Machine Learning HW5 Report

學號：B05705006　系級：資管四　　姓名：李和維

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*

模型架構：

先疊三層 Embedding dimension 為 200、hidden size 為 96 的單向 LSTM（除了最後一層外皆使用 dropout = 0.8），然後再連接兩層 Fully connected layer 把 96 維降至 32 維、2維，最後輸出。

Word embedding 方法：

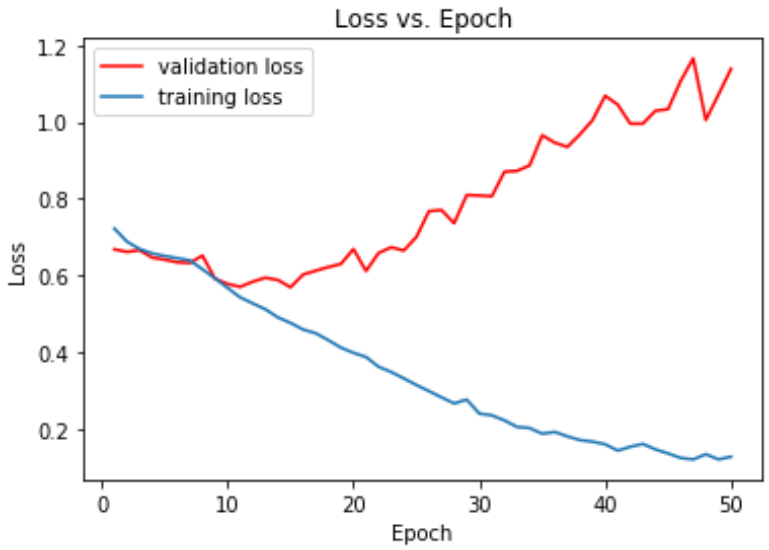
先用 training and testing data 所有的 comments train 出一個 gensim Word2Vec model，然後使用 Spacy 套件的 "en_core_web_sm" 模型對 input 進行斷詞，並將每個 word 丟入 Word2Vec model 得到 200 維的 embedding，並以此 embedding vector 當作 RNN model 的輸入。

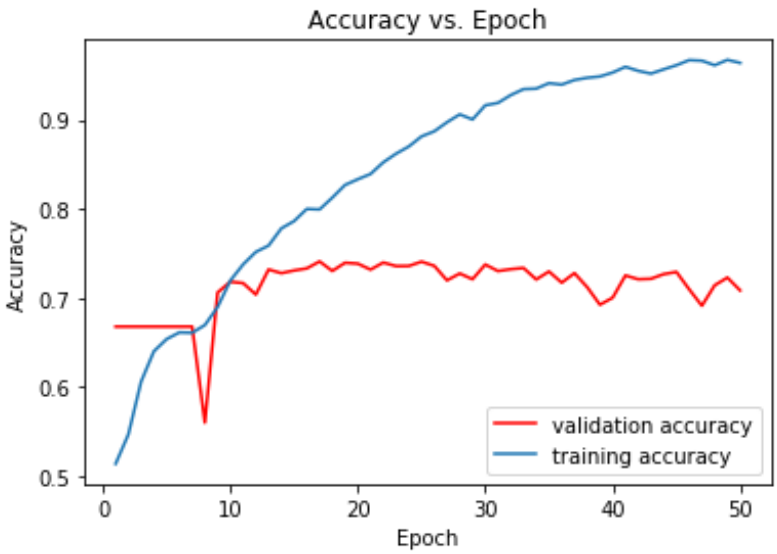
模型正確率：

Public score: 0.78372

Private score: 0.80000

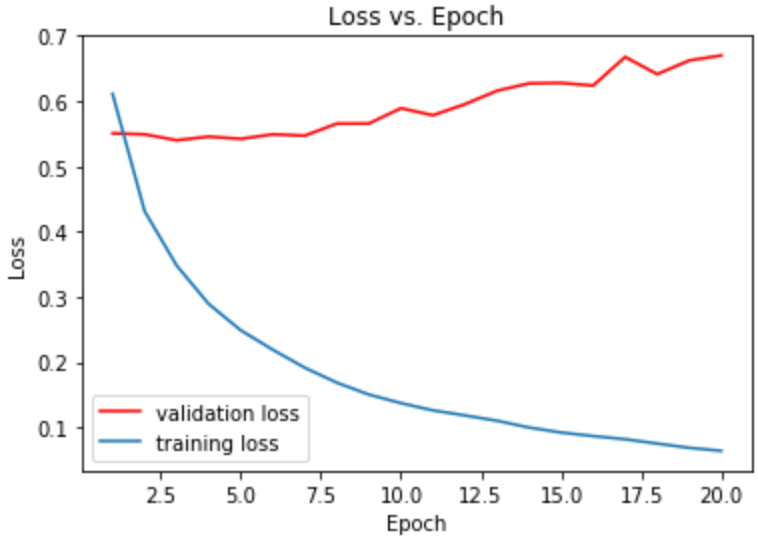
訓練曲線：

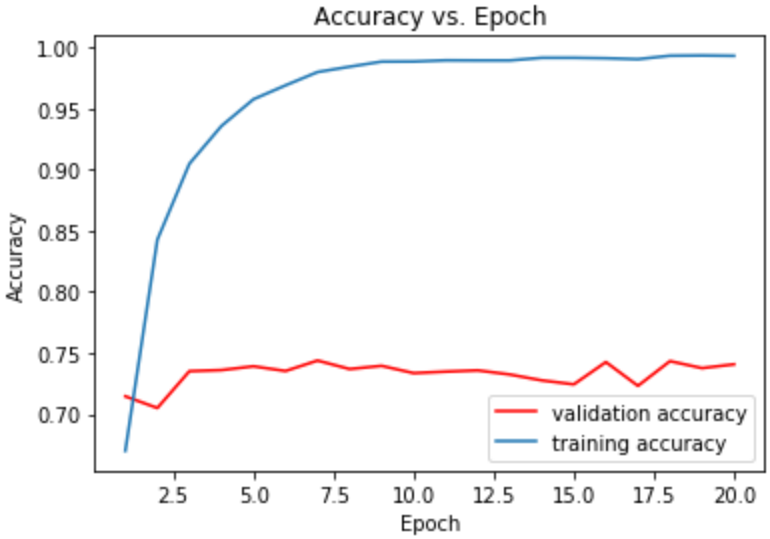




2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

模型架構：
把每筆 comments 轉換成 20388 維的 bag of words vector，傳入 model 後經過三層 fully connected layers，並分別都套上 BatchNorm1d 與 LeakyReLU 再輸出。三層 Linear 把 input vectors 分別降維至 1024, 128 與 2 維，最後比較二維輸出的值做 classification。
模型正確率：
Public score: 0.76279
Private score: 0.80697
訓練曲線：





3. (1%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。

1. Orthogonal initialization

Orthogonal initialization 使得 lstm 上的 weight matrix 全部的 eigenvalues 的絕對值皆為 1，這使得在 training 過程中無論經過多少次 muptiplication 都不會有 exploding 或 vanishing 的問題。

2. 降低 embedding dimension 與 hidden size

Training 過程中大約在 10 個 epoch 以後，training loss 持續下降而 validation loss 卻持續上升，發現嚴重的 overfitting 問題，降維能減低模型複雜度並提升 model generalize 的能力。

3. Early stopping

在 validation loss 停止下降後就停止 training 以避免 overfitting。

4. 提升 Dropout rate

同樣是避免 overfitting，達到 regularization 的效果。

5. 降低 learning rate

使 loss 收斂的更好。

4. (1%) 請比較不做斷詞 (e.g.,用空白分開) 與有做斷詞，兩種方法實作出來的效果差異，並解釋為何有此差別。

	不做斷詞（用空白分開）	用 Spacy 斷詞
Public Score	0.72325	0.78372
Private Score	0.74186	0.80000

對 comments 用 Spacy model 斷詞後 train 出來的 Word2Vec model 效果明顯比直接用 split 斷詞好。會造成此差異的主要原因是不做斷詞的情況下只要單字後面加上任何標點符號或表情符號，就會被視為不同的字，如此一來 train Word2Vec 時的 input 就會非常 sparse，所以容易 train 不好。另外在不做斷詞的情況下不同標點符號就視為不同 embedding 也顯然與直覺不符。

舉例來說，"I feel good" 、 "I feel good!" 、"I feel good." 這三句當中的 good 在不做斷詞的情況下在 Word2Vec model 會被當成不同 word 下去 train，而在有做斷詞的情況下會被拆成 ["good"]、["good", "!"]、["good", "."]，顯然後者對於訓練 embedding 的效果更好也更能正確理解語意。

5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy." 與"I am happy, but today is hot." 這兩句話的分數（model output），並討論造成差異的原因。

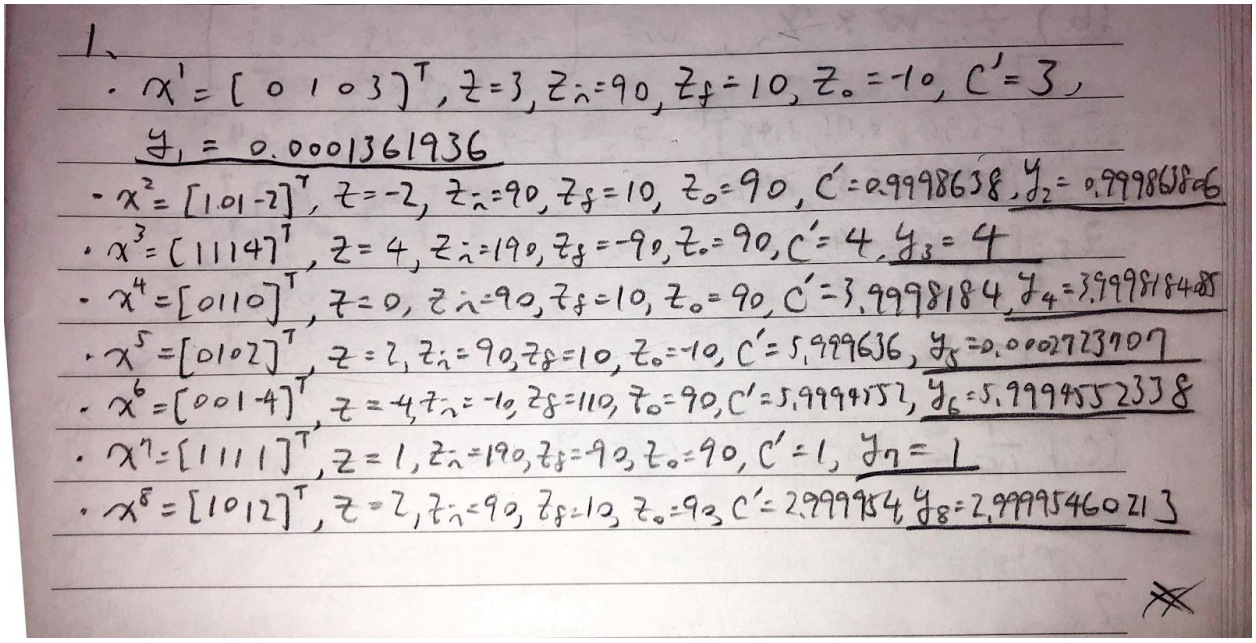
	Today is hot, but I am happy.	I am happy, but today is hot.
RNN	[0.73227582 0.26772418]	[0.71248831 0.28751169]
BOW	[0.90930679 0.09069321]	[0.90930679 0.09069321]

（第一維代表預測為非攻擊性留言的 softmax 機率，第二維代表預測為攻擊性留言的 softmax 機率。）

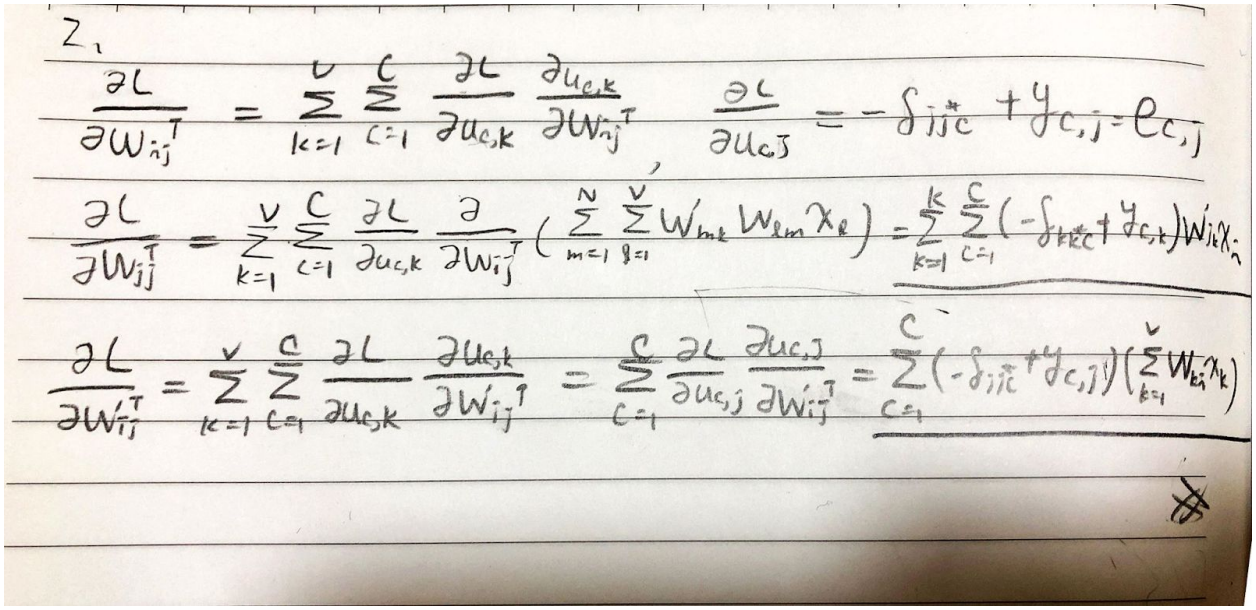
造成差異的原因為 RNN 會因 word 出現的時間順序不同而被 memory 影響導致產生不同 output；而 BOW 是直接把每個 word 出現的次數加總，所以對於這兩句話會有相同的 output。

Math Problems

LSTM Cell (1%)



Word Embedding (1%)



參考資料:
http://www.claudiobellei.com/2018/01/06/backprop-word2vec/?fbclid=IwAR0wrgIHAl2FpM_pJgMbGEzI31EBtLkxrWlBSXfg0X7bSnANOVESx1RqXbk