Building Better Language Models

Colin Raffel
a = []
for w, pos in words:
    if w in self and pos in self[w]:
        # Known word not preceded by a modifier ("good").
        if m is None:
            a.append(dict(w=[w], p=p, s=s, i=i, n=1))
        # Known word preceded by a modifier ("really good").
        if m is not None:
            a[-1]["w"].append(w)
            a[-1]["p"] = max(-1.0, min(p * a[-1]["i"], +1.0))
            a[-1]["s"] = max(-1.0, min(s * a[-1]["i"], +1.0))
            a[-1]["i"] = i
        # Known word preceded by a negation ("not really good").
        if n is not None:
            a[-1]["w"].insert(0, n)
            a[-1]["i"] = 1.0 / a[-1]["i"]
            a[-1]["n"] = -1
polarity = avg([(w, p) for w, p, s, x in a])
How is air traffic controlled? How do you become an air traffic controller? Pick one: these questions are duplicates or not duplicates.

“How is air traffic controlled?” “How do you become an air traffic controller?”

I know that the answer to “What team did the Panthers defeat?” is in “The Panthers finished the regular season [...]”. Can you tell me what it is?

Graffiti artist Banksy is believed to be behind [...]

The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Arizona Cardinals

Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?
The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the internal lighting made them too conspicuous to those outside the cab. The fleet peaked at around 75 cabs, all of which needed to return to the single depot at Lambeth to switch batteries.

Unsupervised pre-training

Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?

Yes
TriviaQA zero-shot performance

From “Language Models are Few-Shot Learners” by Brown et al.
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
  who was frank sinatra? a: an american singer, actor, and producer.

Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
  Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

Natural Language Inference
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
  If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

Summarization
https://blog.nytsoi.net/tag/reddit
  ... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

Pronoun resolution
  Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.
Summarization
The picture appeared on the wall of a Poundland store on Whymark Avenue [...] How would you rephrase that in a few words?

Paraphrase identification
"How is air traffic controlled?"  "How do you become an air traffic controller?"
Pick one: these questions are duplicates or not duplicates.

Question answering
I know that the answer to "What team did the Panthers defeat?" is in "The Panthers finished the regular season [...]". Can you tell me what it is?

Multi-task training
Zero-shot generalization

Natural language inference
Suppose "The banker contacted the professors and the athlete". Can we infer that "The banker contacted the professors"?

Graffiti artist Banksy is believed to be behind [...] Not duplicates Arizona Cardinals Yes
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
From “Crosslingual Generalization through Multitask Finetuning” by Muennighoff et al.
Performance on **held-out** tasks

Average zero-shot accuracy on 13 held-out tasks (%)

0.4B 2B 8B 68B 137B

*Untuned model*

*Instruction tuning*

\( T0 = 11B \) parameters

*From “Fine-Tuned Language Models are Zero-Shot Learners” by Wei et al.*
From "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?" by Wang et al.
From "What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization?" by Wang et al.
Zero-shot
The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1. Translate English to French:
   1. sea otter => loutre de mer
   2. peppermint => menthe poivrée
   3. plush giraffe => girafe peluche
   4. cheese =>
```

Fine-tuning
The model is trained via repeated gradient updates using a large corpus of example tasks.

```
1. sea otter => loutre de mer
   \[\text{gradient update}\]
   \[\text{example } #1\]

1. peppermint => menthe poivrée
   \[\text{gradient update}\]
   \[\text{example } #2\]

\[\text{\ldots}\]

1. plush giraffe => girafe peluche
   \[\text{gradient update}\]
   \[\text{example } #N\]

1. cheese =>
```

From “Language Models are Few-Shot Learners” by Brown et al.
TriviaQA performance

From “Language Models are Few-Shot Learners” by Brown et al.
From “Bidirectional Language Models Are Also Few-shot Learners” by Patel et al.
From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
<table>
<thead>
<tr>
<th>Method</th>
<th>Inference FLOPs</th>
<th>Training FLOPs</th>
<th>Disk space</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Few</td>
<td>1.1e12</td>
<td>2.7e16</td>
<td>4.2 MB</td>
</tr>
<tr>
<td>T0 [1]</td>
<td>1.1e12</td>
<td>0</td>
<td>0 B</td>
</tr>
<tr>
<td>T5+LM [14]</td>
<td>4.5e13</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 6.7B [4]</td>
<td>5.4e13</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 13B [4]</td>
<td>1.0e14</td>
<td>0</td>
<td>16 kB</td>
</tr>
<tr>
<td>GPT-3 175B [4]</td>
<td>1.4e15</td>
<td>0</td>
<td>16 kB</td>
</tr>
</tbody>
</table>

From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022
From "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning", Liu et al. 2022

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Few</td>
<td>75.8%</td>
</tr>
<tr>
<td>Human baseline [2]</td>
<td>73.5%</td>
</tr>
<tr>
<td>PET [50]</td>
<td>69.6%</td>
</tr>
<tr>
<td>SetFit [51]</td>
<td>66.9%</td>
</tr>
<tr>
<td>GPT-3 [4]</td>
<td>62.7%</td>
</tr>
</tbody>
</table>

Table 2: Top-5 best methods on RAFT as of writing. T-Few is the first method to outperform the human baseline and achieves over 6% higher accuracy than the next-best method.
From “Language Models are Few-Shot Learners” by Brown et al.
From “Large Language Models Struggle to Learn Long-Tail Knowledge” by Kandpal et al.
Dante was born in **Florence** in what is now **Italy**. His birth date is unknown, although it is generally believed to be around 1265. This can be deduced from autobiographic allusions in the *Divine Comedy*. Its first part implies that Alighieri was near 35 years old at the time of writing.
From “Large Language Models Struggle to Learn Long-Tail Knowledge” by Kandpal et al.
From “Large Language Models Struggle to Learn Long-Tail Knowledge” by Kandpal et al.
From "Large Language Models Struggle to Learn Long-Tail Knowledge" by Kandpal et al.
Thanks.

Please give me feedback: