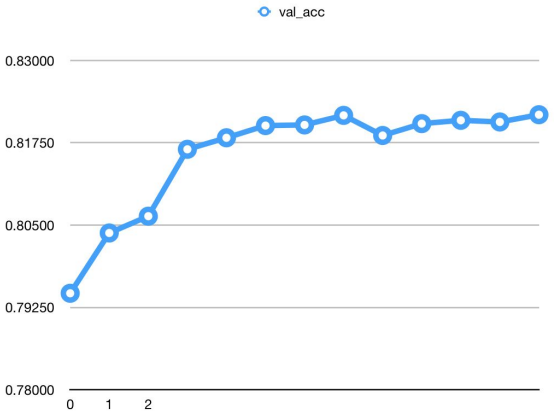
學號：R06725028 系級：資管碩一 姓名：黃于真

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：左下圖為模型架構，右下圖為訓練過程，另外embedding使用word2vec

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 39)	0
embedding_1 (Embedding)	(None, 39, 100)	5931000
bidirectional_1 (Bidirection	(None, 39, 512)	731136
bidirectional_2 (Bidirection	(None, 256)	656384
dense_1 (Dense)	(None, 64)	16448
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 32)	2080
dropout_2 (Dropout)	(None, 32)	0
dense_3 (Dense)	(None, 2)	66



下圖為參數設置

```
parser.add_argument('--batch_size', default=128, type=float)
parser.add_argument('--nb_epoch', default=20, type=int)
parser.add_argument('--val_ratio', default=0.1, type=float)
parser.add_argument('--gpu_fraction', default=1.0, type=float)
parser.add_argument('--vocab_size', default=None, type=int)
parser.add_argument('--max_length', default=39,type=int)

# model parameter
parser.add_argument('--loss_function', default='binary_crossentropy')
parser.add_argument('--cell', default='LSTM', choices=['LSTM','GRU'])
parser.add_argument('-emb_dim', '--embedding_dim', default=300, type=int)
parser.add_argument('-hid_siz', '--hidden_size', default=512, type=int)
parser.add_argument('--dropout_rate', default=0.2, type=float)
parser.add_argument('-lr','--learning_rate', default=0.001,type=float)
parser.add_argument('--threshold', default=0.08,type=float)
```

下圖為準確率

Private Score	Public Score
0.81901	0.82104

https://blog.csdn.net/lovebyz/article/details/77712003

https://github.com/thtang/ML2017FALL/tree/master/hw4

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：

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：

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：除了有無標點符號以外，未做任何前處理，使用sample code的模型架構來做訓練，發現有標點符號的準確率會比較好，可能是因為標點符號其實也算是文字的一種，而且有時候可能更可以代表人的情緒。下圖上為原本的準確率，下為無標點符號的準確率，左為private score、右為public score，可以看出差距約為1%左右，也是有一定的影響。

0.774	0.775
0.766	0.768

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：未做任何前處理，使用sample code的模型架構以及semi-supervised方法做實驗，將訓練好的模型用在unlabel data上，並設置門檻，決定是否有足夠信心給予標記，再一起加入label data中，繼續原本的訓練，如此反覆，在訓練過程中發現，其實val的準確率並不太會有明顯上升，有可能是因為原本使用label data訓練出的模型就不太好的緣故。下圖上為原本的準確率，下為semi-supervised兩輪後的準確率，左為private score、右為public score，可以看出差距很小。

0.77373	0.77452
0.77661	0.77649