

## Machine Learning HW5 Report

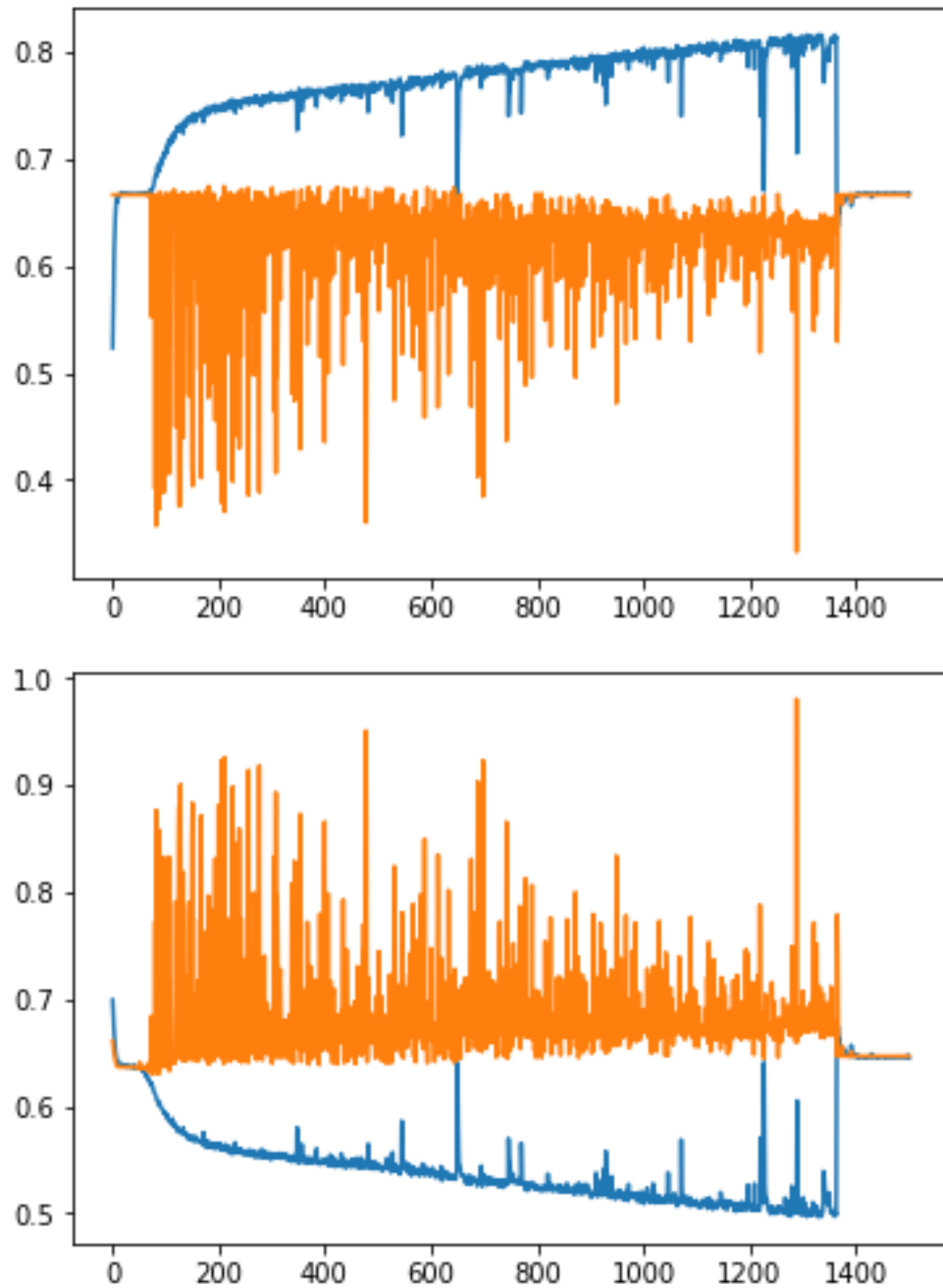
學號：B05702095 系級：會計四 姓名：黃禹翔

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線\*

我的模型是採用 GRU，word-embedding 利用所有的推文搭配 word2vec 去訓練，200 維。GRU 模型及分類器架構如下，學習率為 0.0002，訓練了 1500 個 epoch：

```
GRUmodel(  
  (gru): GRU(200, 200, num_layers=2, batch_first=True, dropout=0.3, bidirectional=True)  
  (classifier): Sequential(  
    (0): Linear(in_features=200, out_features=512, bias=True)  
    (1): Dropout(p=0.5, inplace=False)  
    (2): ReLU()  
    (3): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)  
    (4): Linear(in_features=512, out_features=128, bias=True)  
    (5): Dropout(p=0.5, inplace=False)  
    (6): ReLU()  
    (7): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)  
    (8): Linear(in_features=128, out_features=32, bias=True)  
    (9): Dropout(p=0.5, inplace=False)  
    (10): ReLU()  
    (11): BatchNorm1d(32, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)  
    (12): Linear(in_features=32, out_features=2, bias=True)  
    (13): Softmax(dim=None)  
  )  
)
```

最後 Kaggle 的分數是 0.70465，連 simple baseline 都沒有過。訓練曲線如下：其中藍色的線為 training set 的表現，橘色線為 validation set 的表現。上面的圖縱軸為 accuracy、下面的圖縱軸為 cross entropy loss，橫軸皆為 epoch。可以看到，模型一直都有學到東西，但表現進步得很緩慢，且時不時會出現跳動的情況，並不是很穩定的在收斂。訓練的準確率最後落在 0.80 左右，但驗證集的表現始終沒有提升，一直在 0.68 附近。



2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線\*。

我先對資料去除停用字以及符號，再建立 BOW 分類模型，模型的架構如下：

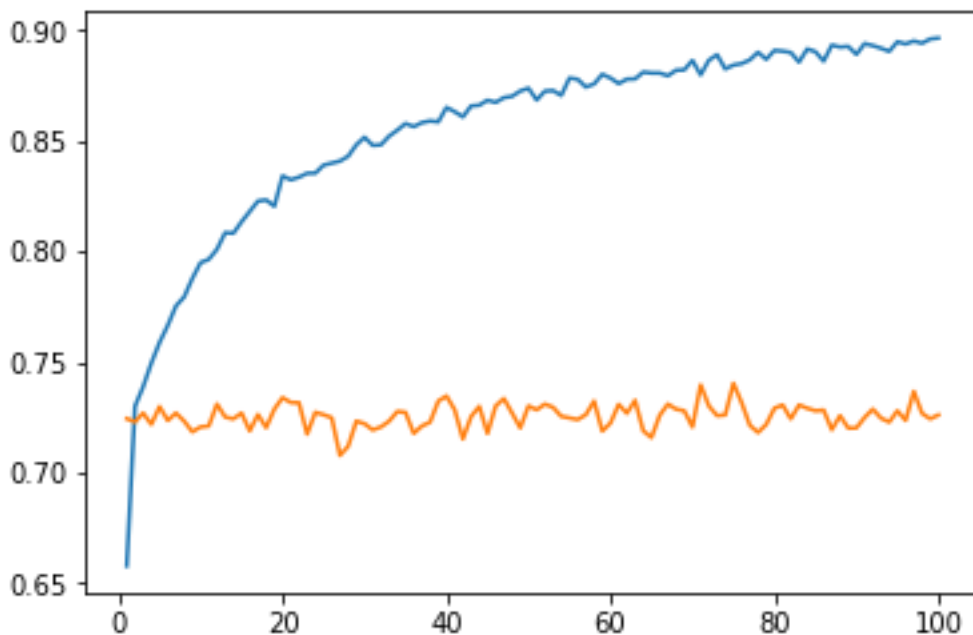
```

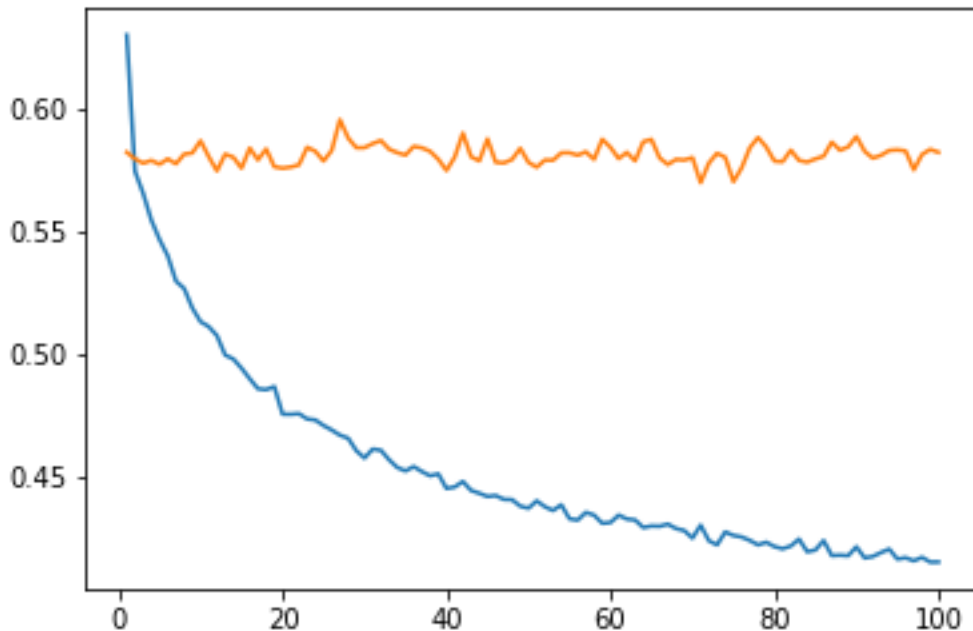
BOWmodel(
  (classifier): Sequential(
    (0): Linear(in_features=17693, out_features=2048, bias=True)
    (1): Dropout(p=0.5, inplace=False)
    (2): ReLU()
    (3): BatchNorm1d(2048, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (4): Linear(in_features=2048, out_features=512, bias=True)
    (5): Dropout(p=0.5, inplace=False)
    (6): ReLU()
    (7): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (8): Linear(in_features=512, out_features=128, bias=True)
    (9): Dropout(p=0.5, inplace=False)
    (10): ReLU()
    (11): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    (12): Linear(in_features=128, out_features=2, bias=True)
    (13): Softmax(dim=None)
  )
)

```

訓練曲線如下：

上圖為 **accuracy**、下圖為 **loss**；藍線為訓練集、橘線為驗證集。





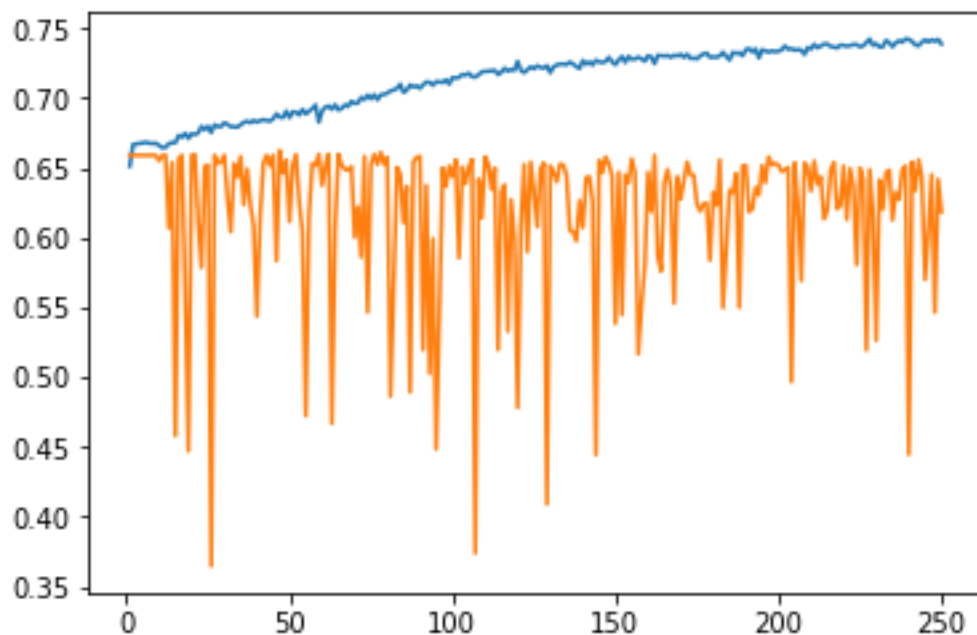
可以看到，訓練集上的表現算是相當不錯，但驗證集的表現一直上不来。即使加了 Dropout 也沒有變好。

3. (1%) 請敘述你如何 improve performance (preprocess, embedding, 架構等)，並解釋為何這些做法可以使模型進步。

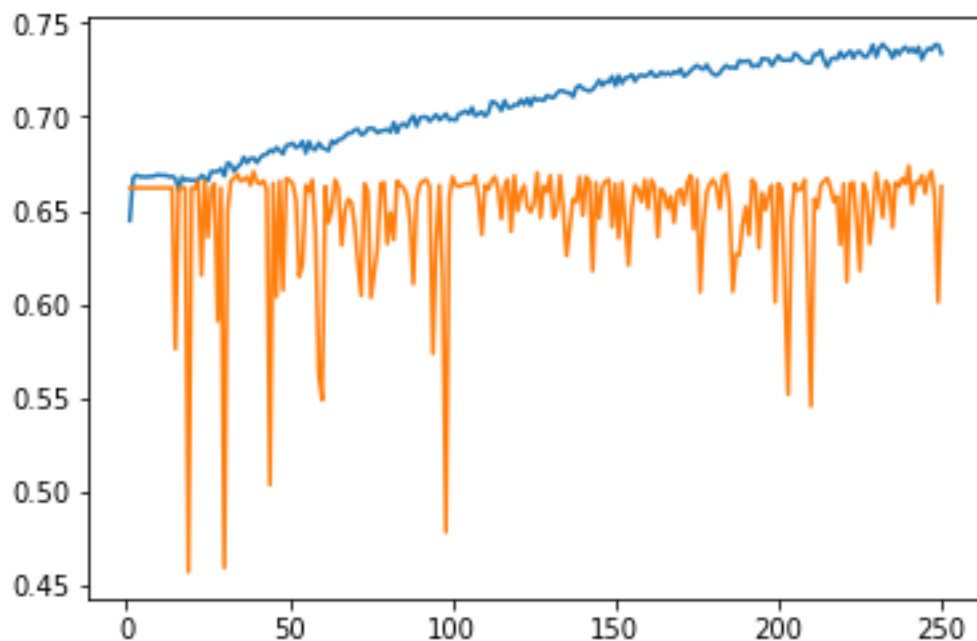
讓模型進步的部分，我有嘗試過不同的 embedding 維度 (100、200)，減少分類器的深度、加入 Dropout、增加 GRU 的 num\_layer、換不同的 learning rate (1e-3、2e-4、1e-4)、嘗試不同的 hidden state 維度(40、200)等等，但不知道是不是因為模型要訓練比較久的關係，在前 250 個 epoch 表現都沒有明顯的不同，且驗證集的表現和訓練集始終有一段落差。這樣的落差在增加 epoch 之後反而增大了：訓練集表現越來越好，測試集表現則一直都沒有明顯的上升。其中算是比較有用的是學習率的調整，降低學習率之後模型的表現變化比較緩慢，但變得相對穩定。

4. (1%) 請比較不做斷詞 (e.g., 用空白分開) 與有做斷詞，兩種方法實作出來的效果差異，並解釋為何有此差別。

有斷詞的表現(accuracy)：



沒作斷詞、只以空格分開的表現(accuracy)：



可以看到，表現的差異並不是那麼明顯，但嚴格來說，同樣訓練 250 個 epoch，同樣的 GRU+分類器模型，訓練集上的表現差了約 2%。顯示有作斷詞的表現仍然稍佳，究其原因，應是由於 word embedding 得到了較好的訓練，去除了包含符號的字(如：Trump...、hi!!!)，這使得模型的訓練效果較好。

5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy." 與 "I am happy, but today is hot." 這兩句話的分數 (model output)，並討論造成差異的原因。

RNN model：前者為 0、後者為 1

BOW model：兩者 output 一樣為 0。

BOW 模型只考慮每個字的出現，而 RNN 模型還考慮的字出現的次序性。故相似的兩句話在兩種模型會有不同的表現。

Math Problem：

HW 5. § 43. -ML

① LSTM cell.  $f(z) = \frac{1}{1+e^{-z}}$

$$z = [0 \ 0 \ 0 \ 1] x^t + 0$$

$$z_i = [100 \ 100 \ 0 \ 0] x^t - 10$$

$$z_f = [-100 \ -100 \ 0 \ 0] x^t + 110$$

$$z_o = [0 \ 0 \ 100 \ 0] x^t - 10$$

$$t=1,$$

$$x^1 = (0 \ 1 \ 0 \ 3)^T \quad c' \equiv f(z_i) \cdot g(z_f) + c f(z_o)$$

$$z^1 = 3 + 0 = 0$$

$$z_i^1 = 100 - 10 = 90 \quad f(z_i^1) \approx 1$$

$$z_f^1 = -100 + 110 = 10 \quad f(z_f^1) \approx 1$$

$$z_o^1 = -10 \quad f(z_o^1) \approx 0$$

$$\Rightarrow c' \approx 1 \cdot 0 + 0 \cdot 1 = 0 \quad y' = 0 \cdot 0 = 0$$

$$t=2$$

$$X^2 = (1 \ 0 \ 1 \ -2)^T$$

$$z^2 = -2$$

$$z_1^2 = 100 - 10 = 90 \quad f(z_1^2) \approx 1$$

$$z_f^2 = -100 + 110 = 10 \quad f(z_f^2) \approx 1$$

$$z_0^2 = 100 - 10 = 90 \quad f(z_0^2) \approx 1$$

$$C^2 = 1 \cdot -2 + 0 \cdot 1 = -2$$

$$y^2 = 1 \cdot -2 = -2$$

$$t=3$$

$$X^3 = (1 \ 1 \ 1 \ 4)^T$$

$$z^3 = 4 + 0 = 4$$

$$z_1^3 = 200 - 10 = 190 \quad f(z_1^3) = 1$$

$$z_f^3 = -200 + 110 = -90 \quad f(z_f^3) = 0$$

$$z_0^3 = 100 - 10 = 90 \quad f(z_0^3) = 1$$



$$C^3 = 1 \cdot 4 + (-2) \cdot 0 = 4.$$

$$y^3 = 1 \cdot 4 = 4.$$

$$t = 4.$$

$$X^4 = (0 \ 1 \ 1 \ 0)^T$$

$$z^4 = 0$$

$$z_i^4 = 100 - 10 = 90 \quad f(z_i^4) = 1$$

$$z_f^4 = -100 + 110 = 10 \quad f(z_f^4) = 1$$

$$z_o^4 = 100 - 10 = 90 \quad f(z_o^4) = 1$$

$$C^4 = 1 \cdot 0 + 4 \cdot 1 = 4$$

$$y^4 = 1 \cdot 4 = 4.$$

$$t = 5$$

$$X^5 = (0 \ 1 \ 0 \ 2)$$

$$z^5 = 2$$

$$z_1^5 = 100 - 10 = 90 \quad f(z_1^5) = 1$$

$$z_f^5 = -100 + 110 = 10 \quad f(z_f^5) = 1$$

$$z_0^5 = 0 - 10 = -10 \quad f(z_0^5) = 0$$

$$C^5 = 1 \cdot 2 + 4 \cdot 1 = 6$$

$$y^5 = 0 \cdot 6 = 0$$

$$t = 6$$

$$X^6 = (0 \quad 0 \quad 1 \quad -4)^T$$

$$z^6 = -4$$

$$z_i^6 = 0 - 10 = -10 \quad f(z_i^6) = 0$$

$$z_f^6 = 110 \Rightarrow f(z_f^6) = 1$$

$$z_0^6 = 100 - 10 = 90 \quad f(z_0^6) = 1$$

$$C^6 = 0 \cdot (-4) + 6 \cdot 1 = 6$$

$$y^6 = 1 \cdot 6 = 6$$

$$t=7.$$

$$X^7 = (1 \ 1 \ 1 \ 1)^T$$

$$z^7 = 1$$

$$z_i^7 = 200 - 10 = 190 \quad f(z_i^7) = 1$$

$$z_f^7 = -200 + 110 = -90 \quad f(z_f^7) = 0$$

$$z_o^7 = 100 - 10 = 90 \quad f(z_o^7) = 1$$

$$C^7 = 1 \cdot 1 + 6 \cdot 0 = 1$$

$$y^7 = 1 \cdot 1 = 1.$$

$$t=8, \quad X^8 = (1 \ 0 \ 1 \ 2)^T$$

$$z^8 = 2$$

$$z_i^8 = 100 - 10 = 90 \quad f(z_i^8) = 1$$

$$z_f^8 = -100 + 110 = 10 \quad f(z_f^8) = 1$$

$$z_o^8 = 100 - 10 = 90 \quad f(z_o^8) = 1$$

$$C^8 = 2 \cdot 1 + 1 = 3$$

$$y^8 = 3 \cdot 1 = 3$$

$t$	1	2	3	4	5	6	7	8
$y^t$	0	-2	4	4	0	6	1	3

②

$$\frac{\partial L}{\partial w'_{ij}} = \sum_{k=1}^V \sum_{c=1}^C \frac{\partial L}{\partial u_{c,k}} \cdot \frac{\partial u_{c,k}}{\partial w'_{ij}}$$

$$\frac{\partial L}{\partial w_{ij}} = \sum_{k=1}^V \sum_{c=1}^C \frac{\partial L}{\partial u_{c,k}} \frac{\partial u_{c,k}}{\partial w_{ij}}$$

$$\frac{\partial L}{\partial u_{c,j}} = -\delta_{jj}^* c + y_{c,j} \triangleq e_{c,j}.$$

$$\Rightarrow \frac{\partial L}{\partial w'_{ij}} = \sum_{k=1}^V \sum_{c=1}^C \frac{\partial L}{\partial u_{c,k}} \frac{\partial u_{c,k}}{\partial w'_{ij}}$$

$$= \sum_{c=1}^C \frac{\partial L}{\partial u_{c,j}} \frac{\partial u_{c,j}}{\partial w'_{ij}}$$

$$= \sum_{c=1}^C (-\delta_{jj}^* c + y_{c,j}) \left( \sum_{k=1}^V w_{ki} x_k \right)$$

$$\frac{\partial L}{\partial w_{ij}} = \sum_{k=1}^V \sum_{c=1}^C \frac{\partial L}{\partial u_{ck}} \frac{\partial}{\partial w_{ij}} \left( \sum_{m=1}^N \sum_{l=1}^V w'_{ml} w_{lm} x_l \right)$$

$$= \sum_{k=1}^V \sum_{c=1}^C (-\delta_{kk}^* + y_{ck}) w_{jk} x_i$$

⇓

$$\frac{\partial L}{\partial w_{ij}'} = \sum_{c=1}^C (-\delta_{jj}^* + y_{cj}) \left( \sum_{k=1}^V w_{ki} x_k \right)$$

$$\frac{\partial L}{\partial w_{ij}} = \sum_{k=1}^V \sum_{c=1}^C (-\delta_{kk}^* + y_{c,k}) w_{jk} x_i$$