Fine-Tuning ERNIE 2.0 for Text Classification

Rajan Singh

1 Introduction

The goal of this experiment is to build a text classifier using the HuggingFace library and fine-tuning the model to replicate high-performing text classifiers. The model that was used from HuggingFace is called ERNIE 2.0 Base, which is a "continual pre-training framework . . . that builds and learns incrementally pre-training tasks through constant multi-task learning" (1). The model was run using Google Colab Pro on a T4 GPU. In this report, the building of this classifier will be described, detailing the models used, test set performance, training information, and error analysis.

2 Model Performance and Comparison

The evaluation metrics used in this experiment were accuracy and loss. The model that was fine-tuned and tested had an accuracy of 86.78% on the test data with loss of 0.309. The paper for this dataset did not propose any metrics, but on the Papers with Code leaderboard, this model had an accuracy of 86.1%, so this fine-tuned model performed similarly to the one on the leaderboard in that metric. Loss is not listed for the leaderboard's model.

3 Training Information

Training took 50 seconds for the first epoch and 47 seconds for the second epoch. Evaluation took 1 second for each epoch. The number of epochs used was 2 to avoid overfitting; the learning rate was started at 2e-5 and adjusted as we trained, and weight decay was 0.01. Dropout was 0.1, there were 12 layers and 12 heads with 768 as the hidden size. The dropouts, layers, and hidden size are kept from the base ERNIE setup. The learning curve for training is included as Figure 1. The downward trend with some upticks shows that training loss decreased as we trained more batches, but there are some issues with stability during the training, causing those sharp increases in loss. When the batch size is increased, the stability increases and the line is smoother, but the accuracy goes down to about 82%.

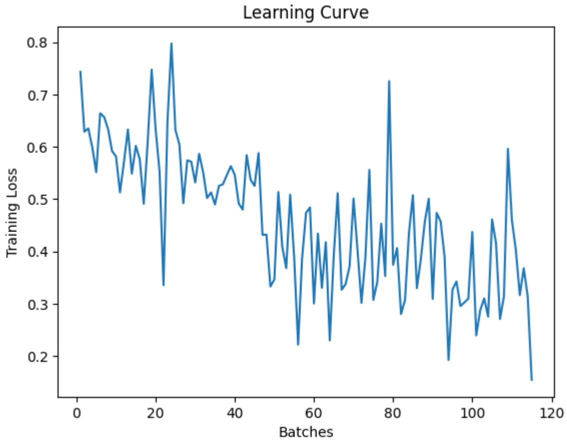

Figure 1: Learning curve of training loss over batches shows a generally downward trending relation

4 Error Analysis

4.1 Incorrect Samples

Incorrect Sample 1:
Sentence 1: The driver, Eugene Rogers, helped to remove children from the bus, Wood said.
Sentence 2: At the accident scene, the driver was "covered in blood" but helped to remove children, Wood said.
Predicted Label: 1 (Confidence: 0.9819)
True Label: 0

Incorrect Sample 2:
Sentence 1: Wal-Mart said it would check all of its million-plus domestic workers to ensure they were legally employed.
Sentence 2: It has also said it would review all of its domestic employees more than 1 million to ensure they have legal status.

Predicted Label: 0 (Confidence: 0.9988)
True Label: 1

Incorrect Sample 3:
Sentence 1: Cooley said he expects Muhammad will similarly be called as a witness at a pretrial hearing for Malvo.
Sentence 2: Lee Boyd Malvo will be called as a witness Wednesday in a pretrial hearing for fellow sniper suspect John Allen Muhammad.
Predicted Label: 1 (Confidence: 0.9992)
True Label: 0

Incorrect Sample 4:
Sentence 1: The top rate will go to 4.45 percent for all residents with taxable incomes above $500,000.
Sentence 2: For residents with incomes above $500,000, the income-tax rate will increase to 4.45 percent.
Predicted Label: 1 (Confidence: 0.9947)
True Label: 0

Incorrect Sample 5:
Sentence 1: The association said 28.2 million DVDs were rented in the week that ended June 15, compared with 27.3 million VHS cassettes.
Sentence 2: The Video Software Dealers Association said 28.2 million DVDs were rented out last week, compared to 27.3 million VHS cassettes.
Predicted Label: 1 (Confidence: 0.9930)
True Label: 0

Incorrect Sample 6:
Sentence 1: No dates have been set for the civil or the criminal trial.
Sentence 2: No dates have been set for the criminal or civil cases, but Shanley has pleaded not guilty.
Predicted Label: 0 (Confidence: 0.6163)
True Label: 1

Incorrect Sample 7:
Sentence 1: Friday, Stanford (47-15) blanked the Gamecocks 8-0.
Sentence 2: Stanford (46-15) has a team full of such players this season.
Predicted Label: 0 (Confidence: 0.9950)
True Label: 1

Incorrect Sample 8:
Sentence 1: The delegates said raising and distributing funds has been complicated by the U.S. crackdown on jihadi charitable foundations, bank accounts of terror-related organizations and money transfers.
Sentence 2: Bin Laden's men pointed out that raising and distributing funds has been complicated by the U.S. crackdown on jihadi charitable foundations, bank accounts of terror-related organizations and money transfers.
Predicted Label: 0 (Confidence: 0.9446)
True Label: 1

Incorrect Sample 9:
Sentence 1: HONG KONG, July 9 Tens of thousands of demonstrators gathered tonight before the legislature building here to call for free elections and the resignation of Hong Kong's leader.
Sentence 2: Tens of thousands of demonstrators gathered yesterday evening to stand before this city's legislature building and call for free elections and the resignation of Hong Kong's leader.
Predicted Label: 1 (Confidence: 0.6344)
True Label: 0

Incorrect Sample 10:
Sentence 1: The results appear in the January issue of Cancer, an American Cancer Society journal, being published online today.
Sentence 2: The results appear in the January issue of Cancer, an American Cancer Society (news - web sites) journal, being published online Monday.
Predicted Label: 0 (Confidence: 0.7174)
True Label: 1

4.2 Suggestions for Improvement

Looking at these samples, numbers seem to cause issues with the classifier. Changing how numbers are represented could improve performance. For example, sample 7 has a team's win-loss record listed, so that is something that could be represented better in the model. Beyond numbers, it seems that additional information can trick the model into thinking that there is a difference. In samples 6 and 10, there is very low confidence and the sentences have almost identical wording, but one sample is longer than the other. If we can reduce how length of strings affects the model's decision making, we could improve performance.

References

[1] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. *ERNIE 2.0: A Continual Pre-training Framework for Language Understanding*. arXiv preprint arXiv:1907.12412, 2019.