學號：R06922117 系級： 資工碩一　姓名：李岳庭
(Collaborators: R05921016)

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？

答：

model 相關參數:

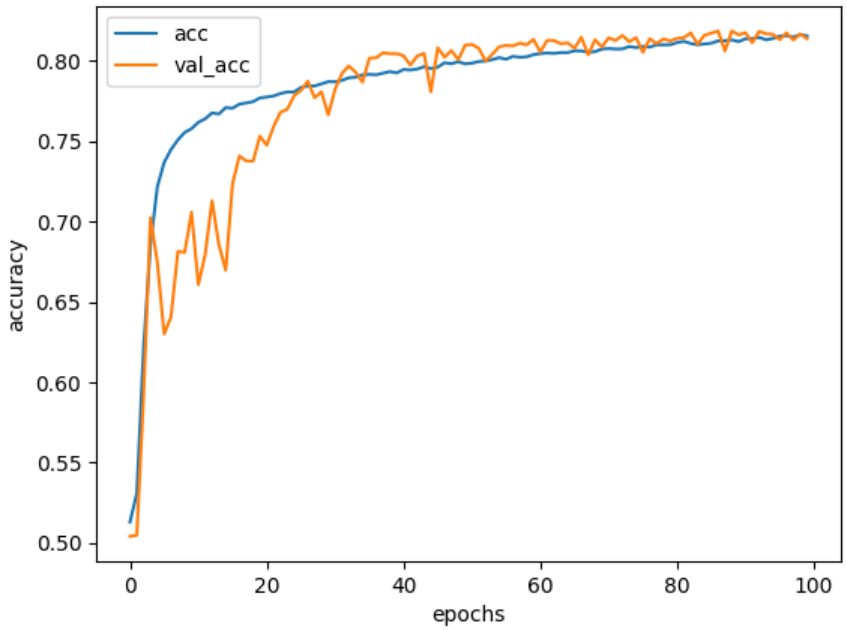
summary

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 40, 256)           525312
_________________________________________________________________
lstm_2 (LSTM)                (None, 40, 128)           197120
_________________________________________________________________
lstm_3 (LSTM)                (None, 64)                49408
_________________________________________________________________
dense_1 (Dense)              (None, 512)               33280
_________________________________________________________________
dropout_1 (Dropout)          (None, 512)               0
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
dropout_2 (Dropout)          (None, 256)               0
_________________________________________________________________
dense_3 (Dense)              (None, 128)               32896
_________________________________________________________________
dropout_3 (Dropout)          (None, 128)               0
_________________________________________________________________
dense_4 (Dense)              (None, 100)               12900
_________________________________________________________________
dropout_4 (Dropout)          (None, 100)               0
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 202
=================================================================
Total params: 982,446
Trainable params: 982,446
Non-trainable params: 0
```

```
opt = Adadelta(lr=0.8, rho=0.95, epsilon=1e-08)
epochs=100
batch_size=1800
loss='categorical_crossentropy'
```

另外還有使用 gensim 的 word2vec 套件，pre-train 出一個 embedding，代替原本 keras 的 embedding layer。

準確率：



kaggle: 0.820015

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？

答：

model 相關參數:

summary

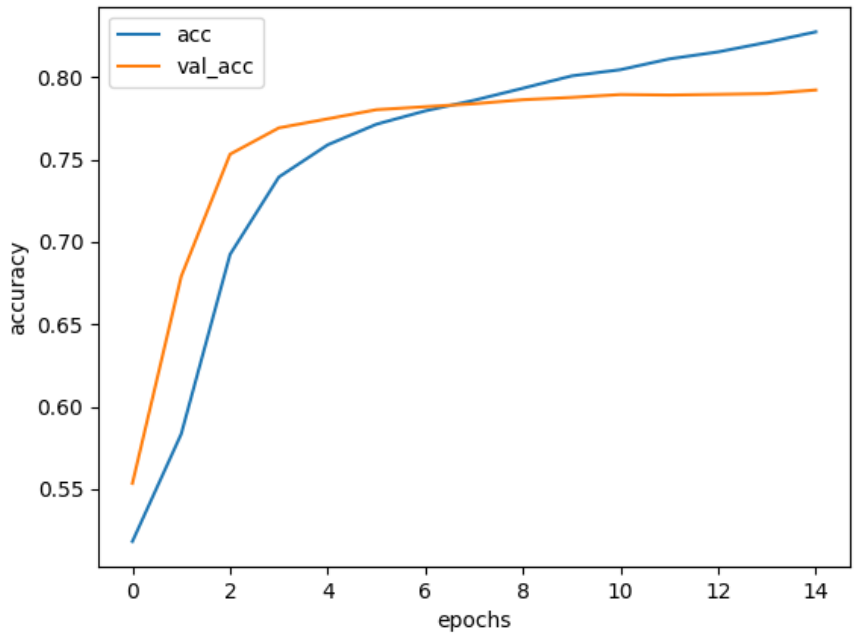
Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 1024)	5447680
dropout_1 (Dropout)	(None, 1024)	0
dense_2 (Dense)	(None, 512)	524800
dropout_2 (Dropout)	(None, 512)	0
dense_3 (Dense)	(None, 256)	131328
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 128)	32896
dropout_4 (Dropout)	(None, 128)	0
dense_5 (Dense)	(None, 100)	12900
dropout_5 (Dropout)	(None, 100)	0
dense_6 (Dense)	(None, 2)	202

Total params: 6,149,806
Trainable params: 6,149,806
Non-trainable params: 0

epoch=15

其餘都參數都相同

準確率：



kaggle: 0.79474

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。**

答：

RNN (probability)

	negative	positive
today is a good day, but it is hot	0.2366273	0.76337272
today is hot, but it is a good day	0.00461505	0.99538499

bag of word (probability)

	negative	positive
today is a good day, but it is hot	0.16152503	0.83847499
today is hot, but it is a good day	0.16152503	0.83847499

由上表可觀察出 bag of word model 只計算每個 word 的出現次數，所以兩句話預測出的機率相同，RNN 有考慮到前後關係的差別，兩句話正負面的程度可從機率看出。

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。

答：

	有標點符號	無標點符號
accuracy	0.820015	0.81587

加入標點符號一起 train 的準確率較高，這點蠻 make sense，標點符號本來就是文字來表達情緒的一種方式。

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。

答：

使用 RNN model 將 unlabeled data 標上 label，我設定的 threshold=0.9，意思是若預測正負面的機率有 0.9 以上才將此 data 納入 training data，但即便加上這些 data，準確率反而下降一點點到達 0.81842，因為資料量龐大，train 一次要太多時間，故沒有嘗試其他 threshold。