GLUE Benchmark Report

Natural Language Processing
Final Project Report
Bradley Reardon, Nigel Martis, Jongchan Kim, Rohan Paul
12/8/21

Table of Contents

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.

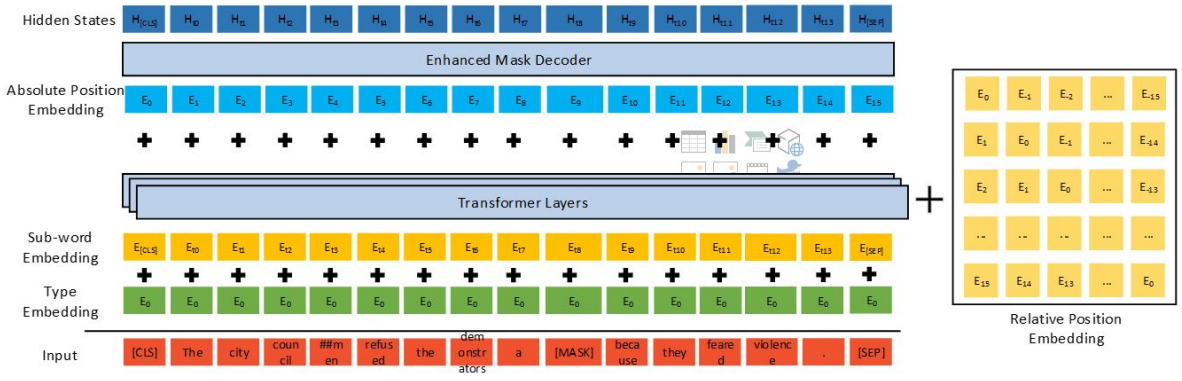
Dataset Overview

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:

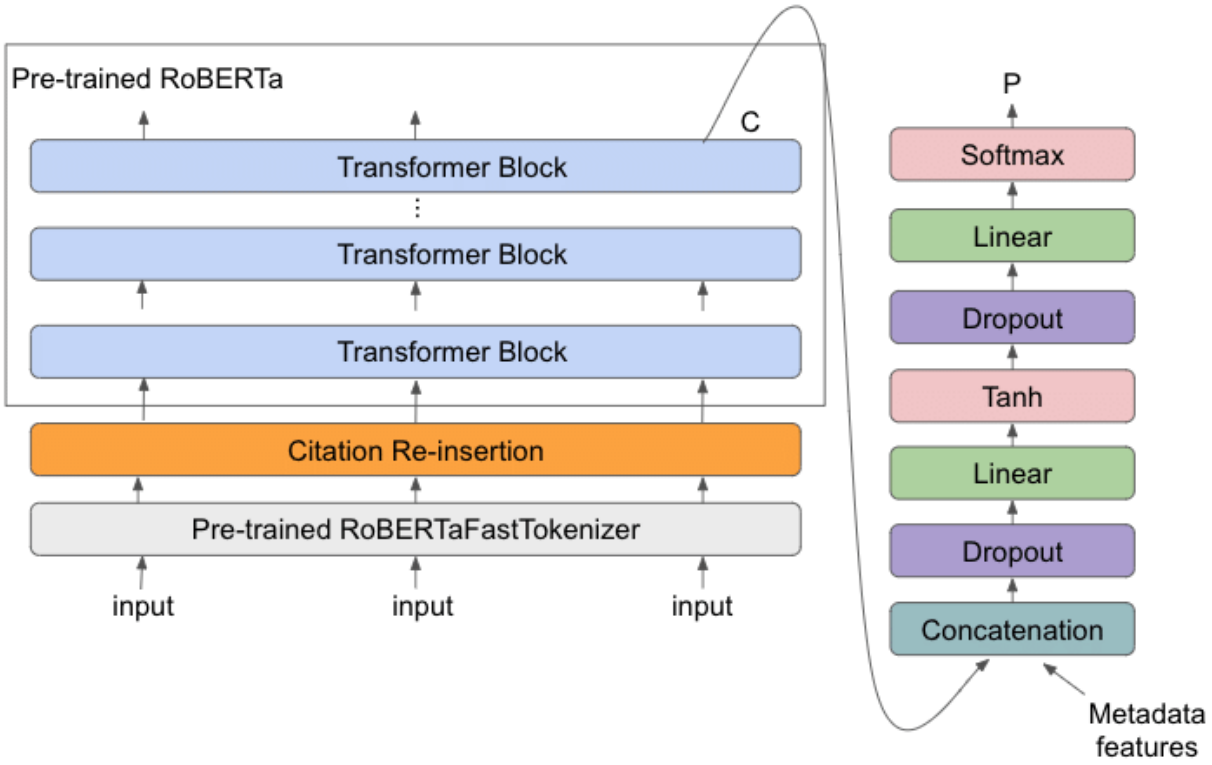
Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (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, with the DeBERTa model being a rendition of the RoBERTa model. RoBERTa iterates on BERT's pretraining procedure, including training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data.DeBERTa (Decoding-enhanced BERT with disentangled attention) improves the BERT and RoBERTa models using two novel techniques. The first is the disentangled attention mechanism, where each word is represented using two vectors that encode its content and position, respectively, and the attention weights among words are computed using disentangled matrices on their contents and relative positions. Second, an enhanced mask decoder is used to replace the output softmax layer to predict the masked tokens for model pretraining. These two techniques significantly improve the efficiency of model pre-training and performance of downstream tasks.



DeBERTa architecture (source)



RoBERTa architecture (source)

Experimental Setup

We experimented with the DeBERTa and RoBERTa transformer models to see which would perform best on the GLUE Benchmark tasks. When testing, we changed both epoch count and batch size in an attempt to improve model performance. We ran into efficiency and memory issues and had to alter epoch and batch size accordingly to allow for our models to process the data. Unfortunately, we did not have the chance to test other hyperparmeters such as optimizer and learning rate due to time constraints and computing power limitations. The RoBERTa

tokenizer was paired with the RoBERTa model and similarly the DeBERTa tokenizer was paired with the DeBERTa model in our experiments.

Results

As shown in the following tables, we ran all models using the AdamW optimizer and a learning rate of 0.0005. This was due to time constraints and a lack of processing power allowing us to A/B test all hyperparameters. The number of epochs and batch size varies per task due to dataset size and sequence length. In order to increase efficiency to allow testing of multiple different models, we had to sacrifice performance by using lower epoch counts and the highest batch sizes our cloud hardware would allow. DeBERTa outperformed RoBERTa in COLA, SST-2, STSB, QQP, MNLI matched, MNLI mismatched, and RTE. The models performed equally on QNLI and WNLI, and RoBERTa outperformed DeBERTa on MRPC.

RoBERTa					
Task	Optimizer	Epoch	Batch Size	Learning Rate	Metrics
CoLA	AdamW	4	32	0.00005	Acc: 0.84, Mat Cor: 0.63
SST-2	AdamW	4	64	0.00005	Acc: 0.94
MRPC	AdamW	4	8	0.00005	Acc: 0.89, F1: 0.92
STS-B	AdamW	4	8	0.00005	Pearson: 0.91, Spearman: 0.90
QQP	AdamW	1	8	0.00005	Acc: 0.89
MNLI-m	AdamW	4	64	0.00005	Acc: 0.88
MNLI-mm	AdamW	4	64	0.00005	Acc: 0.87
QNLI	AdamW	4	8	0.00005	Acc: 0.92
RTE	AdamW	4	8	0.00005	Acc: 0.52
WNLI	AdamW	4	8	0.00005	Acc: 0.56

DeBERTa					
Task	Optimizer	Epoch	Batch Size	Learning Rate	Metrics
CoLA	AdamW	4	32	0.00005	Acc: 0.90, Mat Cor: 0.67
SST-2	AdamW	4	64	0.00005	Acc: 0.95
MRPC	AdamW	4	8	0.00005	Acc: 0.85, F1: 0.89
STS-B	AdamW	4	8	0.00005	Pearson: 0.90, Spearman: 0.90
QQP	AdamW	1	8	0.00005	Acc: 0.90
MNLI-m	AdamW	4	64	0.00005	Acc: 0.89
MNLI-mm	AdamW	4	64	0.00005	Acc: 0.88
QNLI	AdamW	4	8	0.00005	Acc: 0.92
RTE	AdamW	4	8	0.00005	Acc: 0.55
WNLI	AdamW	4	8	0.00005	Acc: 0.56

Transformer Model Comparison		
Model	Task	Score
DeBERTa	CoLA	Acc: 0.90, Mat Cor: 0.67
RoBERTa	CoLA	Acc: 0.84, Mat Cor: 0.63
DeBERTa	SST-2	Acc: 0.95
RoBERTa	SST-2	Acc: 0.94
DeBERTa	MRPC	Acc: 0.85, F1: 0.89
RoBERTa	MRPC	Acc: 0.89, F1: 0.92
DeBERTa	STS-B	Pearson: 0.90, Spearman: 0.90
RoBERTa	STS-B	Pearson: 0.91, Spearman: 0.90
DeBERTa	QQP	Acc: 0.90
RoBERTa	QQP	Acc: 0.89
DeBERTa	MNLI-m	Acc: 0.89
RoBERTa	MNLI-m	Acc: 0.88
DeBERTa	MNLI-mm	Acc: 0.88
RoBERTa	MNLI-mm	Acc: 0.87
DeBERTa	QNLI	Acc: 0.92
RoBERTa	QNLI	Acc: 0.92
DeBERTa	RTE	Acc: 0.55
RoBERTa	RTE	Acc: 0.52
DeBERTa	WNLI	Acc: 0.56
RoBERTa	WNLI	Acc: 0.56

Conclusion

When comparing the results of our experiments, we found that DeBERTa overall outperformed RoBERTa, excelling in all tasks other than MRPC. This was expected considering DeBERTa is essentially an improved version of RoBERTa. Our scores on the validation sets ended up being very close to those of the test scores on the GLUE benchmark leaderboard for the majority of the tasks.

In the future, a few changes we could make when experimenting further are increasing the computing processing power and increased variation in our testing. The increased processing power would allow us to experiment with varying hyper parameters such as number of epochs, learning rate, batch size, and optimizer. Due to time constraints and a lack of resources, we were unable to test as many times as we wanted.

If we had not dealt with computing power limitations and time constraints, we would have liked to test out ensembling multiple models. This would most likely increase our metric scores and decrease the gap between our scores and those at the top of the leaderboards.

Citations:

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
6. He, Pengcheng, et al. "Deberta: Decoding-enhanced bert with disentangled attention." arXiv preprint arXiv:2006.03654 (2020).