Chi Han
November 6, 2022

OpenAI GPT, GPT-2

Technology Review

1. Introduction

GPT (Radford et al., 2018) and GPT-2 (Radford et al., 2019) are two large pre-trained language models proposed by OpenAI. The two papers are pioneering work on training a versatile and efficient text generation model. Not only are the generated text is realistic and fluent, the two papers demonstrated that well-trained language models are capable of few-shot or zero-shot learning across a wide range of tasks. More specifically, when conditioned on a document plus questions, GPT-2 reach a score matching or exceeding the performance of multiple baseline systems without using the training examples.

GPT and GPT-2 have shed light on the direction of pre-training large language models as fundamental architectures (Bommasani et al., 2021) for downstream applications. After these models are pre-trained over a large corpus, the parameters are fixed and can be shared publicly. Then the architecture can be applied to a wide range of tasks without further training. This opens a new possibility of effectively sharing knowledge between researchers.

2. Approach

GPT is a 12-layer transformer model (Vaswani et al., 2017) with an embedding size of 768 and a hidden size of 3072. GPT-2 is a 12~48-layer transformer model with an embedding size of 768~1600. They both belong to the auto-regressive generative language models, where at each time, the model looks at the preceding tokens and predicts the probability distribution over the next token.

The training objective of GPT and GPT-2 are similarly simple: they are both optimized to fit the sentence distribution in the corpus by the following likelihood decomposition:

$$p_{LM}(x) = \prod_{i=1}^{n} p_{LM}(x_i | x_1, \cdots, x_{i-1})$$

When applied to the downstream task, the two models can be used to generate text with a given prompt. For example, when query about the writer of a book "*Origin of Species*", one can use (probably automatically generated) prefix prompt "Who wrote the book the origin of species", and GPT-2 will output the correct answer "Charles Darwin" with 83.4% probability. This is owing to that GPT-2 fits the distribution over corpus that contains this kind of knowledge, so it can memorize these facts.

3. Usage

HuggingFace provides an easy to use platform for multiple NLP tools. For pre-trained language models like GPT and GPT-2, they provide APIs for downloading the weights, building models via PyTorch and calling the functions over these models. In the following part, we take GPT-2 as example and show how to use it.

```
    # Importing tools from HuggingFace library
    from transformers import pipeline, set_seed
    # Building an text-generation pipeline with GPT-2 as model
backbone
    generator = pipeline('text-generation', model='gpt2')
    # Fixing the random seed for reproducible results
    set_seed(42)
    # Prompting GPT-2 to generate 5 sentences with max length 30
    generator("Hello, I'm a language model,", max_length=30,
num_return_sequences=5)
```

If you need to fine-tune the model on any specific data, you can use the following API:

```
    # Getting the GPT-2 backbone and tokenizer to fine-tune its
parameters
    model = generator.model
    tokenizer = generator.tokenizer
    model.train(True)
    tokenized = tokenizer(["Hello, world"], padding=True,
max_length=32, truncation=True)
    loss = model(input_ids=input_ids, labels=input_ids).loss
    loss.backward()
```

```
    # Omitting codes for gradient-based optimization, e.g., with
Adam optimizer…
```

4. Limitation

Despite the convenience and advantages of using GPT and GPT-2, there are several disadvantages that are worth noting:

1. GPT and GPT-2 can generate fluent text, but currently it is hard to conveniently instruct it to generate as the use might wish. Looking for the correct prefix for a task, i.e., prompt engineering, still requires extensive human effort.
2. GPT-2 is more powerful than GPT, and is also more computationally heavy in generation. It is infeasible to use it through a large amount or fine-tune its parameters without GPU.
3. As language models are pre-train on text that are written by humans, it still come with bias prevalent among humans. For example, if you prompt GPT-2 with text: "A doctor went home.", an example generated is "He did not know why…" which assumes that the doctor is a male.
4. Like humans, large language models are susceptible to linguistic biases (Patel et al., 2021). Compared with the context sentence "President Bush stated that", another context sentence "President Bush claimed that" is more likely to guide language model to doubt the truthiness and trustworthiness of President Bush. So one should be careful when using language model in order not to introduce their own biases into the text and mislead GPT-2.

References

1. Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).
2. Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
3. Bommasani, Rishi, et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).
4. Patel, Roma, and Ellie Pavlick. ""Was it "stated" or was it "claimed"?: How linguistic bias affects generative language models." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.