CS-GY 6953 Project 2

Matthew So[1]

[1]Tandon School of Engineering
New York University
m.so@nyu.edu

Abstract

We aim to achieve the highest possible accuracy on an unknown, custom test set. The task is to train a RoBERTa base model, freeze some parameters such that less than 1M parameters are trainable, and train on the AG News classification dataset. We use LoRA parameters found from other public sources and reproduce their findings. We eventually achieve an accuracy of 84.68.

Project code can be found at this link - https://github.com/ms15032/project-2.

Introduction

The base model we used, RoBERTa (Liu et al. 2019), is based on the original BERT model with some slight modifications. It uses byte-level byte-pair encoding (BPE) tokenization, but has the same architecture as BERT. It also improves on pre-training, mainly using larger batches, larger sample sentences by combining samples together, and masking different parts of sentences randomly.

We use LoRA (Hu et al. 2021), a framework for freezing certain parameters of an LLM such that you no longer need to re-train every single parameter when fine-tuning. Not only does this help with model performance, the cost of model training greatly decreases since the memory requirement is much smaller.

We adjust the r value, which represents the rank of the matrix AB, $alpha$, which is the scaling factor for the weights, and $dropout$, which is the dropout proportion for the LoRA layers which prevents overfitting. The only requirement is to adjust these parameters in such a manner that the number of trainable parameters stays below 1 million.

Methodology

We first conducted a search on Hugging Face, a large repository for machine learning models, to find similar experiments to ours. By using the search query ”roberta lora ag news,” we found the following results -

- TransferGraph/ncduy_roberta-imdb-sentiment-analysis-finetuned-lora-ag_news
- TransferGraph/JonatanGk_roberta-base-bne-finetuned-cyberbullying-spanish-finetuned-lora-ag_news
- TransferGraph/robertou2_roberta-base-bne-finetuned-amazon_reviews_multi-finetuned-lora-ag_news
- TransferGraph/rmihaylov_roberta-base-sentiment-bg-finetuned-lora-ag_news
- TransferGraph/cardiffnlp_twitter-roberta-base-2021-124m-finetuned-lora-ag_news
- TransferGraph/roberta-base-finetuned-lora-ag_news
- TransferGraph/cross-encoder_quora-roberta-base-finetuned-lora-ag_news
- TransferGraph/boychaboy_MNLI_roberta-base-finetuned-lora-ag_news
- TransferGraph/navteca_quora-roberta-base-finetuned-lora-ag_news
- TransferGraph/aditeyabaral_finetuned-sail2017-xlm-roberta-base-finetuned-lora-ag_news
- TransferGraph/korca_bae-roberta-base-boolq-finetuned-lora-ag_news
- TransferGraph/cointegrated_roberta-base-formality-finetuned-lora-ag_news
- TransferGraph/JonatanGk_roberta-base-ca-finetuned-hate-speech-offensive-catalan-finetuned-lora-ag_news

These are all models pre-trained on various datasets and fine-tuned on the AG News dataset. The highest achieved accuracy from all these models was 0.9422, which has amongst the highest when comparing various other RoBERTa-based models, and since the project was relatively restrictive in terms of what parameters could be adjusted, we decided to use these implementations as a guide for our own implementation. Every single one of these model implementations had the following parameters -

- learning rate = 4e-4
- training batch size = 24
- Adam optimizer with beta 1 = 0.9, beta 2 = 0.999, and epsilon = 1e-8
- LoRA r = 8, with no bias
- LoRA alpha = 16
- LoRA dropout = 0.1
- 4 epochs

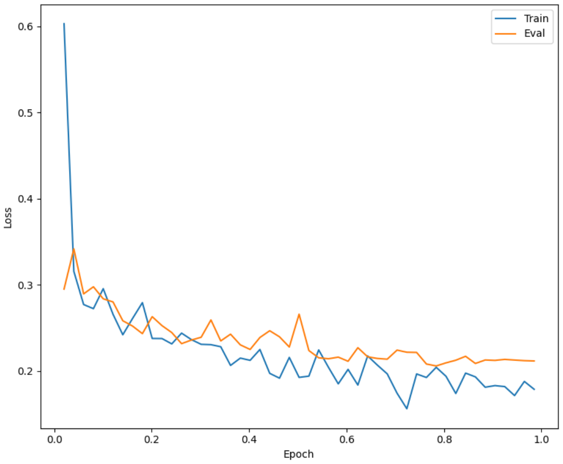

Figure 1: Train and evaluation set loss over 1 epoch.

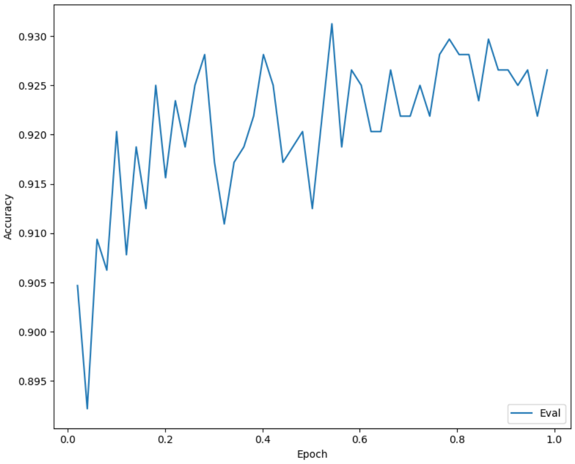

Figure 2: Evaluation set accuracy over 1 epoch.

Our implementation was a faithful reproduction of these implementations, leading to a model with 888,580 trainable parameters out of 125,537,288 total parameters. Our model was written as a Hugging Face transformers module and trained in a Kaggle notebook with an Nvidia P100 GPU. We ran the training multiple times to find the best test result.

Results

The best achieved test accuracy was 92.66 on a randomly selected evaluation set size of 640 samples out of a total 120,000. The resulting accuracy on the unknown test set was 84.15. We observed a sharp decrease in loss within the first few training steps, and only decreased marginally through the rest of training 1. We observed similar behavior with accuracy - accuracy sharply increased at the very beginning, and leveled out through the rest of training 2. We observed decreased test set performance after 1 epoch, presumably due to overfitting, so we only show the results of 1 epoch.

Since it wasn't clear what data is contained in the custom test set, and accuracy on the AG News test set was already quite high for models within this scope, we did not experiment further. Based on our runs, we did not foresee any benefit from additional model parameter changes or data augmentations. Considering the current SoTA (state-of-the-art) accuracy on AG News is from XLNet (Yang et al. 2020) with an accuracy of 95.55, and we achieved an evaluation accuracy of 92.66, we did not experiment any further.

References

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.