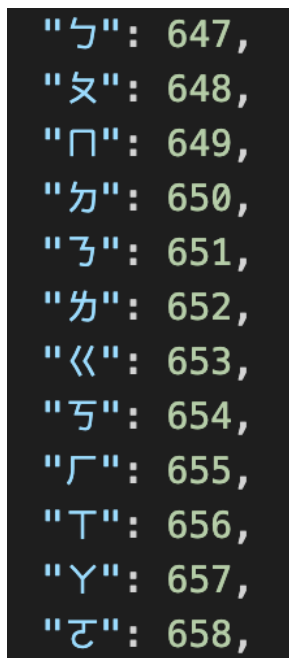


ADL HW2

Question 1. Tokenizer

- 1) Describe in detail about the tokenization algorithm you use. You need to explain what it does in your own ways.

我採用 Hugging Face 內建的 wordpiece 分詞器，只不過在 bert 裡和我們理解的中文分詞不太一樣，中文字是以”字”為基本單位，不同於英文以 subword 為基本單位。在 wordpiece 演算法方面，wordpiece 會先初始化一個包括所有在訓練集出現出現過的字的 vocabulary，並且迭代訓練。不同於 BPE 演算法的地方在於，wordpiece 並不是選擇出現頻率最高的 symbol pair，而是計算加入字典時的 maximize likelihood。



"亅":	647,
"攴":	648,
"冂":	649,
"勹":	650,
"勹":	651,
"勹":	652,
"勹":	653,
"勹":	654,
"勹":	655,
"勹":	656,
"勹":	657,
"勹":	658,

FIGURE 1. Tokenizer in bert-based-chinese.

- 2) After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position?

如果在 predict 的結果中，如果有包含特殊符號的話，就要 map 回原本的文本，此外，我們在預測時會去找在 start position 後的 end position 來確保文本是有效的。

Question 2. Modeling with BERTs and their variants.

使用 hfl/chinese-roberta-wwm-ext 做爲 pretrained model

- 1) describe your model (configuration of the transformer model)

The figure displays two side-by-side JSON configuration files for transformer models. The left configuration is for a context selection model, and the right is for a question answering model. Both configurations are based on the hfl/chinese-roberta-wwm-ext pretrained model.

```

1 {
2   "_name_or_path": "bert-base-chinese",
3   "architectures": [
4     "BertForMultipleChoice"
5   ],
6   "attention_probs_dropout_prob": 0.1,
7   "classifier_dropout": null,
8   "directionality": "bidi",
9   "hidden_act": "gelu",
10  "hidden_dropout_prob": 0.1,
11  "hidden_size": 768,
12  "initializer_range": 0.02,
13  "intermediate_size": 3072,
14  "layer_norm_eps": 1e-12,
15  "max_position_embeddings": 512,
16  "model_type": "bert",
17  "num_attention_heads": 12,
18  "num_hidden_layers": 12,
19  "pad_token_id": 0,
20  "pooler_fc_size": 768,
21  "pooler_num_attention_heads": 12,
22  "pooler_num_fc_layers": 3,
23  "pooler_size_per_head": 128,
24  "pooler_type": "first_token_transform",
25  "position_embedding_type": "absolute",
26  "torch_dtype": "float32",
27  "transformers_version": "4.24.0",
28  "type_vocab_size": 2,
29  "use_cache": true,
30  "vocab_size": 21128
31 }

```

```

1 {
2   "_name_or_path": "question-answering",
3   "architectures": [
4     "BertForQuestionAnswering"
5   ],
6   "attention_probs_dropout_prob": 0.1,
7   "classifier_dropout": null,
8   "directionality": "bidi",
9   "hidden_act": "gelu",
10  "hidden_dropout_prob": 0.1,
11  "hidden_size": 768,
12  "initializer_range": 0.02,
13  "intermediate_size": 3072,
14  "layer_norm_eps": 1e-12,
15  "max_position_embeddings": 512,
16  "model_type": "bert",
17  "num_attention_heads": 12,
18  "num_hidden_layers": 12,
19  "pad_token_id": 0,
20  "pooler_fc_size": 768,
21  "pooler_num_attention_heads": 12,
22  "pooler_num_fc_layers": 3,
23  "pooler_size_per_head": 128,
24  "pooler_type": "first_token_transform",
25  "position_embedding_type": "absolute",
26  "torch_dtype": "float32",
27  "transformers_version": "4.24.0",
28  "type_vocab_size": 2,
29  "use_cache": true,
30  "vocab_size": 21128
31 }

```

FIGURE 2. context selection model and question answering model(hfl/chinese-roberta-wwm-ext)

- 2) describe performance of your model.

在 context selection 的任務中，經過了 3 個 epochs 的訓練後模型 validation set 的 accuracy 可以達到 0.961。

在 question answering 的任務中，經過了 40 個 epochs 的訓練後模型 validation set 的 exact_match 可以達到 0.82。

- 3) describe the loss function you used.

CrossEntropyLoss

- 4) The optimization algorithm (e.g. Adam), learning rate and batch size.

在 context selection 的任務中，AdamW(learning rate=3e-5); batch size=8; gradient_accumulation_step=1

在 question answering 的任務中，AdamW(learning rate=3e-5); batch size=8; gradient_accumulation_step=1

使用 bert-base-chinese 做爲 pretrained model

- 1) describe your model (configuration of the transformer model)

see FIGURE 3.

- 2) describe performance of your model.

在 context selection 的任務中，經過了 3 個 epochs 的訓練後模型 validation set 的 accuracy 可以達到 0.961。

在 question answering 的任務中，經過了 40 個 epochs 的訓練後模型 validation set 的 exact_match 可以達到 0.804。

- 3) The difference between pretrained model.

roberta model 與 bert-based model 相比，訓練時間更長，batch size 更大，training data 更多，訓練的序列更長，此外還加入了動態的 masking 的機制，因此有更好的 performance.

其實上在 question answering 的任務上我們也可以從 exact_match 看到 roberta model 比 bert-based model 有著更好的表現。

```

1 {
2   "_name_or_path": "bert-base-chinese",
3   "architectures": [
4     "BertForMultipleChoice"
5   ],
6   "attention_probs_dropout_prob": 0.1,
7   "classifier_dropout": null,
8   "directionality": "bidi",
9   "hidden_act": "gelu",
10  "hidden_dropout_prob": 0.1,
11  "hidden_size": 768,
12  "initializer_range": 0.02,
13  "intermediate_size": 3072,
14  "layer_norm_eps": 1e-12,
15  "max_position_embeddings": 512,
16  "model_type": "bert",
17  "num_attention_heads": 12,
18  "num_hidden_layers": 12,
19  "pad_token_id": 0,
20  "pooler_fc_size": 768,
21  "pooler_num_attention_heads": 12,
22  "pooler_num_fc_layers": 3,
23  "pooler_size_per_head": 128,
24  "pooler_type": "first_token_transform",
25  "position_embedding_type": "absolute",
26  "torch_dtype": "float32",
27  "transformers_version": "4.24.0",
28  "type_vocab_size": 2,
29  "use_cache": true,
30  "vocab_size": 21128
31 }

```

```

1 {
2   "_name_or_path": "question-answering",
3   "architectures": [
4     "BertForQuestionAnswering"
5   ],
6   "attention_probs_dropout_prob": 0.1,
7   "classifier_dropout": null,
8   "directionality": "bidi",
9   "hidden_act": "gelu",
10  "hidden_dropout_prob": 0.1,
11  "hidden_size": 768,
12  "initializer_range": 0.02,
13  "intermediate_size": 3072,
14  "layer_norm_eps": 1e-12,
15  "max_position_embeddings": 512,
16  "model_type": "bert",
17  "num_attention_heads": 12,
18  "num_hidden_layers": 12,
19  "pad_token_id": 0,
20  "pooler_fc_size": 768,
21  "pooler_num_attention_heads": 12,
22  "pooler_num_fc_layers": 3,
23  "pooler_size_per_head": 128,
24  "pooler_type": "first_token_transform",
25  "position_embedding_type": "absolute",
26  "torch_dtype": "float32",
27  "transformers_version": "4.24.0",
28  "type_vocab_size": 2,
29  "use_cache": true,
30  "vocab_size": 21128
31 }

```

FIGURE 3. context selection model and question answering model(bert-base-chinese)

此外，我也嘗試了 hfl/chinese-roberta-wwm-ext-large，本來預期可能會有更好的表現，即便用 v100-32g 訓練了 2 個 epoch 之後 accuracy 還是只有 0.26 左右，我猜測可能是 effective batch size 不夠大並且 epoch 太少的關係，我想未來有機會了話用分散式系統可能可以達到比較好的效果。

Question 3. Curves Plot the learning curve of your QA model

以 hfl/chinese-roberta-wwm-ext 為 pretrained model 進行 question answering 任務並訓練 40 個 epochs.

1) Learning curve of loss(total loss)

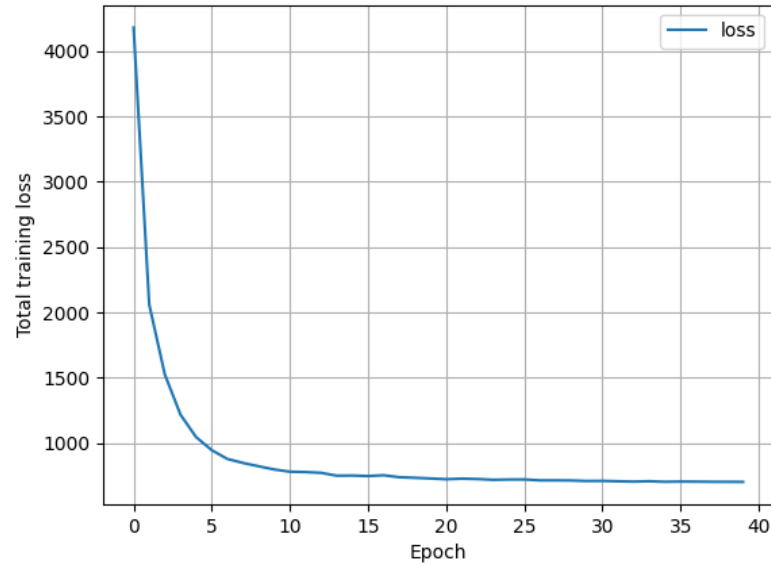


FIGURE 4. Total training loss curve.

2) Learning curve of EM

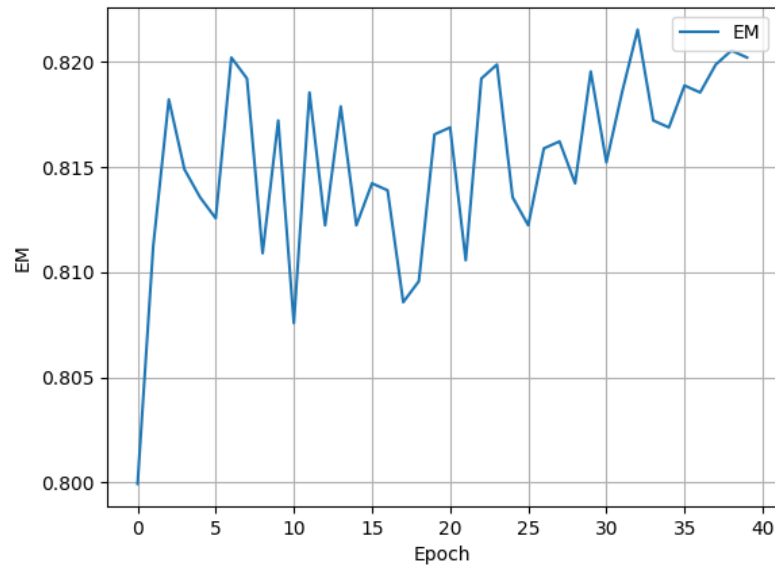


FIGURE 5. EM curve.

Question 4. Pretrained vs Not Pretrained Plot the learning curve of your QA model
我以 bert-base-chinese 的架構並以是否導入 pretrained weight 與否進行模型表現的比較。

1) The configuration of the model

<pre> 1 { 2 "_name_or_path": "bert-base-chinese", 3 "architectures": [4 "BertForMultipleChoice" 5], 6 "attention_probs_dropout_prob": 0.1, 7 "classifier_dropout": null, 8 "directionality": "bidi", 9 "hidden_act": "gelu", 10 "hidden_dropout_prob": 0.1, 11 "hidden_size": 768, 12 "initializer_range": 0.02, 13 "intermediate_size": 3072, 14 "layer_norm_eps": 1e-12, 15 "max_position_embeddings": 512, 16 "model_type": "bert", 17 "num_attention_heads": 12, 18 "num_hidden_layers": 12, 19 "pad_token_id": 0, 20 "pooler_fc_size": 768, 21 "pooler_num_attention_heads": 12, 22 "pooler_num_fc_layers": 3, 23 "pooler_size_per_head": 128, 24 "pooler_type": "first_token_transform", 25 "position_embedding_type": "absolute", 26 "torch_dtype": "float32", 27 "transformers_version": "4.24.0", 28 "type_vocab_size": 2, 29 "use_cache": true, 30 "vocab_size": 21128 31 } </pre>	<pre> 1 { 2 "_name_or_path": "question-answering", 3 "architectures": [4 "BertForQuestionAnswering" 5], 6 "attention_probs_dropout_prob": 0.1, 7 "classifier_dropout": null, 8 "directionality": "bidi", 9 "hidden_act": "gelu", 10 "hidden_dropout_prob": 0.1, 11 "hidden_size": 768, 12 "initializer_range": 0.02, 13 "intermediate_size": 3072, 14 "layer_norm_eps": 1e-12, 15 "max_position_embeddings": 512, 16 "model_type": "bert", 17 "num_attention_heads": 12, 18 "num_hidden_layers": 12, 19 "pad_token_id": 0, 20 "pooler_fc_size": 768, 21 "pooler_num_attention_heads": 12, 22 "pooler_num_fc_layers": 3, 23 "pooler_size_per_head": 128, 24 "pooler_type": "first_token_transform", 25 "position_embedding_type": "absolute", 26 "torch_dtype": "float32", 27 "transformers_version": "4.24.0", 28 "type_vocab_size": 2, 29 "use_cache": true, 30 "vocab_size": 21128 31 } </pre>
---	---

FIGURE 6. context selection model and question answering model(bert-base-chinese)

2) the performance of this model v.s. BERT

下圖為 2 個模型經過 20 個 epochs 訓練後的表現，我們可以明顯地看出導入 pretrained weight 後模型的 performance 會有顯著地提升。

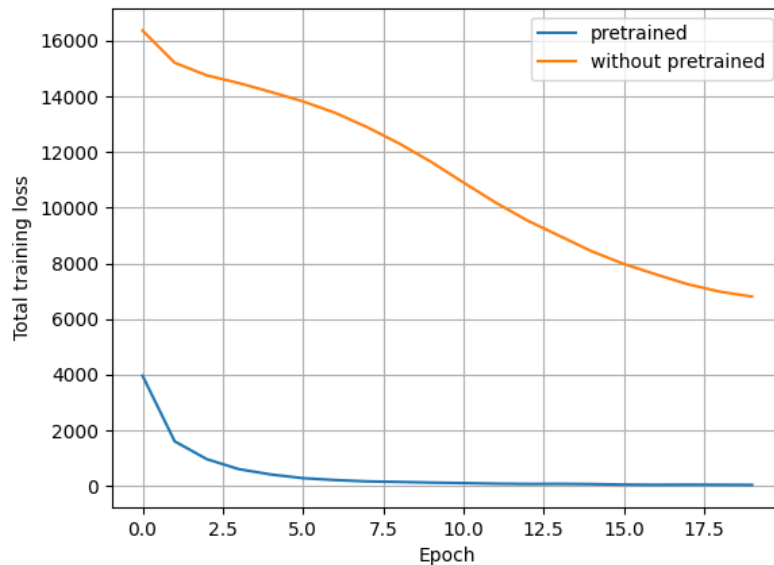


FIGURE 7. Total training loss curve.

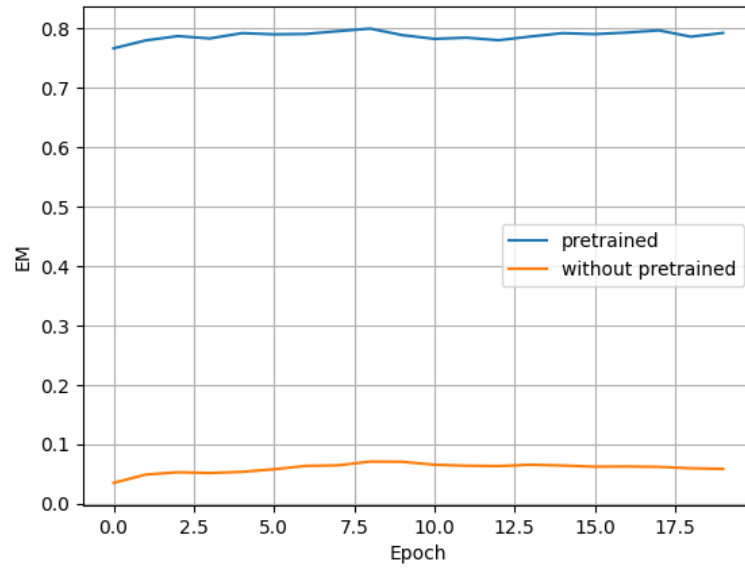


FIGURE 8. EM curve.

R10521601