**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: Cronus6w6

成員: 陳威呈

1. 環境

訓練環境: Google Colab

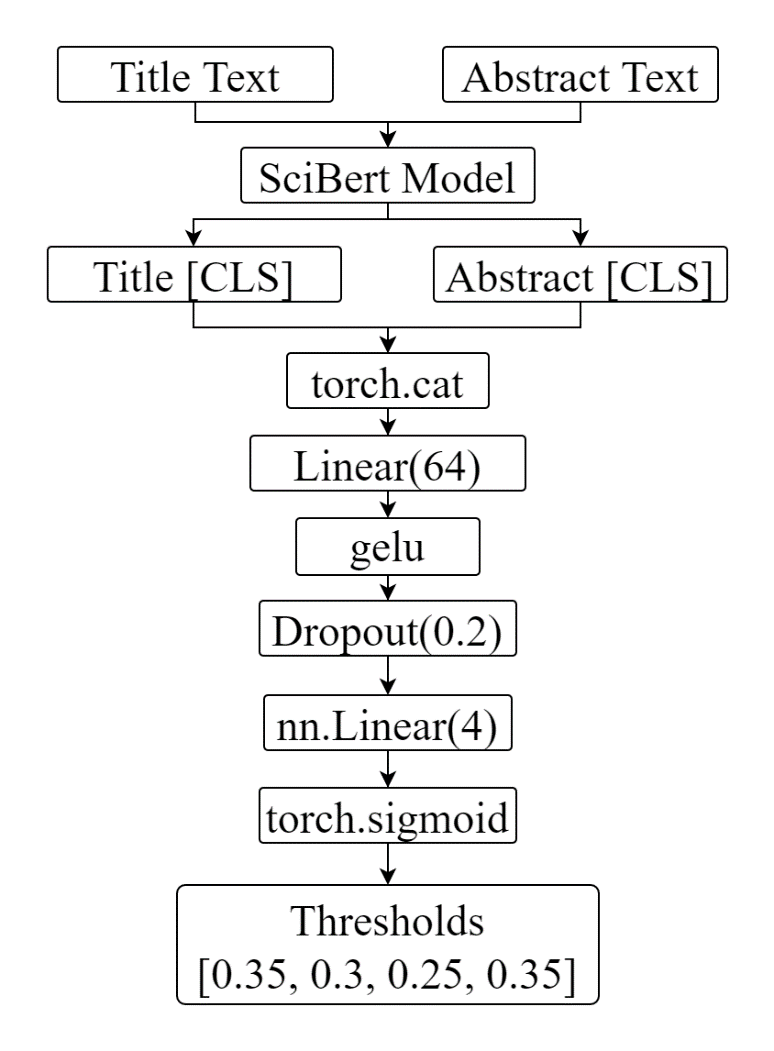
使用套件: transformers==4.2.1 (額外安裝)、torch==1.7.0+cu101、pandas==1.1.5、numpy==1.19.5、tqdm==4.41.1

預訓練模型: scibert\_scivocab\_uncased ( <https://github.com/allenai/scibert> )

1. 資料處理

將Abstract 的 $$$ 全部替換為空格，然後與 Title 個別放入 Tokenizer 做成title\_indices、title\_segments、abstract\_indices、abstract\_segments 四個輸入。訓練驗證比例為9：1。

1. 模型架構

 將 Title 與 Abstract 放入預訓練的 SciBert Model，將兩者 [CLS] 的輸出合併後放入 Classifier 中進行分類，經過 sigmoid 後，其值 > Threshold 分類為 1 ，否則為 0。

1. 訓練方式

使用 AdamW 作為 Optimizer ，因為Label數量不均，訓練回歸不經過 sigmoid ，而在計算Loss使用Positive Weights([1., 1., 1.75, 7.5]) 進行計算，以增加少量 label 的權重。

1. 分析&結論

一開始只使用Abstract 的 Label，後來看到 Dual Input 的做法，但雙Model使用記憶體太大，導致 Batch Size 減少，正確率也沒提升，但使用Siamese Model 的做法，可以將 F1 提高0.1 因此使用此架構做預測，未來可以在Scibert Model輸出加一個 Dropout 也許可以減少訓練後期F1稍微下降的問題。

1. 程式碼

<https://github.com/cronus6w6/AI-CUP-2020>

1. 使用的外部資源與參考文獻

Beltagy, Iz, et al. “SciBERT: A Pretrained Language Model for Scientific Text.”

ArXiv:1903.10676 [Cs], Sept. 2019. arXiv.org, <https://arxiv.org/abs/1903.10676>.

Prithvi Jaunjale. (2020). SciBERT Multi Label Classification - Dual Input. Retrieved from <https://www.kaggle.com/prithvijaunjale/scibert-multi-label-classification-dual-input#Data>