GLUE Benchmark Individual Report

Natural Language Processing
Final Project Individual Report
Nigel Martis
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.

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

Describing individual work:

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. I was responsible for 3 tasks SST-2, MNLI-matched and MNLI-mismatched. 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.

Description of the portion of work:

- All 3 tasks were trained on Google Cloud platform with a 16gb GPU.
- The Batch size was 64 and Epochs were 4 for all tasks.
- Max sequence length for SST-2 was 256 and for MNLI was 128.
- SST-2 took approximately 2 hours to train, where as MNLI took about 5 hours to train.
- For SST-2, `Number of training sentences: 67349 Number of validation sentences: 872`
- For MNLI-matched, `Number of training sentences: 391120 Number of validation sentences: 9714`
- For MNLI-mismatched, `Number of training sentences: 391120 Number of validation sentences: 9832`

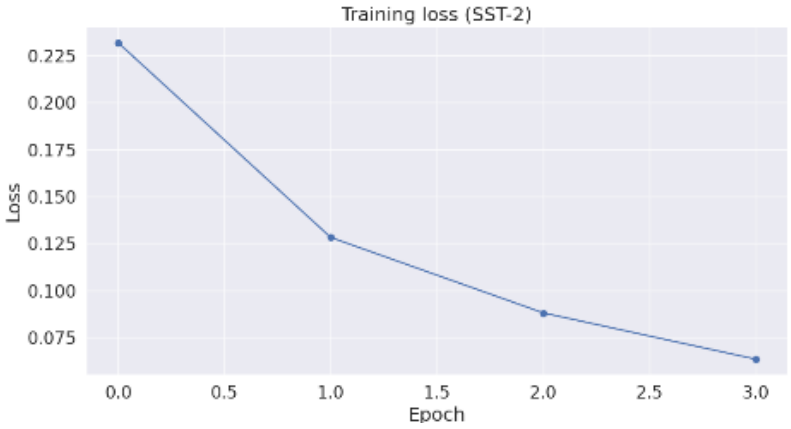
```python
#Deberta
from transformers import DebertaTokenizer, DebertaForSequenceClassification, AdamW, DebertaConfig

print('Loading Deberta tokenizer...')
tokenizer = DebertaTokenizer.from_pretrained('microsoft/deberta-base')

# Tokenize and map the tokens to word IDs.
# And Padding
train = tokenizer(list(train_sentences), padding='max_length', max_length = 256, truncation=True)
train_input_ids = train['input_ids']
train_masks = train['attention_mask']

validation = tokenizer(list(validation_sentences), padding='max_length', max_length = 256, truncation= True)
validation_input_ids = validation['input_ids']
validation_masks = validation['attention_mask']
```

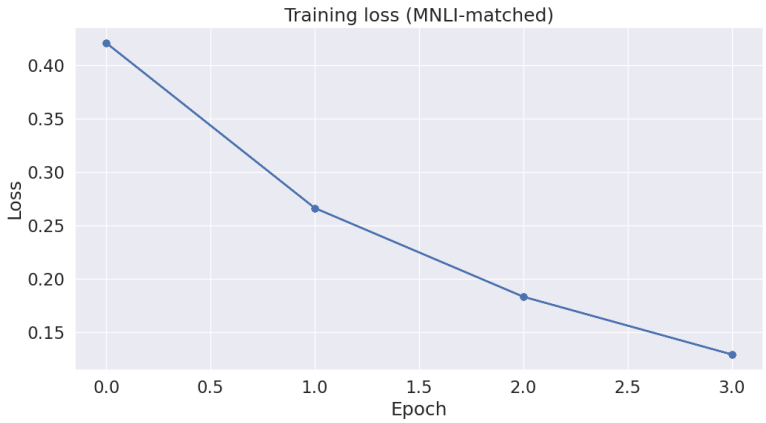
```python
#Deberta
from transformers import DebertaTokenizer, DebertaForSequenceClassification, AdamW, DebertaConfig

print('Loading Deberta tokenizer...')
tokenizer = DebertaTokenizer.from_pretrained('microsoft/deberta-base')

# Tokenize and map the tokens to word IDs.
# And Padding to length 128
train = tokenizer(list(train_sentences1), list(train_sentences2), max_length=128, padding='max_length', truncation = True, return_token_type_ids= True)
train_input_ids = train['input_ids']
train_masks = train['attention_mask']
train_tokentype = train['token_type_ids']

validation = tokenizer(list(validation_sentences1), list(validation_sentences2), max_length=128, padding='max_length', truncation = True , return_token_type_ids= True
validation_input_ids = validation['input_ids']
validation_masks = validation['attention_mask']
validation_tokentype = validation['token_type_ids']
```

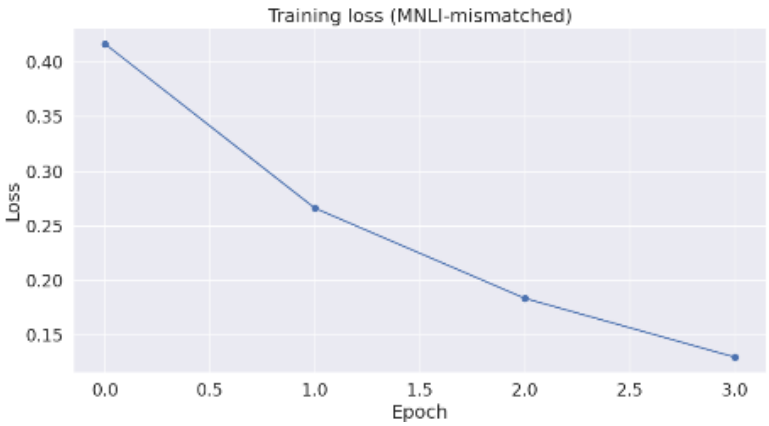
Results:

DeBerta:

1. SST-2: {‘accuracy’: 0.95}
    a. Batch size: 64
    b. Epoch: 4
2. MNLI matched - {'accuracy': 0.89}
    a. Batch size: 64
    b. Epoch: 4
3. MNLI mismatched - {'accuracy': 0.88}
    a. Batch size: 64
    b. Epoch: 4



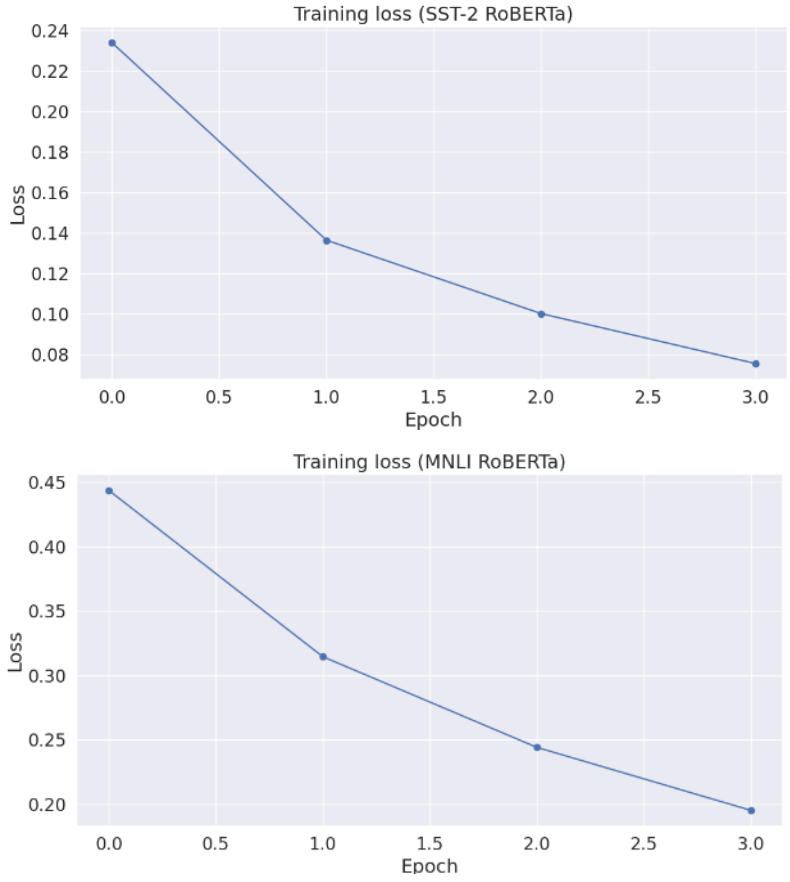




RoBerta:

1. SST-2: {‘accuracy’ : 0.94}
    a. Batch size: 64
    b. Epoch: 4
2. MNLI matched - {'accuracy': 0.88}
    a. Batch size: 64
    b. Epoch: 4
3. MNLI mismatched - {'accuracy': 0.87}

a. Batch size: 64
b. Epoch: 4



Summary and Conclusions:

To summarize, the DeBERTa model performed better on the validation sets of all 3 tasks i.e. SST-2, MNLI-matched and MNLI-mismatched. The scores of all 3 tasks on validation were close to the test scores on the glue benchmark leaderboard.

Percentage of Code written:

The information about DeBERTa and RoBERTa was referred from huggingface. I also referred huggingface for the actual train script for glue. About 35% of the code was referred from the internet.

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