ADL HW1

R13922022 王凱右

Q1-1: Tokenizer

Tokenizer 首先會基於規則和統計將輸入文本分割成單詞。然後在句子內加上五個特殊的 token 來標示句子的開始、結束、填充與分割等位置，最後再把分詞編碼成單詞表。

不同模型也會針對 Tokenizer 進行優化，比如說 chinese-bert-wwm 就改進了 bert-base-chinese 的字分割方式，分詞時不是以字而是以詞作為分割。此外使用了 Whole Word Masking (wwm)技術，在訓練時會將整個詞進行 Masking，以求強化模型的訓練效果。

ex: 十五世紀歐洲認為天動說是世間唯一的真理
字分割: 十 五 世 紀 歐 洲 認 為 天 動 說 是 世 間 唯 一 的 真 理
詞分割: 十五 世紀 歐洲 認為 天動說 是 世間 唯一的 真理
wwm: 十五 世紀 [mask][mask] 認為 天動說 是 世間 唯一的 [mask][maks]

Q1-2: Answer Span(tokenization)

我們可以使用 offset_mapping 內的 token id 找到 answer span 的 start_positions 和 end_positions，如下圖:

```
# help us compute the start_positions and end_positions.
offset_mapping = tokenized_examples.pop("offset_mapping")
```

如果對應回去 answer span 時發現找不到答案或在原本內容的範圍外，就把該樣本的答案標示在 CLS token 位置。如果在內容範圍內找到答案，就更新答案位置。

Q1-2: Answer Span(probability of position)

首先找到最好的 n 個 start_logits 和 end_logits，取出其對應的開始位置與結束位置。

然後從這些點中排除不合理的答案，比如長度太長或是超過文本最大長度等狀況。再來把這些剩下合理的可能答案的 start_logits 與 end_logits 相加，作為他們的分數。最後就是找到分數最高的那一個作為我們的答案。

Q2-1:

Base Model: BERT (pretrained in bert-base-chinese)
Loss Function: CrossEntropy Loss
Optimizer: AdamW
Learning Rate: 3e-5
Scheduler: Linear
Epochs: 1(MC)/ 2(QA)
Total_train_batch_size: 2
Max Sequence Length: 512

Q2-2

我將幾種 BERT 變體包含 Base Model 的訓練結果用下表呈現。
如果沒有特別標註，代表超參數維持相同設定，只有改模型架構。

Model	Val Score (MC:Accuracy/ QA:EM score)	Kaggle Score (Public/Private)
bert-base-chinese	0.9593 / 82.85	0.7461 / 0.7607
hfl/chinese-bert-wwm	0.9551 / 82.85	0.7497 / 0.7607
hfl/chinese-roberta-wwm-ext	0.9601 / 83.28	0.7825 / 0.7909
hfl/chinese-roberta-wwm-ext (QA train 5 epoch)	0.9601 / 83.62	0.7766 / 0.7857
hfl/chinese-lert-base	**0.9624** / 85.31	0.7813 / 0.7946
hfl/chinese-lert-base (QA train 4 epoch)	**0.9624 / 85.74**	**0.7965 / 0.7953**

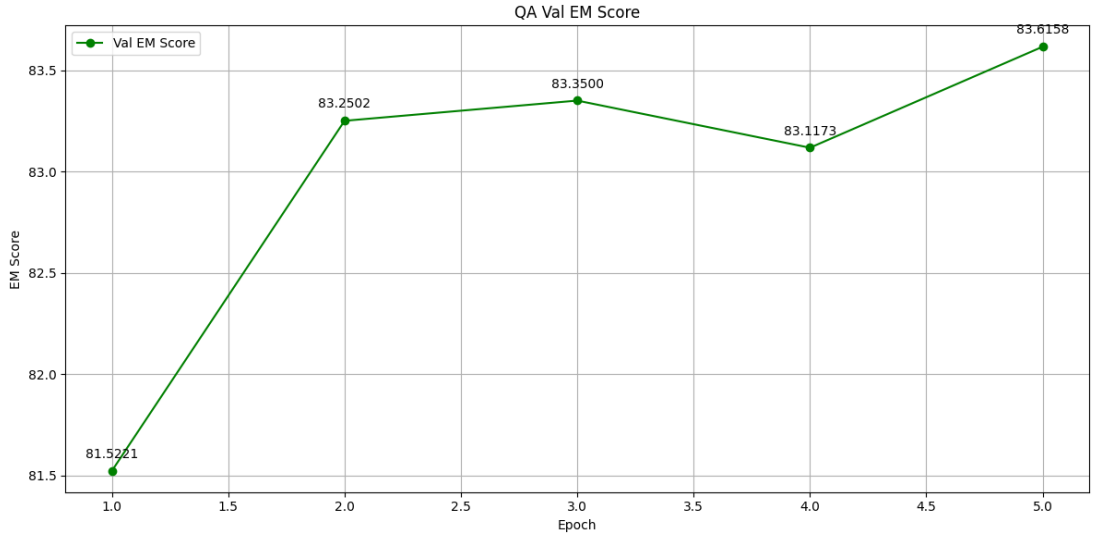
(註:這裡都是以 mc/qa 使用同模型測試之結果。)

Best Model:

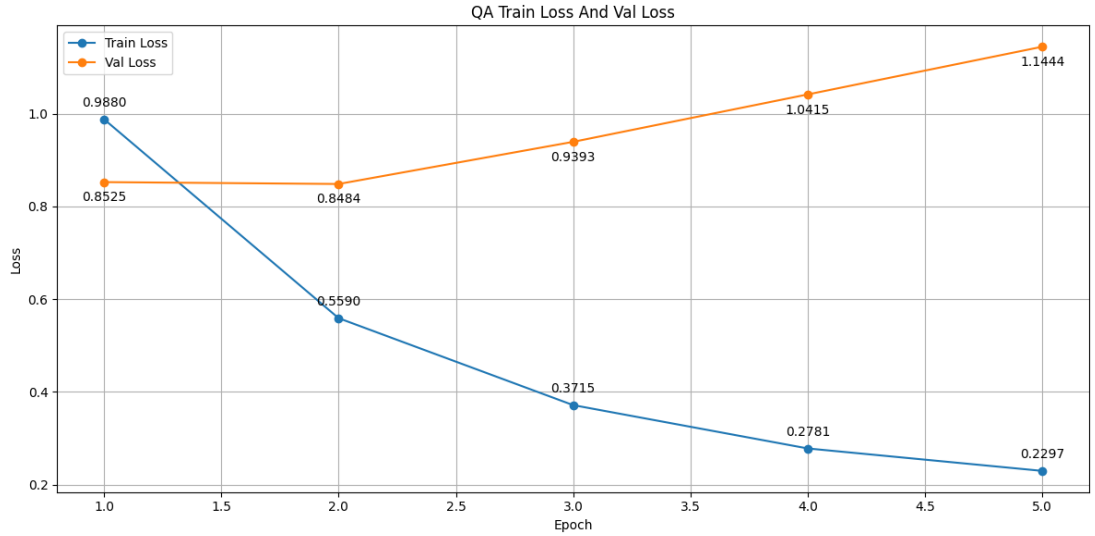
hfl/chinese-lert-base(QA train 4 epoch)

Q3:以下提供 hfl/chinese-roberta-wwm-ext QA Model 的訓練結果

Val EM Score:

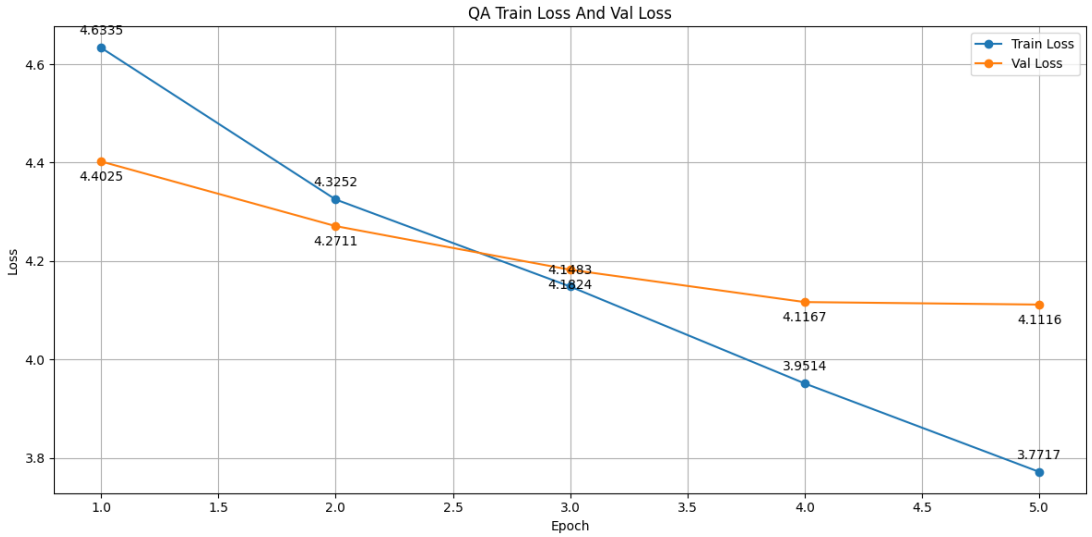


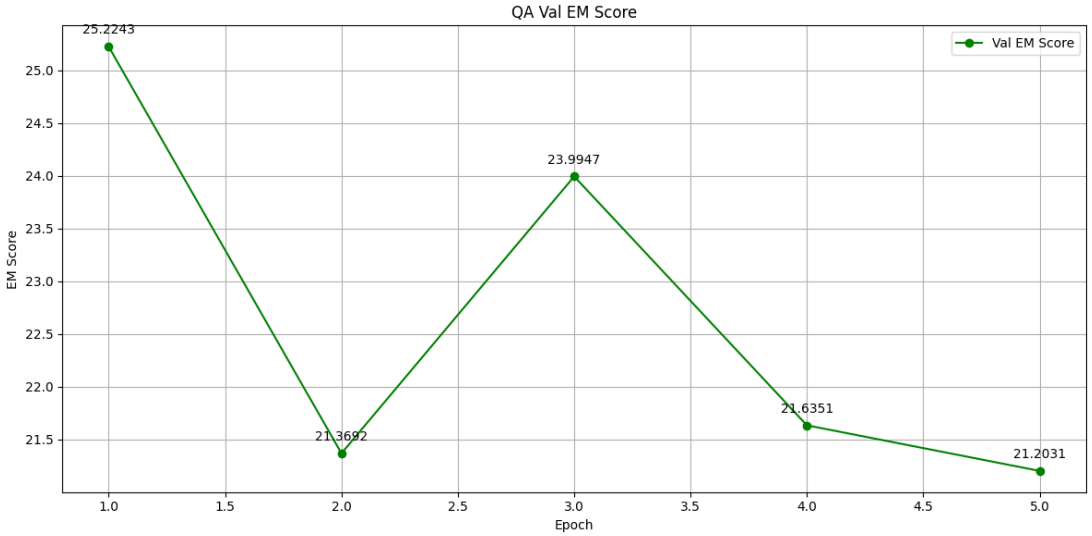
Train Loss & Val Loss:



可以看到其實大概 5 epoch 左右成果還有上升的趨勢。但實際上繳交到 kaggle 上 5 epoch 反而出現 overfitting 的狀況，效果比 4 epoch 的 model 表現得還更差。因此驗證集的效果也並不代表真實的測試效果。

Q4:以下提供 bert-base-chinese（scratch）QA Model 的訓練結果





Model	Loss / EM score
bert-base-chinese(pretrained)	0.9593 / 82.85
bert-base-chinese(scratch)	4.1116 / 21.20

可以看到雖然在訓練的 5 epoch 中 train loss 不斷下降，但是 model 還是訓練不起來，驗證集的 EM score 反而越來越低。結論就是這個資料集的資料量對於 BERT 要訓練還是太少了，如果想從頭訓練的話可能要準備更多資料或是改成其他模型。

參考文獻:

冬于的博客-Transformer/BERT/实战

LERT_GitHub

Chinese-BERT-wwm_GitHub

進擊的 BERT：NLP 界的巨人之力與遷移學習

HuggingFace_MultipleChoice

Claude

ChatGPT