LLM Finetuning Task:
Comparing PEFT Techniques on Gemma-3-1b-it

1 Introduction

The aim of this task is finetuning the google/gemma-3-1b-it large language model (LLM) using two different Parameter-Efficient Fine-Runing (PEFT) techniques. The pure LLM model and two distinct LLM modes with different PEFT techniques are compared in the end in terms of BLEU-4 and ROUGE-L performances together with memory and time consumptions.

For this task, the base LLM is google/gemma-3-1b-it. On top of that two different PEFT techniques are applied. The first technique is Quantized Low-Rank Adaptation (QLoRA). The second one is Low-Rank Adaptation (LoRA). Both methods are fine-tuned for two epochs. For training and test set, 5k and 2k test samples are collected from tatsu-lab/alpaca, allenai/tulu-v2-sft-mixture, and HuggingFaceH4/ultrachat_200k separetely. Final training and test sizes are 15k and 6k.

2 Approach & Methodology

For training and test set, 5k and 2k test samples are collected from tatsu-lab/alpaca, allenai/tulu-v2-sft-mixture, and HuggingFaceH4/ultrachat_200k respectively. Both training and test samples are chosen randomly from each of the train split of the mentioned datasets. There is no overlap between sampled training and test datasets. A fixed random seed is used for reproducibility. Hugging Face datasets is used for for loading, sampling, and mapping/formatting.

Since datasets are different from each other in terms of their structure. They are formatted as system, instruction, input and response. System and input are left empty if there is no corresponding information in the dataset. Then, the formatted samples are mapped into Dataset format. In the end, all 3 training and 3 test samples are combined into training and test datasets.

The base model for this task is google/gemma-3-1b-it. Gemma model uses only user and model roles. Therefore, the dataset is once converted as follows. System, instruction and input are combined as instruction. Response is converted to assistant. Instruction and assistant are chosen in order to obtain a generic format. Gemma model works with a specific $< start\_of\_turn > user/model < end\_of\_turn >$. In fine-tuning phase this conversion is handled by the model itself. For evaluation, the test set converted to this format manually.

For fine-tuning QLoRA (Quantized Low-Rank Adaptation) and QLoRA (Quantized Low-Rank Adaptation) are chosen. LoRA injects trainable low-rank matrices (A and B) into selected layers. The base model stays frozen. Only the small LoRA parameters are updated. It requires the base model to be loaded in full precision (FP16/BF16). QLoRA extends LoRA by quantizing the base model to 4-bit (NF4). It keeps the quantized model frozen but adds LoRA adapters on top. It is trained in FP16. It uses double quantization and paged optimizers to reduce memory even further. LoRA adds adapters to a full-precision model. QLoRA adds adapters to a 4-bit compressed model which enables large-model fine-tuning on very small GPUs.

LoRA boosts instruction-following by training only small low-rank updates while keeping the full model frozen, making it memory-efficient and resistant to overfitting, which supports good generalization. QLoRA applies these same adapters on top of a 4-bit quantized base model, enabling large-model fine-tuning on limited hardware. It offers major memory savings while preserving near full-precision quality, with generally strong generalization aside from minor potential quantization noise on very small datasets.

BLEU-4 and ROUGE-L are used as evaluation metrics. They both compare the generated text (predicted text) to reference texts. BLEU-4 measures n-gram overlaps up to 4-grams. It favors exact matches. ROUGE-L measures measures the longest common subsequence (LCS) again between generated text (predicted text) and reference texts. It is more tolerant of partial matches. While BLEU is precision oriented, ROUGE is recall oriented.

3 Implementation Details

Key libraries are datasets, evaluate, transformers, and trl.

Both the fine-tuning and the evaluation run on a Google Colab instance with a A100 GPU. Inference is slower with T4 GPU. Therefore, A100 is selected. Some library compatibility problems occurred for trl and protobuf but they are fixed.

4 Results & Discussion

The evaluation results of two different fine-tuning techniques are compared with the baseline is shown in Table 1. There is a significant difference between before and after results. Fine-tuning improves the results. While QLoRA gets better results on metric performance, LoRA gets better results for training time. They are similar on peak memory.ROUGE-L scores are lower than BLEU-4 scores as it penalizes paraphrased outputs due to its order matching nature.

Metric	**QLoRA**	**LoRA**
BLEU-4 (Before)	0.2625	0.2625
BLEU-4 (After)	0.5421	0.4482
ROUGE-L (Before)	0.1137	0.1137
ROUGE-L (After)	0.1384	0.1296
Peak Memory (GB)	11.76 GB	11.58 GB
Training Time (Hrs)	2040.94 seconds	1559.89 seconds

Table 1: Comparison of QLoRA and LoRA.

Some examples are provided in the appendix section.

5 Ensemble/Hybrid Proposal

The proposal is combining QLoRA and LoRA as follows. First, use the pretrained gemma model as in the assesment. Then, fine-tune using QLoRA in 4-bit quantized mode for memory efficiency and large model fine-tuning. Afterwards, merge QLoRA updates into the base weights in order to produce a full FP16 model that includes QLoRA knowledge. In the next step apply LoRA adapters on top of this merged model for task-specific adaptation or stylistic control. Finally, orthogonal low-rank projections relative to QLoRA updates can be enforced during LoRA training to maximize complementary information capture.

6 Conclusion

Fine-tunining improves the results significantly. QLoRA and LoRA has their own advantages and drawbacks. Therefore, a hybrid approach is proposed to obtain better results.

7 Appendix

- **Baseline**: Sender:** XYZ Electronics
  Your Order Has Shipped:** You have received your order.
  Tracking Number: 123-456-789

This email provides tracking information for your order.
QLoRA: op has been shipped.
- The tracking number for the package is 123-456-789.
- Please check your email for more detailed updates.
LoRA: new laptop has been shipped. The tracking number for the package is 123-456-789. You can check your email for more details about the shipment.

I hope this helps!
Reference: 123-456-789

- **Baseline**: **Yellow**
4. **Green Lantern** (This is a color, but it's a very warm, glowing color)
5. **Blue**
QLoRA: 3. Green 4. Blue 5. Yellow
LoRA: he colors from the warmest to the coolest is: red, orange, green, blue, yellow. So, the final answer is: red, orange, green, blue, yellow.
Reference: Red, Orange, Yellow, Green, Blue

- **Baseline**: g a bedroom – let's go with a **"Botanical Shadowscape"** concept. It combines natural klasse with a touch of moody, ethereal beauty.

  Here's a breakdown of how to achieve it, broken down into stages:
  QLoRA: er. This can be achieved by placing a comfortable chair, a small table, and a floor lamp in a corner of the room. This corner can be used to read books, magazines, or newspapers, and can be decorated with a soft blanket, pillows
  LoRA: reate the illusion of more space in the room and can also be used to add a touch of glamour to the decor. You can use mirrors to create a more interesting focal point in the room or to add a touch of sophistication to the decor.
  Reference: A creative way to decorate a bedroom is to use wall murals. You can either find pre-made murals or have them custom made for your bedroom. Other ideas include using wallpaper or fabric to create an accent wall, or using art pieces or mirrors to create a focal point in the room.

- **Baseline**: focusing on a multi-faceted approach that tackles the root causes and leverages various technologies and societal shifts. Here' vacancies a layered strategy, combining technological advancements with policy changes and behavioral modifications:

  **1. Transitioning to Renewable Energy Sources – The
  QLoRA: on standards for factories and vehicles. This would require industries to install pollution control technologies such as scrubbers, filters, and combustion improvers to reduce the amount of pollutants released into the air. Additionally, it would require the use of cleaner fuels such as natural gas or electric
  LoRA: y sources such as solar, wind, and hydropower. This will reduce the amount of air pollution caused by burning fossil fuels. Another solution is to improve public transportation and encourage people to use it. This will reduce the number of cars on the road, which will reduce
  Reference: A solution for reducing air pollution is implementing stricter regulations on industries, encouraging the use of renewable energy sources, creating more green spaces in urban areas, improving public transportation, and promoting car-sharing programs.

- **Baseline**: reed" is a **coordinating conjunction**.

  Specifically, it's a **coordinating conjunction** that connects two independent clauses.

  More precisely, it's a **Southern conjunction** (also known as a "and
  QLoRA: reed" is an appropriate conjunction, which is an adverb of Manner. This conjunction is used to describe how the action was discussed or how the agreement was made.
  LoRA: reed" is an adverb of Manner. This type of conjunction is used to describe how something is done or how a situation is presented. In this sentence, the adverb of Manner is used to describe the way in which the issue was discussed.
  Reference: The conjunction used in the sentence is a coordinating conjunction.