Machine Learning HW6 Report

學號：B06902049　系級：資工二　姓名：林首志

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*

   由於我最終的模型是五個model ensemble的結果，這裡我挑其中一個來分析。

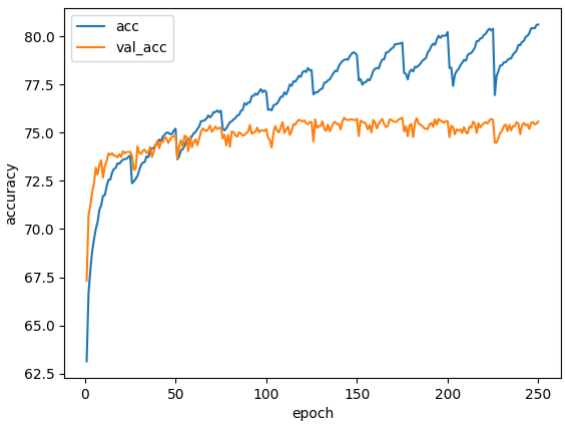
Preprocess: segmentation, to lowercase, pad to seq_length=256
Embedding(num_embeddings=31153, embedding_dim=32, padding_idx=0)
Dropout(p=0.5)
GRU(input_size=32, hidden_size=32, bidirectional=True)
Dropout(p=0.5)
Linear(in_features=64, out_features=1, bias=True)
Sigmoid()

其中Linear的input是Bidirectional GRU跑完整個序列後的hidden states。

Word embedding的方法是先將train_x和test_x傳進gensim word2vec（演算法是CBOW，迭代次數15次，其他參數皆為預設值。）預訓練出初始的參數，之後訓練RNN時再隨著訓練過程微調embeddings的參數。

取val_acc最高的epoch的參數，此RNN模型的public score是0.75370，private score是0.75120。

訓練曲線如下（由於我用SGDR(https://arxiv.org/abs/1608.03983)調整學習率，曲線會有週期性的震盪）：

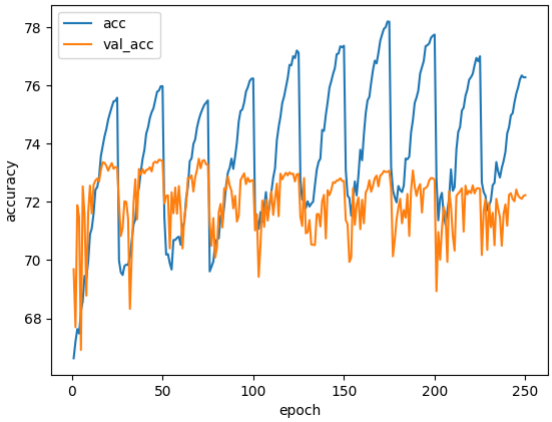


2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

Preprocess: segmentation, to lowercase, pad to seq_length=256, to BOW vector
Linear(in_features=vocab_size(31153), 128, bias=True)
ReLU()
Dropout(p=0.5)
Linear(in_features=128, out_features=1, bias=True)
Sigmoid()

此模型的最高val_acc是73.483448%，對應的public score是0.73490，private score是0.72800。

訓練曲線如下（同上題，由於SGDR的關係，曲線會有週期性的震盪）



3. (1%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。

   Preprocess:這部份除了斷詞之外，對準確率比較有影響的是seq_length(輸入會被pad到多長)。經過多次的嘗試，我發現在合理的範圍內，seq_length越大準確率越高。這可能是因為seq_length越大，對於比較長的句子（儘管數量可能不多）能捕捉到更多的資訊，進而做出更準確的預測。

   架構：最重要的部份可能是Bidirectional GRU。Bidirectional GRU相對於GRU更能掌握序列前段的資訊，因此表現更好。除此之外，我也有對Bidirectional GRU的初始hidden states做訓練，我沒有做過嚴謹的比對，但根據此篇文章(https://r2rt.com/non-zero-initial-states-for-recurrent-neural-networks.html)指出，訓練初始hidden states應該是有助益的。

   Ensemble：由於單個model的表現看起來不夠好，我訓練了五個架構、超參數不一樣的模型，並將他們預測出的labels做unweighted voting。這樣的作法可以降低variance，提昇模型的表現。

4. (1%) 請比較不做斷詞 (e.g., 以字為單位) 與有做斷詞，兩種方法實作出來的效果差異，並解釋為何有此差別。

   有斷詞：public score=0.75370，private score=0.75120。

   不斷詞：public score=0.74970，private score=0.74910。

   由於「詞」才是真正表示語意的單位，用「詞」的效果比用「字」稍微好一點。

5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "在說別人白痴之前，先想想自己"與"在說別人之前先想想自己，白痴" 這兩句話的分數（model output），並討論造成差異的原因。

	第一句	第二句
RNN	0.7082	0.6651
BOW	0.5007	0.5007

可以發現RNN的分數都比BOW高，而且BOW兩句的分數都一樣。RNN分數較高的原因可能是因為模型表現本來就比較好，也有可能單純是訓練的隨機性造成的。BOW兩句分數一樣的原因是，雖然兩句話的字詞排列順序不同，他們轉成BOW向量後是相等的，所以結果相等是合理的。而RNN模型會考慮輸入的順序，因此兩句結果會不同。