AI-CUP 2019

人工智慧論文機器閱讀競賽之論文標註

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然後在每個句子前方加入 [CLS] ，整個abstract最後加上句數及 [SEP] 兩個token作為輸入。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，接著以每個句子前方的 [CLS] token的embedding當作每句的embedding，然後過一層Linear從768維直接降至6維，最後過Sigmoid得到預測logits。

1. 訓練方式

* Optimizer：Adam with linear warmup
* Warmup steps：half of total steps
* Loss function：Binary Cross Entropy
* Positive Weights：[2.0, 2.25, 2.0, 2.87, 4.0, 8.7]
* 訓練參數：learning rate = 1e-5，batch size ≈ 32，epoch = 3

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0.5：0.701
  + 觀察分佈後調整至 [0.56, 0.55, 0.45, 0.59, 0.59, 0.73]：0.731
    - 但是很可能只是過擬合在public test set上。
  + 先皆用0.7，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1：0.735
    - 直觀上想，可以增進預測1而答案確實為1的機率。
  + 不同於task2，各類別是有一定相關性的，例如：RESULTS和CONCLUSIONS很常同時為1。因此對於非OTHERS的五個類別分別另外訂較低的臨界值，當一個句子只被預測一個類別，若其相鄰的類別高於臨界值便也預測為1：0.738
    - 臨界值：[(0, 1): 0.59, (1, 0): 0.59, (1, 2): 0.6, (2, 1): 0.6, (2, 3): 0.6, (3, 2), 0.6, (3, 4): 0.62, (4, 3): 0.62]，其中 ”(a, b): t” 表對於只被預測第a個類別的句子，若第b個類別的臨界值為t。
  + 另外我們發現OTHERS在本機預測的TP遠少於FP及FN，故分析F1分數後可以得到，若一個句子被預測為多類別且包含OTHERS，將OTHERS改為0，F1分數將會提升：0.739
  + 以上即為我們最後的策略，不過我們還有發現，句子間也有相關性，但是加入轉移矩陣進去訓練，抑或是導入後處理，皆沒有讓分數更好，於是同樣就原先模型預測結果計算轉移矩陣，發現幾乎一模一樣。所以我們認為除非有突破性強大的embedding或是截然不同且合理的訓練方式，不然我們的模型應該是SOTA了。
* 投票：最後衝分及降低成績變異的手段。
  + 不同於task2，此task本機和public test set並非明顯正相關或負相關，因此直接訓練95個模型。至於為何是95個，因為由時間與記憶體限制決定。
  + 多數決投票得到最後結果：0.741，但是到private就壞掉了，推測是因為對每個類別都設臨界值而過擬合了。
* 註：以上分數皆為public leaderboard的分數，最後private leaderboard為0.736（第五名）

1. 程式碼

* 訓練：python bert\_finetune\_task1.py [seed] [gpu\_id]
* 預測：python predict\_task1.py [seed] [gpu\_id]
* 投票：直接開vote\_task1.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。