**Report - LLM | Assignment 3**

**Fine Tuning Large Language Models**

**Garv Makkar | 2021530**

**Familiarizing with LoRA and it’s variants**

<https://huggingface.co/docs/peft/en/developer_guides/lora>

<https://huggingface.co/docs/peft/main/en/conceptual_guides/lora>

**Task:** Fine-tune an LLM using QLoRA for Natural Language Inference (NLI).

**Dataset:** [stanfordnlp/snli · Datasets at Hugging Face](https://huggingface.co/datasets/stanfordnlp/snli)

**Training:** 1000 samples | Every 550th sample from a total of 550k.

**Testing**: 100 samples | Every 100th sample from a total of 10k.

**Validation**: 100 samples | Every 100th sample from a total of 10k.

**Model:** [microsoft/phi-2 · Hugging Face](https://huggingface.co/microsoft/phi-2)

Accuracy comparison between the pretrained and fine-tuned models on the test set

**Base model accuracy**: 0.38

**Fine-tuned model accuracy:** 0.82

**Time taken to fine-tune the model using QLoRA**

Training time: 601.44 seconds

**Total parameters in the model and the number of parameters fine-tuned.**

**Base model**

Base model total parameters: 2,648,568,320

**QLoRA model**

Total parameters: 1,398,149,120

Trainable parameters: 7,872,000

Percentage of parameters fine-tuned: 0.56%

**Model Size Analysis**

Base model size: 10119.49 MB

QLoRA model size: 1741.52 MB

Size reduction: 82.79%

**Resources used during fine-tuning.**

Training time: 601.44 seconds

Peak GPU memory usage: 12.32 GB

Final GPU memory usage: 12.50 GB

CPU utilization: 35.4% → 65.7%

RAM usage: 4.27 GB → 4.35 G

**Failure cases of the pre-trained model that were corrected by the fine-tuned model and those that were not corrected and possible explanations for both.**

On the test set, when we check with the base model and with the fine-tuned model,

Number of cases corrected by fine-tuning: 58

Number of cases still wrong after fine-tuning: 4

**=== Corrected Cases (5 examples) ===**

Example 1:

Premise: A woman within an orchestra is playing a violin.

Hypothesis: A woman is playing the violin.

True label: 0

Base prediction: 1

Fine-tuned prediction: 0

Example 5:

Premise: many children play in the water.

Hypothesis: The children are playing mini golf.

True label: 2

Base prediction: 1

Fine-tuned prediction: 2

Example 7:

Premise: A female softball player wearing blue and red crouches in the infield, waiting for the next play.

Hypothesis: the player is flying planes

True label: 2

Base prediction: 1

Fine-tuned prediction: 2

Example 8:

Premise: Workers standing on a lift.

Hypothesis: Workers stand on a lift

True label: 0

Base prediction: 1

Fine-tuned prediction: 0

Example 9:

Premise: Two men in neon yellow shirts busily sawing a log in half.

Hypothesis: Two men are cutting wood to build a table.

True label: 1

Base prediction: 2

Fine-tuned prediction: 1

**Possible Explanations for corrected cases:**

The base model initially predicted a neutral outcome in four corrected cases. Still, after fine-tuning, it accurately identified the relationship as either entailment or contradiction. This improvement can be attributed to the initial probabilities for the neutral class being relatively low—though still higher than the other two classes—in the softmax layer. Following fine-tuning, the distribution of predictions likely shifted in a more favorable direction, allowing the model to make more accurate classifications.

**=== Still Wrong Cases (5 Examples) ===**

Example 4:

Premise: a woman in a black shirt looking at a bicycle.

Hypothesis: A woman dressed in black shops for a bicycle.

True label: 1

Base prediction: 2

Fine-tuned prediction: 0

Example 10:

Premise: A Skier ski-jumping while two other skiers watch his act.

Hypothesis: A skier preparing a trick

True label: 0

Base prediction: 1

Fine-tuned prediction: 1

Example 11:

Premise: Children bathe in water from large drums.

Hypothesis: The kids are wet.

True label: 0

Base prediction: 1

Fine-tuned prediction: 1

Example 87:

Premise: Overly dramatic couple pose for a picture where an "angry "man "chokes" a woman who sticks out her tongue.

Hypothesis: A man chokes a woman

True label: 0

Base prediction: 1

Fine-tuned prediction: 1

**Possible Explanations for Still Wrong Cases:**

Three of the four still-wrong cases resulted in a neutral prediction from the base model. This suggests that the model had a high probability of a neutral outcome even before fine-tuning. It is likely that, after fine-tuning, the distribution in the softmax layer changed. Still, the maximum probability remained for the neutral class (that neutral prediction initially might have been very strong). This phenomenon might have occurred when the model struggles to determine whether the relationship is entailment or contradiction, leading to a reliance on the neutral classification.