CSCI 5541 Natural Language Processing

HW1

Sri Krishna Vamsi Koneru

February 11, 2024

Description of the task and models with references to the original paper and the model repository

Model: Bert-base-cased Huggingface https://huggingface.co/bert-base-cased
Task: Sentiment Analysis
Dataset Link: I used SetFit sst2 as I had found discrepancies while using original sst2 https://huggingface.co/datasets/SetFit/sst2
The Paper Linked:
https://nlp.stanford.edu/sentiment/

What kind of hardware you run your model on

I have used Google Colab Pro with CPU and V100 GPU for compiling.

How do you ensure model has been trained well?

I have plotted the graphs of the training and eval accuracies and losses per epoch using weights and biases. So based on the learning curve and how it changes w.r.t epochs I have determined how well the model is learning from the train dataset. I have evaluated its performance on the validation and test data sets.

The W&B plots are:

The top line left represents evaluation loss followed by training loss. Rightmost graph is the accuracy graph and the second row graph is the eval accuracy which slightly decreases but increases then in latter epoch

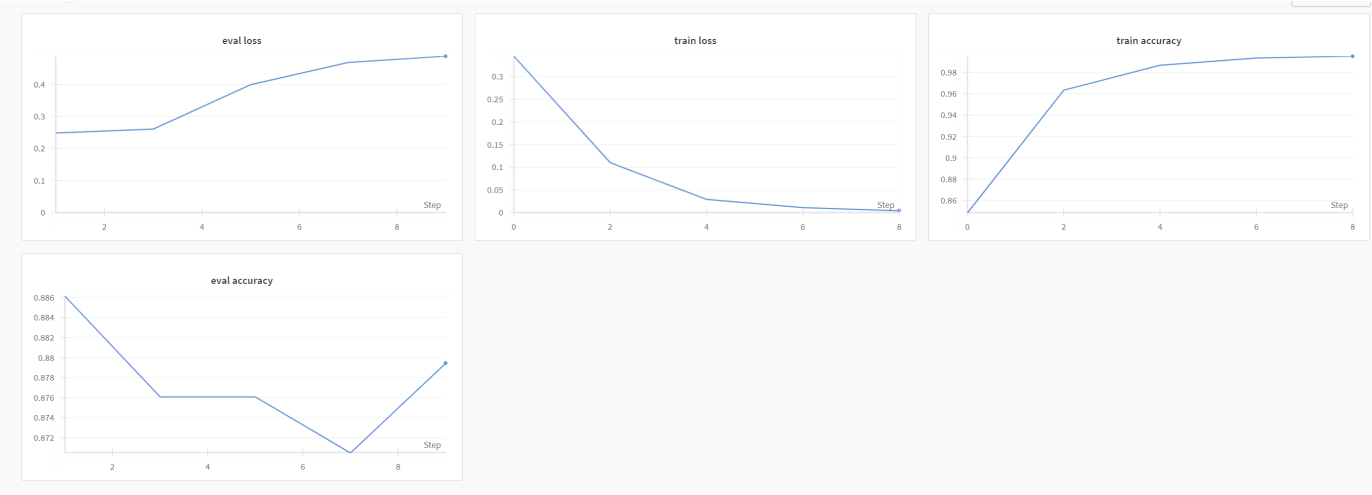

Figure 1: W&B plots

Evaluation Metrics

The evaluation metrics for the model I have used is accuracy. I have chosen this metric as this is the one wihch was reported the most for this task on the leaderboard and the paper as well. Moreover, this seems to be the best measure in this case where we need to ensure the number of both False Positives and False negatives are the least and there is exact match.

Performance on Test Set

I have gotten an accuracy of 90.38% . I have calculated the number of times the label predicted by the model was the same as the ground truth label given in the test dataset and divided it by the number of data samples in the test dataset to give the accuracy score. This is an acceptable measure in my opinion w.r.t the leaderboard as well. While the higher accuracies on the leaderboard were 96above, I feel since I have finetuned pre-trained BERT model which makes sense given my accuracy of over 90. However, I think the difference is in the specific transformations and the optimizations. Some of the papers in the leaderboard have used higher level techniques to achieve these which I am not proficient in.
The reference paper had 80.7% accuracy overall. However the focus of the paper was not just on sentiment analysis and had other areas of interest as well like even in predicting the sentiment they focused on how to deal with negation of negatives and some other peculiar cases which the model I have finetuned doesnt seem to be able to tackle.

Training and Inference Time:

Training time: It took the model 1.928 minutes to complete the training and evaluation epochs specified .
Inference Time: 0.375 minutes. However, I have printed all the values where the model has incorrectly predicted

Hyper Parameters:

The Hyper parameters I have used are the following:
learning_rate=2e-5,
per_device_train_batch_size=32,
per_device_eval_batch_size=32,
num_train_epochs=5,
weight_decay=0.01

Samples my model struggle with

I have represented it in the following table:

Phrase	Predicted label	Ground Truth	Hypothesized Cause	Conf Score
much monkeyfun for all	NEGATIVE	POSITIVE	might be due to the term monkeyfun which it didnt encounter before	0.85
the movie is well shot and very tragic , and one to ponder after the credits roll	NEGATIVE	POSITIVE	This might be due to the term tragic	0.746
a well acted and well intentioned snoozer	POSITIVE	NEGATIVE	considered well as positive and didnt infer snoozer meaning	0.997
george , hire a real director and good writers for the next installment , please	POSITIVE	NEGATIVE	maybe good influenced the decision	0.999
sometimes , nothing satisfies like old-fashioned swashbuckling .	NEGATIVE	POSITIVE	due to oldfashioned and less info on swashbuckling	0.988
despite some gulps the film is a fuzzy huggy .	NEGATIVE	POSITIVE	lack of fuzzy huggy meaning	0.98
her delivery and timing are flawless .	NEGATIVE	POSITIVE	might be influence of trained samples	0.91
not a bad journey at all .	NEGATIVE	POSITIVE	bad might have influenced	0.73
propelled not by characters but by caricatures	POSITIVE	NEGATIVE	lack of inference for caricature	0.998
it almost plays like solaris , but with guns and jokes .	NEGATIVE	POSITIVE	lack of context for solaris	0.996

Table 1: 10 wrongly classified samples

Potential Ideas To Improve Errors

One pattern I observed is that the model is not able to tackle the samples with words that are not as frequent but change the meaning of the sentence. Moreover, I think including diverse speech manners would help in better classifying. I feel this is the reason because as in above table, the ones which fit into my description has high confidence for wrongly predicted label which is attributed to the model not having enough information about the one meaning altering word.
Training on a relatively larger sample might also help the case however again have to ensure the dataset is well represented and also have to make sure overfitting does not occur

Challenging part

This is my first time using Hugging Face and working with remotely anything close to NLP hence, I have learnt a lot of new things. Additionally, this is the first time I have used Overleaf and coded in Latex. Hence it was a bit tough to adapt to this. Please overlook any minor mistakes thank you!!

Annotations

I have attached the spreadsheet in which I have annotated some of the incorrectly predicted values and what I felt were the causes and what could be done to fix those errors. I have made the frequency tables as well like directed.

References

I have given all the references in the first section, additionally I have taken reference while training the model from the finetuning hugging face tutorial taught by the T.A and referred to online sources while debugging some errors.

Thank You