DLHLP HW 4 - 3

GitHub ID: PengWenChen
學號:R08942085 系級:電信丙 姓名:陳芃彣
學號:R08942086 系級:電信丙 姓名:趙達軒
學號:R07942091 系級:電信丙 姓名:許博閔
學號:R08945002 系級:生醫電資 姓名:陳玠妏

1. (2%) Please give some examples predicted correctly and incorrectly respectively. At least one for each case is required. Screenshot recommended.

   回答正確：

   ```
   'id': '5645-11-2', 'question': '中華民國與日軍簽定了哪一個協議以停戰熱河戰役?'
   ```

   ```
   5645-11-2,《塘沽協定》
   ```

   回答錯誤：

   ```
   'id': '1190-17-2', 'question': '東德從什麼時候成為歷史？'
   ```

   ```
   1190-17-2,1961年到1970年柏林圍牆的修建減少了逃亡行為
   ```

2. (3%) Which hyperparameter(s) should be modified in order to reach better performance? (e.g. learning_rate, batch_size, warmup_steps, layer_norm_eps, attention_probs_dropout_prob)

   ➔ 一開始使用助教的sample code，只改了下面幾個參數。

   ```
   --per_gpu_train_batch_size 4 \
   --per_gpu_eval_batch_size 5 \
   --fp16 \
   ```

   ➔ 嘗試把attention_probs_dropout_prob和hidden_dropout_prob從0.1調到0.2，結果變差。

   ➔ 嘗試把pooler_num_fc_layers從3調到5，結果不變。

   ➔ 嘗試把warmup_steps調大，結果變差。

➔ 嘗試用https://github.com/wptoux/albert-chinese-large-webqa ，Kaggle的Public score破0.8。

➔ 嘗試用 https://github.com/huggingface/transformers/tree/master/model_cards/voidful/albert_chinese_xlarge ，為目前嘗試中的最高分。

3. (2%) Bonus - Different models compare bert-base-chinese:

這次使用的是ALBERT chinese xlarge的pretrained model：https://github.com/huggingface/transformers/tree/master/model_cards/voidful/albert_chinese_xlarge

在ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS這篇paper中指出，ALBERT在SQuAD和RACE測試上創造了新的SOTA，並在比賽中以+14.5%的優勢擊敗BERT。

ALBERT主要對BERT做了3點改進，縮小了整體的參數量，加快了訓練速度，增加了模型效果。

第一點，Factorized embedding parameterization，也就是嵌入分解參數化。ALBERT的作者注意到，對於BERT、XLNet和RoBERTa，WordPiece Embedding的大小(E)直接與隱含層大小(H)聯繫在一起。然而，ALBERT的作者指出，WordPiece Embedding是用來學習上下文獨立表示的。隱含層嵌入是為了學習上下文依賴表示的。因此，綁定在不同目的下工作的兩個項目意味著低效的參數。所以作者使用了小一些的E(64、128、256、768)，訓練一個獨立於上下文的embedding(VxE)，之後計算時再投影到隱層的空間(乘上一個ExH的矩陣)，相當於做了一個因式分解。

Model		Parameters	SQuAD1.1	SQuAD2.0	MNLI	SST-2	RACE	Avg	Speedup
BERT	base	108M	90.4/83.2	80.4/77.6	84.5	92.8	68.2	82.3	4.7x
	large	334M	92.2/85.5	85.0/82.2	86.6	93.0	73.9	85.2	1.0
ALBERT	base	12M	89.3/82.3	80.0/77.1	81.6	90.3	64.0	80.1	5.6x
	large	18M	90.6/83.9	82.3/79.4	83.5	91.7	68.5	82.4	1.7x
	xlarge	60M	92.5/86.1	86.1/83.1	86.4	92.4	74.8	85.5	0.6x
	xxlarge	235M	**94.1/88.3**	**88.1/85.1**	**88.0**	**95.2**	**82.3**	**88.7**	0.3x

第二點，Cross-layer parameter sharing跨層數共享，ALBERT通過跨

層共享所有參數進一步提高了參數效率。這意味著前饋網絡參數和注意力參數都是共享的。

因此，與BERT相比，ALBERT從一層到另一層的轉換更平滑，作者發現到這種權值共享有助於穩定網絡參數。

最後，SENTENCE ORDER PREDICTION (SOP)，ALBERT認為，NSP(下一個句子預測)將話題預測和連貫預測混為一談。ALBERT的作者認為句子間的連貫是真正需要關注的任務/損失，而不是主題預測，因此SOP使用了兩個句子，都來自同一個文檔。正樣本測試用例是這兩句話的順序是正確的；負樣本是兩個句子的順序顛倒。有別於NSP的正樣本匹配是第二個句子來自同一個文檔，負樣本匹配是第二個句子來自另一個文檔。ALBERT避免了主題預測的問題，並幫助其學習更好的學習句子間的銜接。