學號：R04546031 系級： 工工碩二 姓名：洪唯凱

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:)

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 256)	394240
dense_1 (Dense)	(None, 256)	65792
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 128)	32896
dropout_2 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

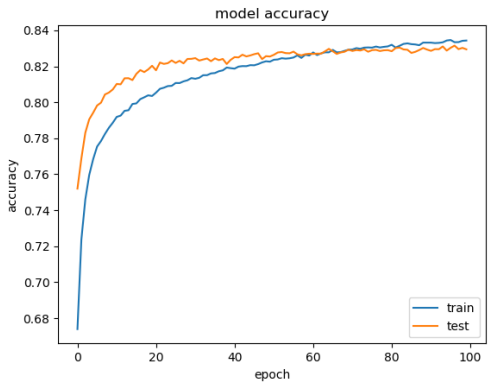
Total params: 493,057

Trainable params: 493,057

Non-trainable params: 0

Train on 180000 samples, validate on 20000 samples

實作過程，先做一個 word2vec 的 fumction，之後直接餵進去 LSTM 模型裡面，我 batch size 為 1024，跑 100 個 epochs，在這過程中，不斷存取有變好的模型，



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:

_____

Layer (type) Output Shape Param #

==========

dense_1 (Dense) (None, 256) 115456

_____

dense_2 (Dense) (None, 128) 32896

_____

dropout_1 (Dropout) (None, 128) 0

_____

dense_3 (Dense) (None, 1) 129

==========

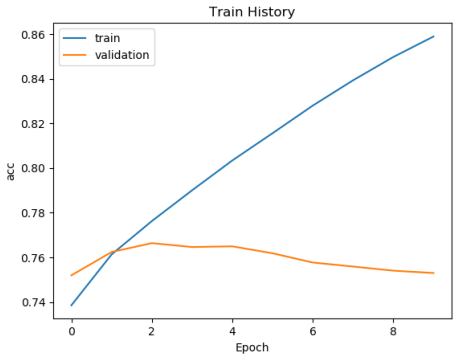
Total params: 148,481

Trainable params: 148,481

Non-trainable params: 0

_____

Train on 180000 samples, validate on 20000 samples



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:)

bag of word 兩個都是 1

RNN 一個是 0，一個是 1

主要原因是 RNN 有考慮到順序的問題

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:)

沒加上標點符號，maxlen，比較小，準確率為 0.82316

有加上標點符號，maxlen，比較大，擔心會有稀疏矩陣的問題，但標點符號更能表示當下的情緒，例如:!可能表示語調略重，表示負面情緒。準確率為: 0.82100

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:)

因為 semi 的最長句子有點長，來到了 228，所以我只取後面的 50 個字當作 testX 去預測，之後再從預測出來的機率，如果高於 85%，我就把他選取來當作我的 trainX，之後再重新 train 我的模型。變差的原因，可能是因為我沒有做機率大於多少的篩選，我就直接 train model，

有做 semi 的準確率為: 0.805

沒有做 semi 的準確率為: 0.82316