NLP Final Project- Individual Report

Rohan Thekanath

1.Introduction:

The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems (source). The GLUE benchmark consists of nine natural language processing (NLP) tasks, all of which contain sentences or sentence-pairs used to assess model performance across a wide range of linguistic phenomena found in natural language. A leaderboard exists to compare model performance against fellow contestants across the globe. As a group, we tested various NLP transformer models with varying hyperparameters to test their capabilities and learn to work with the hugging face *transformers* module.

The datasets used in this project were provided on the GLUE Benchmark website, consisting of a train, test, and validation set per task. The datasets along with their corresponding tasks and metrics are as follows:

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

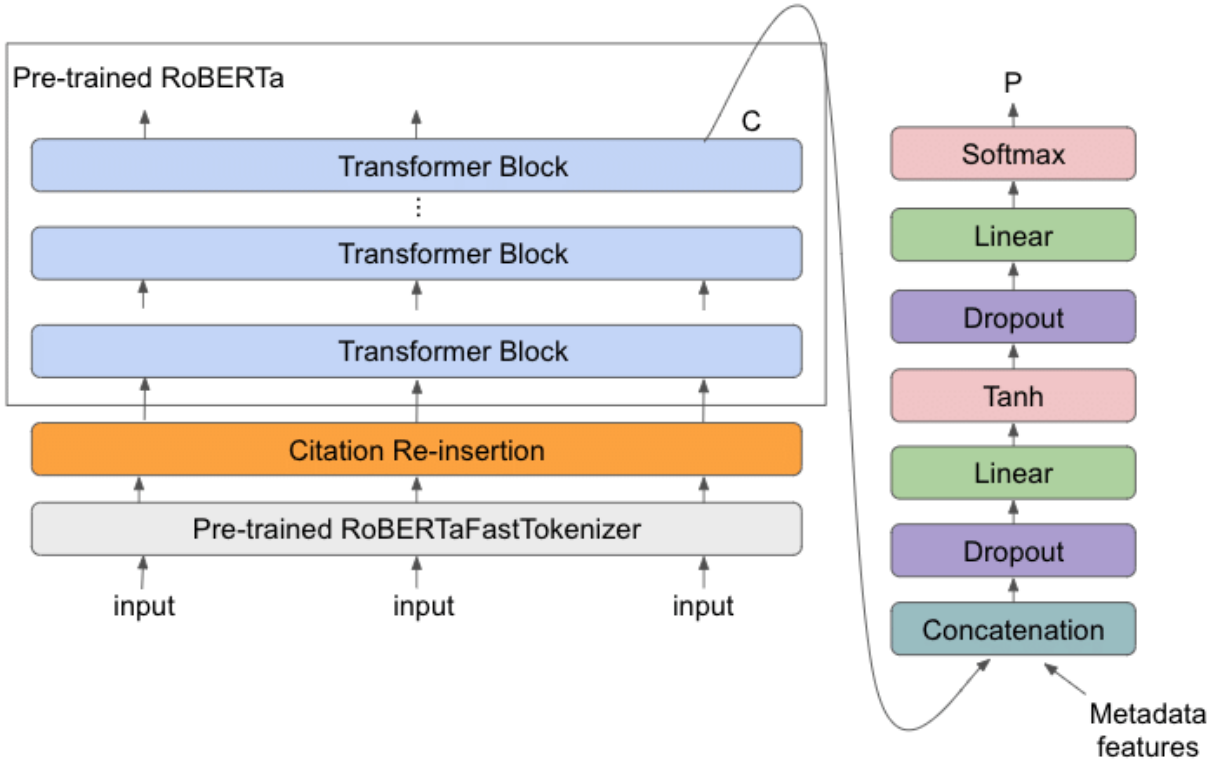
Network and Models

The transformer models we chose to test are the DeBerta and RoBerta models. As the names signify, these transformers' architectures are variations of the Bert transformer model.

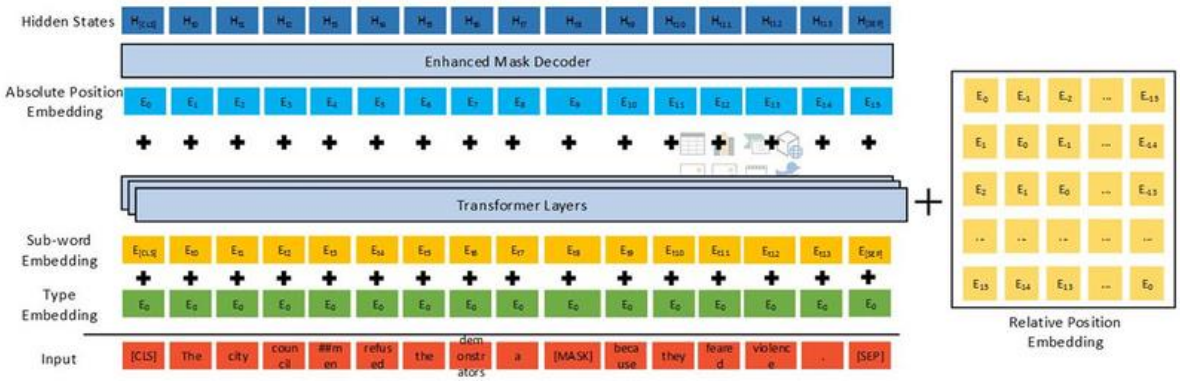
2.

Out of the tasks that are available on the GLUE benchmark, I was responsible for completing The Corpus of Linguistic Acceptability (COLA) and Quora Question Pairs tasks. For both these tasks, I tried, RoBERTa and DeBERTa Transformer models and DeBERTa performed the best.

The below diagram is the architecture of RoBERTa.



The below diagram is the architecture of DeBERTa



3.

The codes to both the tasks are attached below.

training_QQP.py training_COLA.py

4. Results

Using RoBERTa I was able to achieve the below results:

For COLA –

{‘accuracy’: 0.84, Matthew’s correlation coefficient:0.632}

a. Batch size: 32
b. Epoch: 4

For QQP –

QQP - {‘accuracy’: 89}

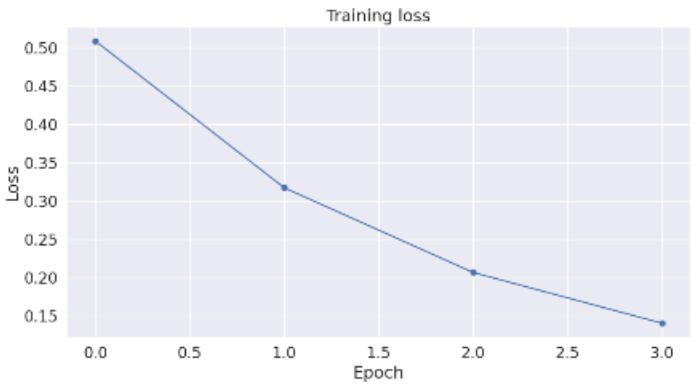
a. Batch size: 8
b. Epoch: 1

Using DeBERTa, I was able to achieve better score which are shown below:

For COLA –

{‘accuracy’: 0.86,’Matthew’s correlation coefficient’:0.67}

a. Batch size: 32
b. Epoch: 4



The above figure shows how the loss reduces considerably across the epochs.

For QQP –

{‘accuracy’: 0.90}
a. Batch size: 8
b. Epoch: 1

Since training was taking a lot of time I had to reduce to number of epochs. I am assuming that if this was not an issue I would be able to increase my score further.

5. Summary/Conclusion:

I learned that using transformer models one can get really good results. After looking at the leaderboard, I learned that we can also ensemble different models to get ever more powerful results. I got a better understanding of PyTorch as both the tasks have been codes in that. More transformer models can also be tried to get better results.

6. About 15% of the code was used from online sources.

7. References:

1. https://gluebenchmark.com/
2. https://huggingface.co/docs/transformers/model_doc/deberta
3. https://huggingface.co/docs/transformers/model_doc/roberta
4. https://github.com/huggingface/transformers/blob/5bfcd0485ece086ebcbed2d008813037968a9e58/examples/run_glue.py#L128
5. https://www.kaggle.com/aryansakhala/bert-pytorch-cola-classification#Evaluate-on-Test-Set