Natural Language Processing - Home Work 1

Name: Mani Deep Cherukuri
StudentId: 5829680
Email: cheru050@gmail.com

Colab link:
https://drive.google.com/file/d/1kDCFPb8v9x6b5XE7obpY8uqzv9eY90mx/view?usp=sharing

Introduction:
In this report, we describe the fine-tuning of the distilbert model('distilbert-base-uncased') provided by the HuggingFace library for the task of Natural Language Inference. We discuss the model being used, the hardware configuration, evaluation metrics, training and inference time, hyperparameters used, incorrectly predicted test samples, potential modelling or representation ideas to improve the errors and the challenging parts of the assignment.

Task and model:
For the assignment, We finetune the **distilbert** model on the **SNLI dataset** which is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the **labels (entailment, contradiction, and neutral),** supporting the **task of natural language inference (NLI).**

Hardware Configurations:
The whole assignment was run on Google Colab with T4 GPU.

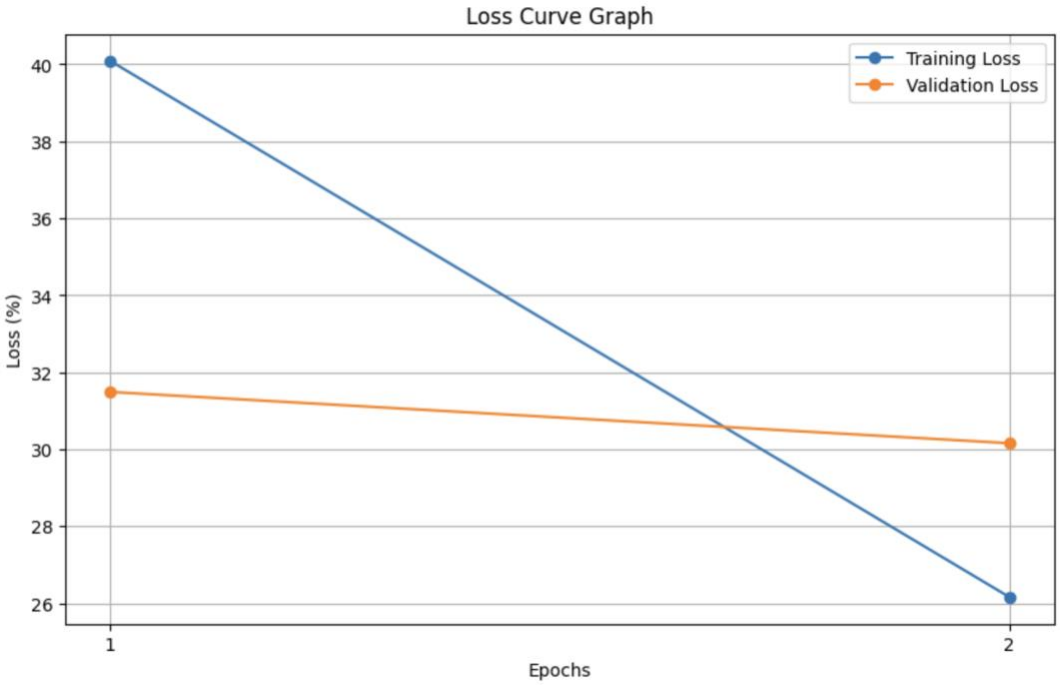
Model Training:

To ensure that the model was trained correctly, we plot the cross entropy loss with respect to the number of epochs. We observe that the validation accuracy crosses the training accuracy indicating that the model needs to be stopped and that it generalises well before overfitting. I also tuned the hyperparameters such as the learning rate and batch size to improve the accuracy. The following are the best accuracies observed in my training:

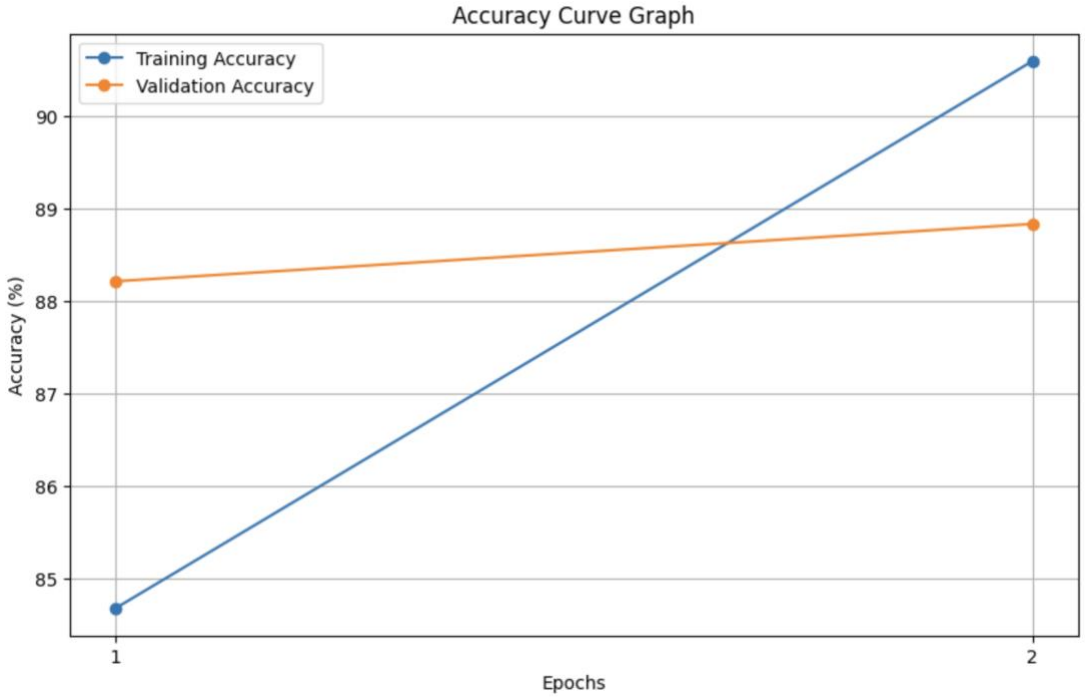
Best Training Accuracy: 90.59%
Best Evaluation Accuracy: 88.83%

```
Training Epoch 0: 100%|██████████| 17168/17168 [39:12<00:00,  7.30batch/s]
Evaluation Epoch 0: 100%|██████████| 308/308 [00:13<00:00, 23.47batch/s]
        Train Loss: 0.401 | Train Acc: 84.68%
        Eval Loss: 0.315 | Eval Acc: 88.21%
Training Epoch 1: 100%|██████████| 17168/17168 [39:07<00:00,  7.31batch/s]
Evaluation Epoch 1: 100%|██████████| 308/308 [00:13<00:00, 23.46batch/s]
        Train Loss: 0.262 | Train Acc: 90.59%
        Eval Loss: 0.302 | Eval Acc: 88.83%
Time: 78.774 minutes
```

Below are the training and evaluation loss curve plots w.r.t. epochs



Below are the training and evaluation accuracy curve plots respectively w.r.t. epochs



Evaluation Metrics:
I used the F1 score, precision, recall and support as my evaluation metrics. Below is the classification report of them.

	precision	recall	f1-score	support
ENTAILMENT	0.92	0.90	0.91	3368
NEUTRAL	0.84	0.87	0.85	3219
CONTRADICTION	0.92	0.90	0.91	3237
accuracy			0.89	9824
macro avg	0.89	0.89	0.89	9824
weighted avg	0.89	0.89	0.89	9824

Test Set performance and comparision:

My accuracy score in on the test set is 88.9 % which is about 1% higher than the validation accuracy

```
100%|██████████| 307/307 [00:13<00:00, 22.56batch/s]
0.8890472312703583
```

Training and Inference Time:
The model took around 80 minutes to train for 2 epoch. The inference took around 13 seconds as seen above.

```
Training Epoch 0: 100%|██████████| 17168/17168 [39:12<00:00,  7.30batch/s]
Evaluation Epoch 0: 100%|██████████| 308/308 [00:13<00:00, 23.47batch/s]
        Train Loss: 0.401 | Train Acc: 84.68%
        Eval Loss: 0.315 | Eval Acc: 88.21%
Training Epoch 1: 100%|██████████| 17168/17168 [39:07<00:00,  7.31batch/s]
Evaluation Epoch 1: 100%|██████████| 308/308 [00:13<00:00, 23.46batch/s]
        Train Loss: 0.262 | Train Acc: 90.59%
        Eval Loss: 0.302 | Eval Acc: 88.83%
Time: 78.774 minutes
```

Hyperparameters:

The hyperparameters used in the model are:

- Number of epoch: 2
- Learning rate: 2e-5
- Weight decay: 0.01
- Hidden size: 768
- Batch size: 32

Incorrectly Predicted Test Samples:

Premise	Hypothesis	Pred	Actual
This church choir sings to the masses as they sing joyous songs from the book at a church.	A choir singing at a baseball game.	1 (neutral)	2 (contradiction)
A woman with a green headscarf, blue shirt and a very big grin.	The woman is young.	2 (contradiction)	1 (neutral)
A woman with a green headscarf, blue shirt and a very big grin.	The woman has been shot.	1 (neutral)	2 (contradiction)
A statue at a museum that no seems to be looking at.	The statue is offensive and people are mad that it is on display.	2 (contradiction)	1 (neutral)
A land rover is being driven across a river.	A sedan is stuck in the middle of a river.	1 (neutral)	2 (contradiction)
A statue at a museum that no seems to be looking at.	Tons of people are gathered around the statue.	1 (neutral)	2 (contradiction)
A land rover is being driven across a river.	A Land Rover is splashing water as it crosses a river.	2 (contradiction)	0 (entailment)
A man playing an electric guitar on stage.	A man is performing for cash.	2 (contradiction)	1 (neutral)
One tan girl with a wool hat is running and leaning over an object, while another person in a wool hat is sitting on the ground.	A tan girl runs leans over an object	2 (contradiction)	0 (entailment)
One tan girl with a wool hat is running and leaning over an object, while another person in a wool	A boy runs into a wall	1 (neutral)	2 (contradiction)

`hat is sitting on the ground.`			

Hypothesis:

The above is the confusion matrix observed for the labels. Looking at it, we can confirm that our model is having difficulty classifying neutral labels. It might be mistaking those for entailment and contradiction at a roughly equal frequency.

Potential Modelling or representation ideas to improve the error:
Since these are human annotated datasets, it might be difficult for an individual to classify them accurately. There might be lot of factors like biasing, cultural differences, demographics etc depending on one's perspective. If these small changes could be fixed, the model might be able to understand the context of the premise and hypothesis to label appropriately during the model training.

Most Challenging part:
The most challenging part was in the customtrainer section and the preprocessing part. Initially, I assumed that the dataset was clean and had just 3 labels. While I was implementing the custom trainer section, the model was running into cuda issue after just few steps. I had to train the model in the cpu mode to get the full traceback error. Debugging further, I realised that the dataset instances which don't have any gold label are marked with -1 label. So, filtered them before starting the training using datasets.Dataset.filter function solved my issue.

References:

https://nlp.stanford.edu/pubs/snli_paper.pdf
https://huggingface.co/datasets/snli
https://nlp.stanford.edu/projects/snli/
https://arxiv.org/abs/1910.01108
ChatGPT