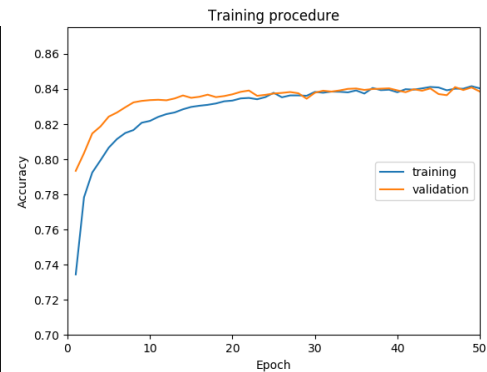


學號：b04901060 系級：電機三 姓名：黃文璽

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: 無) 答：

Layer (type)	Output Shape	Param #
gru_7 (GRU)	(None, 40, 512)	1181184
gru_8 (GRU)	(None, 512)	1574400
dense_10 (Dense)	(None, 512)	262656
dense_11 (Dense)	(None, 512)	262656
dense_12 (Dense)	(None, 1)	513
Total params: 3,281,409		
Trainable params: 3,281,409		
Non-trainable params: 0		



首先大致對文字資料進行下列預處理：

1. 只保留“?! ”三種標點符號和英文、數字。
2. 處理疊字問題。由於文字來源是 Twitter 故有許多疊字出現。
3. 處理基本的 stemming 和詞頻低於一定值的詞。

預處理後利用 gensim 的 word2vec 進行 word embedding，維度為 256。

RNN 架構為： 兩層 GRU: units=512, dropout=0.5, recurrent_dropout=0.5

兩層 Dense: units=512, activation='selu'

訓練時的 optimizer 為 Adam，loss 使用 binary_crossentropy。

由上圖，training 和 validation accuracy 大概都在 25 個 epoch 時收斂到約 0.835。

Kaggle 上的準確率為： Public: 0.83533 Private: ????????

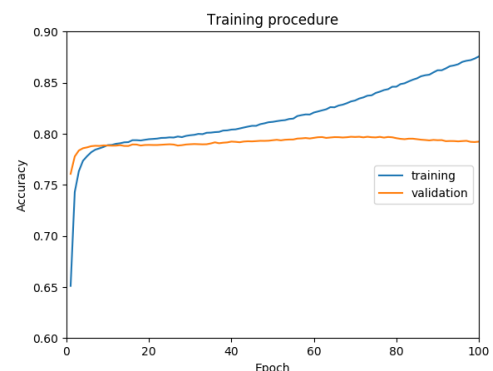
此外還實作了 8 個模型的 ensemble，其他模型和上述的模型參數有些不同。

Kaggle 上的準確率為： Public: 0.83955 Private: ????????

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: 無) 答：

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 64)	273664
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 1)	65
Total params: 273,729		
Trainable params: 273,729		
Non-trainable params: 0		



預處理大致同 RNN，另外還處理了 skipwords，BOW 的維度約為 4000。

BOW 架構為： Dense: units=64, activation='relu'

Dropout: rate=0.2

訓練時的 optimizer 為 Adam，loss 使用 binary_crossentropy。

由上圖，validation accuracy 大概在 5 個 epoch 時收斂到約 0.79。

另外可以觀察到 RNN 相比，BOW 有更明顯的 overfit 情況。

Kaggle 上的準確率為： Public: ??????? Private: ???????

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: 無) 答：

就我們看來，第一個句子稍偏負面，而第二個句子則較明確為正面。

		Prediction	Label
使用 RNN 來預測兩句話，得到的結果為：	第一句：	0.165759	0
	第二句：	0.992874	1
使用 BOW 來預測兩句話，得到的結果為：	第一句：	0.680686	1
	第二句：	0.680686	1

可以觀察到 RNN 對這兩句話的情緒都較為肯定（預測結果很接近 0 或 1）

而 BOW 則顯然無法分辨這兩句話的情緒差異，這是由於兩個句子中的詞頻率相同。對 RNN 來說，會考慮到前後關係，故應能根據 'but' 的語氣轉折預測出正確的結果。而 BOW 則無法判斷前後文，故語氣有轉折的句子就有可能判斷錯誤。

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: 無) 答：

5. (1%) 請描述在你的 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。

(Collaborators: 無) 答：