Report on Fine-Tuning (Phi2) Large Language Model for Natural Language Inference using QLoRA

Sarthak Sharma, MT23083

Introduction

Fine-tuning process of the Phi2 model from Hugging Face is demonstrated here for the task of Natural Language Inference (NLI) using the QLoRA (Quantized Low Rank Adaptation) technique, and the dataset used for this task is the Stanford Natural Language Inference (SNLI) dataset as mentioned in the Assignment 3 document. So, this technique QLoRA is an advanced technique which is particularly useful in reducing the memory usage and computational costs through quantization, enabling efficient fine-tuning of large language models.

Dataset Preparation and Model Used

The SNLI dataset was accessed through Hugging Face. For training, validation, and testing, samples were selected as follows:

- **Training Set**: 1,000 samples were selected by choosing every 550th sample from a total of 550,000 samples.
- **Validation Set**: 100 samples were selected by choosing every 100th sample from a total of 10,000 samples.
- **Testing Set**: 100 samples were selected by choosing every 100th sample from a total of 10,000 samples.

Model Used for Fine Tuning: Microsoft / Phi2 model is used from HuggingFace repository

1. Comparison between Accuracies and Training Time

The total time taken to fine-tune the model using QLoRA was approximately **1 hour and 30 minutes**, each epoch taking approximately 20 minutes each. The model is fine tuned for 5 epochs.

Loss in each Epoch:

Epoch 1, Loss: 0.09430356358562211

Epoch 2, Loss: 0.013739965788621853

Epoch 3, Loss: 0.012722672756433814

Epoch 4, Loss: 0.01225256772021522

Epoch 5, Loss: 0.011845991141967603

The loss is observed to reduce as the epochs increase, indicating better understanding of the model with repeated iterations.

Accuracy:

Pre-trained Model Accuracy: 65.23%

Fine-tuned Model Accuracy after epoch 1: 60.89%

Fine-tuned Model Accuracy after epoch 2: 64.07%

Fine-tuned Model Accuracy after epoch 3: 67.45%

Fine-tuned Model Accuracy after epoch 4: 68.88%

Fine-tuned Model Accuracy after epoch 5: 71.25%

2. Parameters of the Fine-tuned Model:

In the fine-tuning process of the model, the total number of parameters was evaluated as follows:

- Total Parameters in the Model: Approximately 2.80 billion (2,798,033,920)
- Parameters Being Fine-Tuned: Approximately 18.35 million (18,350,080)
- Percentage of Parameters Fine-Tuned: Approximately 0.76%

This indicates that a small fraction of the total model parameters are being adjusted during the fine-tuning process, highlighting the efficiency of the parameter-efficient tuning method employed.

3. Resources Used for Fine Tuning Process

The fine-tuning was performed for 5 epochs using QLoRA. The model was saved after each epoch to ensure that progress could be tracked and any improvements could be retained.

Implementation Details

- **Hardware**: A GPU P100 was used for the fine-tuning process. Initially, the code was tested on CPU to ensure functionality.

- **Memory**: The GPU utilized approximately 16 GB of memory during training.

4. Failure cases of the pretrained model that were corrected by the fine-tuned model

Premise	Hypothesis	Label	Correct Label (By Fine Tuned Model)	Explaination
A woman within an orchestra is playing a violin.	A woman is playing the violin.	0	1	The hypothesis is a generalization of the premise; it can be seen as entailed.
Two men climbing on a wooden scaffold.	Two sad men climbing on a wooden scaffold.	1	2	The emotional state is not supported by the premise.
A woman in a black shirt looking at a bicycle.	A woman dressed in black shops for a bicycle.	1	2	The action of "looking at" vs. "shopping" implies a different level of engagement; remains neutral.
Two men in neon yellow shirts busily sawing a log in half.	Two men are cutting wood to build a table.	1	1	The hypothesis is a reasonable inference from the premise; it accurately describes the action.

Conclusion

In summary, the fine-tuning of the Phi2 model for the NLI task using the QLoRA technique has shown promising results, improving accuracy significantly compared to the pretrained model. The process highlighted both strengths and weaknesses in the model's understanding of language, with certain complexities remaining challenging even after fine-tuning.