a free spirit. I am not a runner, yet this movie encourages me to take the journey of my first few yards without shoes. I plan to watch it over and over again while I practice what I’ve learned. This is an amazing resource guide put to life!

I think this movie is a terrific instructional video for experienced athletes too, who want to halt running injuries (that are discussed in detail in the movie) and find strength and speed that they probably never had.

This movie is filled with information such as (i) discussing “how to” go barefoot, (ii) what kind of shoes to purchase for the toes, (iii) you must have shoes for exercises to help balance and (iv) before getting into running exercises, you must strengthen you feet, and much more. The movie highlights a serious accident that Michael had, which was the catalyst in getting him into barefoot running in the first place. Michael’s story is incredibly inspirational and to see the strength of his muscles up close through the filming, after viewing the footage of what his body had been through prior to being barefoot, is fascinating.

Love it!
Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write too much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcibly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book. Fans of "Beat" Takeshi: his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film.
Positive or negative lengthy movie review?

Negative

- Being a long-time fan of Japanese film, I expected more than this. I can't really be bothered to write too much, as this movie is just so poor. The story might be the cutest romantic little something ever, pity I couldn't stand the awful acting, the mess they called pacing, and the standard "quirky" Japanese story. If you've noticed how many Japanese movies use characters, plots and twists that seem too "different", forcedly so, then steer clear of this movie. Seriously, a 12-year old could have told you how this movie was going to move along, and that's not a good thing in my book. Fans of "Beat" Takeshi; his part in this movie is not really more than a cameo, and unless you're a rabid fan, you don't need to suffer through this waste of film. 2/10
I want this. Not that. Right now.
Learning to Skim Text

LSTM-Jump

\[ p(j_i|h_i; \theta_a) \]

\[ \theta_m \]

\[ f_1(\theta_m) + f_2(\theta_a) \]

[ Yu et al., ACL’17 ]
Learning to Skim Text

\[
\min_{\theta_m, \theta_a} f_1(\theta_m) + f_2(\theta_a)
\]

differentiable

\[
R = \{ 1, -1 \}
\]

\[ f_2(\theta_a) = -\mathbb{E}_{p(N; \theta_a)}[R] \]

policy gradient: REINFORCE

[Yu et al., ACL’17]
Number prediction

Input: \( \{x_t\}_{t=0}^T \)

Output: \( x_{x_0} \)

1. Input: 4, 5, 1, 7, 6, 2  
   Output: 6

2. Input: 2, 4, 9, 4, 5, 6  
   Output: 9

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Token read</th>
<th>Test Time</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-Jump</td>
<td>90%</td>
<td>3.0</td>
<td>19 s</td>
<td>66x</td>
</tr>
<tr>
<td>LSTM</td>
<td>80%</td>
<td>1,000</td>
<td>1,250 s</td>
<td>-</td>
</tr>
</tbody>
</table>

Length = 1,000

[ Yu et al., ACL’17 ]
QA on Children’s Book Test

- Context: 20 contiguous sentences
- Query: A sentence with a word removed.
- Task: Fill the blank from 10 candidates.

Suddenly Henry remembered something: Mother's Day!
Henry and his dad always made a funny lunch for Henry's mother on Mother’s Day.

One year they made a Tomato Snowman.

Another year they baked a Sweet Potato Shoe.

Mother's Day was only a day away.
What would they make for lunch this year?
But to you and me it would have looked just as it did to Cousin Myra — a very discontented and unbecoming scowl.

“I’m awfully glad to see you, Cousin Myra,” explained Frank carefully.

“But Christmas is just a bore – a regular bore.”

That was what Uncle Edgar called things that didn’t interest him, so that Frank felt pretty sure of.

Nevertheless, he wondered uncomfortably what made Cousin Myra smile so queerly.

“Why, how dreadful!”

She said brightly.

“I thought all boys and girls looked upon Christmas as the very best time in the year.”

“We don’t,” said Frank gloomily.

“It’s just the same old thing year in and year out.

We know just exactly what is going to happen.

We even know pretty well what presents we are going to get.

And Christmas Day itself is always the same.

We’ll get up in the morning, and our stockings will be full of things.

Then there’s dinner.

It’s always so poky.

And all the uncles and aunts come to dinner – just the same old crowd, every year.

Aunt Desda always says, ‘Why, Frankie, how you have grown!’

She knows I hate to be called Frankie.

And after dinner they’ll sit round and talk the rest of the day, and that’s all.
But to you and me it would have looked just as it did to Cousin Myra—a very discontented and unbecoming scowl.

“I’m awfully glad to see you, Cousin Myra,” explained Frank carefully.

“But Christmas is just a bore—a regular bore.”

That was what Uncle Edgar called things that didn’t interest him, so that Frank felt pretty sure of.

Nevertheless, he wondered uncomfortably what made Cousin Myra smile so queerly.

“Why, how dreadful!”

She said brightly.

“I thought all boys and girls looked upon Christmas as the very best time in the year.”

“We don’t,” said Frank gloomily.

“It’s just the same old thing year in and year out.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloomily.

We don’t,” said Frank gloo
1. But to you and me it would have looked just as it did to Cousin Myra – a very discontented and unbecoming scowl.
2. “I’m awfully glad to see you, Cousin Myra,” explained Frank carefully.
3. “Christmas is just a bore – a regular bore.”
4. That was what Uncle Edgar called things that didn’t interest him, so that Frank felt pretty sure of.
5. Nevertheless, he wondered uncomfortably what made Cousin Myra smile so queerly.
6. “Why, how dreadful!”
7. She said brightly.
8. “I thought all boys and girls looked upon Christmas as the very best time in the year.”
9. “We don’t,” said Frank gloomily.
10. “It’s just the same old thing year in and year out.
11. We know just exactly what is going to happen.
12. We even know pretty well what presents we are going to get.
13. And Christmas Day itself is always the same.
14. We’ll get up in the morning, and our stockings will be full of things.
15. Then there’s dinner.
16. It’s always so poky.
17. And all the uncles and aunts come to dinner – just the same old crowd, every year.
18. Aunt Desda always says, ‘Why, Frankie, how you have grown!’
19. She knows I hate to be called Frankie.
20. And after dinner they’ll sit round and talk the rest of the day, and that’s all.
1. But to you and me it would have looked just as it did to Cousin Myra – a very discontented and unbecoming scowl.

2. “I’m awfully glad to see you, Cousin Myra,” explained Frank carefully.

3. “Christmas is just a bore – a regular bore.”

4. That was what Uncle Edgar called things that didn’t interest him, so that Frank felt pretty sure of.

5. Nevertheless, he wondered uncomfortably what made Cousin Myra smile so queerly.

6. “Why, how dreadful!”

7. She said brightly.

8. “I thought all boys and girls looked upon Christmas as the very best time in the year.”

9. “We don’t,” said Frank gloomily.

10. “It’s just the same old thing year in and year out.

11. We know just exactly what is going to happen.

12. We even know pretty well what presents we are going to get.

13. And Christmas Day itself is always the same.

14. We’ll get up in the morning, and our stockings will be full of things.

15. Then there’s dinner.

16. It’s always so poky.

17. And all the uncles and aunts come to dinner – just the same old crowd, every year.

18. Aunt Desda always says, ‘Why, Frankie, how you have grown!’

19. She knows I hate to be called Frankie.

20. And after dinner they’ll sit round and talk the rest of the day, and that’s all.
## Result

**Children’s Book Test**

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Infer Time</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-Jump</td>
<td>45.2%</td>
<td>20 s</td>
<td>6x</td>
</tr>
<tr>
<td>LSTM</td>
<td>43.8%</td>
<td>124 s</td>
<td>-</td>
</tr>
</tbody>
</table>

**Other Classification Problems**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Rotten Tomatoes</th>
<th>IMDB</th>
<th>AG News</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-Jump</td>
<td>2x</td>
<td>2.5x</td>
<td>1.3x</td>
</tr>
</tbody>
</table>

Similar or better accuracy as LSTM

[ Yu et al., ACL’17 ]
So Far

RNN / LSTM

1. Skips redundant info

LSTM-Jump

2. Works well on simple tasks
1. Still reads sequentially between jumps

Sequential reading regions

Hard to Compute in Parallel
1. Still reads sequentially between jumps

Sequential reading regions

2. Might not handle complicated tasks
Roadmap

1. Skipping Irrelevant Information [ACL’17]
2. Discarding Recurrence [ICLR’18]
3. Future Work
QANet: Combining Local Convolution and Global Self-Attention for Reading Comprehension

Yu et al.
ICLR 2018
What do RNNs capture?

1. local context

2. global interaction

3. Temporal info

Efficient substitution?
Convolution: Capturing Local Context

The weather is nice today

0.6 0.2 0.9 0.4 0.4
0.2 0.3 0.1 0.1 0.1
0.8 0.1 0.8 0.4 0.6

[Kim, EMNLP’14]; [Johnson and Zhang, NAACL’15]; [Conneau, EACL’17]; [Gehring et al., ICML’17]
Convolution: Capturing Local Context

For the convolution operation, let $k = 2$ and $d = 3$.

The convolution captures the local context by sliding a window over the input data and applying a filter to the window.

[Kim, EMNLP'14]; [Johnson and Zhang, NAACL'15]; [Conneau, EACL'17]; [Gehring et al., ICML'17]
Convolution: Capturing Local Context

- $k = 2$
- $d = 3$

<table>
<thead>
<tr>
<th></th>
<th>0.6</th>
<th>0.2</th>
<th>0.9</th>
<th>0.4</th>
<th>0.4</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.8</td>
<td>0.4</td>
<td>0.6</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.6</td>
<td>1.1</td>
<td>2.5</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Kim, EMNLP'14]; [Johnson and Zhang, NAACL'15]; [Conneau, EACL'17]; [Gehring et al., ICML'17]
Convolution: Capturing Local Context

$k = 2$

$d = 3$

\[
\begin{array}{cccccc}
0.6 & 0.2 & 0.9 & 0.4 & 0.4 & 0.0 \\
0.2 & 0.3 & 0.1 & 0.1 & 0.1 & 0.0 \\
0.8 & 0.1 & 0.8 & 0.4 & 0.6 & 0.0 \\
0.4 & 0.6 & 1.1 & 2.5 & 0.7 & \\
\end{array}
\]

[Kim, EMNLP’14]; [Johnson and Zhang, NAACL’15]; [Conneau, EACL’17]; [Gehring et al., ICML’17]
Convolution: Capturing Local Context

$k = 2 \quad k = 3 \quad k = 3$

$d = 3$

k-gram features

Fully parallel!

[Kim, EMNLP’14]; [Johnson and Zhang, NAACL’15]; [Conneau, EACL’17]; [Gehring et al., ICML’17]
How about Global Interaction?

1. May need $O(\log_k N)$ layers
2. Interaction may become weaker

N: Seq length. k: Filter size.

[ Yu et al., ICLR’18 ]
Self-Attention: Global Interaction
Attention: interaction of two seqs with weighted average

- Life is a box of chocolates, you
- The cat stuck out its tongue and licked its owner

[Bahdanau et al., ICLR’15]; [Luong et al., EMNLP’15]
The weather is nice today

[Vaswani et al., NIPS’17]
The weather is nice today

\[ \text{The} \times 0.6 + \text{weather} \times 0.2 + \text{is} \times 0.9 + \text{nice} \times 0.4 + \text{today} \times 0.4 \]

\[ \text{The} \times 0.2 + \text{weather} \times 0.3 + \text{is} \times 0.1 + \text{nice} \times 0.8 + \text{today} \times 0.6 \]

\[ \text{The} \times 0.8 + \text{weather} \times 0.1 + \text{is} \times 0.8 + \text{nice} \times 0.4 + \text{today} \times 0.6 \]

[Vaswani et al., NIPS’17]
The weather is nice today

\[
\begin{align*}
\text{w}_1, \text{w}_2, \text{w}_3, \text{w}_4, \text{w}_5 &= \text{softmax} \left( \begin{array}{c}
0.6 \\
0.2 \\
0.8 \\
0.1 \\
0.8 \\
\end{array} \right)
\end{align*}
\]

[Vaswani et al., NIPS’17]
The weather is nice today

Fully parallel!

[Vaswani et al., NIPS’17]
### Complexity

<table>
<thead>
<tr>
<th>Operation</th>
<th>Per Unit</th>
<th>Total Per Layer</th>
<th>Sequential Op (Path Memory)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Attn</td>
<td>$O(Nd)$</td>
<td>$O(N^2d)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>Conv</td>
<td>$O(kd^2)$</td>
<td>$O(kNd^2)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td>RNN</td>
<td>$O(d^2)$</td>
<td>$O(Nd^2)$</td>
<td>$O(N)$</td>
</tr>
</tbody>
</table>

- **N**: Seq length.
- **d**: Dim. ($N > d$)
- **k**: Filter size.

[Yu et al., ICLR’18 ]
Explicitly Encode Temporal Info

RNN

Position Embedding

[Vaswani et al., NIPS’17]
QANet Encoder Block

[Yu et al., ICLR’18]
We don’t have recurrence anymore
Lots of QA datasets Available

- SQuAD
- SQuAD 2.0
- QuAC
- CoQA
- HotpotQA
- RACE
- Narrative QA
- TriviaQA
- MS MARCO
In education, teachers **facilitate student learning**, often in a school or academy or perhaps in another environment such as outdoors. A teacher who teaches on an individual basis may be described as a tutor.

**Question:** What is the role of teachers in education?

**Groundtruth:** facilitate student learning

**Prediction 1:** facilitate student learning  
**EM = 1, F1 = 1**

**Prediction 2:** student learning  
**EM = 0, F1 = 0.8**

**Prediction 3:** teachers facilitate student learning  
**EM = 0, F1 = 0.86**

[Yu et al., ICLR’18]
General Framework

- **Start Position**
- **End Position**
- **RNN Encoder**
- **Document-Query Attention**

**Similar architectures:**
- R-Net [Wang et al., ACL’17]
- BiDAF [Seo et al., ICLR’17]
- DCN [Xiong et al., ICLR’17]
- DrQA [Chen et al, ACL’17]

---

- **fusion layer**
- **query-aware**
- **context-aware**
- **raw texts to vectors**

**Document-Query Attention**

- **elmo/cove**
- **char**
- **word**

**Raw texts to vectors**

<!-- Document | Query "[Yu et al., ICLR’18 ]" -->
QANet

1. Leverage existing techniques in image classification for free
stochastic depth, residual connection, squeeze and excitation, ......

2. Fully Feedforward – Fast
train with more data

130 layers

[Yu et al., ICLR’18 ]
Data Augmentation

[Yu et al., ICLR’18]
Data Augmentation

Do Not work for languages

[Yu et al., ICLR’18]
Autrefois, le thé avait été utilisé surtout pour les moines bouddhistes pour rester éveillé pendant la méditation.

Previously, tea had been used primarily for Buddhist monks to stay awake during meditation.

In the past, tea was used mostly for Buddhist monks to stay awake during the meditation.

[Yu et al., ICLR’18 ]
About the SQuAD

- More than 50 teams.
- Hundreds of submissions.

Industry:

Academia:

[Yu et al., ICLR’18]
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>QANet (ensemble)</td>
<td>84.454</td>
<td>90.490</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>r-net (ensemble)</td>
<td>84.003</td>
<td>90.147</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MARS (ensemble)</td>
<td>83.982</td>
<td>89.796</td>
</tr>
<tr>
<td></td>
<td>YUANFUDAO research NLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>QANet (ensemble)</td>
<td>83.877</td>
<td>89.737</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MARS (single model)</td>
<td>83.122</td>
<td>89.224</td>
</tr>
<tr>
<td></td>
<td>YUANFUDAO research NLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>QANet (ensemble)</td>
<td>82.744</td>
<td>89.045</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>MARS (single model)</td>
<td>82.587</td>
<td>88.880</td>
</tr>
<tr>
<td></td>
<td>YUANFUDAO research NLP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Reinforced Mnemonic Reader + A2D (ensemble model)</td>
<td>82.849</td>
<td>88.764</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia &amp; NUDT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>QANet (single)</td>
<td>82.471</td>
<td>89.306</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#1 on SQuAD (Mar-Aug 2018)
<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Human Performance</td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td>Stanford University</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Rajpurkar et al. '16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT (single model)</td>
<td>85.083</td>
<td>91.835</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>nnNet (ensemble)</td>
<td>85.356</td>
<td>91.202</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>nnNet (ensemble)</td>
<td>85.954</td>
<td>91.677</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>QANet (ensemble)</td>
<td>84.454</td>
<td>90.490</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>r-net (ensemble)</td>
<td>84.003</td>
<td>90.147</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>QANet (ensemble)</td>
<td>83.877</td>
<td>89.737</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Inference Speedup

Training Speedup

QANet

The RNN-based competitors on leaderboard

Fast and Accurate!

[Yu et al., ICLR’18]
**Stanford Dawnbench Competition:**

<table>
<thead>
<tr>
<th>Model</th>
<th>Time to 75 F1</th>
<th>Hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td>QANet</td>
<td>0:45:56</td>
<td>1 TPUv2</td>
</tr>
<tr>
<td>BiDAF</td>
<td>7:38:10</td>
<td>1 K80/61 GB/4 CPU</td>
</tr>
<tr>
<td>BiDAF</td>
<td>7:51:22</td>
<td>1 P100/512 GB/56 CPU</td>
</tr>
<tr>
<td>BiDAF</td>
<td>8:43:40</td>
<td>1 K80/30 GB/8 CPU</td>
</tr>
<tr>
<td>BiDAF</td>
<td>10:50:22</td>
<td>60 GB/16 CPU</td>
</tr>
</tbody>
</table>

An End-to-End Deep Learning Benchmark and Competition

10x speedup

[Yu et al., ICLR’18]
Failure cases

Who else did Tesla make the acquaintance of in 1886?

Context

… In late 1886 Tesla met Alfred S. Brown, a Western Union superintendent, and New York attorney Charles F. Peck … Together in April 1887 they formed the Tesla Electric Company with an agreement that profits from generated patents would go 1/3 to Tesla, 1/3 to Peck and Brown, and 1/3 to fund development.

Ground truths

Charles F. Peck

Predictions from 14 models

model 0: Alfred S. Brown
model 1: Alfred S. Brown
model 2: Alfred S. Brown
model 3: Alfred S. Brown
model 4: Alfred S. Brown
model 5: Alfred S. Brown
model 6: Alfred S. Brown
model 7: Alfred S. Brown
model 8: Alfred S. Brown
model 9: Alfred S. Brown
model 10: Alfred S. Brown
model 11: Alfred S. Brown, a Western Union superintendent, and New York attorney Charles F. Peck
model 12: Alfred S. Brown
model 13: Peck and Brown
Failure cases

Question
Where by mass is oxygen a major part?

Context
… Oxygen constitutes 49.2% of the Earth's crust by mass and is the major component of the world's oceans (88.8% by mass). Oxygen gas is the second most common component of the Earth's atmosphere, taking up 20.8% of its volume and 23.1% of its mass (some 10^15 tonnes) …

Ground truths
world's oceans

Predictions from 14 models
model 0: 88.8%
model 1: 88.8%
model 2: world's oceans
model 3: 88.8%
model 4: the world's oceans
model 5: 88.8%
model 6: 88.8%
model 7: 88.8%
model 8: world's oceans
model 9: world's oceans
model 10: 88.8%
model 11: 88.8%
model 12: world's oceans
model 13: world's oceans
QA is far from Solved!
Roadmap

1. Skipping Irrelevant Information [ACL’17]
2. Discarding Recurrence [ICLR’18]
3. Future Work
Future Work

1. Downstream Applications

2. AutoML for QA.

3. Real Human Level Comprehension
Downstream Applications

1. Medical QA
   - Input: doctor-patient dialog.
   - Question: what is the syndrome of the patient?

2. Web QA
   - Query: Why is the sky blue?
AutoML (Neural Architecture Search)

RNN Controller  

network specs

Child Networks

test accuracy  
as reward

Reinforcement Learning
Human Level Comprehension

1. Extremely long texts (Many books)
2. Complicated reasoning (e.g., Medical License Exam)
Summary

RNN / LSTM

LSTM-Jump

QANet
Thank you!