Fine-tuning GPT-2 on Medical Text Data: PubMed QA Dataset

Shraddha Chavan

July 31, 2024

Contents

1 Introduction

In this report, we explore the process of fine-tuning GPT-2 on a medical dataset to enhance its capabilities in handling medical text, specifically focusing on the PubMed QA dataset. The goal is to fine-tune GPT-2 for question answering tasks in the medical domain using the HuggingFace Transformers library.

2 Dataset Overview

The dataset used in this project is the **PubMed QA (PubMed Question Answering)** dataset, which consists of question-answer pairs in the medical domain. The labeled version of the dataset (*pqa_labeled*) is selected for training. The dataset is tokenized using the GPT-2 tokenizer, and the text sequences are padded and truncated to a maximum length of 128 tokens.

3 Methodology

3.1 Model and Tokenizer

We employed the pre-trained **GPT-2 model** from HuggingFace's Transformers library. The model is designed for causal language modeling and is loaded along with its corresponding tokenizer. The tokenizer's end-of-sequence token (`eos_token`) is used as the padding token.

The main steps in the methodology are outlined below:

1. **Data Preprocessing:** The text sequences (questions) from the dataset are tokenized and padded/truncated to ensure a consistent sequence length of 128 tokens.
2. **Fine-Tuning:** The model is fine-tuned on the medical dataset for three epochs. AdamW optimizer is used for weight updates. Mixed precision training (FP16) is enabled to optimize GPU utilization and reduce training time.
3. **Evaluation:** After training, the model is evaluated using its performance on the validation dataset. The evaluation strategy is set to 'epoch', meaning that evaluation is performed after every epoch.

3.2 Training Setup

Training Arguments: The model was trained with the following configurations:

- Learning Rate: 2e-5
- Number of Epochs: 3
- Weight Decay: 0.01
- Batch Size (Train/Eval): 8
- Mixed Precision: Enabled (FP16)

4 Evaluation and Results

The model was trained for 3 epochs, and after training, it was evaluated on the validation dataset. The results include metrics such as loss and accuracy, as shown below:

```
Loss:  0.435865 after epoch 3
Perplexity: 1.5463
```

The loss value of the model shows the error during training and evaluation, while the perplexity metric indicates how well the model predicts the next word in a sequence.

5 Conclusion

In this report, we demonstrated how to fine-tune GPT-2 on a domain-specific medical dataset (PubMed QA). The fine-tuned model can now be used for medical question-answering tasks. Future work could involve experimenting with other transformer-based models such as BERT, LLaMA, or even larger versions of GPT models.