Artificial Intelligence 2

HW 4

Στέφανος Μπακλαβάς 1115201700093

For the purposes of this hw exercise 1 and 2 were completed.

Exercise A.

1.1 Data Pre-Proccessing

The data pre-proccessing is the same as of hw3.First we remove every whitespace bigger than two spaces .Then we tun capital letters into small letters. An extra step was done by removing all numbers and emojis of the tweets but then it was found that our model performs better if we leave them and replace the with the mean vector of the tweet that they come from.

1.2 Tokenization

For the tokenization of the sequences `AutoTokenizer.from_pretrained('distilbert-base-uncased')` was used.The tokenization proccess was done in batches of size 16 .At every batch we find the longest sequence and we set max_padding = length_of_longest_sequence. If this length is bigger than 512 we set max_padding = 512.

1.3 Model

The model that was used is the `AutoModelForSequenceClassification.from_pretrained("distilbert-base-uncased”)` with batch size = 16.The batch size was set equal to 16 because the limitations in RAM memory did not allow a bigger one. There were tests with three different leaning rates because those were presented at the paper .Those are lr = 2e-5, lr = 3e-5 and lr = 5e-5.

1.4 Results

In all three tests the batch size remains equal to 16 and epochs = 4.

1. learning rate = 2e-5

F1 score: 0.7498224480514987 Precision score: 0.7527190498996266 in 1st epoch

2. learning rate = 3e-5

F1 score: 0.7483531378437176 Precision score: 0.7502920947344444 in 2nd epoch

3. learning rate = 5e-5

F1 score: 0.7584132380185677 Precision score: 0.7584025347075629 in 2nd epoch

As it seems above lr = 5e-5 worked a little bit better in this problem but with insignificant better performance compared to the other two learning rates. In the final .ypnb file we have a little lower scores because I was forced to run again the notebook in order to reproduce the results of example 3.

Exercise B.

2.1 Dataset

For the purposes of this exercise I used SQUAD 2 dataset for question answering. The dataset was then tokenized in order to be used by the model.

2.2 Tokenization

For the tokenization of the sequences `AutoTokenizer.from_pretrained('distilbert-base-uncased')` was used. The max length of each sequence was set to 384 and the overlap between two sequences can be up to 184. This happens because if a sequence has length bigger than max length it brakes into smaller overlapped sequences that may answer the same question.

2.3 Model

The model that was used is the `DistilBertForQuestionAnswering.from_pretrained("distilbert-base-uncased")` with batch size = 24.The batch size was set equal to 24 because the paper that was proposed to read used and produced better results with this batch size. There were tests with three different leaning rates because those were presented at the paper as well .Those are lr = 2e-5, lr = 3e-5 and lr = 5e-5.

1.4 Results

In all three tests the batch size remains equal to 24 and epochs = 1.Unfortunately because of limitations on the colab GPU usage I could not run an example for more than one epochs.

1. learning rate = 2e-5

F1 score: 0.6304063034661433

2.learning rate = 3e-5

F1 score: 0.6349413420387469

3. learning rate = 5e-5

F1 score: 0.6594578834746686

As it seems above lr = 5e-5 worked a little bit better in this compared to the other two learning rates.