學號：R06922128 系級： 資工碩一 姓名：楊碩碉

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？

答：

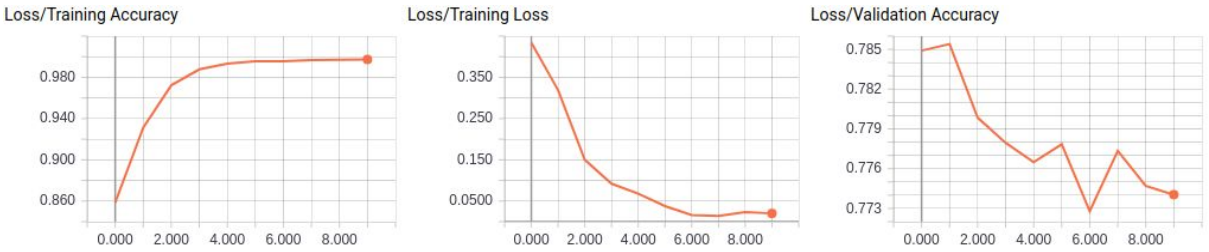

```
input_x = tf.placeholder(tf.float32, [None, 39, 128])
input_y = tf.placeholder(tf.int64, [None])
keep_prob = tf.placeholder(tf.float32)

lstm_cell = tf.contrib.rnn.BasicLSTMCell(128)
lstm_cell = tf.contrib.rnn.DropoutWrapper(lstm_cell, output_keep_prob=keep_prob)
state = lstm_cell.zero_state(256, tf.float32)

rnn_input = tf.nn.dropout(input_x,keep_prob)

with tf.variable_scope('LSTM'):
    for t in range(39):
        if t>0: tf.get_variable_scope().reuse_variables()
        (rnn_output, state) = lstm_cell(rnn_input[:, t, :], state)

fc1 = fc(rnn_output, [39, 1024], tf.nn.relu)
fc2 = fc(fc1, [1024, 2])
```

模型架構如上圖，輸入部份把每一個句子轉為一串id，再把id轉成embedding餵進去network。訓練embedding的資料包含label、unlabel跟testing，採用skip-gram。最後的kaggle準確率為0.82多一點。(kaggle傳的是ensemble版本)

訓練過程採用AdamOptimizer，learning rate固定為1e-3，dropout rate=0.5，10個epoch，loss function為cross entropy。

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？

答：




```
input_x = tf.placeholder(tf.float32, [None, dict_size])
input_y = tf.placeholder(tf.int64, [None])

fc1 = fc(self.input_x, [dict_size, 1024], tf.nn.relu)
fc2 = fc(fc1, [1024, 2])
```

BOW model只用2層fc，可以看到出現了over fitting的現象，kaggle準確率為0.77。

訓練過程採用AdamOptimizer，learning rate固定為1e-3，沒有dropout rate，10個epoch。

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。

BOW:0.939, 0.997　RNN:0.337, 0.965

這2句話在BOW裡面差別極小，可推測BOW分不太出來。RNN model則可以明顯分別出差異，由此觀察出單字的順序對語意是有影響的。

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。

答：本題比較基準為第一題的model。再有標點的情況下準確率為0.82，沒有標點的準確率則為0.805，可以看出tokenizer對RNN影響是很大的。因此合理推測，標點符號例如~!@#$%^&*()_+<>?,.等等對於句子的語意預測是不可或缺的。另一方面，從人類的觀點來看，書寫時「！」「？」常常會加強句子的語氣「，」用做斷句，若缺少這些符號，對於人類的辨識也會有影響。

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。

semi-supervised訓練方法採用self training，threshold定在0.995(定太小如0.99會造成一次新增的量太多)。在每一次計算threshold時把大於threshold的標上label，若小於則去掉該筆資料的label。但是從結果看來做semi-supervise learning對結果沒有變好，有

時候在validation set上甚至變差。因此unlabel data就只有在訓練embedding時才有用到