學號：B03505031 系級： 工海四 姓名：邱昱軒

1.(1%) 請說明你實作的 RNN model, 其模型架構、訓練過程和準確率為何？(Collaborators:)

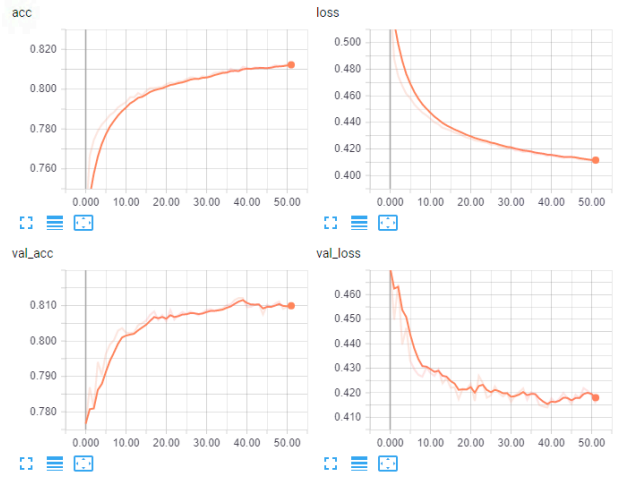
答：

原本利用 glove 訓練 vector，但效果不佳，所以改用 word2vec 訓練。

dim=300, min_count=15, alpha=0.005

Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 50, 300)	77678400
conv1d_1 (Conv1D)	(None, 48, 150)	135150
dropout_1 (Dropout)	(None, 48, 150)	0
lstm_1 (LSTM)	(None, 48)	38208
dropout_2 (Dropout)	(None, 48)	0
dense_1 (Dense)	(None, 1)	49
activation_1 (Activation)	(None, 1)	0

epoch=70
batch=500
optimizer=binary_crossentropy
loss fuction=rmsprop
每個句子的最大字數為 50
conv1d(150,kernel_size=3)
LSTM(48,activation=tanh)
Dense(1,activation=sigmoid)



Earlystop 的 val_acc=0.81217
以該 checkpoint 的模型預測，
kaggle public= 0.80740
val_acc 大概在前 20 個 epoch 就會趨近於收斂，過程中雖然大致是穩定上升，但並不太平滑，應該是跟 RNN 本身會有許多 error 坡度非常陡峭的地方有關。

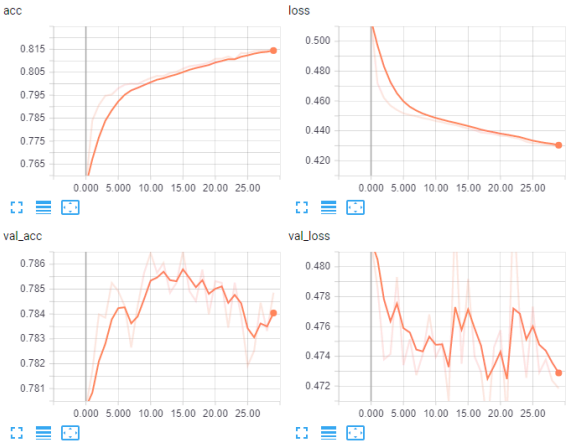
2.(1%) 請說明你實作的 BOW model, 其模型架構、訓練過程和準確率為何？(Collaborators:)

答：

因為 BOW 無序的特性，先去除標點符號，再篩選出出現次數大於 15 次的字建立字典。找到約 4000 個單詞，對 training data 建立 np array，再進 DNN 訓練。

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 256)	1019136
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 64)	16448
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65
activation_1 (Activation)	(None, 1)	0

epoch=30
batch=128
optimizer=binary_crossentropy
loss fuction=rmsprop



最好的 val_acc = 0.78652
kaggle public = 0.78290
在訓練過程中，validation 的進步情況很不好，嘗試了幾次都是會在大約第 10~15 個 epoch 時達到最高值，隨後又會稍微下降。

3.(1%) 請比較 bag of word 與 RNN 兩種不同 model 對於"today is a good day, but it is hot"與"today is hot, but it is a good day"這兩句的情緒分數，並討論造成差異的原因。
(Collaborators:)
答：

	today is a good day, but it is hot	today is hot, but it is a good day
BOW	0.40205988	0.40205988
RNN	0.22926138	0.94598347

由於 BOW 無序的特性，這兩個句子在 BOW 中的表示都是一樣的，因此沒辦法辨別出由於語序所造成的情感差異。

4.(1%) 請比較"有無"包含標點符號兩種不同 tokenize 的方式, 並討論兩者對準確率的影響。
(Collaborators:)
答:

捨棄的標點符號包含!@#$%+-*/=()[],.

有標點的 tokenizer，kaggle **準確率** 0.80740

沒標點的 tokenizer，kaggle **準確率** 0.80314

因為標點符號（尤其是驚嘆號、問號等帶有情感的符號）也隱含了發文者當下的情緒。若捨棄標點符號，僅使用文字判斷，可能不足以準確的判斷。

5.(1%) 請描述在你的 semi-supervised 方法是如何標記 label, 並比較有無 semi-surpervised training 對準確率的影響。
(Collaborators:)
答：

載入第一次訓練好的 model，對 unlabel data 做 predict，**將結果 0.97 以上的標記為 1，0.03 以下的標記為 0**。第一次標記 label 約標記出 300000 筆資料。加上 semi-supervised 的步驟後，**準確率反而下降（kaggle 0.80740 -> 0.80534）**。猜測是因為使用第一次訓練出的模型來標記 no label，這些標記出來的資料本身就已經代表這個模型本身（因為可以貼合這個模型），繼續訓練可能會有 overfit 此模型的狀況。