學號：R08922167 系級：資工碩一 姓名：曾民君

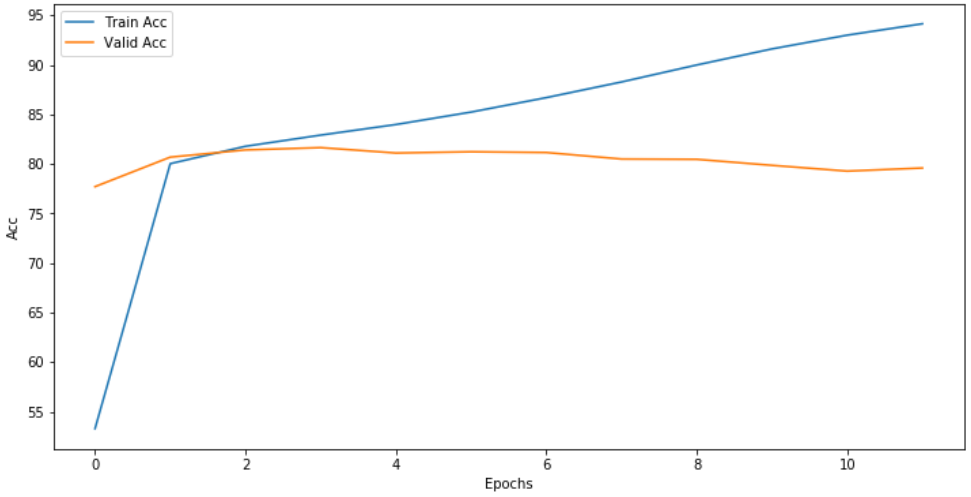
1. (1%) 請說明你實作的RNN的模型架構、word embedding 方法、訓練過程(learning curve)和準確率為何？ (盡量是過public strong baseline的model)

   Ans:
   RNN的模型架構：用兩層 LSTM layers，第一層是bidirectional lstm 搭配 hidden_dimension 為 150，dropout_rate = 0.5，第一層one-directional lstm 搭配 300 的hidden_dimension，最後 classifier 即一層dropout (rate = 0.5)，一層 dense layer 最後再接一層softmax。

   word embedding： word to vector

   訓練參數: sen_len = 40, batch_size = 128, epoch = 12, lr = 0.0015

   最終上傳結果 public 準確率為: 0.82272



2. (2%) 請比較BOW+DNN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"這兩句的分數(過softmax後的數值)，並討論造成差異的原因。

   Ans:

Model	"today is a good day, but it is hot"	"today is hot, but it is a good day"
BOW+DNN	0.4898	0.4898
RNN	0.5261	0.9589

首先 BOW 的含意是某個句子中，我們事先建立的字典中的自個出現幾次，也因為只在乎出現過給，而不在乎出現的先後順序，導致 "today is a good day, but it is hot" 及 "today is hot, but it is a good day" 這兩句話，經 BOW 得到的分數會是一樣。
然而 RNN 則是會考慮每個字詞出現的先後順序，也就是說在語意部份的分析比 BOW 多了一個時間上的維度，所以對於 "today is a good day, but it is hot" 及 "today is hot, but it is a good day" 這兩句話，RNN有能力判別兩者的不同，所以分數才會不一樣。

3. (1%) 請敘述你如何 improve performance（preprocess、embedding、架構等等），並解釋為何這些做法可以使模型進步，並列出準確率與improve前的差異。（semi supervised的部分請在下題回答）

   Ans: 在 preprocess 部份，將 “?”, “.”, “?”, “^” 等符去除，另外將 “ ‘ “ 以及前後文字合併，例如資料中出現的 can ' t ，改成 can't，又或者是，I ' m 合併成 I'm 。以上部分影響不大，大概只會改善 0.1 ~ 0.2% 的 accuracy，model 部分改用 bidirectional lstm，這部分大約能穩定提供 0.5 % 左右的 accuracy。

4. (2%) 請描述你的semi-supervised方法是如何標記label，並比較有無semi-supervised training對準確率的影響並試著探討原因（因為semi-supervise learning 在 labeled training data 數量較少時，比較能夠發揮作用，所以在實作本題時，建議把有 label 的training data從 20 萬筆減少到 2 萬筆以下，在這樣的實驗設定下，比較容易觀察到semi-supervise learning所帶來的幫助）。

   Ans: 將 non-label-training data 餵進事先訓練好的模型，若 output 結果 > 0.8 則加入訓練資料集並賦予其 label 為 1，另一方面若 output 結果 < 0.2 則加入訓練資料集並賦予其 label 為 0，之後再依照擴充後的訓練集重新訓練一個新的 model。實驗結果這次的 semi-supervised learning 會提升大約 0.5 ~ 1% 的準確度。
   其中一個有無使用 semi-supervised 的影響為，沒有使用在 validation accuracy 是 81.594 ，在使用後得到 validation accuracy 為 82.089 的版本。以及其他數次測試 (更換training data 與 validation data 以及不同參數的模型)，都能有一些改善。