Machine Learning HW6 Report

學號：b05901063 系級：電機三 姓名：黃世丞

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*

   用jieba斷詞後，使用gensim訓練word_embedding matrix作為embedding layer, size=100, epochs=30, min_count=3。

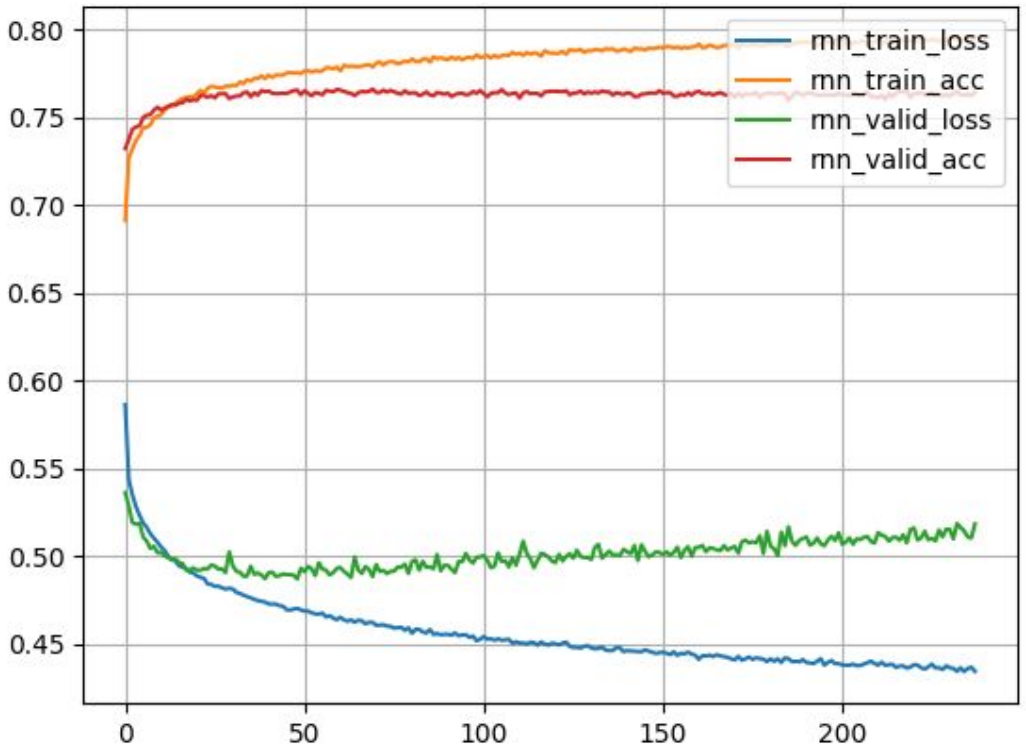
   模型以keras實現，使用2層Bidirectional LSTM，max_time_step=100, dropout_rate=0.5, 並加上Attention Layer，使用peephole connection。

   DNN為output -> 256 -> LeakyReLU(0.2) -> 1，最後過sigmoid，loss使用binary crossentropy。

   正確率：

   public: 0.76980　　private: 0.76610

   最後best使用ensemble讓正確率提升到 public: 0.77380　　private: 0.76680



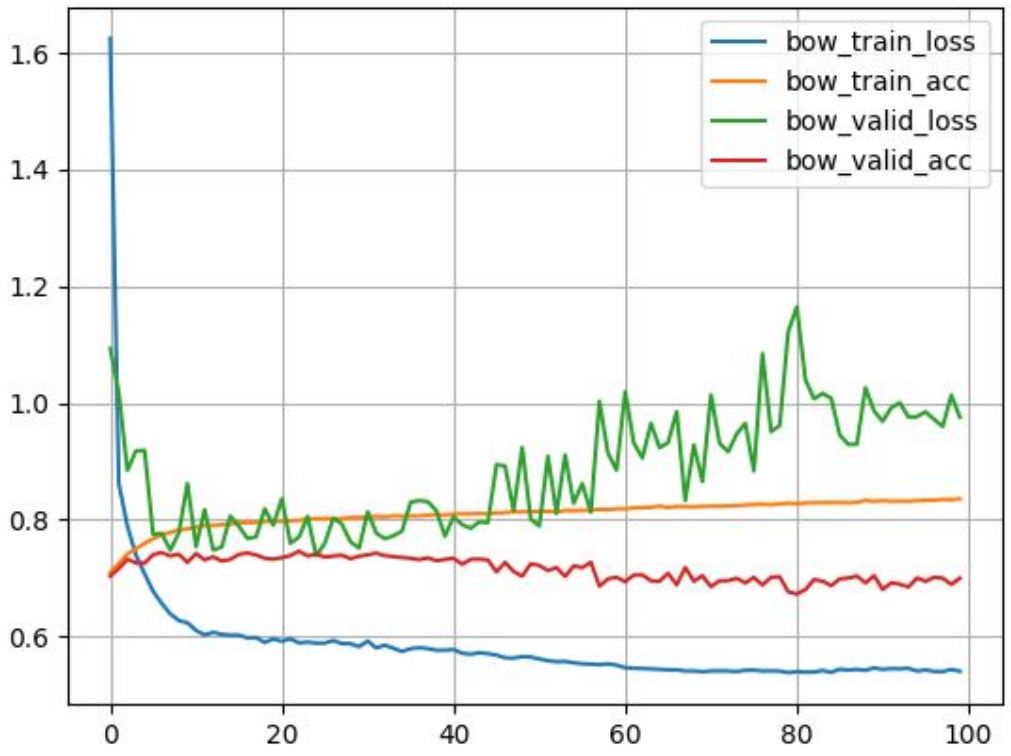
2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

   Preprocess: 把BOW的句子每句分別除以該句字彙出現次數的最大值，使BOW的vector值落在0~1之間

Input size: (45449,) (min_count = 3)
Dense(1024) with regularization(kernel: l2,0.01; bias: l1, 0.01)
->LeakyReLU(0.02)->BatchNormalization()
->Dense(512)->LeakyReLU(0.2)->BatchNormalization()
->Dense(256)->LeakyReLU(0.2)->BatchNormalization()
->Dense(128)->LeakyReLU(0.2)->BatchNormalization()
->Dense(64)->LeakyReLU(0.2)->BatchNormalization()
->Dense(1)->Sigmoid()

正確率：
public: 0.7429　　private: 0.7418

訓練曲線如下圖：



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。
(1) 使用雙層的LSTM，並加上dropout，dropout rate=0.5，雙層LSTM能讓模型更強大，但是LSTM很容易overfit，故加上dropout

(2) 加上Attention Layer，許多seq2seq的模型使用attention都有非常好的效果，能讓模型focus在較為重要的詞彙上。
(3) 加上peephole，peephole是把LSTM的c vector也傳到下一個cell的輸入中，在peephole的架構下模型更能掌握前後文之間的關係
(4) 我的模型在embedding dim=100或200時皆能獲得不錯的效果，但更小或更大的embedding dim都會爛掉
(5) max time step=100，我把斷完詞的句子長度印出來看過，大約有75~80%的句子都不到40字，但是前1%的句子全都超過600字，因此斟酌選擇了100字當作句子最大長度，超過的部份直接砍掉，適度保留有用的資訊和控制模型參數。

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。
有做斷詞：
public: 0.76980　　private: 0.76610
沒做斷詞：
public: 0.73220　　private: 0.72890
沒有斷詞的情況表現很明顯非常差，因為中文的字組合成的辭彙意思可以有非常大的差異，然而沒有斷詞的情況下，embedding無法將這種差異表現出來，故會有很多沒辦法正確判斷的情況。

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。

	在說別人白痴之前，先想想自己	在說別人之前先想想自己，白痴
RNN	2.37820562	1.78930421
BOW	0.45623911	0.45623911

RNN的分數會隨著句子結構產生差異，但BOW只考慮單詞出現頻率，故兩者分數皆相同。