Overview of OpenAI's GPT models

Blake Jones
blakeaj2@illinois.edu

Summary

While inroads have been made to generalized, multitask machine learning models, many ML tasks are still narrowly scoped. Current ML models are focused on a single objective and task - require significant amounts of data associated with the task in question. Attempts to expand a general ML model to multiple tasks results in skewing of model output - models are too sensitive to the training data and cannot easily generalize to multiple objectives. Multitask training is an alternative to this, whereby training is defined in dataset, task pairs. Models are then trained on these pairs, thereby creating many model modalities which can be selected from during inference. While this approach solves our narrow scope issue, it requires significant amounts of curated data to properly train.

OpenAI's aims to solve the multitask problem via GPT - a pre-trained transformer model which can be fine-tuned to a specific objective. This enables a generalized model trained on large amounts of data to be released, which can then be selected and fine-tuned for a user's specific task on a miniscule amount of data. Effectively, this language model then rewrites the language model training function from **P(output|input)** to **P(output|input, task)**. The underlying theory behind this model is that if a model can predict it's teacher's outputs, it inherently becomes an unsupervised multitask model - and with a large enough language model, the model can infer tasks and accurately predict them.

GPT2

GPT2 is the embodiment of a generalized multitask model. In order to pretrain the model, researchers used Reddit outlinks to obtain a large variety of diverse documents with relatively high quality. This diverse pretrained set enables GPT2 to be unrestricted in eligible tasks for fine-tuning. For GPT2's vocabulary, the researchers used byte-pair encoding, which incorporates both word and byte encoding. This enabled the model to retain contextual variations of words (e.g. punctuation marks), while still encompassing word-centric meanings. Detokenizers were employed to refine the scraped training dataset and remove character artifacts that would otherwise distort the models vocabulary and output (e.g. extra punctuation). In order to (mostly) prevent the testing data from being polluted by the scraped training data, the researchers used 8-grams to determine dataset overlap.

GPT2 performed well on a variety of language tasks, such as predicting choices for omitted words (Children's Book Test), predicting the end of long sentences (LAMBADA),

and resolving ambiguities in text. Interestingly, GPT2 was shown to encode concepts of tasks inherently, as when summarizing content a “task hint” greatly improved results. This implies that GPT2 works well in partitioning concepts into distinct concepts. Additionally, GPT2 appeared to “memorize” common quotes, and could reconnect these quotes to their original source material if reasonably short.

GPT3

GPT3 is the successor to GPT2, where this iteration significantly improves upon the original intent of GPT2 by removing the need for significant fine-tuning of the model. GPT3 is a significantly larger, pretrained model which is significantly more generalized, and given a text input can effectively respond in the text’s domain. As a result, GPT3 removes the need for curating domain-specific datasets and task specific fine-tuning, making GPT3 widely applicable to NLP tasks.

GPT3 exploits the recent improvements in transformer language models to support billions of parameters. This explosion in model capacity allows GPT3 to retain context-specific skills during model training (also known as meta-learning). The number of parameters in a model generally correlates with a model’s ability to retain this context knowledge, which enables quick reasoning on the request task. Similarly to GPT3 training, the team used web content (obtained from Common Crawl), which it then filtered for high quality documents and performed de-duplication. The resulting GPT3 model performed significantly well on many NLP benchmark tests, and even adapted to different tasks such as understanding and performing arithmetic. This further solidifies GPT3’s strength in context-specific learning.

Conclusion

Overall, large scale language models are incredibly powerful at performing a variety of NLP tasks when trained on proper datasets. GPT2 proved the concept that high-parameter transformer models can be pre trained to encode general language patterns and then tuned to include task-specific knowledge. GPT3 expanded on this and showed that massive transformer models trained on a diverse dataset can accurately encode tasks within the base model itself, without a need for fine-tuning. As a result, these transformer based models excel at a variety of NLP tasks, from text summarization, text inference, and question answering.

Sources

Radford, Alec, et al. “Language Models Are Unsupervised Multitask Learners Bibtex, Language Models Are Unsupervised Multitask Learners.” ***Language Models Are***

Unsupervised Multitask Learners,
https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Brown, Tom B., et al. "Language Models Are Few-Shot Learners." ***ArXiv.org*****, 22 July 2020,**
https://arxiv.org/abs/2005.14165.