Fine-tune LLM for Language Translation

Prathamesh Pawar
Matriculation Number - 428966
RPTU Kaiserslautern, Department of Computer Science

Note: *This report contains a project documentation and reflection on the portfolio task submitted for the lecture Engineering with Generative AI in WiSe 2024-25. This report is an original work and will be scrutinised for plagiarism and potential LLM use.*

1

1 Research Phase

1.1 Dataset selection : The dataset I have chosen for the project is **OPUS100**. This is a popular and widely used dataset for machine translation tasks. The **OPUS100** dataset consists of parallel text pairs in various languages including German , French and covers a wide areas of topics which I have used for this project. The dataset size and the quality of translation pairs it contains makes it ideal choice for fine-tuning tasks.

1.2 Model selection : I have chosen GPT-Neo 1.3B model. This is a general purpose model with 1.3 billion parameters which inspite of being a large model is easier to fine tune on a comparatively smaller dataset due to its transformer based architecture. The model is an ideal choice for our task as it satisfies all the computational constraints and the criteria of a model having atleast 1 billion parameters.

2 Design Phase

2.1 Fine-Tuning Approach : The fine tuning approach I have used is **LoRA** (Low rank adaption) approach. This is a very efficient fine-tuning method from the hugging face's PEFT library as it works on freezing the model's weights and adding trainable low-rank matrices, this reduces memory usuage. And making it a good choice for training a larger model like gpt-neo. Training times are also significantly lower which saves up on resource constraints.

2.2 Design of the prompt : For the designing of prompt. I have mentioned a role based definition, along with that I have asked it to cover all kinds of topics like business and travel. I have provided a few examples in the prompts to make sure the model understands the proper format and the accuracy of the data i am asking it to generate.

3 Implementation Phase

3.1 Resources used for implementation : I have used Google Colab notebook with a T4 GPU(15 GB VRAM, 12.7 GB RAM)

3.2 Installing all the necessary packages : Started with installing all the necessary packages required for the implementation.

3.3 Loading Dataset A : Loaded the OPUS100 dataset from the datasets library. Randomly selected 1000 pairs from the dataset and then splited them into 80:20(train:test) ratio.

3.4 Pre-trained model : Loaded the EleutherAI/gpt-neo-1.3B as Model A. And set up the tokenizer as well.

3.5 Evaluating Model A : Defined a function " Generate translation " and passed the Model A, tokenizer and test data to generate the predictions. Evaluted the model with BLEU score by passing predictions from the model A and reference french sentences to obtain a low BLEU score as it is not fine tuned.

3.6 Fine-Tuning Set up : Prepared the model for LoRA by setting up LoRA configuration. In the LoRA config, i have specified the rank(r=16), lora alpha(=64), target modules, dropout rate(=0.1), bias(none) and task type parameters.

3.7 Fine Tuning Model A on train data A : Fine tuning the Model A on train dataset A using LoRA to create Model B. Training arguments used are fp16, data text field, max sequence length(=256), batch size(=4), optim, learning rate(1e-4), number of epochs(=2).

3.8 Evaluting Model B : Evaluating the model B on test data A using the Bleu metric.

3.9 Generating Dataset B : Now using the designed prompt and asking a larger model (I have used QWEN 32B) to generate a synthetic dataset B twice as size of the train dataset A.

3.10 Creating Model C : Fine Tuning Model A using the LoRA configuration on newly generated dataset B to create Model C.

3.11 Evalute Model C : Evaluting Model C on Test dataset A using Bleu score.

3.12 Generating Dataset C: Concatenating the Dataset A train and Dataset B to create Dataset C.

3.13 Creating Model D : Fine Tuning Model A using the LoRA configuration on dataset C to create Model D.

3.14 Evalute Model C : Evaluting Model D on Test dataset A using Bleu score.

3.15 Plotting the model performances : Plotting the bleu scores of all the models using Bar charts.

4 Testing and Evaluation

4.1 Evaluation and Analysis

Following are evaluation approaches I have implemented:

- **BLEU Scores:** I have calculated bleu scores of each model and plotted to visualize the scores. In Figure 1
- **Training Loss and Validation :** I have calculated the training loss of each model. In Figure 2 Training loss Model A
- **Training Loss and Validation** I have calculated the training loss of each model. In Figure 3 Training loss Model B
- **Batch processing:** Evaluations are performed in batches (size=4) to reduce memory usage.

4.2 Testing split dataset : The testing split was a split to ensure better and efficient evaluation while keeping in mind the statistical validity. 80 = training split for fine-tuning and 20= testing for evaluation and to balance the computational constraints.

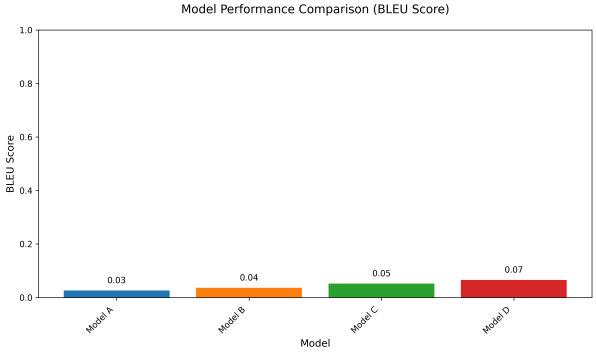

Figure 1: Bleu Scores

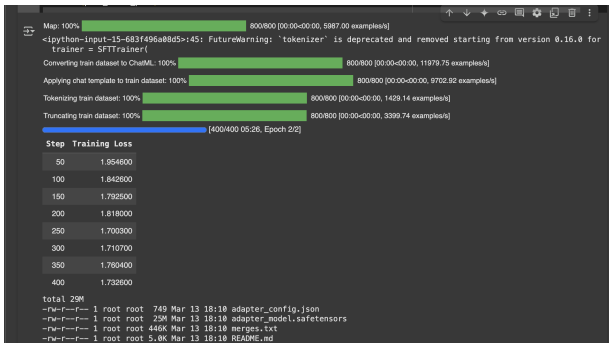

Figure 2: Training loss

5 Reflection

1. **What was the most interesting thing you learned while working on the portfolio? What aspects did you find interesting or surprising?**

   The most interesting thing I learnt while working on my portfolio task was the various fine-tuning approaches I came across and I was surprised to know the difference in the evaluation scores(BlEU scores) after the models were fine tuned. Even a little change in parameters made a huge impact on the scores.This factors have developed a profound interest for me in learning and exploring the domain of advanced NLP techniques , Large-Language models and prompt engineering. During this project, we were asked to use the minimum resources and even the minimum resources were enough to fine tune such a huge mode like GPT-Neo. I am eager to work with even bigger and better models and know the various areas where I can explore and apply these fine-tuning techniques and derived optimized results.

2. **Which part of the portfolio are you (most) proud of? Why? What were the challenges you faced, and how did you overcome them?**

   I am actually proud of that I was able to explore and learn such interesting fine-tuning

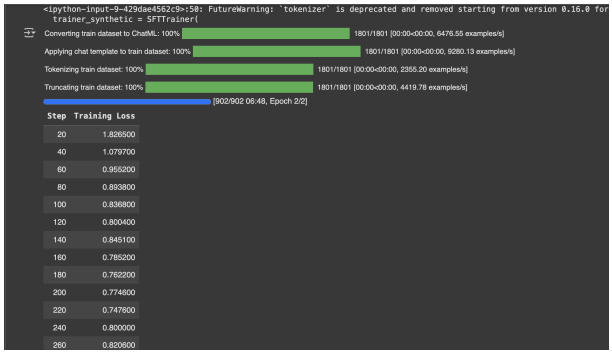

Figure 3: Training loss 2

technique, evaluation metrics and prompt engineering with implementing and hands on working on these techniques. The evaluation pipeline and fine-tuning approaches which provided a better understanding of the different approaches used. Apparently, I faced a few challenges while managing the GPU memory and optimizing over it. I overcame this challenge why carefully looking into my training parameters and optimizing them according to the GPU usuage while maintaining the performance and efficiency. Generating the synthetic dataset from a bigger model was another sort of challenge for me due to the model's size and the prompt designing for a better and accurate dataset which I wanted. I overcame this by using hugging face's API inteference to load the model which downloading it locally and I used a realistic approach towards prompt which I provided to this bigger model.

3. **What adjustments to your design and implementation were necessary during the implementation phase? What would you change or do differently if you had to do the portfolio task a second time? What would be potential areas for future improvement?**

   Many trail and errors were required during each and every stage of the project. At first, I tried figuring out various datasets which would be ideal for the task but could really find it easily as I had to make sure the dataset consisted of various german-french translation pairs and covered a wide variety of topics to make the translation more generalised. Then I came across OPUS100 dataset which had all the properties which i was looking for in the dataset. The model selection part required adjustments as it was a bit difficulty to find a model with atleast 1B parameter which was not a translation model but a general purpose model. If I had to do the portfolio task for the second time, I would try to generate more generic and domain specific synthetic dataset. Areas of potential improvements include implementing different fusion techniques to combine multiple fine-tuned adapters with exploring different multilingual models. Another approach i would like to try would be a better and more optimzied model to see how better are the results.

4. **Include a brief section on ethical considerations when using these models on language translation tasks.**

Considering the impressive and developing capabilities of these language translation models, several ethical considerations are often raised. Firstly, misuse of this models should be considered. As these models can be trained and used for carrying out illegal and dangerous purposes. These model trained in such way can be used for spreading harmful content. Another aspect should considered is of the biases. During the training we should make sure that the model generates a generalised data and not a biased data towards anything. This can lead to severe gender based conflicts. While working on with these translation models, the developers must priortize ethical considerations and transparency about the model's limitations and should go through trail and error of various diverese datasets and data sources.

5. **From the lecture/course including guest lectures, what topic excited you the most? Why? What would you like to learn more about and why?**

   The topic which excited me the most among the course was of the parametering basied fine-tuning methods. The idea of accessing a billion parameters of models to implement any kind of specific task by using a very small percentage of the trainable parameters is just mind blowing advancements in the AI domain. This idea helops us in saving up on our environmental impact by training such large models with better fine-tuning approaches with very less computational power usuage. Moving ahead I would like to learn more about optimization of the these larger models. Along with expanding the knowledge beyond text processing and exploring the images, audio and other aspects in which how these models will perform is what I am curious about and looking forward. Sustainable AI is also a domain which interests me as its focus is more on reducing the environmental damages caused by the LLMs and optimizing energy and resources during training phases.

6. **How did you find working with DIFY platform during the course work? Would you recommend using DIFY in learning Generative AI technologies and why? What is the best start for learning Generative AI either by Python code or No-code platforms and why?**

   During the course work, working with DIFY platform provided a hands experience while working on Generative AI workflows which manual extensive coding. Working on tasks like generating synthetic dataset which was very helpful during this portfolio task, prompt engineering which was very much simplified and organised in a easy to understand way by DIFY's visual platform. DIFY platform is very helpful especially for beginners in the IT industry who focus more on practical exposure compared to the theortical knowledge. DIFY should used as a supplementary tool to practically try out ideas and pair them with learning fundamental coding. Code based platforms offer a deeper understanding of the advanced skills that required in the industry whereas No-code platforms provides accessibility and prototyping methods.

7. **How did you find the assignments and exercises in the course and how they help you in portfolio exam?**

   The assignments and exercises covered in this course were very comprehensive and throughly carried out with excellent knowledge and was provided with clarity. Each and every aspects during the exercises even though it was a very basic thing was explained and documented

in very well manner.The hands on nature of the assignments and exercises was valuable during this portfolio exam as many of the techniques implemented were already introduced in the exercises and assignments which helped me understand the portfolio exam task better. Specifically, the LoRA fine-tuning method which was discussed and explained in the exercise 3, helped in carrying out the implementation phase with ease. The overall mix of the theortical and practical implementation of this course made sure that i am aligned with both the technical skills and the implementation as well. Assignments and exercises were very well crafted to provide a complete overview and understanding of the Generative AI domain and played a very crucial part in this portfolio exam.

All of the resources used by the student to complete the portfolio task should be organised in the references section. **Note that the Reference section does not count towards the number of pages of the report.** Example references are given below [1] [2] [3]. **If you are using a reference manager like Zotero, you can export your Zotero library as a .bib file and use it on Overleaf. As you cite the article/technology/library in your main text, the References section will automatically update accordingly.** Please include a full list of references found. If students are using Zotero for their research paper management, a bibTeX will help them during citation which automatically adds references to the report.

References

[1] Albert Einstein. Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]. *Annalen der Physik*, 322(10):891–921, 1905.

[2] Donald Knuth. Knuth: Computers and typesetting.

[3] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1993.