學號：R05921120 系級： 電機碩二 姓名：黃浩恩

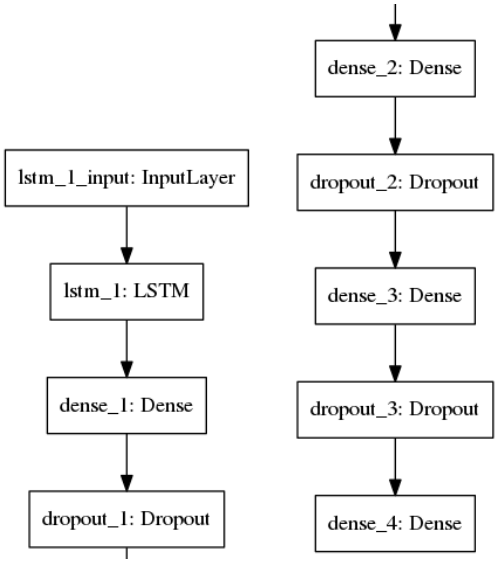
1. **(1%) 請說明你實作的 RNN model，其模型架構、訓練過程和準確率為何？ (Collaborators: 歐靖 R06921089、賴棹元 D05921011)**

答：

模型架構：

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 512)	1665024
dense_1 (Dense)	(None, 256)	131328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 256)	65792
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	257

Total params: 1,928,193
Trainable params: 1,928,193
Non-trainable params: 0



訓練過程：

先透過 gensim 做出字典(skip-gram)，其中 pre-train 的資料來自 no-label 之 data set，並透過該字典先將文章句子轉換成 vector，之後直接味進 LSTM，進行訓練。

使用 epoch = 50，optimizer = adam，loss fuction = binary_crossentropy

準確率：

Kaggle score	**Public**	**Private**
Point	**0.82648**	**0.82539**

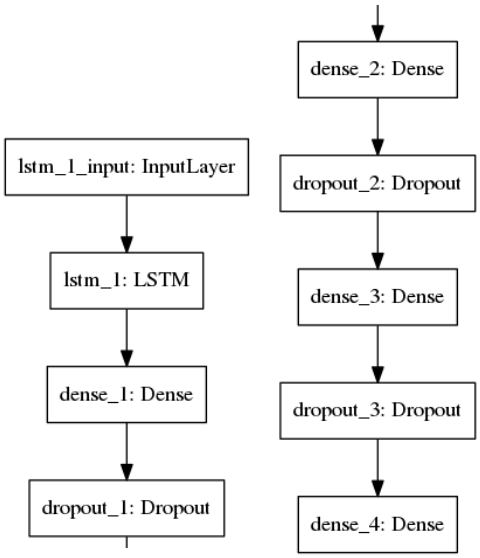
2. **(1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？ (Collaborators: 歐靖 R06921089、賴棹元 D05921011)**

答：

模型架構：

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 512)	1665024
dense_1 (Dense)	(None, 256)	131328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 256)	65792
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 256)	65792
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	257

Total params: 1,928,193
Trainable params: 1,928,193
Non-trainable params: 0



訓練過程：

將透過 gensim 做出字典之方法進行改變，使得產生字典的型態變成 BOW，並透過該字典先將文章句子轉換成 vector，之後直接味進 LSTM，進行訓練，故神經模型架構不變。

使用 epoch = 20，optimizer = adam，loss fuction = binary_crossentropy

準確率：

Kaggle score	**Public**	**Private**
Point	**0.81555**	**0.81347**

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: 歐靖 R06921089、賴棹元 D05921011)
   答：

	LSTM	**BOW**
today is a good day, but it is hot	**0.09302395**	**0.14945504**
today is hot, but it is a good day	**0.99076968**	**0.99829847**

討論：

Word2Vec 模型中，主要有 Skip-Gram 和 CBOW 兩種模型， Skip-Gram 是給定 input word 來預測上下文。而 CBOW 是給定上下文，來預測 input word。

圖 CBOW

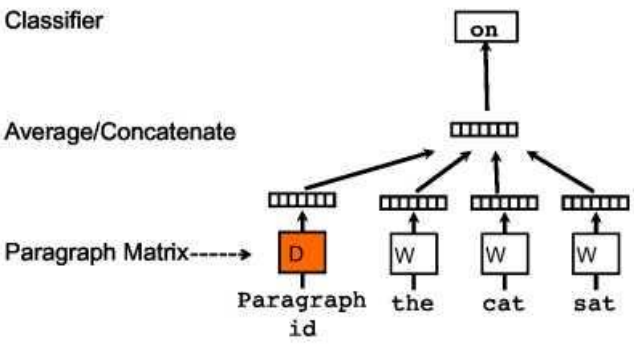
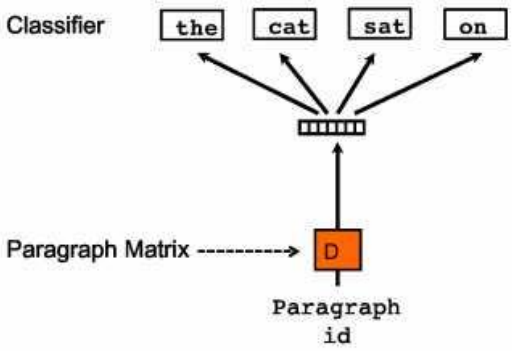


圖 Skip-Gram



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: 歐靖 R06921089、賴棹元 D05921011)
   答：

	Kaggle Public
「有」標點符號之 RNN	**0.82648**
「沒」標點符號之 RNN	**0.79095**

討論：

於英文句子內，標點符號可能扮演著邏輯分段之重要角色，使得在有標點符號之 RNN 預測比較準確，而沒有標點符號之 RNN，可能使某些易混淆的句子判斷錯誤。

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: 歐靖 R06921089、賴棹元 D05921011)
   答：

	Kaggle Public
「無」semi-surpervised training	**0.82649**
「有」semi-surpervised training	**0.83255**

討論：

設計透過一開始的 RNN，對 no-label 做預測，將預測機率大於 0.8 結果設為 1，且將預測機率小於 0.2 結果設為 0，再將這些字句存入原先之 training data，重複 training，再製作出新 model。其結果明顯上升，透過 training set 變多，讓 model 受更多訓練，達到預測能力上升之結果。