Recall: BERT

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

ELMo (Language Modeling)

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

BERT

- 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 with random
- For 10%, keep the same (why?)

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 with random
- For 10%, keep the 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

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

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)

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

Infilling is longer spans than masking

Infilling is all-around a bit better than masking or deletion

Final system: combination of infilling and sentence permutation
Results on GLUE not better than RoBERTa

Lewis et al. (2019)

Frame many problems as sequence-to-sequence ones:

Raffel et al. (2019)

Pre-training: similar denoising scheme to BART

Raffel et al. (2019)
We haven’t hit the limit of bigger data being useful for pre-training: here we see stronger MT results from the biggest data.

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.

<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>$2^{20}$</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>79.75</td>
<td>69.97</td>
<td>27.62</td>
<td>39.71</td>
<td>27.53</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>

Raffel et al. (2019)

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.

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!

Kaplan et al. (2020)
These scaling laws suggest how to set model size, dataset size, and training time for big datasets.

GPT-3

- This is the “normal” way of doing learning in models like GPT-2.

- 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.”

GPT-3: Few-shot Learning

- In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.
Key observation: few-shot learning only works with the very largest models!

Results on other datasets are equally mixed — but still strong for a few-shot model!

Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad

Prompt Engineering

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 $x$:

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

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

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

Repeat:

- Use these models to "vote" on labels for unlabeled data
- Retrain each prompt model on this dataset

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

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

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

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

- “Toxic degeneration”: systems that generate toxic stuff

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

Stochastic Parrots

- 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

Stochastic Parrots

- 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

In the News

- Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.
- Google Researcher Says She Was Fired Over Paper Highlighting Bias in A.I.
- Timnit Gebru was fired from Google — then the harassers arrived
What’s Next?

*Small subset of the models available now