NetID: komminen

Tech Review – Language models are Few-Shot Learners (GPT-3)

Introduction

GPT-3 stands for "Generative Pre-trained Transformer 3" which was released by OpenAI in 2020 and is the follow-up to a previous generation large language learning model known as GPT-2. While GPT-2 and earlier language models were well known in the research community, GPT-3 was the first language model to garner widespread coverage in non-technical press including the New York Times[1], and even garnered an article in Nature[2] about the risks of language-generating AI. The API to access GPT-3 and perform natural language tasks is now deployed in thousands of applications[3] across major companies including Microsoft, IBM, SalesForce, Cisco and Intel.

In this tech review, I dive into and summarize the academic paper[4] published by researchers alongside GPT-3 in 2020. This is a long, dense paper (40 pages for the main text, 75 with appendices) and I aim to summarize only the key points in the interest of brevity and information absorption. While I also read the GPT-2 paper[5] and a follow-on paper on InstructGPT models[6], I have omitted those from this tech review in the interest of length.

Concepts and Approach

There has been significant progress made in recent years towards challenging NLP tasks including reading comprehension, question answering, etc. using pre-trained recurrent or transformer language models that are fine-tuned without the need for task-specific architectures. The issue with this approach is that the task-agnostic architecture still needs task-specific fine-tuning with datasets of thousands to hundreds of thousands of labeled examples. This is not scalable to the wide range of language tasks. The fine-tuning on narrow tasks can lead to good performance on specific benchmarks but poor generalization to the real-world application of those tasks. Human beings also do not require these large, labeled datasets to carry out a new task – either a simple natural language description or a small handful of examples are sufficient to teach them to perform a new task (e.g., with the Mechanical Turn service). It is desirable for NLP systems to have similar fluidity and generality on the ability to perform language related tasks.

Meta-learning is an approach that aims to achieve this generality with a large pre-training phase to develop a broad set of skills and pattern recognition abilities and the use of these abilities at inference time to recognize and perform a desired task. The GPT-2 paper from the same set of authors attempts to do this via "in-context learning" where a pretrained language model is conditioned with a natural

[1] New York Times, November 2020 - https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html
[2] Nature, March 2021 - https://www.nature.com/articles/d41586-021-00530-0
[3] OpenAPI.com - https://openai.com/api/
[4] Arxiv.org - https://arxiv.org/abs/2005.14165
[5] Openai.com - https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[6] Arxix.org - https://arxiv.org/abs/2203.02155

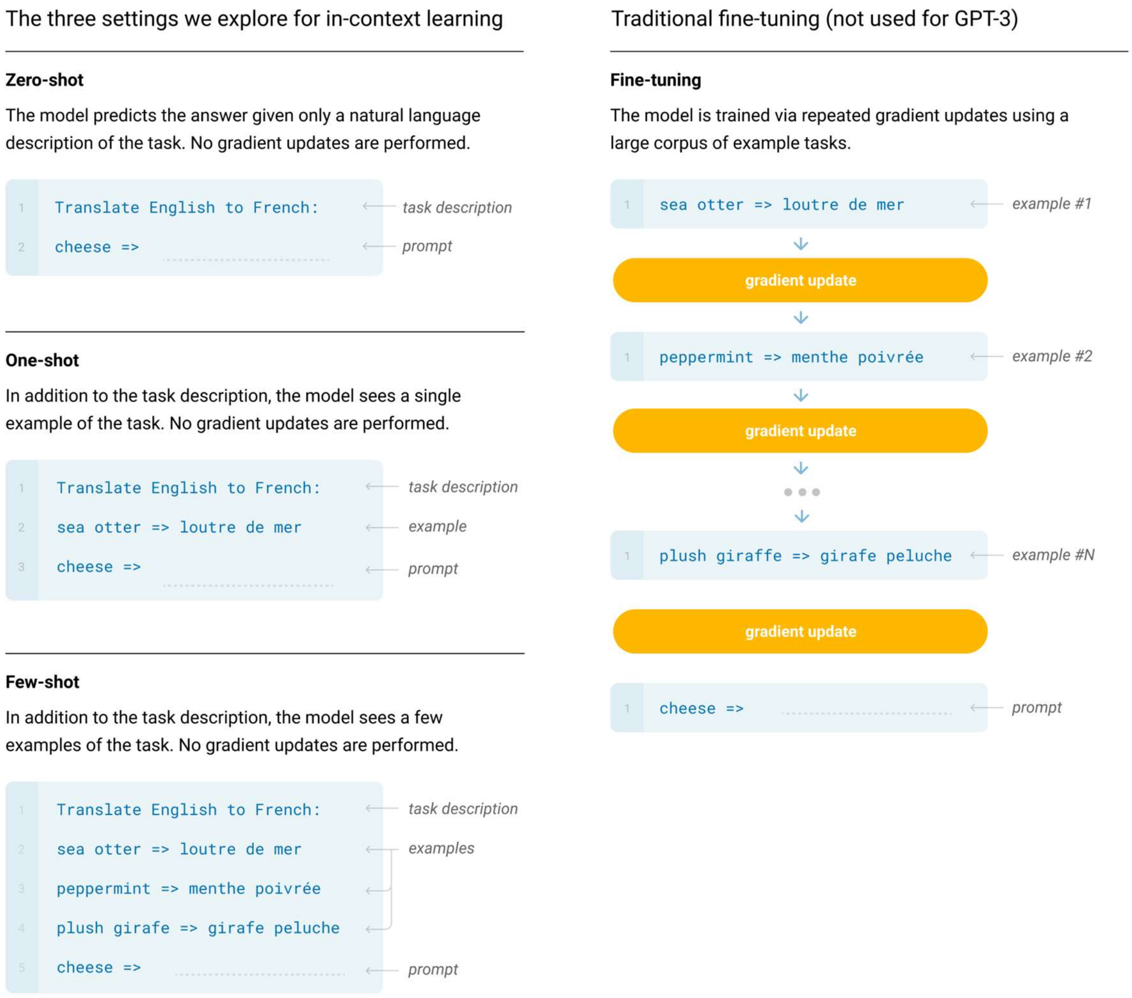
language instruction and zero or more examples of the task at inference time. The issue with GPT-2 is that it is still quite inferior to the fine-tuning approach.

A path forward from GPT-2 to GPT-3 is suggested by the increase in capacity of transformer language models from 1.5 billion parameters (GPT-2) to 17 billion parameters in later research. The paper trains a much larger language model with 175 billion parameters (called GPT-3) and measures it abilities. There are 3 conditions under which GPT-3 is evaluated:

- Zero-shot learning – No examples are allowed and only the natural language instruction is given to the model
- One-shot learning – A single example is allowed along with the natural language instruction
- Few-shot learning – Allow K examples (typically 10-100) that fit into the model’s context window

The results from these evaluations are compared with Fine-tuning methods which are the current state of-the-art with many labeled examples and best performance on the benchmarks. GPT-3 is not fine-tuned since the focus is on task-agnostic performance.

Examples of the different conditions and fine-tuning are shown in the figure below, which is taken from the paper:



The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

One-shot

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

```
Translate English to French:
sea otter => loutre de mer
cheese =>
```

Few-shot

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

```
Translate English to French:
sea otter => loutre de mer
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>
```

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

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

```
sea otter => loutre de mer
```

```
peppermint => menthe poivrée
```

```
plush giraffe => girafe peluche
```

```
cheese =>
```

The basic approach including the model, data and training is similar to the process described in the GPT-2 paper with a similar use of in-context learning. The same model and architecture as GPT-2 are used with some enhancements similar to the Sparse Transformer paper[7]. The major difference vs GPT-2 is the scale of the model…where GPT-2 was trained on 1.5 billion parameters, the model in this paper is trained on 8 different sizes ranging from 125 million parameters to 175 billion parameters (this last one is labeled as “GPT-3” in the paper and in this technical review).

The datasets used for training are a filtered and modified (to improve data quality) version of the Common Crawl dataset[8], and added high-quality datasets including WebText, two internet-based books corpora and English language Wikipedia.

Results

From a broad perspective, the authors of the paper find through their evaluation that GPT-3 achieves very promising results on NLP tasks, with the few-shot setting sometimes competitive or even surpassing the state-of-the-art (SOTA) set by fine-tuned models.

On traditional tasks like language modeling and related tasks, GPT-3 is much better than the previous GPT-2 model. The few-shot approach is even better than the SOTA fine-tuned approaches on datasets like LAMBADA while being quite good at completion prediction datasets like StoryCloze and HellaSwag.

On closed book question answering, GPT-3 few shot exceeds the SOTA on TriviaQA but lags on NaturalQS and WebQS datasets. On translation, few shot GPT-3 outperforms previous unsupervised work when translating into English though it lags behind SOTA supervised work. On common sense reasoning tasks, GPT-3 exceeds the best recorded score on PIQA but falls short on the other benchmarks. On reading comprehension tests, GPT-3 on CoQA, a free-form conversational dataset and performs worst on QuAC, a dataset which requires modeling structured dialog acts.

On synthetic and qualitative tasks, GPT-3 is very good at simple addition and subtraction when the number of digits is small (<=2) and tails off with larger numbers of digits with some displayed capacity to generalize to larger numbers of digits. Small models do poorly on all these tasks, indicating the benefits of the large parameter GPT-3. GPT-3 is also able to generate synthetic “news articles” that are difficult for human evaluators to distinguish from human-generated articles. These were short (~200 word) articles where the mean human accuracy at detecting the articles generated by GPT-3 was barely better than chance at 52%. As well, GPT-3 was proven to be proficient at using novel (i.e., made-up) words in a sentence plausibly and was also quite good at correcting English grammar.

Limitations

Despite strong improvements described in the earlier section, especially over GPT-2, GPT-3 still has notable gaps on comparison tasks such as determining if two words are used in the same way in a sentence or if one sentence implies another (WIC and ANLI). Qualitatively, it also has issues with text synthesis (semantic repetition, losing coherence over long passages, non-sequitur sentences) and “common sense physics” (the example cited is “If I put cheese into the fridge, will it melt”)

[7] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

[8] https://commoncrawl.org/the-data/

GPT-3 will likely also run into limits of large pretrained language models not being grounded in other domains of experience; for example, not knowing about knowledge from other modalities such as images, video, real-world physical interaction, etc.

Sample efficiency during pre-training is still poor with GPT-3; humans are much better at doing more with far less text than GPT-3 is trained with. It is also unclear whether the model is learning new tasks "from scratch" at inference time or simply recognizing and identifying tasks learned during training. For instance, tasks like word scrambling are likely learned from scratch while translation is likely learned during pre-training.

Broader Considerations

The paper discusses two potential categories of issues with improved language models like GPT-3:

1. **Misuse of language models**: These are hard to anticipate, but 3 factors are discussed
    a. Misuse applications: Examples provided are those which typically bottleneck on human beings to write high quality text including misinformation, spam, phishing, fraudulent writing and abuse of governmental processes. The misuse potential increases as the quality of text synthesis improves and the authors posit that the accuracy of GPT-3 on generalized tasks represents a concerning milestone
    b. Threat actor analysis: The paper posits that there is limited risk from low and medium-skill actors based on monitoring forums/chat groups and consulting with threat analysts around the earlier GPT-2 model.
    c. External incentive structures: Lowering cost of deployment and ease of use are cited as significant incentives to use language models for the misuse applications described above combined with improved language models. The authors observe that at this point, a human is still needed to filter outputs from language models like GPT-3 reducing the scalability of any mis-use operation.
2. **Bias, fairness and representation**: The authors focus on biases and limitations in GPT-3 around gender, race and religion with the broad finding that the model tended to reflect stereotypes present in the training data. The paper shows the top 10 male and female biased words, sentiment for different racial groups (Asian with high sentiment and Black with low sentiment) and sentiment for religions (including showing the co-occurrence of negative sentiment words with certain religions)

On both of the above topics, the paper urges the need for more research and proactive interventions and model changes to prevent misuse and biases from occurring.

The paper also covers the issue of energy usage (presumably given similar concerns and problems in the crypto space). The news here is more positive in that the large-scale pretraining, while very energy - intensive is amortized over the lifetime any many applications of a pre-trained model like GPT-3. They observe that tasks in the inference phase can be very efficient, on the order of <1 KWh or a few cents in energy costs.

Conclusion

This was a fascinating paper to read and digest/summarize in this tech review through the lens of the learning in CS410. While it was quite dense and there was material I didn't fully understand, the introduction from the course (especially the second half of the course) helped me parse and understand a surprisingly large part of the paper.

It is clear that advances in research combined with massive cloud-based computational power for training are resulting in models that are far beyond the state of the art of just a couple of years ago. OpenAI is at the bleeding edge of these models as demonstrated by GPT-3 and DALL-E and the resulting applications of the models in many previous “creative/human-driven” applications including software programming with GitHub CoPilot[9] and photography with stock image creation[10]. While progress is inevitable and likely to lead to large societal benefits, it is necessary to also consider the potential for misuse and bias, especially given the negative changes in society driven by earlier technological changes like mobile devices and social networks. I am both excited and apprehensive about the changes to come over the next decade from the advances in NLP and AI.

[9] Github.com - https://github.com/features/copilot

[10] Shutterstock.com - https://www.shutterstock.com/press/20435