CS388: Natural Language Processing

Lecture 19:
Pre-training 2

Greg Durrett

Texas
The University of Texas at Austin
Presentation day announcements posted
Recall: BERT

- How to learn a “deeply bidirectional” model? What happens if we just replace an LSTM with a transformer?

- Transformer LMs have to be “one-sided” (only attend to previous tokens), not what we want
Recall: Masked Language Modeling

- How to prevent cheating? Next word prediction fundamentally doesn't work for bidirectional models, instead do *masked language modeling*

- BERT formula: take a chunk of text, predict 15% of the tokens
  - For 80% (of the 15%), replace the input token with [MASK]
  - For 10%, replace w/random
  - For 10%, keep same (why?)

Devlin et al. (2019)
Recall: BERT Architecture

- BERT Base: 12 layers, 768-dim, 12 heads. Total params = 110M
- BERT Large: 24 layers, 1024-dim, 16 heads. Total params = 340M
- Positional embeddings and segment embeddings, 30k word pieces
- This is the model that gets **pre-trained** on a large corpus

Devlin et al. (2019)
Recall: What can BERT do?

- CLS token is used to provide classification decisions
- Sentence pair tasks (entailment): feed both sentences into BERT
- BERT can also do tagging by predicting tags at each word piece

Devlin et al. (2019)
This Lecture

- BART/T5
- GPT-3
- Ethical considerations of large language models
BART/T5
BART

- BERT is good for “analysis” tasks, GPT is a good language model
- What to do for seq2seq tasks?
- Sequence-to-sequence BERT variant: permute/make/delete tokens, then predict full sequence autoregressively
- Uses the transformer encoder-decoder we discussed for MT (decoder attends to encoder)

Lewis et al. (2019)
BERT vs. BART

- BERT: only parameters are an encoder, trained with masked language modeling objective
  - No way to do translation or left-to-right language modeling tasks
- BART: both an encoder and a decoder
  - Typically used for enc-dec tasks but also can use either for analysis

Lewis et al. (2019)
Infilling is longer spans than masking

Lewis et al. (2019)
Infilling is all-around a bit better than masking or deletion

Final system: combination of infilling and sentence permutation

Lewis et al. (2019)
<table>
<thead>
<tr>
<th></th>
<th>SQuAD 1.1 EM/F1</th>
<th>SQuAD 2.0 EM/F1</th>
<th>MNLI m/mm</th>
<th>SST Acc</th>
<th>QQP Acc</th>
<th>QNLI Acc</th>
<th>STS-B Acc</th>
<th>RTE Acc</th>
<th>MRPC Acc</th>
<th>CoLA Mcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>84.1/90.9</td>
<td>79.0/81.8</td>
<td>86.6/-</td>
<td>93.2</td>
<td>91.3</td>
<td>92.3</td>
<td>90.0</td>
<td>70.4</td>
<td>88.0</td>
<td>60.6</td>
</tr>
<tr>
<td>UniLM</td>
<td>-/-</td>
<td>80.5/83.4</td>
<td>87.0/85.9</td>
<td>94.5</td>
<td>-</td>
<td>92.7</td>
<td>-</td>
<td>70.9</td>
<td>-</td>
<td>61.1</td>
</tr>
<tr>
<td>XLNet</td>
<td>89.0/94.5</td>
<td>86.1/88.8</td>
<td>89.8/-</td>
<td>95.6</td>
<td>91.8</td>
<td>93.9</td>
<td>91.8</td>
<td>83.8</td>
<td>89.2</td>
<td>63.6</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>88.9/94.6</td>
<td>86.5/89.4</td>
<td>90.2/90.2</td>
<td>96.4</td>
<td>92.2</td>
<td>94.7</td>
<td>92.4</td>
<td>86.6</td>
<td>90.9</td>
<td>68.0</td>
</tr>
<tr>
<td>BART</td>
<td>88.8/94.6</td>
<td>86.1/89.2</td>
<td>89.9/90.1</td>
<td>96.6</td>
<td>92.5</td>
<td>94.9</td>
<td>91.2</td>
<td>87.0</td>
<td>90.4</td>
<td>62.8</td>
</tr>
</tbody>
</table>

- Results on GLUE not better than RoBERTa
BART for Summarization

This is the first time anyone has been recorded to run a full marathon of 42.195 kilometers (approximately 26 miles) under this pursued landmark time. It was not, however, an officially sanctioned world record, as it was not an "open race" of the IAAF. His time was 1 hour 59 minutes 40.2 seconds. Kipchoge ran in Vienna, Austria. It was an event specifically designed to help Kipchoge break the two hour barrier.

Kenyan runner Eliud Kipchoge has run a marathon in less than two hours.

PG&E stated it scheduled the blackouts in response to forecasts for high winds amid dry conditions. The aim is to reduce the risk of wildfires. Nearly 800 thousand customers were scheduled to be affected by the shutoffs which were expected to last through at least midday tomorrow.

Power has been turned off to millions of customers in California as part of a power shutoff plan.

- Good results on dialogue, summarization tasks. Will discuss more when we get to these problems

Lewis et al. (2019)
Frame many problems as sequence-to-sequence ones:

- "translate English to German: That is good."
- "cola sentence: The course is jumping well."
- "stsbg sentence 1: The rhino grazed on the grass. sentence 2: A rhino is grazing in a field."
- "summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."
- "Das ist gut."
- "not acceptable"
- "3.8"
- "six people hospitalized after a storm in attala county."

Raffel et al. (2019)
Pre-training: similar denoising scheme to BART

Original text:
Thank you for inviting me to your party last week.

Inputs:
Thank you <X> me to your party <Y> week.

Targets:
<X> for inviting <Y> last <Z>
T5

<table>
<thead>
<tr>
<th>Number of tokens</th>
<th>Repeats</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>★ Full dataset</td>
<td>0</td>
<td>83.28</td>
<td>19.24</td>
<td>80.88</td>
<td>71.36</td>
<td>26.98</td>
<td>39.82</td>
<td>27.65</td>
</tr>
<tr>
<td>$2^{29}$</td>
<td>64</td>
<td>82.87</td>
<td>19.19</td>
<td>80.97</td>
<td>72.03</td>
<td>26.83</td>
<td>39.74</td>
<td>27.63</td>
</tr>
<tr>
<td>$2^{27}$</td>
<td>256</td>
<td>82.62</td>
<td>19.20</td>
<td>79.78</td>
<td>69.97</td>
<td>27.02</td>
<td>39.71</td>
<td>27.33</td>
</tr>
<tr>
<td>$2^{25}$</td>
<td>1,024</td>
<td>79.55</td>
<td>18.57</td>
<td>76.27</td>
<td>64.76</td>
<td>26.38</td>
<td>39.56</td>
<td>26.80</td>
</tr>
<tr>
<td>$2^{23}$</td>
<td>4,096</td>
<td>76.34</td>
<td>18.33</td>
<td>70.92</td>
<td>59.29</td>
<td>26.37</td>
<td>38.84</td>
<td>25.81</td>
</tr>
</tbody>
</table>

- Colossal Cleaned Common Crawl: 750 GB of text
- We still haven't hit the limit of bigger data being useful for pre-training: here we see stronger MT results from the biggest data

Raffel et al. (2019)
GPT-3
Question: what are the scaling limits of large language models?

NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2), showed lower perplexity from this

Didn’t catch on and wasn't used for much
Scaling Laws

- Each model is a different-sized LM (GPT-style)
- With more compute, larger models get further down the loss “frontier”
- Building a bigger model (increasing compute) will decrease test loss!
These scaling laws suggest how to set model size, dataset size, and training time for big datasets. (Kaplan et al. 2020)

**Figure 1** Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.
GPT-3

- GPT-2 but even larger: 1.3B -> 175B parameter models

<table>
<thead>
<tr>
<th>Model Name</th>
<th>(n_{\text{params}})</th>
<th>(n_{\text{layers}})</th>
<th>(d_{\text{model}})</th>
<th>(n_{\text{heads}})</th>
<th>(d_{\text{head}})</th>
<th>Batch Size</th>
<th>Learning Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3 Small</td>
<td>125M</td>
<td>12</td>
<td>768</td>
<td>12</td>
<td>64</td>
<td>0.5M</td>
<td>(6.0 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 Medium</td>
<td>350M</td>
<td>24</td>
<td>1024</td>
<td>16</td>
<td>64</td>
<td>0.5M</td>
<td>(3.0 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 Large</td>
<td>760M</td>
<td>24</td>
<td>1536</td>
<td>16</td>
<td>96</td>
<td>0.5M</td>
<td>(2.5 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 XL</td>
<td>1.3B</td>
<td>24</td>
<td>2048</td>
<td>24</td>
<td>128</td>
<td>1M</td>
<td>(2.0 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 2.7B</td>
<td>2.7B</td>
<td>32</td>
<td>2560</td>
<td>32</td>
<td>80</td>
<td>1M</td>
<td>(1.6 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 6.7B</td>
<td>6.7B</td>
<td>32</td>
<td>4096</td>
<td>32</td>
<td>128</td>
<td>2M</td>
<td>(1.2 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 13B</td>
<td>13.0B</td>
<td>40</td>
<td>5140</td>
<td>40</td>
<td>128</td>
<td>2M</td>
<td>(1.0 \times 10^{-4})</td>
</tr>
<tr>
<td>GPT-3 175B or “GPT-3”</td>
<td>175.0B</td>
<td>96</td>
<td>12288</td>
<td>96</td>
<td>128</td>
<td>3.2M</td>
<td>(0.6 \times 10^{-4})</td>
</tr>
</tbody>
</table>

- Trained on 570GB of Common Crawl

- 175B parameter model’s parameters alone take >400GB to store (4 bytes per param). Trained in parallel on a “high bandwidth cluster provided by Microsoft”

Brown et al. (2020)
This is the “normal way” of doing learning in models like GPT-2
GPT-3: Few-shot Learning

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

1. Translate English to French:
2. sea otter => loutre de mer
3. peppermint => menthe poivrée
4. plush giraffe => girafe peluche
5. cheese => ..................................
Key observation: few-shot learning only works with the very largest models!
Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad

Results on other datasets are equally mixed — but still strong for a few-shot model!
**Yelp** For the Yelp Reviews Full Star dataset (Zhang et al., 2015), the task is to estimate the rating that a customer gave to a restaurant on a 1-to 5-star scale based on their review’s text. We define the following patterns for an input text $a$:

$P_1(a) = \text{It was} \ldots \text{a }$  
$P_2(a) = \text{Just} \ldots! \parallel a$  
$P_3(a) = \text{a. All in all, it was} \ldots$  
$P_4(a) = \text{a }\parallel \text{In summary, the restaurant is} \ldots$

We define a single verbalizer $v$ for all patterns as:

$v(1) = \text{terrible }$  
$v(2) = \text{bad }$  
$v(3) = \text{okay }$  
$v(4) = \text{good }$  
$v(5) = \text{great }$

“verbalizer” of labels

```
Patterns
```

Fine-tune LMs on initial small dataset (note: uses smaller LMs than GPT-3)

Repeat:

Use these models to “vote” on labels for unlabeled data

Retrain each prompt model on this dataset

Schick and Schutze et al. (2020)
Open Questions

1) How much farther can we scale these models?

2) How do we get them to work for languages other than English (discussing this later)

3) Which will win out: prompting or fine-tuning?
Ethical Considerations
Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- Grover-MEGA: $25,000
- XLNet (BERT variant): $30,000 — $60,000 (unclear)

This is for a single pre-training run...developing new pre-training techniques may require many runs

Fine-tuning these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

Pre-Training Cost (with Google/AWS)

- GPT-3: estimated to be $4.6M. This cost has a large carbon footprint
  - Carbon footprint: equivalent to driving 700,000 km by car (source: Anthropocene magazine)
  - (Counterpoints: GPT-3 isn’t trained frequently, equivalent to 100 people traveling 7000 km for a conference, can use renewables)

- BERT-Base pre-training: carbon emissions roughly on the same order as a single passenger on a flight from NY to San Francisco

Strubell et al. (2019)
https://lambdalabs.com/blog/demystifying-gpt-3/
Stochastic Parrots

- **Claim 1**: environmental cost is disproportionately born by marginalized populations, who aren’t even well-served by these tools

- **Claim 2**: massive data is fundamentally challenging to audit, contains data that is biased and is only a snapshot of a single point in time

- **Claim 3**: these models are not grounded in meaning — when they generate an answer to a question, it is merely by memorizing cooccurrence between symbols

Bender, Gebru, McMillan-Major, Shmitchell (2021)
Bias and Toxicity

- “Toxic degeneration”: systems that generate toxic stuff

GENERATION OPTIONS:

Model: GPT-2
Prompt: I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....

Toxicity: Work Safe, Toxic, Very Toxic

⚠️ Toxic generations may be triggering.

- System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

https://toxicdegeneration.allenai.org/
We are likely to assume the model is producing factual information and presenting it in a coherent way, but this is our interpretation we project on the model.

Risks: medical diagnosis (“What do I have if I have X, Y, and Z symptoms?”) could seem possible but cause serious harm.

Bender, Gebru, McMillan-Major, Shmitchell (2021)
Claim 1: environmental cost is disproportionately born by marginalized populations, who aren’t even well-served by these tools

Counterpoint: train it once and it can be reused widely

Claim 2: massive data is fundamentally challenging to audit, contains data that is biased and is only a snapshot of a single point in time

Counterpoint: a language model is “descriptive” about how language is used; to detect hate speech, these models need to be pre-trained on hate speech

Claim 3: these models are not grounded in meaning — when they generate an answer to a question, it is merely by memorizing cooccurrence between symbols

Counterpoint: these models can be fine-tuned to learn meaning from labeled data, and they enable few-shot learning

Bender, Gebru, McMillan-Major, Shmithell (2021)
Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.

Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.

Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.

https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html

Timnit Gebru was fired from Google — then the harassers arrived

Even three months after Gebru’s controversial termination from the AI Ethics team, the sustained campaign of aggressive tweets and emails keeps coming

https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean
*Small subset of the models available now

By Xiaochi Wang & Zhengyan Zhang @THUNLP