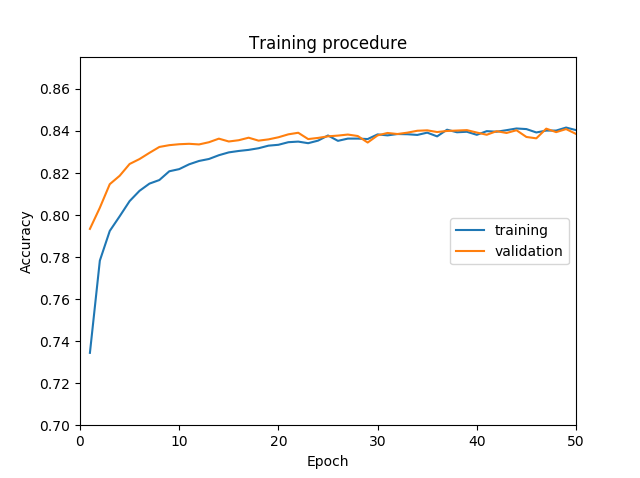
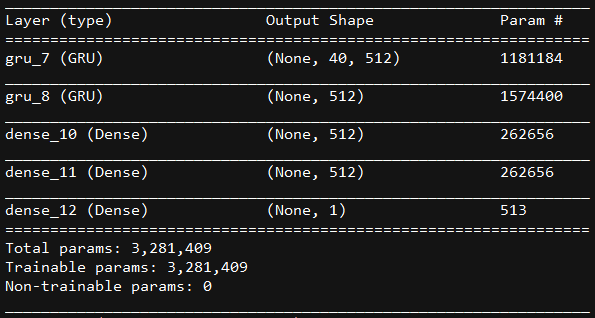
學號：b04901060 系級：電機三 姓名：黃文璁

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: 無) 答：**



首先大致對文字資料進行下列預處理：

1. 只保留 “.?!” 三種標點符號和英文、數字。

2. 處理疊字問題。由於文字來源是Twitter故有許多疊字出現。

3. 處理基本的stemming和詞頻低於一定值的詞。

預處理後利用gensim的word2vec進行word embedding，維度為256。

RNN架構為： 兩層GRU: units=512, dropout=0.5, recurrent\_dropout=0.5

兩層Dense: units=512, activation=‘selu’

訓練時的optimizer為Adam，loss使用binary\_crossentropy。

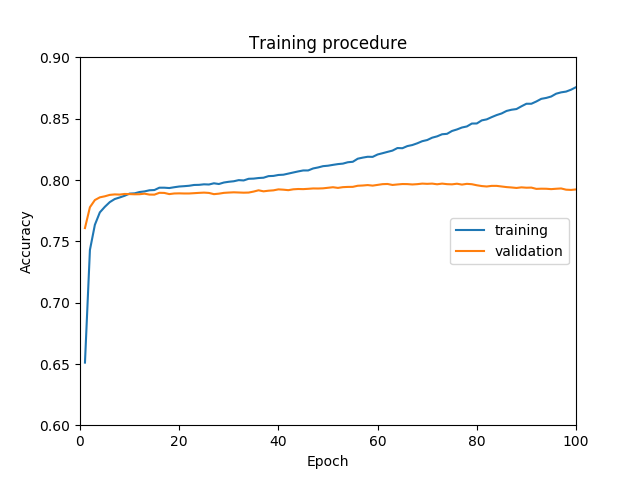
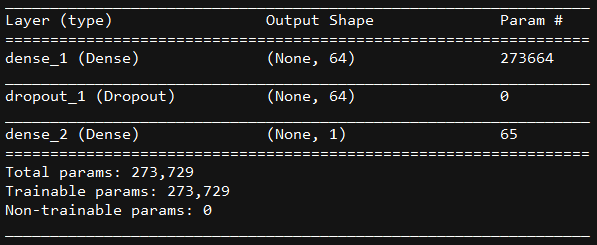
由上圖，training 和validation accuracy大概都在25個epoch時收斂到約0.835。

　　Kaggle上的準確率為： Public: 0.83533 Private: **???????**

此外還實作了8個模型的ensemble，其他模型和上述的模型參數有些不同。

　　Kaggle上的準確率為： Public: 0.83955 Private: **???????**

1. **(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: 無) 答：**



預處理大致同RNN，另外還處理了skipwords，BOW的維度約為4000。

BOW架構為： Dense: units=64, activation=‘relu’

Dropout: rate=0.2

訓練時的optimizer為Adam，loss使用binary\_crossentropy。

由上圖，validation accuracy大概在5個epoch時收斂到約0.79。

另外可以觀察到RNN相比，BOW有更明顯的overfit情況。

　　Kaggle上的準確率為： Public: **???????** Private: **???????**

1. **(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: 無) 答：**

　　就我們看來，第一個句子稍偏負面，而第二個句子則較明確為正面。

Prediction Label

使用RNN來預測兩句話，得到的結果為： 第一句： 0.165759 0

第二句： 0.992874 1

使用BOW來預測兩句話，得到的結果為： 第一句： 0.680686 1

第二句： 0.680686 1

可以觀察到RNN對這兩句話的情緒都較為肯定（預測結果很接近0或1）

而BOW則顯然無法分辨這兩句話的情緒差異，這是由於兩個句子中的詞頻率相同。

對RNN來說，會考慮到前後關係，故應能根據 ‘but’ 的語氣轉折預測出正確的結果。

而BOW則無法判斷前後文，故語氣有轉折的句子就有可能判斷錯誤。

1. **(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: 無) 答：**
2. **(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: 無) 答：**