Applied Deep Learning, Spring 2022 - Homework 2

郭柏志 R09521205

Q1: Data processing

1. Tokenizer:

Tokenize 的目標是把輸入的句子切成一個個 token，每個 token 有完整的語意，一方面可以將沒看過的字詞切成多個 vocabulary 中存在的 token，一方面方便學習 embedding，hugging-face 提供的 Tokenizer 會將 token encode 成 ids，再透過 embedding layer 將 token 轉為帶有詞語意義的詞向量給模型學習。我使用的 pretrained model 為 bert-base-chinese，bert 的 tokenize algorithm 用的是 WordPiece，與 Byte-Pair Encoding (BPE)類似，會先標準化處理字串，如 unicode 轉換、切割標點符號等，再切割成數個 subword 建立基礎詞彙表，中文句子的 subword 單位為單一中文字，因此基礎詞彙表包含了所有中文字以及非中文字元，接著根據基礎詞彙表，找出能夠最大化訓練集似然度的兩個基礎 A,B 詞彙合併，使得合併的 A, B 有最大的$\frac{P(AB)}{P(A)P(B)}$ 。

2. Answer Span:

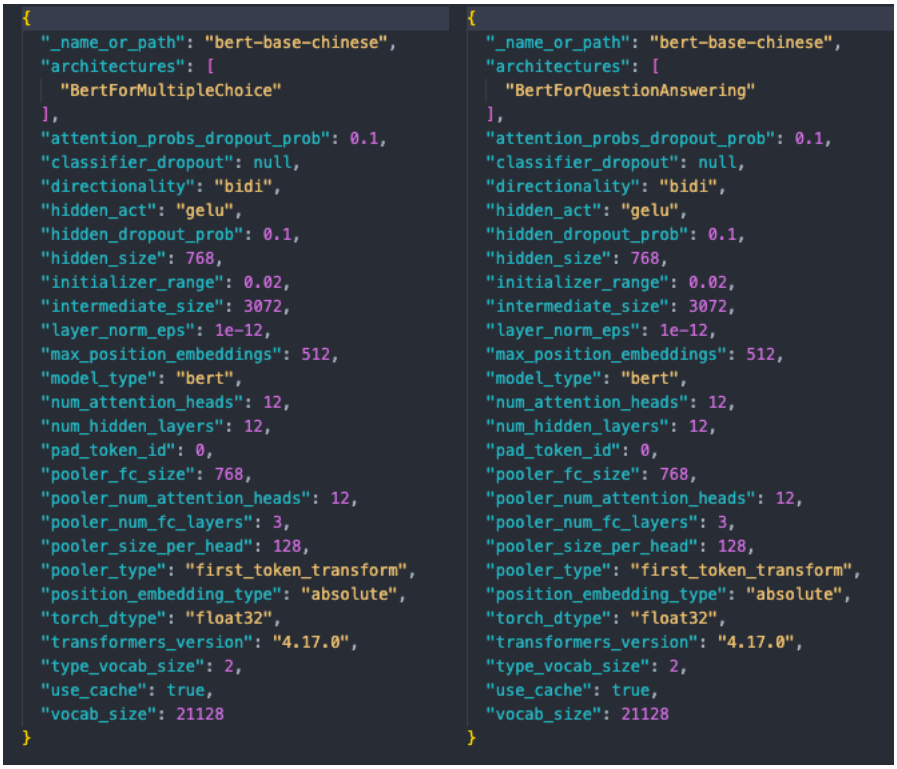
(a) hugging-face 提供的 Tokenizer 若設定 return_offsets_mapping 為 True Tokenized data 會包含每個 token 對應 question 或 context 的(char start position, char end position)，逐個 token 檢查其 char start position, char end position 便能知道哪個 answer span 的起點與終點對應到哪兩個 token。

(b) 一個 QA pair 可能因為字數過長被切成多段，預測時將所有分段的 data 的 start_logit, end_logit 存起來，同一個 QA pair 只留下 start_logit + end_logit 最高的前 n_best_size 組，剔除 start_position 大於 end_position 的組合，挑選分數最高者做為答案。

Q2: Modeling with BERTs and their variants

1.

a. Model: bert-base-chinese




```
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForMultipleChoice"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "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",
  "torch_dtype": "float32",
  "transformers_version": "4.17.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

```
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForQuestionAnswering"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "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",
  "torch_dtype": "float32",
  "transformers_version": "4.17.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```


b. Performance

Context selection accuracy: 0.973
Question Answering EM (kaggle): 0.744

c. Loss function

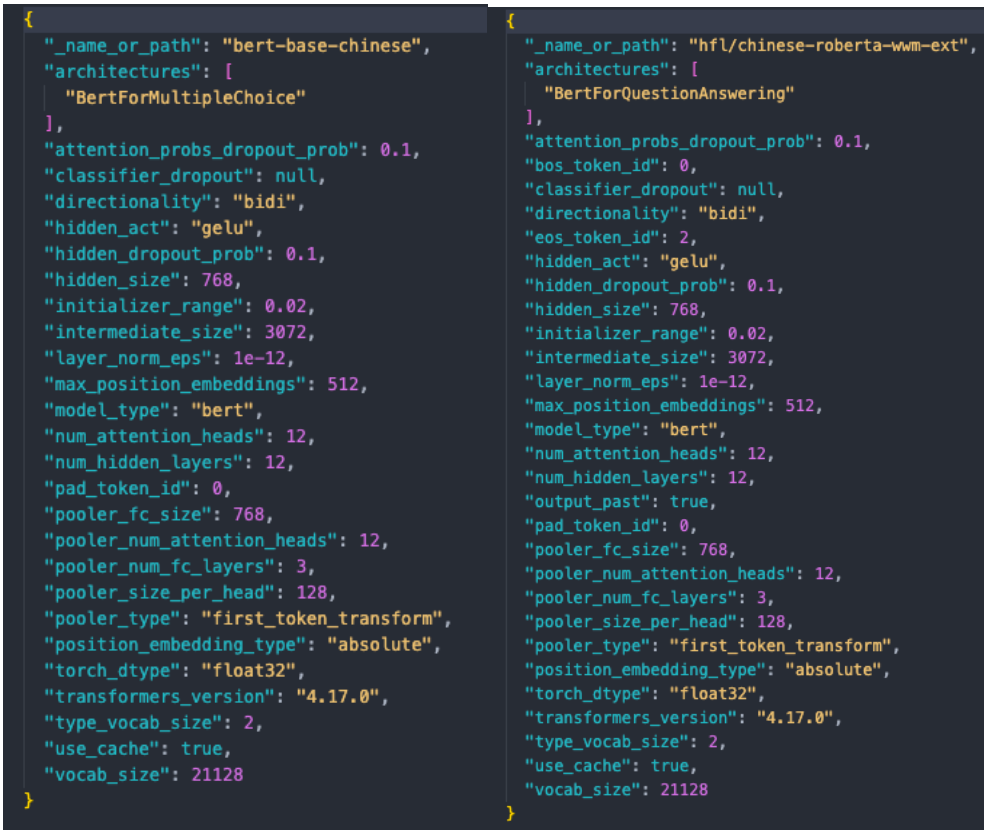
Cross entropy loss

d. Optimization algorithm, Learning rate, Batch size

AdamW with leaning rate = 0.00003,
per_device_train_batch_size = 4,
gradient_accumulation_steps = 4

2.

a. Model: hfl/chinese-roberta-wwm-ext




```
{
  "_name_or_path": "bert-base-chinese",
  "architectures": [
    "BertForMultipleChoice"
  ],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "directionality": "bidi",
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "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",
  "torch_dtype": "float32",
  "transformers_version": "4.17.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```

```
{
  "_name_or_path": "hfl/chinese-roberta-wwm-ext",
  "architectures": [
    "BertForQuestionAnswering"
  ],
  "attention_probs_dropout_prob": 0.1,
  "bos_token_id": 0,
  "classifier_dropout": null,
  "directionality": "bidi",
  "eos_token_id": 2,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "output_past": true,
  "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",
  "torch_dtype": "float32",
  "transformers_version": "4.17.0",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 21128
}
```


b. Performance

Context selection accuracy: 0.973
Question Answering EM (kaggle): 0.761

c. Loss function

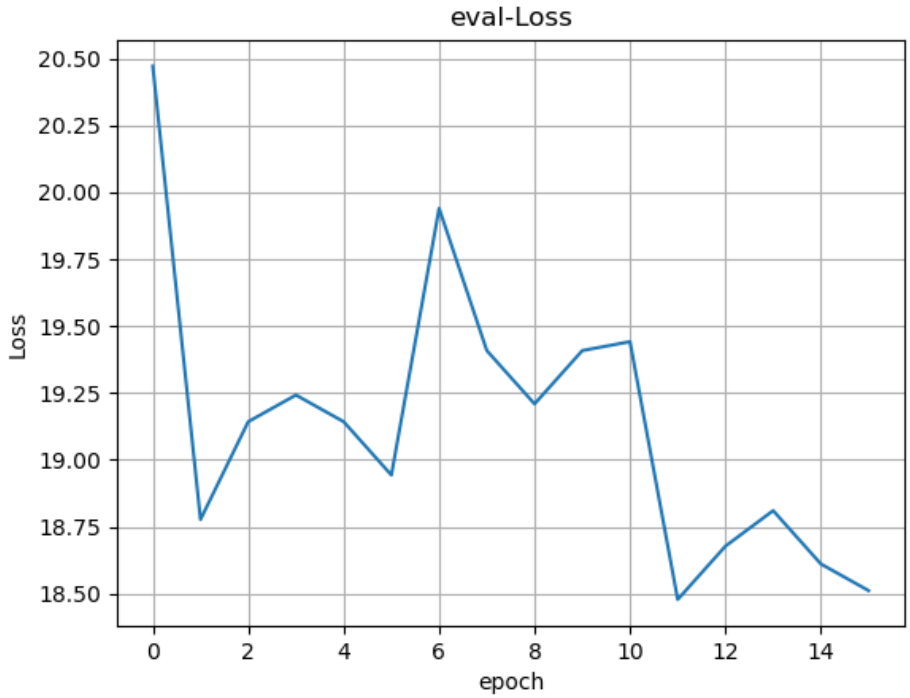
Cross entropy loss

d. Optimization algorithm, Learning rate, Batch size

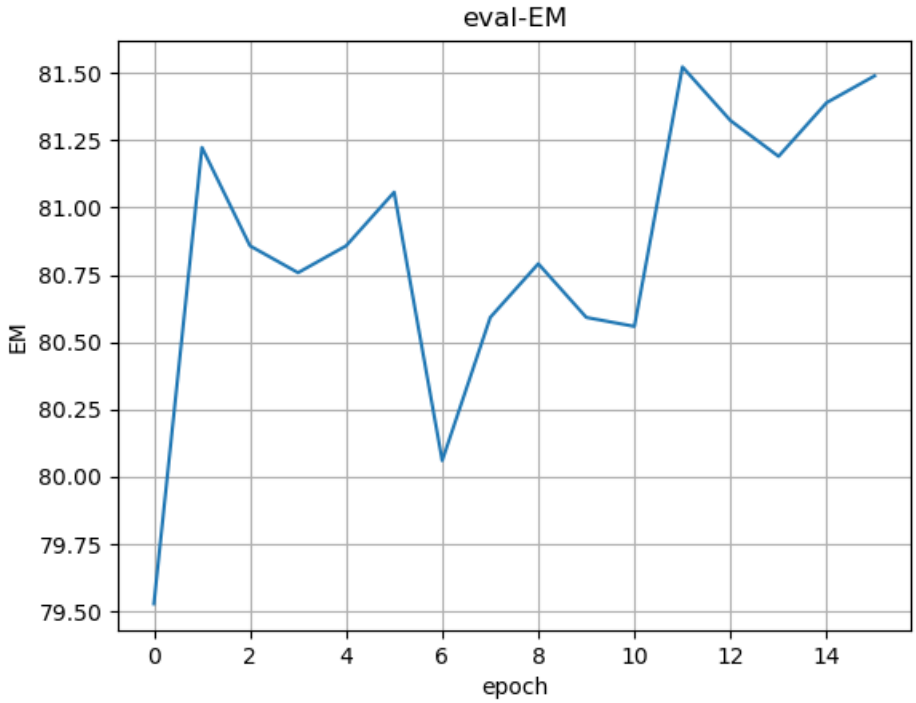
AdamW with leaning rate = 0.00003,
per_device_train_batch_size = 4,
gradient_accumulation_steps = 4

Q3: Curves

a. Learning curve of loss



b. Learning curve of EM



Q4: Pretrained vs Not Pretrained

減少 hidden_dim: 784 → 128，attention_head: 12 → 8，batch_size： 4 → 1，只從新訓練 QA 任務，local 端得到 EM = 0.066, 由於 Trainsformer 架構龐大，需要大量訓練資料與時間才能訓練得起來，此任務的訓練資料與時間都與 pretrained 相差甚遠，因此若要從 0 開始訓練 bert model 會非常困難。