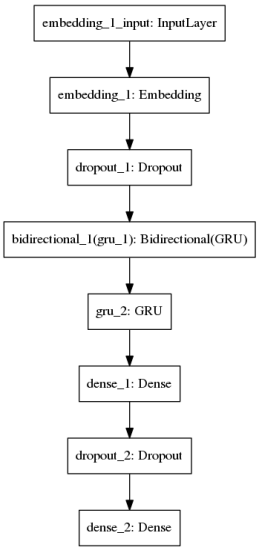
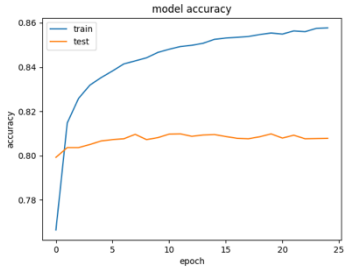
學號：R06944049 系級： 網媒碩一 姓名：黃敬庭

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: 吳政軒 R06922118, 倪溥辰 R06944032)

答：

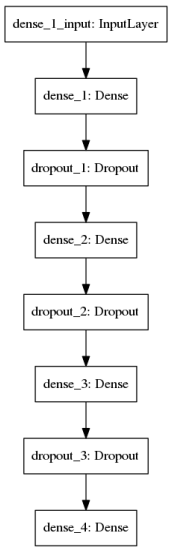


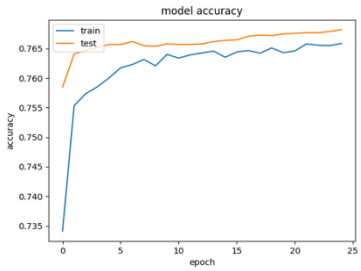




我的 model 依序是 word embedding, gru(bidirectional), gru ,dense, dense, 然後在各層都有使用 dropout(0.5)以防止 overfitting, batch size 為 128, optimizer 使用 adam(learning rate 為 0.01), 並且在 validation accuracy 沒有進步的時候會將 learning rate 減半 ,若在沒有進步就會直接停止訓練, 使用 label data 的最後 10000 筆作為 validation data, 最後在 validation set 上面拿到 80.9(圖片裡 test 就是 validation)左右的成績, 上傳 kaggle 之後拿到 81.1

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: 吳政軒 R06922118, 倪溥辰 R06944032)

答：




model 為 bow 的資料直接丟入四層 dense layer(256, 128, 64, 2)裡面, 並在前三層後面都加上 dropout(0.25), 訓練過程和 rnn 的 model 基本相同, 都利用減半 learning rate 和 early stopping 的方式, 最後 validation 的成績大概 76.5, 上傳 kaggle 之後拿到 76

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: 吳政軒 R06922118, 倪溥辰 R06944032)

答：

bow 對兩句判斷出來的結果是一樣的, 因為他並不考慮字的先後順序並且對兩句都判斷為正面, 我推測可能是因為看到了 good, 而 rnn 對兩句則能正確的判斷出來, 所以字的先後順序造成的語意轉折在這次作業看來是有很大的影響

| rnn | 第一句是正面機率 | 第二句是正面機率 |
|---|---|---|
| | 0.37423262 | 0.98545432 |
| bow | 第一句是正面機率 | 第二句是正面機率 |
| | 0.73488575 | 0.73488575 |

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: 吳政軒 R06922118, 倪溥辰 R06944032)
   答：

整體來講如果在 tokenize 的時候將除了\n\t 濾掉的話, 在只考慮 supervised 的 model 的情況下效能會變差大概 1 到 0.5(81 vs 80), 造成的原因可能是因為有些標點符號對整體的語義影響(例如驚嘆號或問號之類的)也會對句子的架構造成影響, 所以包含標點符號的 model 表現會比較好

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: 吳政軒 R06922118, 倪溥辰 R06944032)
   答：

我先用一個準確率 81 的 model 替沒有 label 的資料加上 label, 然後選用 softmax 之後機率達到 85(對正面和負面分別挑)以上的 data 加入原本就有的 training data 一起去訓練, 讓總體的 data 數量大概達到 90 萬多筆左右, 加入之後準確率大概從 81 左右上升到 81.8, 在開始 semi-supervised 的 training 之前會將 model 重置, 其餘的 training procedure 和第一題差不多