AI-CUP 2019

人工智慧論文機器閱讀競賽之論文分類

隊伍: SDML\_SimpleBaseline

成員: 楊濟宇，黃奎鈞，楊之郡

1. 環境

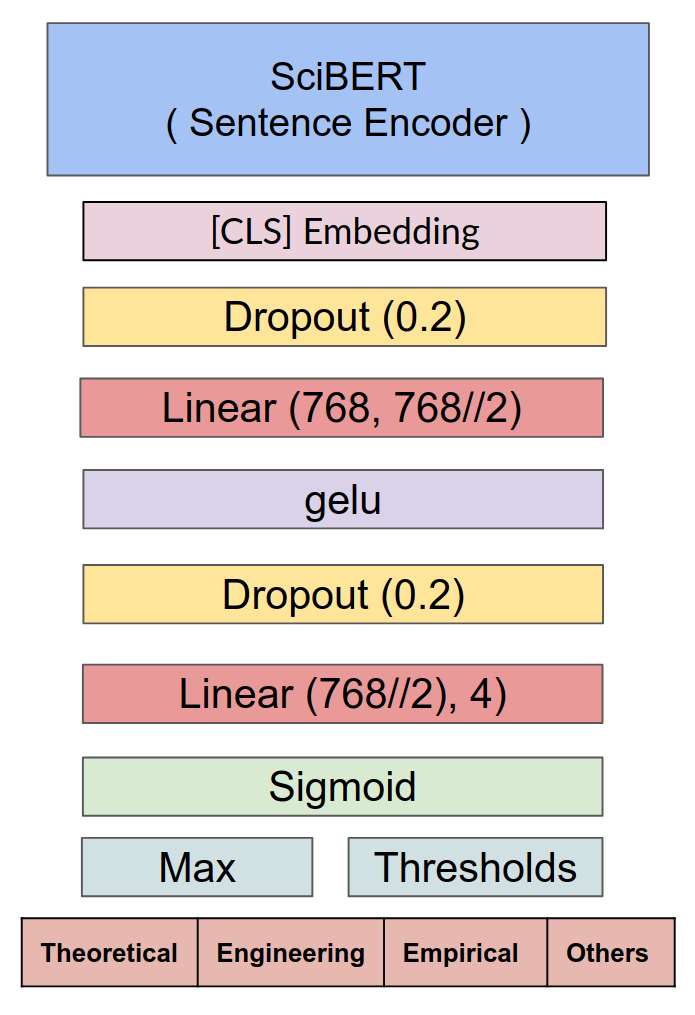
作業系統：Ubuntu 18.04；語言：Python 3.6.8；套件：torch 1.2.0，transformers 2.0.0，sklearn 0.22，預訓練模型：[SciBERT](https://github.com/allenai/scibert)。

1. 資料處理

拿abstract過 SciBERT的tokenizer然後加入 [CLS] 和 [SEP] 的special token，最後使用StratifiedShuffleSplit切90%當作訓練資料，10%用來驗證。另外試過接title或分別過encoder再接起來，甚至加入領域資料，皆沒有變好，因此只拿abstract做訓練資料。

1. 模型架構

先過預訓練的SciBERT encoder得到contextualized embedding，接著以 [CLS] token的embedding當作整句的embedding，然後過一層Linear從768維降至384維，再過另一層Linear從384維降至4維，最後過Sigmoid得到預測logits。其中，兩層Linear前皆加上Dropout及activation function為gelu。



1. 訓練方式

* Optimizer：Adam with linear warmup
* Warmup steps：half of total steps
* Loss function：Binary Cross Entropy
* Positive Weights：[1.0, 1.0, 1.75, 7.5]
* 訓練參數：learning rate = 1e-5，batch size = 32，epoch = 3，Dropout rate = 0.2

1. 分析&結論

* 預訓練模型選擇：SciBERT因預訓練在科技論文上，所以理論上會表現比其他模型好。經過實驗，確實贏過BERT-large，XLNet及RoBERTa等SOTA模型。
* 參數選擇：以測試結果調整，其中僅有warmup steps和epoch較會影響訓練結果（過擬合速度），因此以上參數幾乎為最一開始直覺下的。
* 預測0，1之臨界值調整：這是此比賽要取得好成績的重大關鍵。
  + 皆用0.5：0.708
  + 觀察分佈後調整至 [0.35, 0.3, 0.25, 0.35]：0.729
    - 但是很可能只是過擬合在public test set上。
  + 先皆用0.35，會得到較少的1，因此另外將原先預測0的類別裡，若其為所有類別最大值，亦預測為1：0.733
    - 直觀上想，可以增進預測1而答案確實為1的機率。
* 投票：最後衝分及降低成績變異的手段。
  + 篩選本機F1分數大於0.726的亂數種子得到41個模型
  + 多數決投票得到最後結果：0.735
* 註：以上分數皆為public leaderboard的分數，最後private leaderboard為0.751（第一名）

1. 程式碼

* 訓練：python bert\_finetune.py [seed] [gpu\_id]
* 預測：python predict.py [seed] [gpu\_id]
* 投票：直接開vote\_task2.ipynb然後run all cells
* 註：seed為亂數種子，可固定結果；gpu\_id指定欲使用之gpu。

1. 使用的外部資源及參考文獻

無。