Homework 5 : Text Sentiment Classification

學號：b05902008　系級：資工二　　姓名：王行健

1. **(1%)請說明你實作的RNN model, 其模型架構，訓練過程和準確率?**

Data preprocessing (Handcrafted cleaning, remove duplicate char, stemmer)
|
W2V word embedding (embed_size 50, min count 1, window size 5)
| pad to length 30
biderectional LSTM*2 (hidden_size 100, dropout 0.2)
|
dense (200,100)
| dropout 0.5, activation swish, batchnorm
dense (100,64)
| dropout 0.5, activation swish
dense (64,2)
| softmax
CrossEntropyLoss

**Early stopping applied
batch size : 100
learning rate : 1e-4

Vald Acc : 0.82~0.84

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?**

Data preprocessing (Handcrafted cleaning, remove duplicate char, stemmer)
|
W2V word embedding (embed_size 50, min count 1, window size 5)
|
Average of word vectors
|
dense (50,100)
| dropout 0.5, activation swish, batchnorm
dense (100,64)
| dropout 0.5, activation swish
dense (64,2)
| softmax
CrossEntropyLoss

**Early stopping applied
batch size : 100
learning rate : 1e-4

Vald Acc : 0.75~0.77

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。**

RNN : Sentence1: [0.6428 , 0.3572]
Sentence2: [0.2922, 0. 7078]
BOW : Sentence1: [0.2132, 0.7868]
Sentence2: [0.2132, 0.7868]

RNN較有能力處理序列性的資料，換句話說，句中but的前後語句兌換可以被注意到，BOW只能區分句子中有哪些詞，因此語句順序完全不被考慮

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。**

   有標點符號：Vald Acc 0.831
   無標點符號：Vald Acc 0.838

   兩者並沒有顯著差別（無標點符號稍微好一點），推測是每個人使用的習慣差異太大，使其指標性不足

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。**

   labeled_train_data early stop
   |
   add in non_labeled_train_data with certainty 0.7<P<0.8
   |
   repeat process for 2 rounds

   Vald Acc 0.82~0.84
   之所以只softmax後其中一項0.8<P<0.9的資料是考慮到取太低等同於加入雜訊，太高則無法幫助模型繼續學習，然而整體看下來並沒有對結果造成顯著影響，有可能需要實驗更多種threshold才能得到好的結果