Overview of OpenAI's GPT, GPT2 and GPT3

Boyun Deng | 2021-11

boyund2@illinois.edu

OpenAI's GPT (Generative Pre-trained Transformer) and it's successors GPT-2, GPT3 are generative language models that can be used to perform various NLP tasks by predicting the next word given the previous words in context. The key idea here is to use a semi-supervised approach that combines unsupervised pre-training and supervised fine-tuning: first train a Transformer model on a large dataset, and then fine-tune the model on a smaller dataset for each specific task.

Task agnostic model

Unsupervised learning does not require manual labeling data which opens up the opportunity for such a generalized model: label creation requires a large amount of manual effort, especially if there's a large amount of training data; labels often need to be optimized for specific tasks, which puts restrictions to training a more generalized model.

By performing unsupervised training on large amounts of data and generating a large number of parameters, GPT is able to make the LM flexible and powerful to adapt to more specific tasks. The caveat of such a large model is that it also requires a large amount of computing power, which is a reasonable tradeoff with manual effort.

Fine-tuned tasks

OpenAI has experimented fine-tuning GPT with supervised learning for the following tasks: natural language inference, question answering, semantic similarity, and text classification. When comparing the performance of the fine-tuned GPT LM with other state-of-the-art methods like ELMo, the fine-tuned GPT LM has out-performed in accuracy in most cases (9 out of 12 datasets).

Since the launch of GPT, there have been many interesting applications that were built based on it: summarizing customer feedback, generating conversations for game characters, customer support reply etc.

GPT-2

Released in 2019, GPT-2 is trained on 8 million web pages (WebText dataset) which increases significantly in diversity as compared to the GPT experiment which was trained on the BooksCorpus dataset. GPT-2 has 1.5 billion parameters, 10x compared to GPT, making the model more flexible and has been demonstrated by OpenAI to have resulted in a great improvement in accuracy and reduction in perplexity.

Input: GPT-2 uses Byte Pair Encoding (BFE) which allows the model to generate any Unicode string, instead of preprocessing the input using tokenization, lower-casting, and increases the number of modelable strings.

Training: GPT-2 is trained using a zero-shot setting, and is not trained to the specific datasets for fine-tuning.

GPT-3

After GPT-2 has demonstrated the improvement of increasing model capacity on accuracy and on zero-shot/few-shot setting, GPT-3 is trained to have even more parameters: there are 175 billion parameters. GPT-3 uses few-shot setting, where performance increases more rapidly than one-shot setting, when the number of parameters increases.

GPT-3 is trained on a mixture of datasets, mainly a refined version of the CommonCrawl dataset where filtering and fuzzy deduplications were applied and high quality conference data was added.

The same limitations are still seen throughout GPT, GPT-2 and GPT-3, for example, due to the ungrounded domain of training data, there are still failures in real-world-modeling; another failure mode observed is the repetition and loss of coherence.

References

- Language Understanding Paper - OpenAI https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

- Language Models are Unsupervised Multitask Learners - OpenAI
  https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- GPT-3 Powers the Next Generation of Apps - OpenAI
  https://openai.com/blog/gpt-3-apps/
- Language Models are Few-Shot Learners - OpenAI
  https://arxiv.org/abs/2005.14165