ADL Hw2 report

R10723059 財金碩二 胡祖望

November 10, 2022

1 Q1: Data processing

1.1 Tokenizer

a. 使用 wordpiece 的 Tokenizer，此 Tokenizer 的原理類似於 BPE Tokenizer，把字分開再加 prefix，不同的點是 wordpiece 在挑選配對的字詞是用 score 來做挑選而不是出現頻率，還有 WordPiece 只存最後的 vocabulary, 而不是 merge rules 學到的.

score=(freq of pair)/(freq of first element×freq of second element)
score 高代表這對組合很常出現，要把他合在一起

做法如下:
1. 設定一個可容許的 vocabulary 大小
2. 將詞切成一個一個字元 (包含結束符號)，存成字元資料庫
3. 將資料中的字元皆放入 vocabulary 資料庫中
4. 選擇 score 最高的 pair 放入 vocabulary 資料庫，🄵將字元資料庫中的 pair 合🄵
5. 反覆做 4 直到 vocabulary 資料庫的大小到達設定標准
6. 將 vocabulary 資料庫中配對好的字元透過 language model pretraining 得到 token

1.2 Answer Span

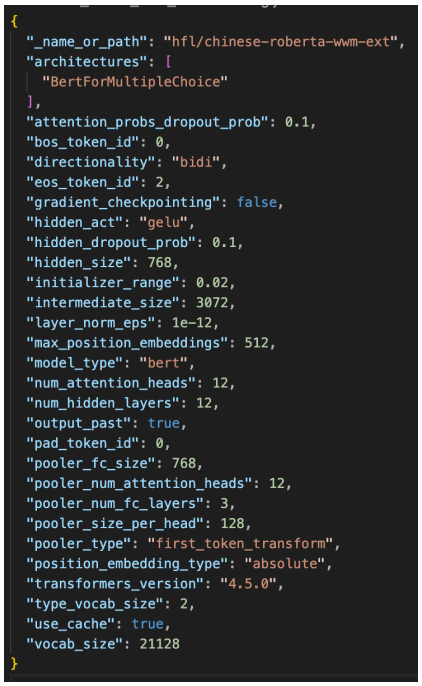
a. 使用 BatchEncoding 中的 char to token() 函式，可以將原本使用一個一個字元分割的位置編號改成輸入經過 Tokenization 後對應的位置編號

b. 把每對估計出的 start/end 的機率相乘，保留機率最大的，但要如果有 start 位置大於 end 位置情形就要跳過。最後估計出的答案代回 input ids 再套用 tokenizer.decode() 把他轉回來得到預測目標

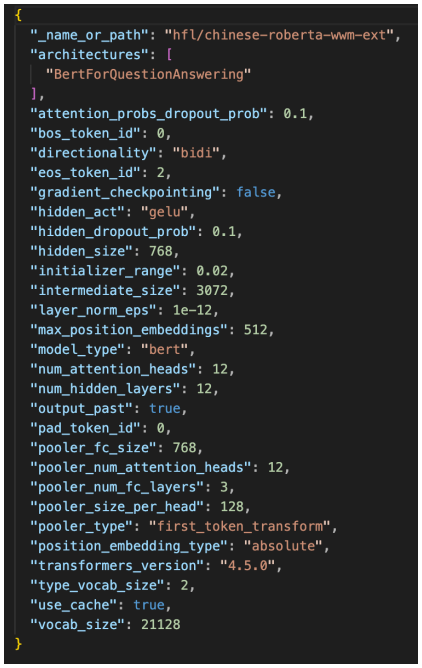
2 Q2: Modeling with BERTs and their variants

2.1 Describe

a. my model:
”hfl/chinese-roberta-wwm-ext”
Multiple Choice - Config:



QA - Config:



b. performance of model, public score：0.73327

c. Loss function: Cross entropy loss.

d. Training argument

Multiple choice:

optimization algorithm: AdamW

accumulation steps = 5

learning rate: 5e−5(使用 lr scheduler.StepLR， step size=100,gamma=0.95)

batch size:16

Question answering:

optimization algorithm: AdamW

accumulation steps = 5

learning rate: 5e−5(使用 lr scheduler.StepLR， step size=100,gamma=0.95)

batch size: 16

2.2 Try another type of pretrained model and describe

a. my model:
”luhua/chinese pretrain mrc roberta wwm ext large”
Multiple Choice - Config:


```
{
  "_name_or_path": "luhua/chinese_pretrain_mrc_roberta_wwm_ext_large",
  "architectures": [
    "BertForMultipleChoice"
  ],
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 1024,
  "initializer_range": 0.02,
  "intermediate_size": 4096,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 16,
  "num_hidden_layers": 24,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "transformers_version": "4.5.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

QA - Config:

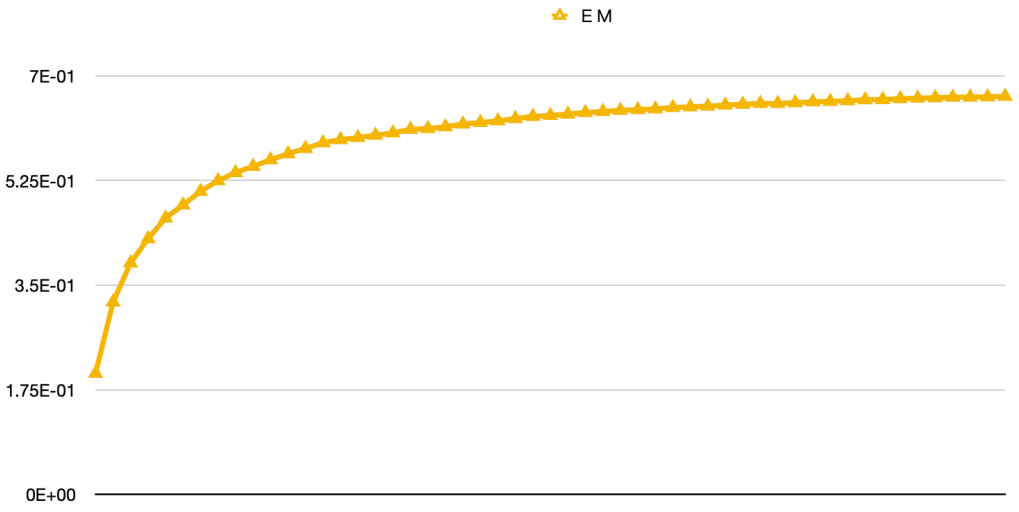
b. performance of model, public score：0.73327

c. luhua 是將 roberta wwm ext large 再作進一步的改良，context>1024 的舍去、question>64 的舍去
還有加强模型對樣本變🄴能力像是
對於每個問題，随機從數據中取 context，🄴保留 title 作为負樣本; 對於每個問題，將其正樣本中答案出现的句子删除，以此作为負樣本

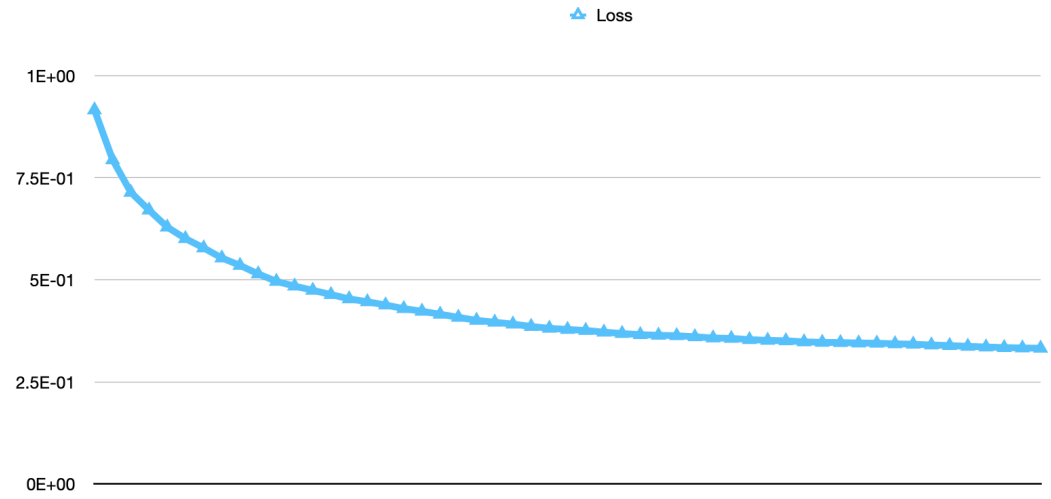
d. roberta-wwm-ext –> luhua/chinese pretrain mrc roberta wwm ext large

3 Q3: Curves

a.Learning curve of loss
every 100 steps Loss

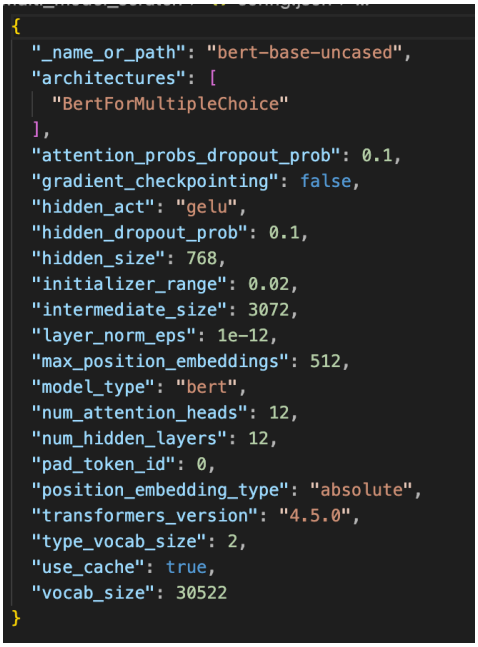


b.Learning curve of EM
every 100 steps EM



4 Q4: Pretrained vs Not Pretrained

Multiple Choice Config:



QAConfig:

先使用 init weight() 函式去除原本的 weight 由於資料量相比 pretrained model 少上許多，認F調整方向可調整模型的 configuration，可調整項目包括 hidden layer, hidden size

performance: public score：0.01807

相比 pretrained 好的 BERT 模型，訓練使用的資料量非常少，訓練時間也非常少，可看出在此方法下很難訓練出一個很好的模型，由此可得知訓練模型准確率到可實際使用的程度，需要更大的資料量與時間。