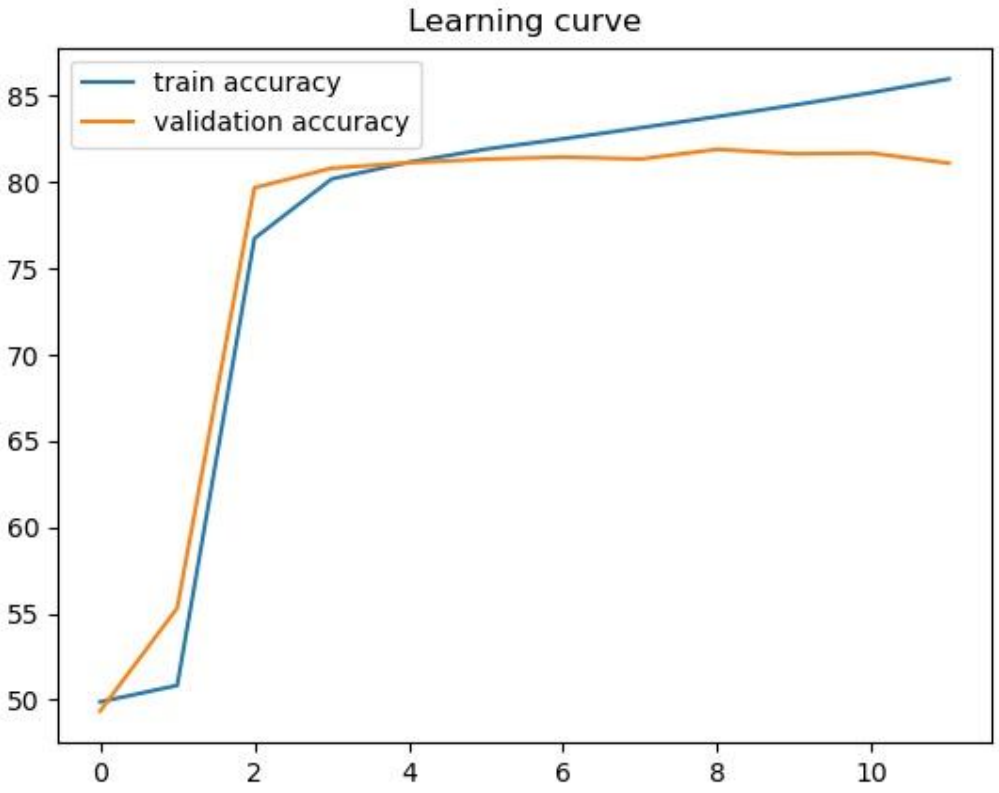
學號：B06902053　　系級：資工三　　姓名：張維哲

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)

   我的RNN是由兩層Bidirectional LSTM再加上兩層神經網路所組成，Embedding size = 300，word embedding是用word2vec實作的，比較特別的是我也將test_data 一同進行embedding。可以看到epoch過多時，有點overfitting的感覺，只有training的有在上升，validation甚至有點下降。
   Training accuracy 可以到85甚至更高，但validation最高則到82左右，過多的epoch表現甚至會變差。



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。

   RNN: [0.6146, 0.9971]
   BOW+DNN :[0.7974, 0.7974]
   RNN 有考慮進順序的問題。而 BOW 只有記錄每個字出現的次數，因此這兩句由相同字組成的話，對 BOW 來說都是一樣，因此兩句話的分數也是一樣。這樣看來RNN比較能考慮上下文，BOW+RNN則是看出現的字彙。

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）

   Preprocessing: 將一個單字內的’ 去除，can’ t->cant. I’ m->im…等，使模型去考慮字跟字之間的關係，以免此符號出現在其他地方，影響到判斷的精準度。

   Embedding:採用word2vec將testing data 一同進行embedding，推測是資料量變大，更能找出字之間的相關程度。

   架構: 將sen_len 調整至40，考慮更長的句子全面分析句子的意義，learning rate 調至0.0005，避免收斂的太快。將Lstm轉成Bilstm，就能考慮到整個句子，而不是只有前文而已。

   在調整前準確率大概80左右，經調整之後可以到82左右。

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）。

   首先用和前面相同的參數進行10個epoch後將最好的model存下來，接著取原本的後20000筆為training data、前20000筆為validation，在下去跑5個epoch後，根據此model下去標記data，threshold為0.99，獲得data後，在下去跑5個epoch，觀察結果。

   尚未加入training data時，validation accuracy 最高為: 79.130

加入semi-supervise learning label的data後， validation accuracy可以來到: 79.359。

推測是因為，semi-supervise learning 所標註的training data在threshold的限制下，這些data都有很明確的分類方向，使得原model原本分不太清楚的那些data，能夠更明確的分辨屬於兩端的哪一邊。