ECE60146: Homework 10

Manish Kumar Krishne Gowda, 0033682812

(Spring 2024)

1 Introduction

In this report, we use BERT, a powerful large language language tool created by Google. Our task involved fine-tuning a pre-trained BERT model to perfom Q&A analysis on a subset of the The Stanford Question Answer Dataset (SQuAD). To help us with this, we used a framework called Hugging Face transformers.

2 Methodology

Pickle library was used to load the train, test and, eval dataset. To fine-tune the BERT model for the question and answer task, we needed to make some adjustments. Specifically, we added extra layers to the existing BERT model to tailor it for this purpose.

First, we specified the base BERT model we wanted to use. In our case, we chose the 'bert-base-uncased' model. This model is already pre-trained on a vast amount of text data. Next, we imported a special tool called BertForQuestionAnswering from the Hugging Face transformers library. This tool helps us fine-tune the BERT model specifically for question and answer tasks.This means we started with the pre-trained weights and architecture of the BERT base model, which we can then fine-tune for our specific question and answer task. With these steps, we prepared our BERT model to tackle the task of question answering, setting the stage for further fine-tuning and analysis.

```
#First, initialize a model. We use BertForQuestionAnswering to initialise the
                                                model
from transformers import BertForQuestionAnswering

model_name = 'bert-base-uncased'
model = BertForQuestionAnswering.from_pretrained(model_name)
```

The Bert model used consists of multiple layers, including embeddings, encoder, and output layers.

1. **Embeddings Layer** : This layer handles the transformation of input tokens into dense vectors. It includes sub-layers for word embeddings, positional embeddings, and token type embeddings. Word embeddings convert each input token into a fixed-size vector representation, while positional embeddings encode the position of each token in the input sequence. Token type embeddings distinguish between different segments of input text, such as question and context in a question answering task. Additionally, layer normalization and dropout are applied to enhance model robustness and generalization.

2. **Encoder Layer** : The encoder consists of multiple BertLayer modules stacked on top of each other. Each BertLayer contains several sub-modules, including BertAttention, BertIntermediate, and BertOutput. The BertAttention module performs multi-head self-attention, allowing the model to focus on different parts of the input sequence simultaneously. The BertIntermediate module applies a feed-forward neural network to transform the attention output into a higher-dimensional representation. Finally, the BertOutput module combines

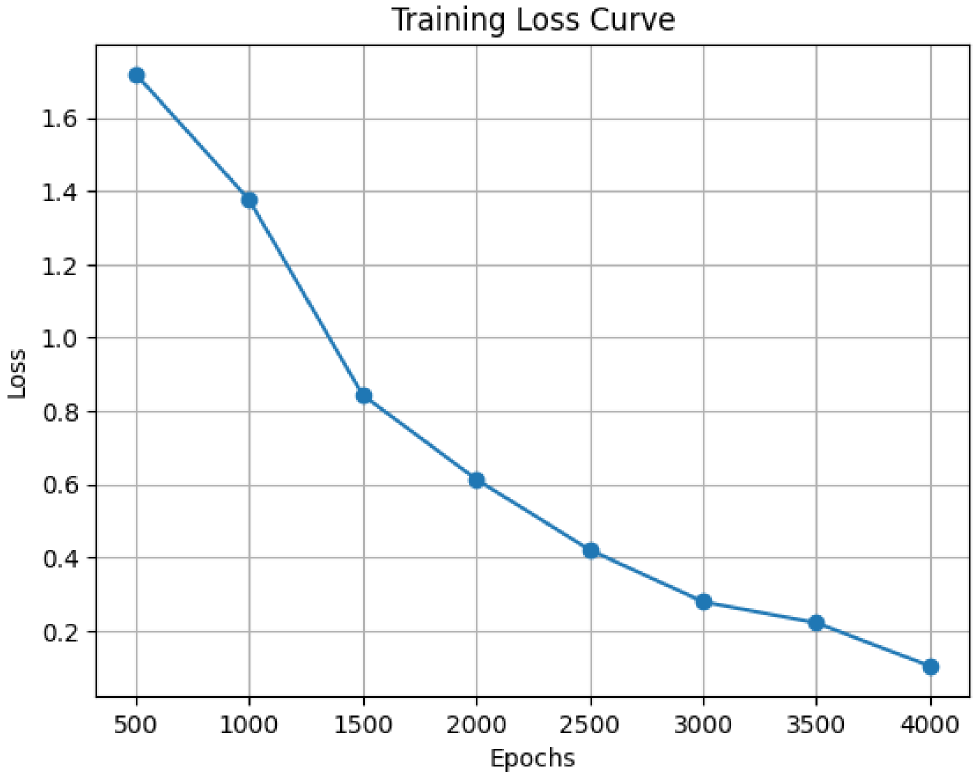


Figure 1: Loss curve of Bert model for 5 Epochs

the intermediate representation with the original input through residual connections and layer normalization.

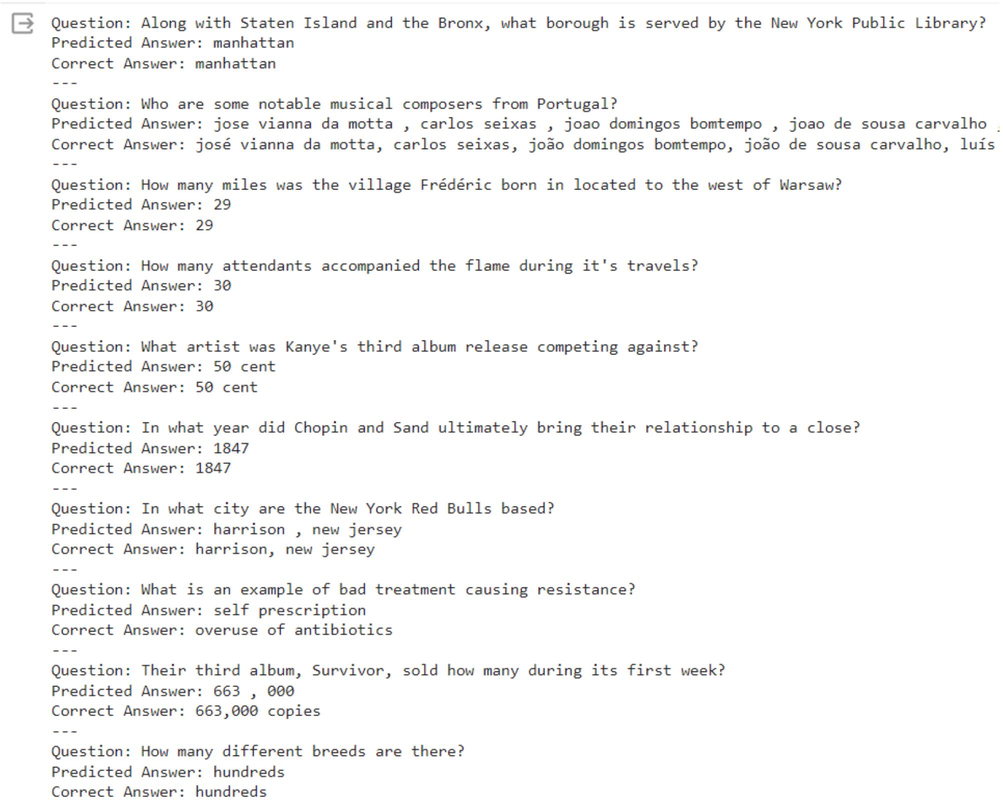
3. **Output Layer** : The output layer is responsible for producing the final predictions for the question answering task. It includes a linear transformation (Linear) that maps the output of the last layer in the encoder to a 2-dimensional vector representing the probability distribution over start and end positions of the answer span. This linear transformation is followed by softmax activation, which normalizes the output scores to represent probabilities

Training for a model is done using the same Hugging Face transformers library. First we define the training arguments like the output directory, with the model training for 5 epochs. It sets the batch size per device during training and evaluation to 8. Then, we prepares the datasets for training, evaluation, and testing. The model is then moved to the appropriate device (GPU if available). Finally, a Trainer object is created with the model, training arguments, and datasets, and training is initiated with the trainer.train() method. This is the usual training method carried out for other deep learning homework tasks.The training loss is shown in Figure 1. (The x axis here is the number of epochs*step_size, i.e. the total number of iterations)

2.1 Testing the Bert Model

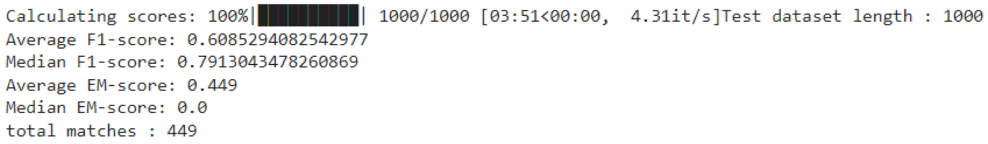
To test the trained model, 10 answers were looked at and qualitatively analysed. The sample output is shown in Figure 2.

Here are some general comments on the predictions.



```
Question: Along with Staten Island and the Bronx, what borough is served by the New York Public Library?
Predicted Answer: manhattan
Correct Answer: manhattan
---
Question: Who are some notable musical composers from Portugal?
Predicted Answer: jose vianna da motta , carlos seixas , joao domingos bomtempo , joao de sousa carvalho ,
Correct Answer: josé vianna da motta, carlos seixas, joão domingos bomtempo, joão de sousa carvalho, luís
---
Question: How many miles was the village Frédéric born in located to the west of Warsaw?
Predicted Answer: 29
Correct Answer: 29
---
Question: How many attendants accompanied the flame during it's travels?
Predicted Answer: 30
Correct Answer: 30
---
Question: What artist was Kanye's third album release competing against?
Predicted Answer: 50 cent
Correct Answer: 50 cent
---
Question: In what year did Chopin and Sand ultimately bring their relationship to a close?
Predicted Answer: 1847
Correct Answer: 1847
---
Question: In what city are the New York Red Bulls based?
Predicted Answer: harrison , new jersey
Correct Answer: harrison, new jersey
---
Question: What is an example of bad treatment causing resistance?
Predicted Answer: self prescription
Correct Answer: overuse of antibiotics
---
Question: Their third album, Survivor, sold how many during its first week?
Predicted Answer: 663 , 000
Correct Answer: 663,000 copies
---
Question: How many different breeds are there?
Predicted Answer: hundreds
Correct Answer: hundreds
```

Figure 2: Sample Predicted Answers - BertforQ&A



```
Calculating scores: 100%|██████████| 1000/1000 [03:51<00:00,  4.31it/s]Test dataset length : 1000
Average F1-score: 0.6085294082542977
Median F1-score: 0.7913043478260869
Average EM-score: 0.449
Median EM-score: 0.0
total matches : 449
```

Figure 3: F1 and EM Score Output of BertforQ&A model

In some cases, the predicted answers are accurate and closely match the correct answers. For example, predicting ”Manhattan” for the borough served by the New York Public Library and ”50 Cent” for the artist competing against Kanye’s third album release demonstrates good accuracy.

Further The model shows decent performance in predicting numerical values, such as miles, number of attendants, or album sales, although it occasionally misses details like commas or units.The model often provides detailed answers, listing multiple examples or providing specific numbers where necessary. For instance, in the question about notable musical composers from Portugal, the predicted answer includes several names, showing an understanding of the breadth of the question.

However, there are inconsistencies in the predictions. For instance, in the question about the New York Red Bulls, the model correctly identifies ”Harrison, New Jersey” but adds unnecessary detail with ”harrison” being lowercase. Similarly, in the question about Chopin and Sand’s relationship, while the predicted year is correct, it lacks punctuation and capitalization.

The relevance of some predictions may vary. For instance, in the question about bad treatment causing resistance, the predicted answer ”self prescription” is relevant but not as specific as the correct answer ”overuse of antibiotics.”Further the model shows decent performance in predicting numerical values, such as miles, number of attendants, or album sales, although it occasionally misses details like commas or units.

Overall, the predictions from the Bert model seem to capture the essence of the correct answers in many cases, but they also exhibit some inaccuracies and inconsistencies.

2.1.1 Quantitative Analysis

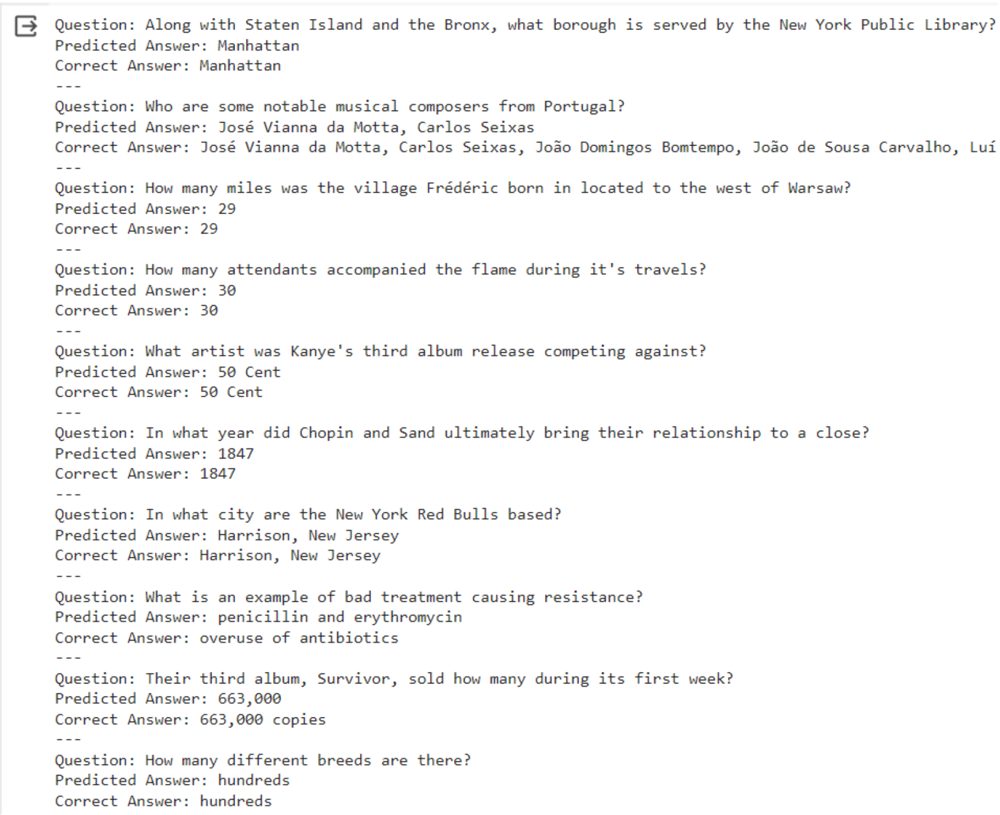
The F1-score measures the model’s ability to balance precision and recall in its predictions, while the Exact Match (EM) score evaluates whether the model’s predictions exactly match the ground truth answers. The F1 and EM score output is shown in Figure 3

1. Average F1-score: With an average F1-score of approximately 0.609, the model achieves moderate performance in terms of precision and recall. This indicates that, on average, the model’s predictions capture about 60.9% of the relevant information compared to the ground truth answers.

2. Median F1-score: The median F1-score of approximately 0.791 suggests that there is significant variability in the model’s performance across different examples. While some predictions achieve high precision and recall, others may fall short.

3. Average EM-score: The average EM-score of 0.449 indicates that, on average, the model’s predictions match the ground truth answers about 44.9% of the time. This suggests that the model struggles to produce exact matches for the correct answers.

4. Median EM-score: The median EM-score of 0.0 indicates that there is a considerable number of examples where the model fails to produce exact matches for the correct answers. This sug-



```
Question: Along with Staten Island and the Bronx, what borough is served by the New York Public Library?
Predicted Answer: Manhattan
Correct Answer: Manhattan
---
Question: Who are some notable musical composers from Portugal?
Predicted Answer: José Vianna da Motta, Carlos Seixas
Correct Answer: José Vianna da Motta, Carlos Seixas, João Domingos Bomtempo, João de Sousa Carvalho, Luí
---
Question: How many miles was the village Frédéric born in located to the west of Warsaw?
Predicted Answer: 29
Correct Answer: 29
---
Question: How many attendants accompanied the flame during it's travels?
Predicted Answer: 30
Correct Answer: 30
---
Question: What artist was Kanye's third album release competing against?
Predicted Answer: 50 Cent
Correct Answer: 50 Cent
---
Question: In what year did Chopin and Sand ultimately bring their relationship to a close?
Predicted Answer: 1847
Correct Answer: 1847
---
Question: In what city are the New York Red Bulls based?
Predicted Answer: Harrison, New Jersey
Correct Answer: Harrison, New Jersey
---
Question: What is an example of bad treatment causing resistance?
Predicted Answer: penicillin and erythromycin
Correct Answer: overuse of antibiotics
---
Question: Their third album, Survivor, sold how many during its first week?
Predicted Answer: 663,000
Correct Answer: 663,000 copies
---
Question: How many different breeds are there?
Predicted Answer: hundreds
Correct Answer: hundreds
```

Figure 4: Sample Predicted Answers - Distilbert Model

gests that there may be specific types of questions or contexts where the model's performance is particularly challenging.

The fact that there are 449 total matches indicates that the model successfully produces exact matches for a subset of the examples. However, the relatively low average and median EM-scores suggest that the model's overall performance in achieving exact matches is limited.

2.2 Distilbert-base-cased-distilled-squad model

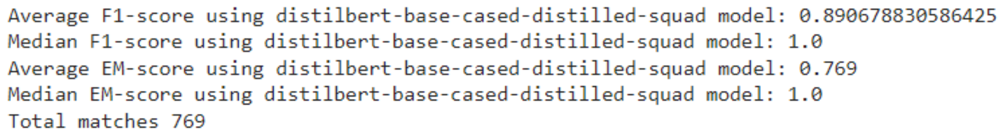
This model from Hugging Face, is a distilled version of BERT fine-tuned on SQuAD dataset.

2.2.1 Qualtitative Analysis

The DistilBERT predictions demonstrate overall strong performance, with several correct answers matching the ground truth closely. The Predicted Answers are shown in Figure 4.

Many predicted answers closely match the correct answers. For instance, predicting "Manhattan" for the borough served by the New York Public Library and "50 Cent" as the artist competing against Kanye's third album release shows accurate understanding of the questions.The predicted answers often capture key details, such as specific composers or numerical values.

However, in some cases, the predictions are slightly less comprehensive compared to the correct answers. For example, in the question about notable musical composers from Portugal, the predicted answer includes fewer names compared to the correct answer.



```
Average F1-score using distilbert-base-cased-distilled-squad model: 0.890678830586425
Median F1-score using distilbert-base-cased-distilled-squad model: 1.0
Average EM-score using distilbert-base-cased-distilled-squad model: 0.769
Median EM-score using distilbert-base-cased-distilled-squad model: 1.0
Total matches 769
```

Figure 5: F1 and EM Score Output of DistilBERT model model

The predicted answers generally remain relevant to the questions asked. However, in some instances, there are deviations. For example, in the question about bad treatment causing resistance, while the predicted answer mentions antibiotics, it misses the specific point about overuse.The model shows good performance in predicting numerical values, such as miles, number of attendants, or album sales, with close matches to the correct answers.The consistency of predictions varies, with some questions yielding highly accurate responses and others slightly less so. However, overall, the predictions maintain a satisfactory level of consistency.

2.2.2 Quantitative Analysis

The performance metrics for the DistilBERT model demonstrate a significant improvement compared to the previous model. The F1 and EM score output is shown in Figure 5

1. Average F1-score: With an average F1-score of approximately 0.891, the DistilBERT model achieves high performance in terms of balancing precision and recall. This indicates that, on average, the model's predictions capture about 89.1% of the relevant information compared to the ground truth answers.

2. Median F1-score: The median F1-score of 1.0 suggests that the majority of the model's predictions achieve perfect precision and recall. This indicates consistent high-quality performance across different examples.

3. Average EM-score: The average EM-score of 0.769 indicates that, on average, the model's predictions exactly match the ground truth answers about 76.9% of the time. This suggests that the model's performance in producing exact matches is quite strong.

4. Median EM-score: The median EM-score of 1.0 indicates that the majority of the model's predictions achieve exact matches for the correct answers. This suggests consistent high-quality performance in providing exact matches.

The fact that there are 769 total matches further reinforces the strong performance of the DistilBERT model. It indicates that the model consistently produces accurate predictions across a wide range of examples.

3 Conclusion

In short, our report looked into different models for answering questions using BERT, a smart language tool. We found that while the basic BERT model did okay, the DistilBERT model, especially the distilbert-base-cased-distilled-squad version, did much better. It gave more accurate and consistent answers. This shows that choosing the right model is important for getting good results. The DistilBERT model is lighter and faster but still very effective. In the future, we can make these models even better for understanding language in all kinds of situations.

References

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016