Fine-Tuning and Distilling GPT-2 for Language Modeling Task

Zixi Wang
M.S. student in Applied Data Science
Viterbi School of Engineering
wangzixi@usc.edu

Abstract

The task of language modeling, which plays a fundamental role in the realm of natural language processing (NLP), involves the prediction of the succeeding word in a sentence or a sequence of words. The subsequent text seeks to explicate a method for fine-tuning a pretrained GPT-2 model on a particular dataset, and following that, employing distillation on the fine-tuned model to produce a smaller model, in this case, DistilBERT.

The GPT-2 model, a brainchild of OpenAI, is a large transformer-based language model that holds a reputation for generating text that is quite similar to that produced by humans. Despite the impressive results shown by GPT-2, it is quite demanding in terms of computational resources owing to its size. A common strategy that is often used to address this problem involves knowledge distillation, a method that enables the transfer of capabilities of the larger model to a smaller one, leading to more computational efficiency.

1 Summary of Methods

A. In the methodology section, we started by detailing the fine-tuning process of the GPT-2 model. Our primary focus was the preprocessing aspect, which is a critical part of the training process. Preprocessing is all about converting the raw text data into a format that the machine learning model can understand and learn from. For this task, we leveraged the Wikitext-2 dataset, a large collection of sentences taken from Wikipedia. This dataset is well-regarded in the natural language processing community for its utility in language modeling tasks. The first step in preprocessing this data is tokenization, wherein the text is broken down into individual tokens or words. We utilized a tokenizer from the Hugging Face library, which is specifically designed for the GPT-2 model. This tokenizer not only breaks down the text into tokens, but also maps these tokens to their respective IDs that the model can understand.

In the next part of the fine-tuning process, we covered the configuration and instantiation of the GPT-2 model. The GPT-2 model has multiple versions, each differing by the number of parameters they contain. For this task, we utilized the 'gpt2' version. The Hugging Face library provides an easy-to-use interface for loading this pretrained model. Once loaded, the model is ready to be fine-tuned on our task-specific data.

The next crucial aspect of the fine-tuning process is the setting up of the training arguments and initializing the trainer. The TrainingArguments class from the Hugging Face library allows us to specify a variety of training parameters such as the learning rate, weight decay, and evaluation strategy. The initialization of the trainer involves passing the model, training arguments, and the datasets to the Trainer class. This setup allows the model to be trained and evaluated in a streamlined manner.

B. Dataset Distillation

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Dataset distillation is a novel method that leverages the power of large-scale pretrained models to create smaller, task-specific datasets. The distilled dataset essentially contains the same knowledge as the original dataset but in a more condensed format.

The process of dataset distillation involves generating synthetic examples and their corresponding labels from the pretrained model. These synthetic examples are designed to be challenging for the model, thereby pushing it to learn and generalize better. The generated examples and their corresponding labels form the distilled dataset. The advantage of this process is that the distilled dataset is much smaller in size compared to the original dataset, which leads to faster training times while still maintaining similar performance.

The main component of the distillation process is the generation of synthetic examples. We use the model's token probabilities to guide the creation of these examples. The token probabilities represent the likelihood of each possible next token given the current input. By selecting tokens with lower probabilities, we ensure that the synthetic examples are challenging for the model.

Following the generation of synthetic examples, we produce the corresponding labels. The labels are essentially the token IDs of the synthetic examples, and they guide the model during the fine-tuning process. The synthetic examples and their labels together form the distilled dataset.

In summary, our methodology involves fine-tuning a GPT-2 model on a large-scale pretraining dataset and then using this model to perform dataset distillation. The resulting distilled dataset is smaller and more task-specific, leading to faster training times and potentially similar or even superior performance.

2 Performance

The expected performance of the fine-tuned model, and particularly the influence of dataset distillation on the results, is a significant focus of this project. Given the innovative nature of dataset distillation, we anticipate some interesting outcomes.

When assessing the performance of a language model like GPT-2, common metrics include Perplexity, BLEU score, and ROUGE score. Perplexity is a measurement of how well a probability model predicts a sample, and in the context of language modeling, a lower perplexity score indicates better performance. BLEU and ROUGE scores, on the other hand, are metrics used to measure the quality of text which has been machine-generated or translated. Higher BLEU and ROUGE scores denote better text generation.

The primary expectation is that the fine-tuned model should perform better than the baseline GPT-2 model on these metrics. Fine-tuning tailors a model to a specific task by further training it on task-related data, which in our case is the Wikitext-2 dataset. This additional layer of training should enable the model to generate more coherent and contextually appropriate text, thereby improving its performance on the evaluation metrics.

As for the influence of dataset distillation, the expected results are multi-fold. Firstly, due to the smaller size of the distilled dataset, we anticipate faster training times. This speed-up can be particularly beneficial in iterative development processes where models are continuously updated and improved.

Secondly, we expect the distilled dataset to maintain, if not enhance, the performance of the fine-tuned model. This expectation stems from the nature of the synthetic examples in the distilled dataset. As these examples are generated to be challenging for the model, they should push the model to learn more effectively, leading to improved generalization.

However, it is important to note that the performance gains from dataset distillation can be dependent on various factors such as the quality of the synthetic examples and the model's capacity to learn from them. There could also be a risk of overfitting if the distilled dataset is too small or not diverse enough.

Ultimately, we believe that the combination of fine-tuning and dataset distillation could provide a promising approach to enhance the efficiency and effectiveness of language models. With the advanced computational capabilities of GPT-2 and the novel concept of dataset distillation, we are

hopeful for notable improvements in model performance. But we also understand the importance of empirical validation, and as such, we are excited to put our methods to the test.

References

[1]Zhao, B. & Mopuri, K. R. & Bilen, H. (2021, March 29). Dataset condensation with gradient matching. University of Edinburgh Research Explorer. https://www.research.ed.ac.uk/en/publications/dataset-condensation-with-gradient-matching

[2]Zhao, B. & amp; Bilen & H. (2023, January 7). Dataset condensation with distribution matching. University of Edinburgh Research Explorer. https://www.research.ed.ac.uk/en/publications/dataset-condensation-with-distribution-matching