Natural Language Processing with Deep Learning
CS224N

The Future of Deep Learning + NLP
Kevin Clark
Deep Learning for NLP 5 years ago

- No Seq2Seq
- No Attention
- No large-scale QA/reading comprehension datasets
- No TensorFlow or Pytorch
- ...

...
Future of Deep Learning + NLP

• **Harnessing Unlabeled Data**
  - Back-translation and unsupervised machine translation
  - Scaling up pre-training and GPT-2

• **What’s next?**
  - Risks and social impact of NLP technology
  - Future directions of research
Why has deep learning been so successful recently?

Gentlemen, our learner overgeneralizes because the VC-Dimension of our kernel is too high. Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin.

Neural networks:

Stack more layers.
Why has deep learning been so successful recently?

1980s and 1990s

Accuracy

Scale (data size, model size)

neural networks

other approaches
Big deep learning successes

- Image Recognition: Widely used by Google, Facebook, etc.
- Machine Translation: Google translate, etc.
- Game Playing: Atari Games, AlphaGo, and more
Big deep learning successes

- **Image Recognition:** ImageNet: 14 million examples

- **Machine Translation:** WMT: Millions of sentence pairs

- **Game Playing:**
  - 10s of millions of frames for Atari AI
  - 10s of millions of self-play games for AlphaZero
NLP Datasets

- Even for English, most tasks have 100K or less labeled examples.

- And there is even less data available for other languages.
  - There are thousands of languages, hundreds with > 1 million native speakers
  - <10% of people speak English as their first language

- Increasingly popular solution: use **unlabeled** data.
Using Unlabeled Data for Translation
Machine Translation Data

- Acquiring translations required human expertise
  - Limits the size and domain of data

- Monolingual text is easier to acquire!
Pre-Training

1. Separately Train Encoder and Decoder as Language Models

2. Then Train Jointly on Bilingual Data
Pre-Training

- English -> German Results: 2+ BLEU point improvement

Ramachandran et al., 2017
Self-Training

• Problem with pre-training: no “interaction” between the two languages during pre-training
• Self-training: label unlabeled data to get noisy training examples

I traveled to Belgium
Translation: Je suis étudiant

MT Model

train

Je suis étudiant
Self-Training

- Circular?

I traveled to Belgium

Translation: Je suis étudiant

MT Model

Je suis étudiant

I already knew that!
Back-Translation

- Have two machine translation models going in opposite directions (en -> fr) and (fr -> en)

I traveled to Belgium

\[ \text{Je suis étudiant} \]

Translation: I traveled to Belgium

\[ \text{Je suis étudiant} \]
Back-Translation

- Have two machine translation models going in opposite directions (*en* -> *fr*) and (*fr* -> *en*)

- No longer circular

- Models never see "bad" translations, only bad inputs

I traveled to Belgium

\[ \text{en} \rightarrow \text{fr} \quad \text{Model} \quad \text{Je suis étudiant} \]

\[ \text{Je suis étudiant} \quad \text{Translation: I traveled to Belgium} \]

\[ \text{fr} \rightarrow \text{en} \quad \text{Model} \]

train
Large-Scale Back-Translation

- 4.5M English-German sentence pairs and 226M monolingual sentences

<table>
<thead>
<tr>
<th>Citation</th>
<th>Model</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shazeer et al., 2017</td>
<td>Best Pre-Transformer Result</td>
<td>26.0</td>
</tr>
<tr>
<td>Vaswani et al., 2017</td>
<td>Transformer</td>
<td>28.4</td>
</tr>
<tr>
<td>Shaw et al, 2018</td>
<td>Transformer + Improved Positional Embeddings</td>
<td>29.1</td>
</tr>
<tr>
<td>Edunov et al., 2018</td>
<td>Transformer + Back-Translation</td>
<td>35.0</td>
</tr>
</tbody>
</table>
What if there is no Bilingual Data?
What if there is no Bilingual Data?
Unsupervised Word Translation
Unsupervised Word Translation

- Cross-lingual word embeddings
  - Shared embedding space for both languages
  - Keep the normal nice properties of word embeddings
  - But also want words close to their translations
- Want to learn from monolingual corpora
Unsupervised Word Translation

- Word embeddings have a lot of structure

- Assumption: that structure should be similar across languages
Unsupervised Word Translation

- Word embeddings have a lot of structure

- Assumption: that structure should be similar across languages
Unsupervised Word Translation

- First run word2vec on monolingual corpora, getting words embeddings $X$ and $Y$
- Learn an (orthogonal) matrix $W$ such that $WX \sim Y$
Unsupervised Word Translation

- Learn $W$ with adversarial training.
- Discriminator: predict if an embedding is from $Y$ or it is a transformed embedding $Wx$ originally from $X$.
- Train $W$ so the Discriminator gets “confused”

Discriminator predicts: is the circled point red or blue?

- Obviously red

- Other tricks can be used to further improve performance, see Word Translation without Parallel Data
Unsupervised Machine Translation
Unsupervised Machine Translation

- Model: **same** encoder-decoder used for both languages
- Initialize with cross-lingual word embeddings

I am a student

Je suis étudiant

Je suis étudiant
Unsupervised Neural Machine Translation

- Training objective 1: de-noising autoencoder
Unsupervised Neural Machine Translation

- Training objective 2: back translation
  - First translate $fr \rightarrow en$
  - Then use as a “supervised” example to train $en \rightarrow fr$

I am student

Je suis étudiant <EOS>
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point

```
I
am
a
student
<En>
```

```
I
am
a
student
<EOS>
```
Why Does This Work?

- Cross lingual embeddings and shared encoder gives the model a starting point

```
I am a student
Je suis étudiant
```
Why Does This Work?

• Cross lingual embeddings and shared encoder gives the model a starting point

I am a student

I am a student

Je suis étudiant
Why Does This Work?

• Objectives encourage language-agnostic representation

Auto-encoder example

I am a student ➔ Encoder vector ➔ I am a student

Back-translation example

Je suis étudiant ➔ Encoder vector ➔ I am a student
Why Does This Work?

• Objectives encourage language-agnostic representation

Auto-encoder example

I am a student $\rightarrow$ Encoder vector $\rightarrow$ I am a student

Back-translation example

Je suis étudiant $\rightarrow$ Encoder vector $\rightarrow$ I am a student

need to be the same!
Unsupervised Machine Translation

- Horizontal lines are unsupervised models, the rest are supervised

Lample et al., 2018
Attribute Transfer

- Collector corpora of “relaxed” and “annoyed” tweets using hashtags
- Learn ununsupervised MT model

<table>
<thead>
<tr>
<th>Relaxed ↔ Annoyed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relaxed</td>
</tr>
<tr>
<td>Annoyed</td>
</tr>
<tr>
<td>Annoyed</td>
</tr>
<tr>
<td>Relaxed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male ↔ Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
</tbody>
</table>

Lample et al., 2019
Not so Fast

• English, French, and German are fairly similar

• On very different languages (e.g., English and Turkish)...
  • Purely unsupervised word translation doesn’t work very. Need *seed dictionary* of likely translations.
    • Simple trick: use identical strings from both vocabularies

• UNMT barely works

<table>
<thead>
<tr>
<th>System</th>
<th>English-Turkish BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>~20</td>
</tr>
<tr>
<td>Word-for-word unsupervised</td>
<td>1.5</td>
</tr>
<tr>
<td>UNMT</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Hokamp et al., 2018
Not so Fast
Cross-Lingual BERT
Cross-Lingual BERT

Lample and Conneau., 2019
Cross-Lingual BERT

Lample and Conneau., 2019
Unsupervised MT Results

<table>
<thead>
<tr>
<th>Model</th>
<th>En-Fr</th>
<th>En-De</th>
<th>En-Ro</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNMT</td>
<td>25.1</td>
<td>17.2</td>
<td>21.2</td>
</tr>
<tr>
<td>UNMT + Pre-Training</td>
<td>33.4</td>
<td>26.4</td>
<td><strong>33.3</strong></td>
</tr>
<tr>
<td>Current supervised State-of-the-art</td>
<td><strong>45.6</strong></td>
<td><strong>34.2</strong></td>
<td>29.9</td>
</tr>
</tbody>
</table>
Huge Models and GPT-2
## Training Huge Models

<table>
<thead>
<tr>
<th>Model</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-sized LSTM</td>
<td>10M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90M</td>
</tr>
<tr>
<td>GPT</td>
<td>110M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5B</td>
</tr>
</tbody>
</table>
## Training Huge Models

<table>
<thead>
<tr>
<th>Model</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-sized LSTM</td>
<td>10M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90M</td>
</tr>
<tr>
<td>GPT</td>
<td>110M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5B</td>
</tr>
<tr>
<td>Honey Bee Brain</td>
<td>~1B synapses</td>
</tr>
</tbody>
</table>
# Training Huge Models

<table>
<thead>
<tr>
<th>Model</th>
<th># Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium-sized LSTM</td>
<td>10M</td>
</tr>
<tr>
<td>ELMo</td>
<td>90M</td>
</tr>
<tr>
<td>GPT</td>
<td>110M</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>320M</td>
</tr>
<tr>
<td>GPT-2</td>
<td>1.5B</td>
</tr>
<tr>
<td>Honey Bee Brain</td>
<td>~1B synapses</td>
</tr>
</tbody>
</table>

![Graphic showing comparison between an apple and an orange]


This is a General Trend in ML
Huge Models in Computer Vision

Large Scale GAN Training for High Fidelity Natural Image Synthesis

Andrew Brock
Heriot-Watt University
ajb5@hw.ac.uk

Jeff Donahue
DeepMind
jeffdonahue@google.com

Karen Simonyan
DeepMind
simonyan@google.com

- 150M parameters

See also: thispersondoesnotexist.com
Huge Models in Computer Vision

- 550M parameters

GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism

Yanping Huang
Google Brain
huangyp@google.com

Youlong Cheng
Google Brain
ylc@google.com

Dehao Chen
Google Brain
dehao@google.com

HyoukJoong Lee
Google Brain
hyouklee@google.com

Jiquan Ngiam
Google Brain
jngiam@google.com

Quoc V. Le
Google Brain
qvl@google.com

Zhifeng Chen
Google Brain
zhifengc@google.com

ImageNet Results

Top-1 Accuracy vs. Number of Parameters (Millions)
Training Huge Models

- Better hardware
- Data and Model parallelism
GPT-2

- Just a really big Transformer LM
- Trained on 40GB of text
  - Quite a bit of effort going into making sure the dataset is good quality
  - Take webpages from reddit links with high karma
So What Can GPT-2 Do?

- Obviously, language modeling (but very well)!
- Gets state-of-the-art perplexities on datasets it’s not even trained on!

<table>
<thead>
<tr>
<th></th>
<th>LAMBADA (PPL)</th>
<th>LAMBADA (ACC)</th>
<th>CBT-CN (ACC)</th>
<th>CBT-NE (ACC)</th>
<th>WikiText2 (PPL)</th>
<th>PTB (PPL)</th>
<th>enwik8 (BPB)</th>
<th>text8 (BPC)</th>
<th>WikiText103 (PPL)</th>
<th>1BW (PPL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA</td>
<td>99.8</td>
<td>59.23</td>
<td>85.7</td>
<td>82.3</td>
<td>39.14</td>
<td>46.54</td>
<td>0.99</td>
<td>1.08</td>
<td>18.3</td>
<td>21.8</td>
</tr>
<tr>
<td>117M</td>
<td>35.13</td>
<td>45.99</td>
<td>87.65</td>
<td>83.4</td>
<td>29.41</td>
<td>65.85</td>
<td>1.16</td>
<td>1.17</td>
<td>37.50</td>
<td>75.20</td>
</tr>
<tr>
<td>345M</td>
<td>15.60</td>
<td>55.48</td>
<td>92.35</td>
<td>87.1</td>
<td>22.76</td>
<td>47.33</td>
<td>1.01</td>
<td>1.06</td>
<td>26.37</td>
<td>55.72</td>
</tr>
<tr>
<td>762M</td>
<td>10.87</td>
<td>60.12</td>
<td>93.45</td>
<td>88.0</td>
<td>19.93</td>
<td>40.31</td>
<td>0.97</td>
<td>1.02</td>
<td>22.05</td>
<td>44.575</td>
</tr>
<tr>
<td>1542M</td>
<td>8.63</td>
<td>63.24</td>
<td>93.30</td>
<td>89.05</td>
<td>18.34</td>
<td>35.76</td>
<td>0.93</td>
<td>0.98</td>
<td>17.48</td>
<td>42.16</td>
</tr>
</tbody>
</table>

Radford et al., 2019
So What Can GPT-2 Do?

• **Zero-Shot Learning**: no supervised training data!
  • Ask LM to generate from a prompt

• **Reading Comprehension**: <context> <question> A:

• **Summarization**: <article> TL;DR:

• **Translation**: 
  <English sentence1> = <French sentence1>
  <English sentence 2> = <French sentence 2>
  ..... 
  <Source sentence> =

• **Question Answering**: <question> A:
GPT-2 Results
How can GPT-2 be doing translation?

- It’s just given a big corpus of text that’s almost all English
How can GPT-2 be doing translation?

- It’s just given a big corpus of text that’s almost all English

“I’m not the cleverest man in the world, but like they say in French: Je ne suis pas un imbecile [I’m not a fool].

In a now-deleted post from Aug. 16, Soheil Eid, Tory candidate in the riding of Joliette, wrote in French: ”Mentez mentez, il en restera toujours quelque chose,” which translates as, "Lie lie and something will always remain."

“I hate the word ‘perfume,’” Burr says. ‘It’s somewhat better in French: ‘parfum.’

If listened carefully at 29:55, a conversation can be heard between two guys in French: “-Comment on fait pour aller de l’autre côté? -Quel autre côté?”, which means “- How do you get to the other side? - What side?”. If this sounds like a bit of a stretch, consider this question in French: As-tu aller au cinéma?, or Did you go to the movies?, which literally translates as Have-you to go to movies/theater?

“Brevet Sans Garantie Du Gouvernement”, translated to English: “Patented without government warranty”.

GPT-2 Question Answering

- Simple baseline: 1% accuracy
- GPT-2: ~4% accuracy
- Cherry-picked most confident results

<table>
<thead>
<tr>
<th>Question</th>
<th>Generated Answer</th>
<th>Correct</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who wrote the book the origin of species?</td>
<td>Charles Darwin</td>
<td>✓</td>
<td>83.4%</td>
</tr>
<tr>
<td>Who is the founder of the ubuntu project?</td>
<td>Mark Shuttleworth</td>
<td>✓</td>
<td>82.0%</td>
</tr>
<tr>
<td>Who is the quarterback for the green bay packers?</td>
<td>Aaron Rodgers</td>
<td>✓</td>
<td>81.1%</td>
</tr>
<tr>
<td>Panda is a national animal of which country?</td>
<td>China</td>
<td>✓</td>
<td>76.8%</td>
</tr>
<tr>
<td>Who came up with the theory of relativity?</td>
<td>Albert Einstein</td>
<td>✓</td>
<td>76.4%</td>
</tr>
<tr>
<td>When was the first star wars film released?</td>
<td>1977</td>
<td>✓</td>
<td>71.4%</td>
</tr>
<tr>
<td>What is the most common blood type in sweden?</td>
<td>A</td>
<td>X</td>
<td>70.6%</td>
</tr>
<tr>
<td>Who is regarded as the founder of psychoanalysis?</td>
<td>Sigmund Freud</td>
<td>✓</td>
<td>69.3%</td>
</tr>
<tr>
<td>Who took the first steps on the moon in 1969?</td>
<td>Neil Armstrong</td>
<td>✓</td>
<td>66.8%</td>
</tr>
<tr>
<td>Who is the largest supermarket chain in the uk?</td>
<td>Tesco</td>
<td>✓</td>
<td>65.3%</td>
</tr>
<tr>
<td>What is the meaning of shalom in english?</td>
<td>peace</td>
<td>✓</td>
<td>64.0%</td>
</tr>
<tr>
<td>Who was the author of the art of war?</td>
<td>Sun Tzu</td>
<td>✓</td>
<td>59.6%</td>
</tr>
<tr>
<td>Largest state in the us by land mass?</td>
<td>California</td>
<td>X</td>
<td>59.2%</td>
</tr>
</tbody>
</table>
What happens as models get even bigger?

• For several tasks performance seems to increase with $\log(\text{model size})$
What happens as models get even bigger?

- But trend isn’t clear
GPT-2 Reaction
Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter:
GPT-2 Reaction

Elon Musk-founded OpenAI builds artificial intelligence so powerful it must be kept locked up for the good of humanity

Jasper Hamill  Friday 15 Feb 2019 10:06 am

Machine-generated text is about to break the internet

OpenAI built a text generator so good, it’s considered too dangerous to release

Zack Whittaker  @zackwhittaker  /  3 weeks ago
GPT-2 Reaction

Just wanted to give you all a heads up, our lab found an amazing breakthrough in language understanding, but we also worry it may fall into the wrong hands. so we decided to scrap it and only publish the regular *ACL stuff instead. Big respect for the team for their great work.

10:08 AM - 15 Feb 2019

[Discussion] Should I release my MNIST model or keep it closed source fearing malicious use?

Discussion

Today I trained a 23064 layer ResNet and it got 99.6% accuracy on MNIST. I would love to share the model but I fear it being used maliciously. What if it is used to read documents by the Russians? What are your thoughts?
GPT-2 Reaction

OpenAI: Please Open Source Your Language Model
19.FEB.2019

Hugh Zhang
Stanford University

OpenAI Shouldn’t Release Their Full Language Model
03.MAR.2019

Eric Zelikman
GPT-2 Reaction

Some arguments for release: 

Some arguments against:
GPT-2 Reaction

Some arguments for release:
• This model isn’t much different from existing work
• Not long until these models are easy to train
  • And we’re already at this point for images/speech
• Photoshop
• Researchers should study this model to learn defenses
• Dangerous PR Hype
• Reproducibility is crucial for science
• ...

Some arguments against:
• Danger of fake reviews, news comments, etc.
  • Already done by companies and governments
• Precedent
  • Event if this model isn’t dangerous, later ones will be even better
• Smaller model is being released
• ....
Today's meta-Twitter summary for machine learning:
None of us have any consensus on what we're doing when it comes to responsible disclosure, dual use, or how to interact with the media.
This should be concerning for us all, in and out of the field.
GPT-2 Reaction

• Should NLP experts be the ones making these decisions?
  • Experts on computer security?
  • Experts on technology and society?
  • Experts on ethics?

• Need for more interdisciplinary science

• Many other examples of NLP with big social ramifications, especially with regards to bias/fairness
High-Impact Decisions

• Growing interest in using NLP to help with high-impact decision making
  • Judicial decisions
  • Hiring
  • Grading tests

• Plus side: can quickly evaluate a machine learning system for some kinds of bias

• However, machine learning reflects or even amplifies bias in training data
  • ...which could lead to the creation of even more biased data
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

Intelligent Machines

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.
High-Impact Decisions

Ben Zimmer @bgzimmer · 2 Jul 2018
This gobbledygook earns a perfect grade from the GRE's automated essay scoring system. Algorithms writing for algorithms. npr.org/2018/06/30/624...

"History by mimic has not, and presumably never will be precipitously but blithely ensconced. Society will always encompass imaginativeness; many of scrutinizations but a few for an amanuensis. The perjured imaginativeness lies in the area of theory of knowledge but also the field of literature. Instead of enthralling the analysis, grounds constitutes both a disparaging quip and a diligent explanation."
Chatbots

- Potential for positive impact

- But big risks

AI ROBOTS LEARNING RACISM, SEXISM AND OTHER PREJUDICES FROM HUMANS, STUDY FINDS
What did BERT “solve” and what do we work on next?
<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag-of-Vectors</td>
<td>58.6</td>
</tr>
<tr>
<td>BiLSTM + Attention</td>
<td>63.1</td>
</tr>
<tr>
<td>BiLSTM + Attention + ELMo</td>
<td>66.5</td>
</tr>
<tr>
<td>GPT</td>
<td>72.8</td>
</tr>
<tr>
<td>BERT-Large</td>
<td>80.5</td>
</tr>
<tr>
<td>Human</td>
<td>87.1</td>
</tr>
</tbody>
</table>
The Death of Architecture Engineering?

Some SQuAD NN Architectures
The Death of Architecture Engineering?

Some SQuAD NN Architectures

Attention Is All You Need
The Death of Architecture Engineering?

- 6 months of research on architecture design, get 1 F1 point improvement
- ... Or just make BERT 3x bigger, get 5 F1 points
- Top 20 entrants on the SQuAD leaderboard all use BERT
Harder Natural Language Understanding

• Reading comprehension...
  • On longer documents or multiple documents
  • That requires multi-hop reasoning
  • Situated in a dialogue

• Key problem with many existing reading comprehension datasets: *People writing the questions see the context*
  • Not realistic
  • Encourages easy questions
QuAC: Question Answering in Context

• Dialogue between a student who asks questions and a teacher who answers
  • Teacher sees Wikipedia article on the subject, student doesn’t

---

Choi et al., 2018
**QuAC: Question Answering in Context**

- Still a big gap to human performance

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>F1</th>
<th>HEQQ</th>
<th>HEQD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance (Choi et al. EMNLP '18)</td>
<td>81.1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td><em>NTT Media Intelligence Labs</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT w/ 2-context (single model)</td>
<td>64.9</td>
<td>60.2</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td><em>Anonymous</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GraphFlow (single model)</td>
<td>64.9</td>
<td>60.3</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td><em>Anonymous</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>FlowQA (single model)</td>
<td>64.1</td>
<td>59.6</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td><em>Allen Institute of AI</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BERT + History Answer Embedding (single model)</td>
<td>62.4</td>
<td>57.8</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td><em>Anonymous</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BiDAF++ w/ 2-Context (single model)</td>
<td>60.1</td>
<td>54.8</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td><em>baseline</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BiDAF++ (single model)</td>
<td>50.2</td>
<td>43.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td><em>baseline</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
HotPotQA

- Designed to require *multi-hop reasoning*
- Questions are over *multiple documents*

Paragraph A, Return to Olympus:
[1] *Return to Olympus* is the only album by the alternative rock band Malfunkshun. [2] It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990. [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:

Q: What was the former band of the member of Mother Love Bone who died just before the release of “Apple”?
A: Malfunkshun

**Supporting facts:** 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HOTPOTQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Zang et al., 2018
### HotPotQA

- **Human performance is above 90 F1**

<table>
<thead>
<tr>
<th>Model</th>
<th>Code</th>
<th>Ans</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. QFE (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>53.86 68.06</td>
</tr>
<tr>
<td>2. GRN (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>52.92 66.71</td>
</tr>
<tr>
<td>3. DFGN + BERT (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>55.17 68.49</td>
</tr>
<tr>
<td>4. BERT Plus (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>55.84 69.76</td>
</tr>
<tr>
<td>5. Baseline Model (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>45.60 59.02</td>
</tr>
<tr>
<td>6. DecompRC (single model)</td>
<td><img src="image" alt="Code" /></td>
<td>55.20 69.63</td>
</tr>
</tbody>
</table>
Multi-Task Learning

- Another frontier of NLP is getting one model to perform many tasks. GLUE and DecaNLP are recent examples.

- Multi-task learning yields improvements on top of BERT

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GLUE Human Baselines</td>
<td>GLUE Human Baselines</td>
<td></td>
<td>87.1</td>
</tr>
<tr>
<td>2</td>
<td>王玮</td>
<td>ALICE large (Alibaba DAMO NLP)</td>
<td></td>
<td>83.0</td>
</tr>
<tr>
<td>3</td>
<td>Microsoft D365 AI &amp; MSR AI MT-DNNv2 (BigBird)</td>
<td></td>
<td></td>
<td>83.0</td>
</tr>
<tr>
<td>4</td>
<td>Jason Phang</td>
<td>BERT on STILTs</td>
<td></td>
<td>82.0</td>
</tr>
<tr>
<td>5</td>
<td>Jacob Devlin</td>
<td>BERT: 24-layers, 16-heads, 1024-hid</td>
<td></td>
<td>80.5</td>
</tr>
</tbody>
</table>

BERT + Multi-task
Low-Resource Settings

- Models that don’t require lots of compute power (can’t use BERT)!
  - Especially important for mobile devices

- Low-resource languages

- Low-data settings (few shot learning)
  - Meta-learning is becoming popular in ML.
Interpreting/Understanding Models

• Can we get explanations for model predictions?
• Can we understand what models like BERT know and why they work so well?

• Rapidly growing area in NLP
• Very important for some applications (e.g., healthcare)
Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models “know”

- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task
Diagnostic/Probing Classifiers

- Popular technique to see what linguistic information models “know”

- Diagnostic classifier takes representations produced by a model (e.g., BERT) as input and do some task

- Only the diagnostic classifier is trained
Diagnostic/Probing Classifiers

- Diagnostic classifiers are usually very simple (e.g., a single softmax). Otherwise they could learn to do the tasks without looking at the model representations.

- Some diagnostic tasks

<table>
<thead>
<tr>
<th>POS</th>
<th>The important thing about Disney is that it is a global [brand]₁. → NN (Noun)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constit.</td>
<td>The important thing about Disney is that it [is a global brand]₁. → VP (Verb Phrase)</td>
</tr>
<tr>
<td>Depend.</td>
<td>[Atmosphere]₁ is always [fun]₂ → nsubj (nominal subject)</td>
</tr>
<tr>
<td>Entities</td>
<td>The important thing about [Disney]₁ is that it is a global brand. → Organization</td>
</tr>
<tr>
<td>SRL</td>
<td>[The important thing about Disney]₂ [is]₁ that it is a global brand. → Arg1 (Agent)</td>
</tr>
<tr>
<td>SPR</td>
<td>[It]₁ [endorsed]₂ the White House strategy... → {awareness, existed...}</td>
</tr>
<tr>
<td>Coref.⁰</td>
<td>The important thing about [Disney]₁ is that [it]₂ is a global brand. → True</td>
</tr>
<tr>
<td>Coref.⁰</td>
<td>Characters₂ entertain audiences because [they]₁ want people to be happy. → True</td>
</tr>
<tr>
<td>Coref.⁰</td>
<td>Characters entertain [audiences]₂ because [they]₁ want people to be happy. → False</td>
</tr>
<tr>
<td>Rel.</td>
<td>The [burst]₁ has been caused by water hammer [pressure]₂. → Cause-Effect(e₂, e₁)</td>
</tr>
</tbody>
</table>
Diagnostic/ Probing Classifiers: Results

- Lower layers of BERT are better at lower-level tasks
NLP in Industry

- NLP is rapidly growing in industry as well. Two particularly big areas:
  - Dialogue
    - Chatbots
    - Customer service
  - Healthcare
    - Understanding health records
    - Understanding biomedical literature
Conclusion

• Rapid progress in the last 5 years due to deep learning.

• Even more rapid progress in the last year due to larger models, better usage of unlabeled data
  • Exciting time to be working on NLP!

• NLP is reaching the point of having big social impact, making issues like bias and security increasingly important.
Good luck with your projects!