Applied Deep Learning (2022 Spring)
Homework 2

資管四 B07705015 劉鎮霆

Q1: Data processing

1. Tokenizer

- Tokenization algorithm: Multiple choice 與 question answering 皆使用 transformers 套件中的 AutoTokenizer 進行 subword tokenization
- Pretrained model: hfl/chinese-roberta-wwm-ext [1]

2. Answer Span

- Convert the answer span start/end position on characters to position on tokens after BERT tokenization

(1) 使用 Q1.1 的方法 tokenize 後，每個句子會得到一組 offsets mapping，紀錄每個 token 開頭與結尾字元的位置 [2]，將 context 第 i 個 token 的 offset 表示為 (s_i, e_i) [3]

(2) 將答案的字元位置表示為 (s, e)，選擇滿足 $s_i > s$ 且順序最前面的 token 作為 start token，滿足 $t_i < t$ 且順序最後面的 token 作為 end token

(3) 以 start/end token 在整段句子 (合併問題與 context) 中的位置作為 start/end position on tokens

- Determine the final start/end position

(1) 針對每個句子，模型會輸出對應到各個 token 的 start logit 與 end logit

(2) 選出 start logit 最大的 20 個 token 與 end logit 最大的 20 個 token

(3) 對於所有 start/end token 可行的排列組合計算分數： [4]

$$o_i^s + o_j^e$$

o_i^s 為第 i 個 start token 的 start logit，o_j^e 為 j 個 end token 的 end logit

(4) 找出分數最高的組合，使用 offsets mapping 轉換為字元位置，並從 context 中取出答案，選擇 start/end position 的目標式可表示為：

$$\max_{i,j} o_i^s + o_j^e$$

---

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext

[2] 「我/是/學生」的「我」會對應到 (0,1)，表示是第 0 個字元；「學生」則會對應到 (2,4)，表示是從第 2 到第 3 個字元

[3] 由於問題與 context 會被合併成一個句子，如「[CLS]QQQ[SEP]CCC[SEP]」(QQQ 為問題，CCC 為 context)，因此 context 的第 i 個 token 不一定是整個句子的第 i 個 token

[4] 如 start token 的順序在 end token 前面，兩個 token 皆在 context 中，不超過設定的最大長度

Q2: Modeling

1. Describe

- Model configuration
  - Multiple choice (config.json)

Pretrained model	hfl/chinese-roberta-wwm-ext
Architecture	BertForMultipleChoice
Dropout ratio (attention probability)	0.1
Directionality	bidirectional
Hidden activation layer	GELU
Dropout ratio (hidden layer)	0.1
Hidden size	768
Initializer range	0.2
Intermediate size	3072
Layer norm eps	1e-12
Max position embeddings	512
Number of attention heads	12
Number of hidden layers	12
Pooler fc size	768
Pooler number of attention heads	12
Pooler number of fc layers	3
Pooler size per head	128
Pooler type	first token transform
Position embedding type	absolute
Vocab size	21128

  - Question answering (config.json)

Pretrained model	hfl/chinese-roberta-wwm-ext
Architecture	BertForQuestionAnswering
Dropout ratio (attention probability)	0.1
Directionality	bidirectional
Hidden activation layer	GELU
Dropout ratio (hidden layer)	0.1
Hidden size	768
Initializer range	0.2
Intermediate size	3072
Layer norm eps	1e-12
Max position embeddings	512
Number of attention heads	12
Number of hidden layers	12
Pooler fc size	768
Pooler number of attention heads	12
Pooler number of fc layers	3
Pooler size per head	128
Pooler type	first token transform
Position embedding type	absolute
Vocab size	21128

- Performance
  - Public score on Kaggle：0.78571

- Loss function
  - Multiple choice: cross entropy loss
  - Question answering: cross entropy loss on start logits and end logits

- 將 o 表示為模型的輸出，o_s 為 start logits，o_e 為 end logits
- 將 t 表示為答案的位置，t_s 為 start position，t_e 為 end position
- 將 $CE(\cdot,\cdot)$ 表示為 cross entropy loss

Question answering 模型的損失函數 $\ell(o,t)$ 為：

$$\ell(o,t) = \frac{CE(o_s,t_s) + CE(o_e,t_e)}{2}$$

- Optimization algorithm, learning rate and batch size
  - Multiple choice

Optimization algorithm	Learning rate	Batch size
AdamW	5e-5	16

  - Question answering

Optimization algorithm	Learning rate	Batch size
AdamW	2e-5	8

2. Another type of pretrained model

將 multiple choice 的 pretrained model 改為 ckiplab/albert-tiny-chinese [5]，並與 Q2.1 的模型進行比較 (除了 pretrained model 以外其他部分的設定皆相同)

- Model configuration (config.json)

Pretrained model	ckiplab/albert-tiny-chinese
Architecture	AlbertForMultipleChoice
Dropout ratio (attention probability)	0.0
Dropout ratio (classifier)	0.1
Down scale factor	1
Embedding size	128
Gap size	0
Hidden activation layer	GELU
Dropout ratio (hidden layer)	0.0
Hidden size	312
Initializer range	0.02
Number of inner groups	1
Intermediate size	1248
Layer norm eps	1e-12
Max position embeddings	512
Number of attention heads	12
Number of hidden groups	1
Number of hidden layers	4
Number of memory blocks	0
Position embedding type	absolute
Type vocab size	2
Vocab size	21128

- Performance

	RoBERTa	ALBERT
Accuracy (multiple choice, validation)	**0.962**	0.931
Loss (multiple choice, validation)	**0.169**	0.260
Exact Match (with QA task, validation)	**0.796**	0.774

[5] https://huggingface.co/ckiplab/albert-tiny-chinese

Training time	3 hrs 5 mins	**25 mins**

由上表可以看出，RoBERTa 在準確度上表現較佳，而 ALBERT 則可以花較短的時間訓練出準確度不錯的模型。

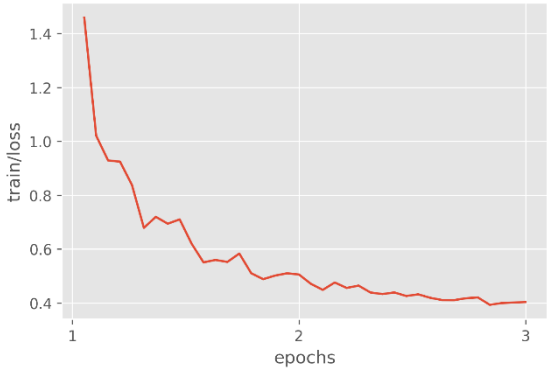
- Difference (architecture, pretraining loss, etc.)
  - RoBERTa 在資料集的選擇與前處理上有所不同
    - 使用更多的資料集進行訓練
    - Dynamic masking：相比 BERT 在訓練前對句子進行 masking，RoBERTa 則在訓練時使用 10 種方法對各個句子進行 masking
  - ALBERT 在模型架構與訓練方式上皆有所不同
    - Factorized embedding parameterization：ALBERT 不直接將 vocabulary 轉換成 hidden layer 的大小，而是透過矩陣分解拆為兩個相乘的矩陣 (矩陣大小為 $V \times E$ 與 $E \times H$，V 為 vocabulary size；E 為 WordPiece embedding size；H 為 hidden layer size)
    - Cross-layer sharing：所有的 recurrent encoder block 皆共用同組參數，因而減少模型的參數量
    - Inter-sentence coherence loss：訓練時進行 sentence order prediction 的任務，讓模型更專注於學習句子的順序關係
    - 增加訓練資料
    - 去除 dropout (參數量的減少可以避免 overfitting，因此不需要額外使用 dropout)

Q3: Curves

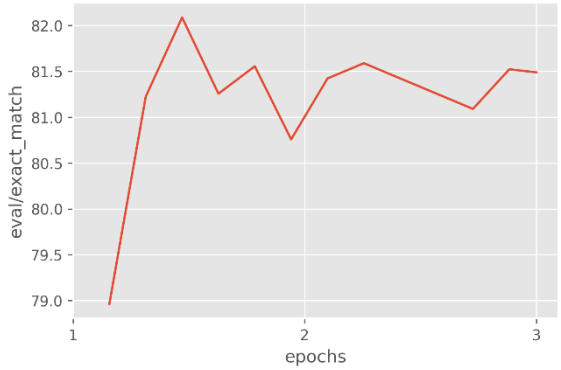
1. Learning curve of QA model

由於 QA model 只訓練 3 個 epoch，因此在每個 epoch 的途中紀錄 training dataset 的 loss 與 validation dataset 的 exact match

- Learning curve of loss (training dataset)



- Learning curve of EM (validation dataset)



Q4: Pretrained vs Not Pretrained

- Task: 使用 multiple choice 進行實驗
- The configuration of the model and how to train this model
  除了不使用 pretrained weights，其餘設定皆與 Q2 的 RoBERTa 模型相同
- The performance of this model vs BERT

	Pretrained	Not Pretrained
Accuracy (multiple choice, validation)	**0.962**	0.532
Loss (multiple choice, validation)	**0.169**	1.010
Exact Match (with QA task, validation)	**0.796**	0.465

由上表可以看出不使用 pretrained weights 時，較難訓練出準確的模型

Q5: HW1 with BERTs

- Model
  - Intent classification

Pretrained model	bert-base-cased
Architecture	BertForSequenceClassification
Dropout ratio (attention probability)	0.1
Dropout ratio (classifier)	null
Hidden activation layer	GELU
Dropout ratio (hidden layer)	0.1
Hidden size	768
Initializer range	0.02
Intermediate size	3072
Layer norm eps	1e-12
Max position embeddings	512
Number of attention heads	12
Number of hidden layers	12
Position embedding type	absolute
Problem type	single label classification
Type vocab size	2
Vocab size	28996

  - Slot tagging

Pretrained model	bert-base-cased
Architecture	BertForTokenClassification
Dropout ratio (attention probability)	0.1
Dropout ratio (classifier)	null
Hidden activation layer	GELU
Dropout ratio (hidden layer)	0.1
Hidden size	768
Initializer range	0.02

Intermediate size	3072
Layer norm eps	1e-12
Max position embeddings	512
Number of attention heads	12
Number of hidden layers	12
Position embedding type	absolute
Type vocab size	2
Vocab size	28996

- Performance (intent classification, slot tagging)
  - Intent classification

	RNN (hw01)	BERT
Accuracy (validation)	**0.918**	0.917
Accuracy (testing, public)	0.916	**0.918**
Accuracy (testing, private)	**0.922**	0.919

  - Slot tagging

	RNN (hw01)	BERT
Token accuracy (validation)	0.962	**0.974**
Join accuracy (validation)	0.773	**0.846**
Join accuracy (testing, public)	0.788	**0.846**
Join accuracy (testing, private)	0.795	**0.849**

本實驗中，兩種模型在 intent classification 上的表現皆差不多，而在 slot tagging 上，BERT 則表現得比 RNN 的模型更好。

- Loss function
  - Intent classification: cross entropy loss
  - Slot tagging: cross entropy loss on all token logits

- Optimization algorithm, learning rate and batch size
  - Intent classification

Optimization algorithm	Learning rate	Batch size
AdamW	2e-5	32

  - Slot tagging

Optimization algorithm	Learning rate	Batch size
AdamW	5e-5	8