ML 2017 Fall HW4 Report

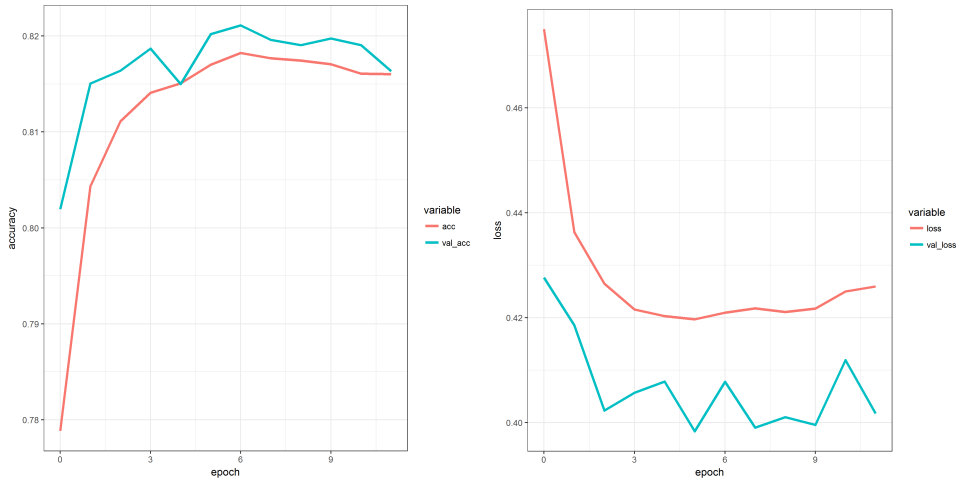
學號：D05921027 系級： 電機博一 姓名：張鈞閔

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: None)

答：使用 3 層 bidirectional LSTM [256, 256, 128] 後接 1 層 fully connected layer 後接 output layer (2 classes)。Dropout = 0.3。

```
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 39)                0
_________________________________________________________________
embedding_1 (Embedding)      (None, 39, 100)           24990500
_________________________________________________________________
bidirectional_1 (Bidirection (None, 39, 256)           234496
_________________________________________________________________
bidirectional_2 (Bidirection (None, 39, 256)           394240
_________________________________________________________________
bidirectional_3 (Bidirection (None, 128)               164352
_________________________________________________________________
dense_1 (Dense)              (None, 64)                8256
_________________________________________________________________
dropout_1 (Dropout)          (None, 64)                0
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 130
=================================================================
Total params: 25,791,974
Trainable params: 801,474
Non-trainable params: 24,990,500
```

訓練跟驗證資料比例為 9:1，訓練資料有 18 萬筆。下方左圖為 accuracy 右圖為 loss 隨著 epoch 的變化。 紅: training 綠: validation。採用 early stopping (patience=5)，最好 val_loss 表現在第 6 個 epoch = 0.82105。



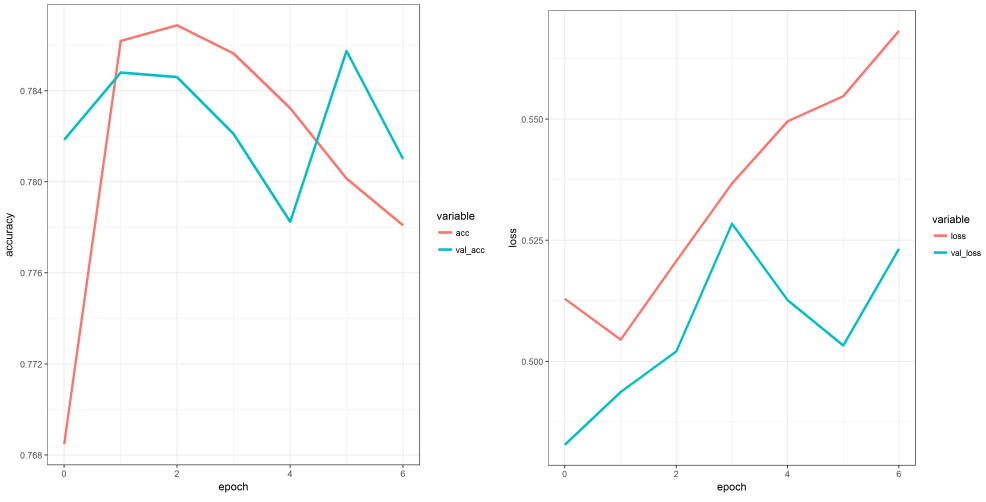
2. (1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？ (Collaborators: None)

答：為了避免 input dimension 太大而使得模型訓練不易，選擇 top 10000 最常出現的字建立字典，並且將用 word count 當作 BOW 的表示。和 RNN 使用相同的層數 (3 層 hidden layers) 與單位數量 (256, 256, 128)，只是從 LSTM 改為 fully connected layer。Dropout = 0.3 亦同，但是很快就 overfitting。

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 10000)	0
dense_1 (Dense)	(None, 256)	2560256
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32896
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8256
dropout_4 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 2)	130

Total params: 2,667,330
Trainable params: 2,667,330
Non-trainable params: 0

下方左圖為 accuracy 右圖為 loss 隨 epoch 的變化。 紅: training 綠: validation



3. (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: None)

答：下方顯示這兩個句子得到的情緒分數不同。另外，比較使用 **but** 和 **and** 發現，以 **and** 為連接詞不會造成情緒分數翻轉，由此可知 RNN 可以學習到 **but** 後面的句子才是主題句。

RNN	0	1
today is a good day , **but** it is hot	**0.74350005**	0.25649992
today is hot , **but** it is a good day	0.00219005	**0.99780995**
today is a good day , **and** it is hot	0.16622847	**0.83377153**
today is hot , **and** it is a good day	0.00262943	**0.99737060**

因為 BOW 沒有掌握字詞順序，兩個句子的情緒分數一模一樣，失去此順序後對於情緒分數的預測也較差，例如：對於 0,1 的預測值較接近 0.5，相較於 RNN 模型可能會達 0.99 的預測值，BOW 難有較高的預測值。

BOW	0	1
today is a good day , **but** it is hot	0.34811926	**0.65188074**
today is hot , **but** it is a good day	0.34811926	**0.65188074**
today is a good day , **and** it is hot	0.26126096	**0.73873907**
today is hot , **and** it is a good day	0.26126096	**0.73873907**

4. (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: None)

答：在訓練 word2vec 模型時，會把標點符號當作 unique token。因此在此比較，只在 testing 時是否去除掉標點符號的 tokenizer 使用。比較結果如下：

	Public score	Private score
有標點符號	**0.81962**	**0.81984**
無標點符號	0.81625	0.81714

有標點符號的 tokenize 方式會得到較佳的表現，但影響相對不大(約 0.2-0.3%)

5. (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: None)

答：在實作 semi-supervised 方法時，是將預測 class probability 大於 0.9 直接視為該 class，例如: today is hot , but it is a good day 預測 class 1 機率為 0.9978 (> 0.9) 故將此句子加入 training dataset 中，並標註為 class 1。
先使用 early stopping (patience = 5) 訓練至收斂後才進行 semi-supervised learning，僅嘗試加入兩次資料。
發現加入 semi-supervised learning 後對於結果沒有太顯著的影響，並且加入的品質參差不齊，對於 performance 的影響時好時壞，可能在 criteria 上面需要多加設計才有辦法穩定的成長。結果比較如下：

	Original	Semi 1 (add 48596)	Semi 2 (add 36722)
Private	0.80397	0.80694	0.80276
Public	0.80476	0.80682	0.80533