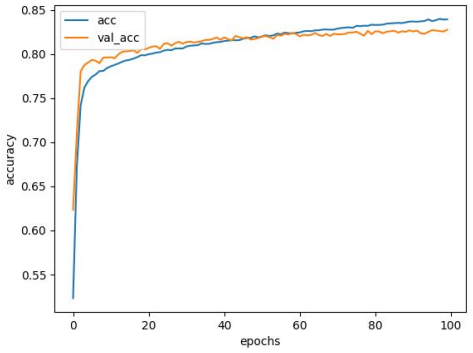
學號：R05921016 系級： 電機碩二 姓名：傅鈞笙

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：
使用genSim 套件事先訓練word2vec (僅使用training、testing data 句子進行訓練)
100 epochs, optimizer: Adadelta(lr=0.8, rho=0.95, epsilon=1e-08),
loss function : 'categorical_crossentropy'
model 架構(summery):

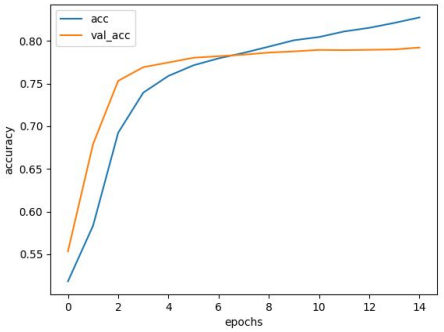
```
Layer (type)                 Output Shape              Param #
=================================================================
lstm_1 (LSTM)                (None, 40, 256)           525312
_________________________________________________________________
lstm_2 (LSTM)                (None, 40, 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
_________________________________________________________________
```



準確率	training acc.	testing public acc.	testing private acc.
RNN model	0.84325	0.82614	0.82434

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：
使用genSim.corpora 建立字典(僅使用training data)，並且濾除太常見與太不常見的文字。15 epochs, optimizer: Adadelta(lr=0.8, rho=0.95, epsilon=1e-08),
loss function : 'categorical_crossentropy'
model 架構(summery):

```
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
_________________________________________________________________
```



準確率	training acc.	testing public acc.	testing private acc.
RNN model	0.8276	0.79050	0.78990

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：

BOW:

情緒分數	today is a good day, but it is hot	today is hot, but it is a good day
NEGATIVE	0.13506301	0.13506301
POSITIVE	0.86493701	0.86493701

RNN:

情緒分數	today is a good day, but it is hot	today is hot, but it is a good day
NEGATIVE	0.22978413	0.00650758
POSITIVE	0.77021587	0.993492421

對於BOW model 而言，兩句話會被轉成一模一樣的 bag of words，所以可想而知兩者獲得的output 也會相同。而RNN model 則能分辨兩句話文字順序的差異，例如第一句話語意略為負面，獲得的positive 分數就比第二句話來的低一點。

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：

準確率	testing public acc.	testing private acc.
無標點符號	0.82614	0.82434
有標點符號	0.82960	0.82768

加入標點符號進行tokenize 後，準確率又提高了 0.3% ，估計是因為標點符號對於語意表達其實有很大影響，加入標點符號可以幫助理解語意。

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：

我使用無標點符號的tokenizing 與第一題的RNN model.
先使用supervised learning 訓練40 個epoch，再進行semi-supervised learning.

Semi-supervised learning 時，先將unlabeled data 進行預測，並將信心度高的unlabeled data 加入training data 中 (prob>0.9 or prob<0.1)。由於unlabeled data 資料量很龐大，因此在train 時需要train_on_batch，一次只將一個batch size 的 unlabeled data 轉成vector 與進行padding，不然CPU 會爆炸。

Semi-supervised learning 訓練60個epoch，結果比較如下：

準確率	training acc.	testing public acc.	testing private acc.
supervised RNN model	0.84325	0.82614	0.82434
semi-supervised RNN model	--	0.82447	0.82396

就結果而言，其實加入semi-supervised learning 後準確率反而下降了一點，雖然training accuracy 已經達到90% 以上，但這主要是來自unlabeled data。

可能原因是train的還不夠久，還沒能正確找到data 的分布。