Announcements

- Two more talks this semester:
  - Shinji Watanabe (CMU): Friday 11/4 11am GDC 6.302
  - Colin Raffel (UNC): Friday 11/18 11am GDC 6.302
- A5 due Tuesday

Recap: BERT

- Seq2seq pre-trained models (BART, T5): how can we leverage the same kinds of ideas we saw in BERT for seq2seq models like machine translation?
- GPT-2/GPT-3: scaling language models further
- Prompting: a new way of using large language models without taking any gradient steps

Seq2seq Pre-trained Models: BART, T5
How do we pre-train seq2seq models?

- LMs $P(w)$: trained unidirectionally
- Masked LMs: trained bidirectionally but with masking
- How can we pre-train a model for $P(y|x)$?
  - Well, why was BERT effective?
    - Predicting a mask requires some kind of text “understanding”:

- What would it take to do the same for sequence prediction?

---

**BART**

- Token Masking
- Sentence Permutation
- Document Rotation
- Token Deletion
- Text Infilling

- Several possible strategies for corrupting a sequence are explored in the BART paper

---

**BART**

- Sequence-to-sequence Transformer trained on this data: permute/make/delete tokens, then predict full sequence autoregressively
**BERT vs. BART**

- **BERT**: only parameters are an encoder, trained with masked language modeling objective. Cannot generate text or do seq2seq tasks.
  - BERT: only parameters are an encoder, trained with masked language modeling objective. Cannot generate text or do seq2seq tasks.

- **BART**: both an encoder and a decoder. Can also use just the encoder wherever we would use BERT.
  - BART: both an encoder and a decoder. Can also use just the encoder wherever we would use BERT.

---

**BART for Summarization**

- **Pre-train** on the BART task: take random chunks of text, noise them according to the schemes described, and try to “decode” the clean text.
  - Pre-train on the BART task: take random chunks of text, noise them according to the schemes described, and try to “decode” the clean text.

- **Fine-tune** on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset).
  - Fine-tune on a summarization dataset: a news article is the input and a summary of that article is the output (usually 1-3 sentences depending on the dataset).

- Can achieve good results even with few summaries to fine-tune on, compared to basic seq2seq models which require 100k+ examples to do well.
  - Can achieve good results even with few summaries to fine-tune on, compared to basic seq2seq models which require 100k+ examples to do well.

---

**BART for Summarization: Outputs**

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.

---

**BART for Summarization: Outputs**

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.
Pre-training: similar denoising scheme to BART (they were released within a week of each other in fall 2019)


Colossal Cleaned Common Crawl: 750 GB of text

We haven't hit the limit of bigger data being useful for pre-training: here we see stronger MT results from the biggest data set.
UnifiedQA

- Multiple choice
- Yes/no

- Past work: different architectures for every QA formulation. (Span selection, answer generation, multiple choice, …)
- Now: one 11B parameter T5 model

Takeaways

- BART and T5 are useful for all sorts of seq2seq tasks involving language — so if you were going to use a seq2seq model, use one of these.
- (Caveat: need specialized models for things like language-to-code, but there’s PLBART and CodeT5)

- UnifiedQA suggests that big generative models are good at generalizing across tasks and even to new tasks (although QA results have a long way to go)

- If we have a strong enough pre-trained model and train on enough tasks, can we generalize to new tasks?

- How do we specify those new tasks if they’re not close to tasks we’ve already run on?

- Answer: prompting. But to do that well, we’ll need to scale up further

OpenAI GPT/GPT2

- Very large language models using the Transformer architecture
- Straightforward left-to-right language model, trained on raw text
- GPT2: trained on 40GB of text

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Layers</th>
<th>$d_{model}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>117M</td>
<td>12</td>
<td>768</td>
</tr>
<tr>
<td>345M</td>
<td>24</td>
<td>1024</td>
</tr>
<tr>
<td>762M</td>
<td>36</td>
<td>1280</td>
</tr>
<tr>
<td>GPT-2</td>
<td>48</td>
<td>1600</td>
</tr>
</tbody>
</table>

- By far the largest of these models trained when it came out in March 2019
- Because it’s a language model, we can generate from it

Approximate size of BERT
OpenAI GPT2

Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today.

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back.

Scroll down for video

Shoplifting: Miley Cyrus was caught shoplifting from Abercrombie and Fitch on Hollywood Boulevard today (pictured)

Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- 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)


Pushing the Limits: GPT-3

- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly $10M

Questions

1) How novel is the stuff being generated? (Is it just doing nearest neighbors on a large corpus?) How can we find out?

2) Can we use this model for things beyond story generation?

3) What harms might come from this model? (OpenAI pursued a “staged release” strategy and didn’t release biggest model)
### GPT-3

**Story completion demo**

### Pre-GPT-3: Fine-tuning

- Fine-tuning: this is the “normal way” of doing learning in models like GPT-2
  - Requires computing the gradient and applying a parameter update on every example
  - This is super expensive with 175B parameters

### GPT-3: Few-shot Learning

- GPT-3 proposes an alternative: in-context learning. Just uses the off-the-shelf model, no gradient updates
  - This procedure depends heavily on the examples you pick as well as the prompt ("Translate English to French")

### GPT-3

- Key observation: few-shot learning only works with huge models!

---

Brown et al. (2020)
GPT-3

<table>
<thead>
<tr>
<th>Model</th>
<th>SuperGLUE Average</th>
<th>BoolQ Accuracy</th>
<th>CB Accuracy</th>
<th>CB F1</th>
<th>COPA Accuracy</th>
<th>RTE Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuned SOTA</td>
<td>89.0</td>
<td>91.0</td>
<td>96.9</td>
<td>93.9</td>
<td>94.8</td>
<td>92.5</td>
</tr>
<tr>
<td>Fine-tuned BERT-Large</td>
<td>69.0</td>
<td>77.4</td>
<td>83.6</td>
<td>75.7</td>
<td>70.6</td>
<td>71.7</td>
</tr>
<tr>
<td>GPT-3 Few-Shot</td>
<td>71.8</td>
<td>76.4</td>
<td>75.6</td>
<td>52.0</td>
<td>92.0</td>
<td>69.0</td>
</tr>
</tbody>
</table>

- Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad
- Results on other datasets are equally mixed — but still strong for a few-shot model!

PaLM

- “Pathways Language Model” from Google — 540B parameters!
- Much of the paper is about data curation and datacenter networking

<table>
<thead>
<tr>
<th>Model</th>
<th>Layers</th>
<th># of Heads</th>
<th>$d_{model}$</th>
<th># of Parameters (in billions)</th>
<th>Batch Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaLM 8B</td>
<td>32</td>
<td>16</td>
<td>4096</td>
<td>8.63</td>
<td>256 → 512</td>
</tr>
<tr>
<td>PaLM 62B</td>
<td>64</td>
<td>32</td>
<td>8192</td>
<td>62.50</td>
<td>512 → 1024</td>
</tr>
<tr>
<td>PaLM 540B</td>
<td>118</td>
<td>48</td>
<td>18432</td>
<td>540.35</td>
<td>512 → 1024 → 2048</td>
</tr>
</tbody>
</table>

- Another big jump over GPT-3, but other advancements meant that new systems were even better

Prompts

- Prompts can help induce the model to engage in certain behavior

- In the GPT-2 paper, “tl;dr:” (too long; didn’t read) is mentioned as a prompt that frequently shows up in the wild indicating a summary

- tl;dr is an indicator that the model should “switch into summary mode” now — and if there are enough clean instances of tl;dr in the wild, maybe the model has been trained on a ton of diverse data?

- Good prompt + a few training examples in-context = strong task performance?
**Prompting**

- Current training: GPT-3/PaLM trained on the web
- Current testing: feed in a very specific prompt and/or a set of in-context examples
- Two goals:
  1. Unify pre-training and testing phases
  2. Exploit data for downstream tasks — why are we trying to do question answering while ignoring all of the existing QA datasets?
- Instruction tuning: fine-tune on supervised tasks after pre-training
- Let’s see how an instruction-tuned GPT-3 works

**Prompting demo**

**Instruction Tuning**

**Task Generalization: T0**

- T0: tries to deliver on the goal of T5 and do many tasks with one model
- Crowdsourced prompts: instructions for how to do the tasks

Brown et al. (2020)

Sanh et al. (2021)
**Task Generalization**

- **Train**: a collection of tasks with prompts. **This uses existing labeled training data**
- **Test**: a new task specified only by a new prompt. **No training data in this task**

---

**Frontiers**

- **FLAN-PaLM (October 20, 2022)**: 1800 tasks, 540B parameter model fine-tuned on many tasks after pre-training
  - **Instruction finetuning**
    - Please answer the following question: What is the boiling point of nitrogen?
  - **Chain-of-thought finetuning**
    - The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

---

**FLAN-PaLM (October 20, 2022)**: 1800 tasks, 540B parameter model

- **MMLU task (Hendrycks et al., 2020)**: 57 high school/college/professional exams:
  - Conceptual Physics
    - When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
      - (A) 9.8 m/s²
      - (B) more than 9.8 m/s²
      - (C) less than 9.8 m/s²
      - (D) Cannot say unless the speed of throw is given.
    - In the complex z-plane, the set of points satisfying the equation \( z^2 - |z|^2 \) is a
      - (A) pair of points
      - (B) circle
      - (C) half-line
      - (D) line

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

---

**FLAN-PaLM (October 20, 2022)**: 1800 tasks, 540B parameter model

- **MMLU task (Hendrycks et al., 2020)**: 57 high school/college/professional exams:
  - Random
  - Average human rater
  - Average human expert
  - GPT-3 5-shot
  - Chinchilla 5-shot
  - PaLM 5-shot
  - Flan-PaLM 5-shot
  - Flan-PaLM 5-shot: CoT + SC

Chung et al. (2022)
Takeaways

- Pre-trained seq2seq models and generative language models can do well at lots of generation tasks.
- Prompting is a way to harness their power and learn to do many tasks with a single model. Can be done without fine-tuning.
- Instruction-tuned models are by far the best models we have for most generation and very complex language understanding tasks today. However, pre-trained models like BERT can still do well for classification.
- Biggest best models (text-davinci-002, FLAN-PaLM) are closed-source.