Mengke Li
Email: mengkel@clemson.edu
Course: CPSC 8430 - Deep Learning
Instructor: Dr. Luo
Github Link: https://github.com/mengkel/CPSC-8430-DP-HW/tree/main/HW3

Homework - #1
ID: C28207748
Term: Spring 2023
Due Date: 1st April, 2023

Problem 1-1: Train a Question and Answer Model Using Spoken-Squad Dataset (Based on pre-trained model from hugging face)

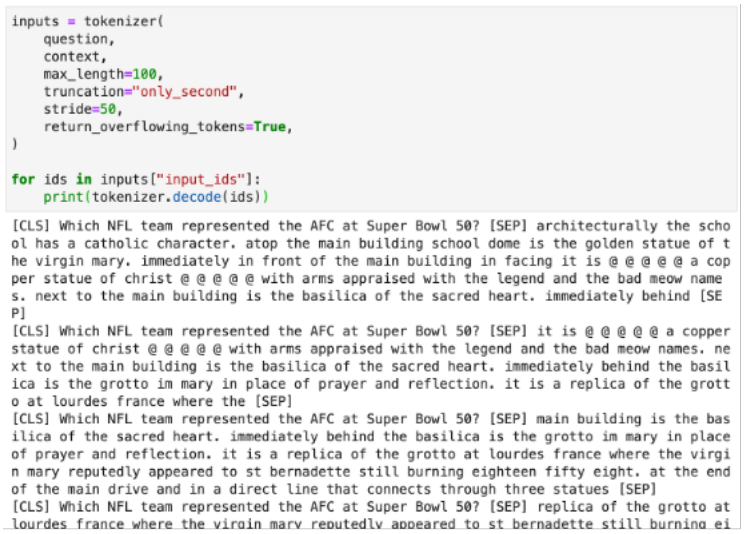
Solution:

- First pre-trained model: bert-base-cased
- Raw data: Spoken_Squad with training sample 37111 and testing sample 5351
- Train epoch: 3

Problem 1-2: Why Long Paragraph is an Issue? How to deal with it?

Solution:

1 Total sequence length = question length + paragraph length + 3 (special tokens). Therefore, if the context is too long, we need to find some ways to process the whole paragraph. If some of the examples in the dataset exceed the maximum length, we can set the internals of the question-answering pipeline, what I did is dealing with contexts by creating several training features from one sample of our dataset, with a sliding window between them.



```
inputs = tokenizer(
    question,
    context,
    max_length=100,
    truncation="only_second",
    stride=50,
    return_overflowing_tokens=True,
)

for ids in inputs["input_ids"]:
    print(tokenizer.decode(ids))
```

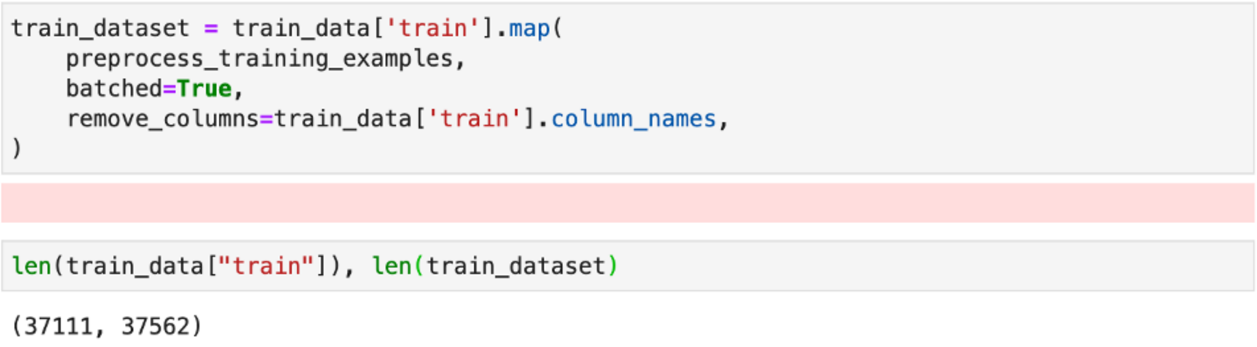
```
[CLS] Which NFL team represented the AFC at Super Bowl 50? [SEP] architecturally the scho
ol has a catholic character. atop the main building school dome is the golden statue of t
he virgin mary. immediately in front of the main building in facing it is @ @ @ @ @ a cop
per statue of christ @ @ @ @ @ with arms appraised with the legend and the bad meow name
s. next to the main building is the basilica of the sacred heart. immediately behind [SE
P]
[CLS] Which NFL team represented the AFC at Super Bowl 50? [SEP] it is @ @ @ @ @ a copper
statue of christ @ @ @ @ @ with arms appraised with the legend and the bad meow names. ne
xt to the main building is the basilica of the sacred heart. immediately behind the basil
ica is the grotto im mary in place of prayer and reflection. it is a replica of the grott
o at lourdes france where the [SEP]
[CLS] Which NFL team represented the AFC at Super Bowl 50? [SEP] main building is the bas
ilica of the sacred heart. immediately behind the basilica is the grotto im mary in place
of prayer and reflection. it is a replica of the grotto at lourdes france where the virgi
n mary reputedly appeared to st bernadette still burning eighteen fifty eight. at the end
of the main drive and in a direct line that connects through three statues [SEP]
[CLS] Which NFL team represented the AFC at Super Bowl 50? [SEP] replica of the grotto at
lourdes france where the virgin mary reputedly appeared to st bernadette still burning ei
```

Problem 1-3: What if answer is near the boundary of windows or across windows?

Solution: We can set a sliding window between them which is included in above figure.

Problem 1-4: How to prevent model from learning something it should not learn during training?

Solution: Some examples in this dataset have extra spaces at the beginning and the end that don't add anything, so we could removed those extra spaces. To apply this function to the whole training set, we use the Dataset.map() method with the batched=True flag to process the training and testing datasets which are included in my code and a screenshot is shown below.



```
train_dataset = train_data['train'].map(
    preprocess_training_examples,
    batched=True,
    remove_columns=train_data['train'].column_names,
)
```

```
len(train_data["train"]), len(train_dataset)
```

```
(37111, 37562)
```

Problem 1-5: In post-processing, implement automatic mixed precision (setting fp16 = True), linear decay learning rate in the training process

Solution: All done. The details can be seen from the code on github.

Problem 1-6: Try to use other pre-trained models except for the base one.

Solution:

- Pre-trained model: bert-base-cased: F1 score = 51.04
- Pre-trained model: bert-base-uncased: F1 score = 59.21
- Pre-trained model: hfl/chinese-electra-180g-small-discriminator: F1 score = 39.20
- Pre-trained model: hfl/chinese-electra-180g-base-discriminator: F1 score = 38.54
- Pre-trained model: hfl/chinese-bert-wwm-ext: F1 score = 38.9
- Pre-trained model: bert-base-chinese: F1 score = 29.42