Fine-Tuning a BERT-Based Model for IELTS Essay Scoring

Shusrith S

PES1UG22AM155

24 March 2025

1 Introduction

This report details the process of fine-tuning a BERT-based classification model to predict IELTS writing task scores. The goal of this lab is to automate the evaluation of essays based on their quality and coherence. By leveraging LoRA (Low-Rank Adaptation), we aim to fine-tune the model efficiently while reducing computational costs. This report covers dataset preparation, model architecture, training process, evaluation metrics, and the implementation of inference for band score prediction.

2 Dataset Preparation

The dataset used in this lab consists of three main components: the essay prompt (Question), the student's response (Essay), and the assigned IELTS band score (Overall Band Score). To prepare the data, the Question and Essay were merged into a single text input to provide a coherent structure for the model. The input was formatted as:

```
Question: [essay prompt]
Answer: [student response]
```

To ensure compatibility with the classification model, the band scores, originally ranging from 1 to 9, were shifted to start from 0. This transformation was necessary for mapping the scores to appropriate classification labels.

Unnecessary columns, such as examiner comments and sub-score categories, were removed from the dataset. After preprocessing, the data was shuffled to maintain randomness and was split into training and testing sets, with 90% used for training and 10% reserved for testing. A stratified split was applied to maintain balance across score classes.

3 Tokenization

For this lab, `bert-base-uncased` was chosen as the tokenizer and model backbone. Tokenization was handled using Hugging Face's Transformers library, ensuring that each essay-question pair was converted into input tokens compatible with the BERT model. The maximum token length was set to 512 to accommodate longer essays while preventing overflow errors. Padding and truncation were applied uniformly to maintain consistent input shapes.

4 Custom Dataset Class and DataLoader

To streamline data handling, a custom PyTorch `Dataset` class was implemented. This class ensures that each data point includes tokenized input (`input_ids` and `attention_mask`) and corresponding labels representing the band score. The dataset was then fed into a DataLoader for efficient batch processing. The training data was shuffled to enhance generalization, while the test data was left ordered for evaluation purposes. The batch size was set to 16 for both training and testing.

5 Model Fine-Tuning with LoRA

The fine-tuning process leveraged LoRA to efficiently adapt the BERT model for sequence classification. The base model used was `bert-base-uncased` with a classification head added for prediction. LoRA was configured with a rank reduction factor of 8, a scaling factor (`lora_alpha`) of 32, and a dropout rate of 0.4. The adaptation was applied specifically to attention layers (`query` and `key`) to minimize computational overhead. Hugging Face's PEFT library was utilized to integrate LoRA into the model, significantly reducing the number of trainable parameters while maintaining performance.

The model contained 109M parameters and the LoRA config had 350k additional parameters to be learnt, meaning only 27.8% of the parameters had to be learnt in order to fine tune the model. This shows how efficiently LoRA can produce good results.

6 Training with PyTorch Lightning

The model was trained using PyTorch Lightning, which provided a structured framework for defining training logic. The `training_step` and `validation_step` functions were implemented to handle loss computation and metric logging. The loss function used was CrossEntropyLoss, appropriate for multi-class classification tasks.

An AdamW optimizer with a learning rate of 5×10^{-5} was chosen for training. To prevent overfitting and ensure steady convergence, a ReduceLROnPlateau scheduler was implemented, which reduces the learning rate if the validation loss plateaus. The model was trained for 30 epochs with a batch size of 16. Training was conducted using a GPU where available to speed up computations.

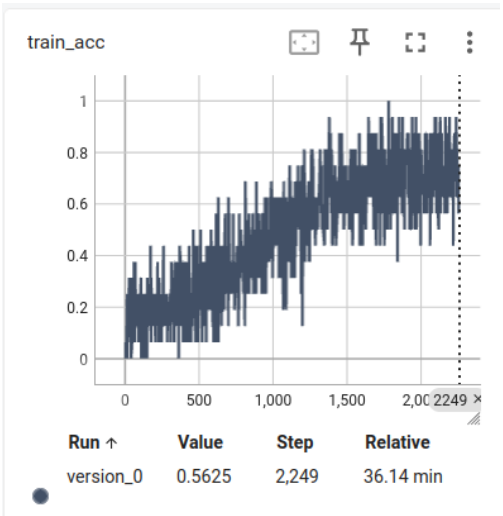

Figure 1: Training Accuracy Curve

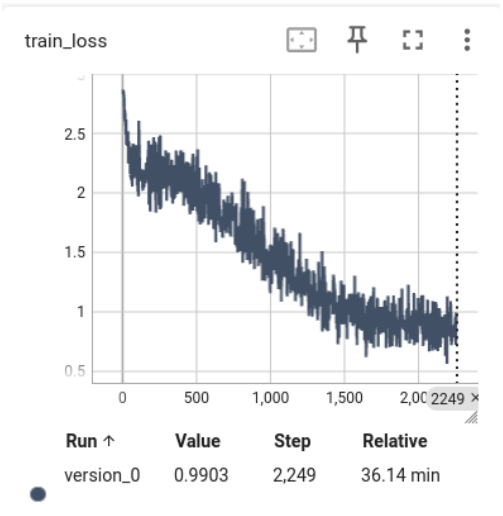

Figure 2: Training Loss Curve

7 Model Evaluation

After training, the model was evaluated on the validation set. The final validation accuracy achieved was **68.75%**, indicating a reasonable level of performance given the complexity of the IELTS scoring task. The loss and accuracy metrics were logged at each step to monitor training progress.

A confusion matrix was plotted to analyze the model's performance across different band scores. This matrix helps in understanding where the model is making correct and incorrect predictions, offering insights into potential areas for improvement.

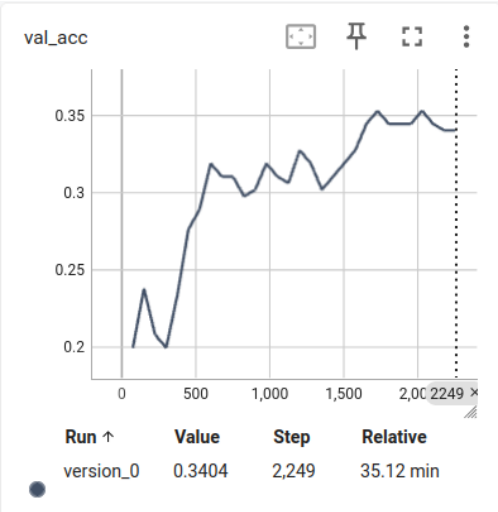

Figure 3: Validation Accuracy Curve

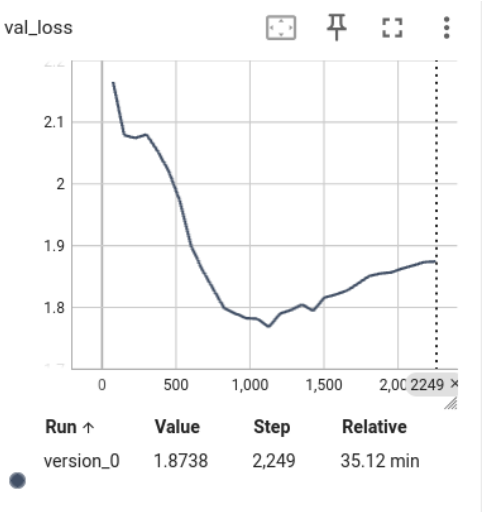

Figure 4: Validation Loss Curve

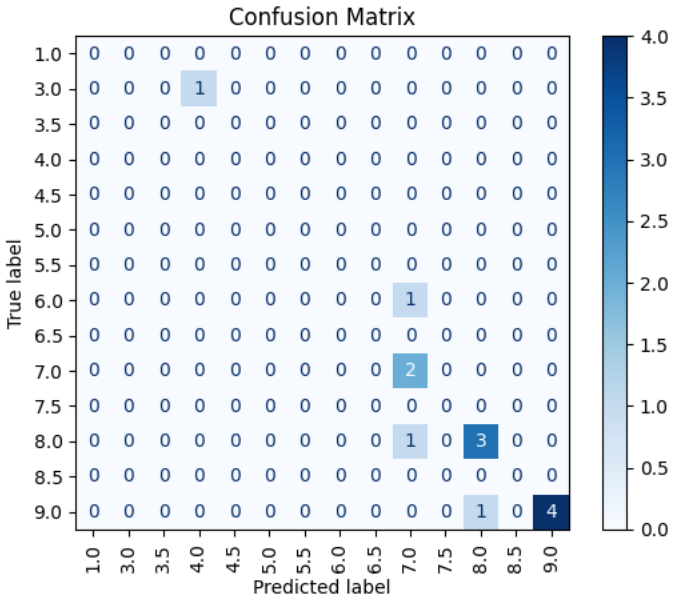

Figure 5: Confusion matrix of model predictions

8 Inference for Band Score Prediction

To enable real-world application, an inference function was implemented. This function takes a new essay-question pair, tokenizes the input, and passes it through the trained model to generate a predicted band score.

Input	**True Labels**	**Predicted Labels**
Question: The graphs below show the numbers...	7.0	6.5
Question: Should people spend a lot on weddings...	8.0	8.0
Question: Some people feel that the design of...	7.0	7.0
Question: Some people think that physical strength...	4.0	4.5
Question: Write about the following topic. The issue...	8.0	8.0

Continued on next page

Input	True Labels	Predicted Labels
Question: The graph below shows information on employment...	6.0	6.0
Question: People are having more and more sugar...	7.0	7.0
Question: The diagram shows a process of making...	8.5	8.0
Question: The charts show survey results concerning why...	6.5	6.5
Question: The table below gives information about the...	6.0	6.0
Question: The chart shows the number of mobile...	6.5	6.5
Question: The bar charts below show the number...	7.0	7.0
Question: The diagram below shows the structure of...	5.5	6.0
Question: Some people think that it is important...	7.0	7.0
Question: The pie graphs show the nutritional consistency...	6.5	6.0
Question: Write about the following topic: Some...	6.5	6.5

Table 1: Predictions over one batch of validation data

9 Conclusion

BERT-based model was successfully fine-tuned for IELTS essay scoring, achieving a validation accuracy of 68.75%. By utilizing LoRA, computational costs were reduced while maintaining effective fine-tuning. A deployable inference function was developed, making the model practical for real-world use.

While the current implementation demonstrates promising results, future improvements could further enhance its accuracy and generalizability. Potential areas for development include using more sophisticated NLP techniques, leveraging larger pre-trained models, and refining data augmentation strategies.