教育部全國大專校院人工智慧競賽(AI CUP)

機器閱讀紀錄-課程挑戰賽

隊伍: 結果尚未公布

成員: 吳亦振

壹、 環境

作業系統 : ubuntu 18.04

語言 : python 3.6.9

套件 :

pandas==1.1.4

numpy==1.17.2

torch==1.3.0

torchtext==0.4.0

transformers==2.10.0

tqdm==4.53.0

預訓練模型 : allenai/scibert_scivocab_uncased

貳、 資料處理

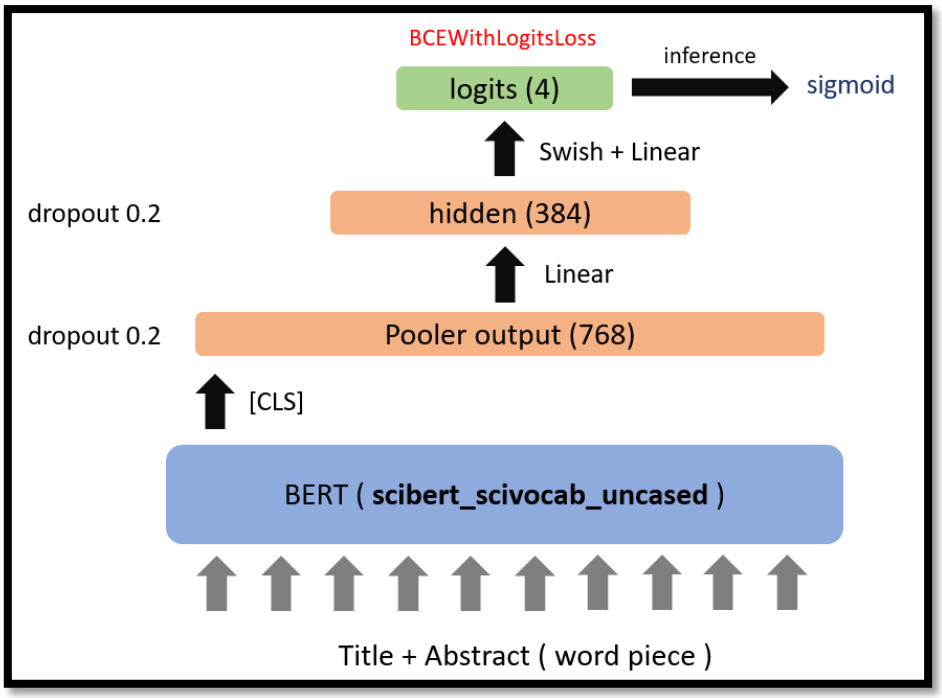
針對 trainset 及 testset 的「Title」、「Abstract」做前處理。

1. 以空白符號取代「Abstract」 中的「$$$」 符號。
2. 「Title」及「Abstract」中的全形字元轉換為半形。
3. 「Title」及「Abstract」中的大寫轉換為小寫。
4. 新增「content」 欄位，為「Title」和 「Abstract」之合併。
5. 將 trainset 以 9:1 切割成 train data 及 valid data。

參、 模型架構

1. 使用 BERT 預訓練模型(allenai/scibert_scivocab_uncased) 再加上兩層 Linear。
2. 激活函數 : swish : $x \cdot sigmoid(x)$。
3. 模型架構如圖(一)。

圖(一)



肆、 訓練方式

1. 訓練參數:

   Optimizer : torch.optim.Adam (learning rate : 1e-05)

   Loss Function : torch.nn.BCEWithLogitsLoss

	pos weight
ENGINEERING	0.98
EMPIRICAL	1.44
THEORETICAL	1.03
OTHERS	3.55

   Dropout : 0.2

2. 根據 valid data 的 loss 做 early stop，防止模型過度擬合。

伍、 分析&結論

由於此任務為科學論文之分類，因此選擇使用大量科學論文訓練的 BERT (allenai/scibert_scivocab_uncased) 作為預訓練模型。而論文類別占比不一致，因此在損失函數中針對不同類別設定權重，其權重為類別總數除以各類別個數再取自然對數。若能進一步做 cross validation 和蒐集更多外部資料來訓練模型和推論，其結果會更加穩定。

陸、 程式碼

詳見附檔

柒、 使用的外部資源與參考文獻

- huggingface