Q1: Data Processing

1.Tokenizer:
使用transformers內建的BertTokenizerFast.from_pretrained，作法是在paragraph的開頭tag [CLS]以及在中間和結尾tag [SEP]，而truncation為only_second，句子長度自動補滿到512，mask則是採用把整個字詞cover掉的方式，即任一paragraph內若有某個字詞分割出來的token被cover，則從該字詞分出來的其他tokens也會同樣被cover掉。

2.Answer span:

- 設置return_offset_mapping=True, return_overflow_token=True來分別取得token、slice和原文章位置的映射關係，先找到start&end position後，若answer的起始和結尾被paragraph包含在內，代表為effective position，回傳對應的位置，否則就回傳一對[CLS]的index
- 在最後traverse test datasets的過程中，設置topk=3來統計3*3組中最高的top logits，取得best index composition，而當start position > end position時，因為不符合條件限制即自動忽略current loop。

Q2: Modeling with Berts and their variants

1.Original use of pretrained model
Models: Mengzi-bert for multiple choice，Roberta-wwm-ext for question answering

Parameters:

- Max length = 512
- Batch size = 32
- Real batch size = 2
- Learning rate = 0.00004
- Scheduler = ReduceLROnPlateau
- Loss function = CrossEntropyLoss
- Optimizer = AdamW
- Epoch = 10

Performance(validation):
Multiple choice: 0.973, Question answering: 0.798

2.Another type of pretrained model
Models: Mengzi-bert for multiple choice, Chinese-macbert for question answering

Parameters:

- Max length = 512
- Batch size = 32
- Real batch size = 2
- Learning rate = 0.00004
- Scheduler = ReduceLROnPlateau
- Loss function = CrossEntropyLoss
- Optimizer = AdamW
- Epoch = 10

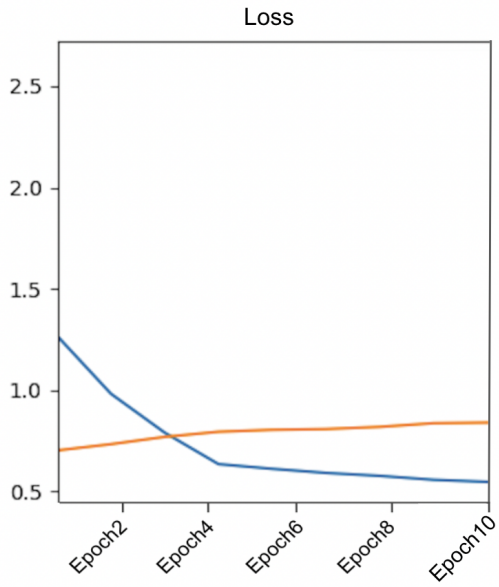
Performance(validation):
Multiple choice: 0.973, Question answering: 0.804,
Score on Kaggle: 0.783(public)

Difference between pretrained models:
Macbert: 使用static masking，這代表一個句子相同的部分在每個epoch都會被cover掉，在Macbert中15%的token會被隨機挑選並且mask，採用連續mask n個詞的方式且1gram~4gram的機率依序為40%,30%,20%,10%。
Roberta: 使用dynamic masking，在每個epoch都挑選句子中不同的part去mask，而sentence order prediction的能力也讓模型隨機挑選context中連續的兩個paragraph，並透過交換句子順序來強壯模型的後續預測能力。

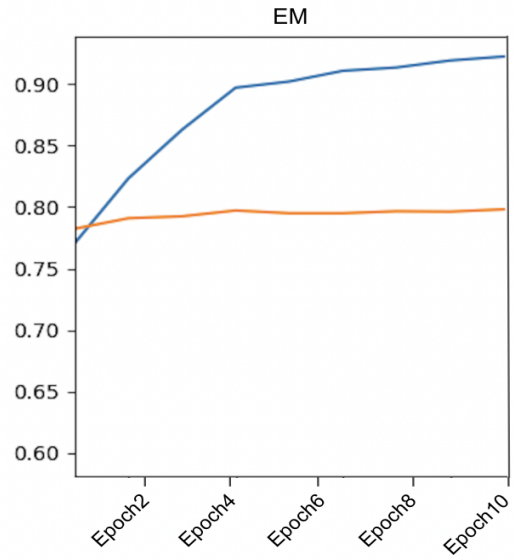
Q3: Curves

Learning curves of loss and exact match:

Loss(blue for train , orange for validate):



Exact match(blue for train , orange for validate):



Q4: Pretrained vs not pretrained

Dataset: question answering

Configuration: AutoConfig's macbert-base

Model: BertForQuestionAnswering(no pretrained)

Parameters: Epoch=5, others are the same as Q2

Performance: 在not pretrained model的訓練過程中，training loss下降的非常慢，大概train到第五個epoch的時候loss才從9.4多降到8.5左右，而validation loss則是沒什麼較大的變動，一直維持在4.5上下，至於exact match的部分，training是從第一個epoch為0.003開始、validation則為0.004，可以很明顯感覺到模型train不太起來，上升的很慢，到了最後第五個epoch時都差不多升到0.016左右。

Q5: HW1 with Berts