學號：B06901066 系級：電機三 姓名：孟妍

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)

一開始將原本的LSTM改成bidirectional，並把sentence length加長成40就可以過strong baseline，word embedding用genism的word2vec，用所有data（both labeled and unlabeled training data and testing data）一起train。

訓練過程如下：

一張含有 地圖 的圖片

自動產生的描述一張含有 文字, 地圖 的圖片

自動產生的描述

一張含有 地圖 的圖片

自動產生的描述一張含有 文字, 地圖 的圖片

自動產生的描述

validation的accuracy一開始就蠻高的大約0.79左右，最好會到0.82多，但很快就會overfit，所以雖然train 10個epoch但大概最好的結果會第6,7個的時候。 kaggle上public的分數為0.82430。

1. (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。

(1) 使用BOW+DNN，這兩句話的分數會一樣一張含有 鳥 的圖片

自動產生的描述

由上圖，使用BOW+DNN兩句話皆會被歸類為正面，因為BOW是不考慮文法以及詞的順序，這兩句的詞都一樣只是順序有點不同，所以用BOW作為embedding會得到一樣的分數。

(2) 使用RNN，兩者output結果差很多

一張含有 標誌, 相片, 黑色, 街道 的圖片

自動產生的描述

由上圖，"today is a good day, but it is hot" 分數很低會被歸類為負面，而 "today is hot, but it is a good day"分數很高會被歸類為正面，實際來看這樣的結果應該是合理的。跟BOW+DNN會有很大差別是因為RNN有可以由前後文判斷語意能力，因此即使兩句話裡只有詞的順序不同，通過softmax後的分數就可能有很大不同，像這裡兩句話就得到完全相反的結果。

1. (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）

一開始有做各種preprocess，有用一些sentiment analysis的方法，例如移掉標點符號，移掉一些stop words 之類的但做了很多preprocess去測試結果都比較差，感覺可能因為每句都蠻短的，所以標點符號也算是佔了判斷正負很重要的部分。所以後來就沒有做preprocess，先把LSTM改成bidirectional，因為改成bidirectional，所以也同時把sentence length加長，這樣就可以過strong baseline。改成bidirectional效果不錯的原因是因為這樣可以不只從，前文推導關係，也可以由後文推導，可以學得更完整。之後再把LSTM換成GRU，accuracy跟kaggle的分數都有再上升一些到0.82553。

1. (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）。

我semi-supervised設定 pos\_threshold = 0.85，取前2萬筆training data做training，再去predict unlabeled data的label，然後加入新預測的較有信心的data(>0.85 or <0.15)一起做training。只用20000筆labeled data去train，validation accuracy大概只能到77-78％左右。

一張含有 文字 的圖片

自動產生的描述

[training with 20000 labeled data]

一張含有 坐, 黑色 的圖片

自動產生的描述

[self training with pos\_threshold 0.85]

(忽略Epoch旁邊的loss和accuracy，不知道為什麼印出怪怪的東西QQ)

加入新data後training accuracy變很高，loss很低，train 10個epoch loss有微微下降，training accuracy 一開始就蠻高的所以也是微微上升。在沒有做semi-supervised training之下其實validation accuracy上升幅度不大，但做了semi-supervised training validation accuracy還是跟本來差不多，loss還變大。可能因為labeled跟unlabeled data數量真的差很多，原本的model accuracy也不是非常高，或是threshold設的不夠高，導致predict出的label錯誤的數量不少，無法fit在有ground truth的unseen data上。