Machine Learning HW6 Report

學號：R07942091　系級：電信碩一　　姓名：許博閔

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*

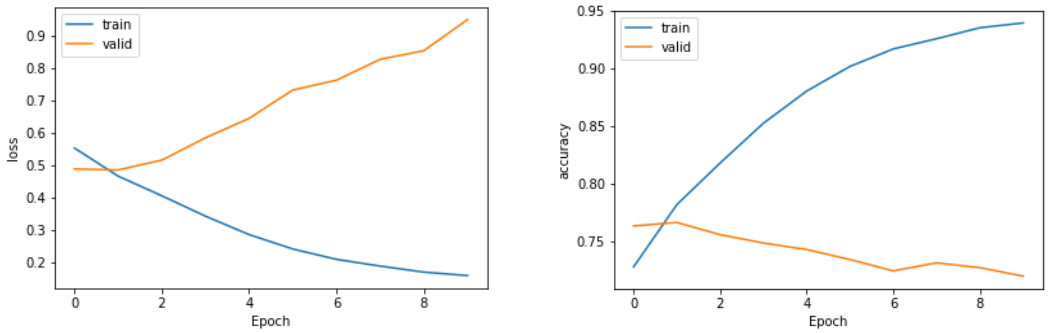
   我的 RNN 模型用了一層 hidden size 等於 300 的 embedding，並且有先用 gensim 預先訓練好的 word2vec 參數的當作 embedding 的初始參數，接著傳到 hidden size 等於 200，兩層且 bidirectional 的 GRU，之後將第二層 GRU 兩個方向最後的 hidden state 的結果 concat 起來傳進 2 層的 fully connected，最後得到預測的結果，有加 dropout 避免 overfitting，輸入的句子長度皆為 128。

   單一 model 在 kaggle public 的分數大約在 0.755~0.759 之間，final submission 是用 3 個 model 做 ensemble 後的結果。

   kaggle public : 0.76400，kaggle private : 0.77110

   :

   左圖:RNN loss　　右圖:RNN accuracy

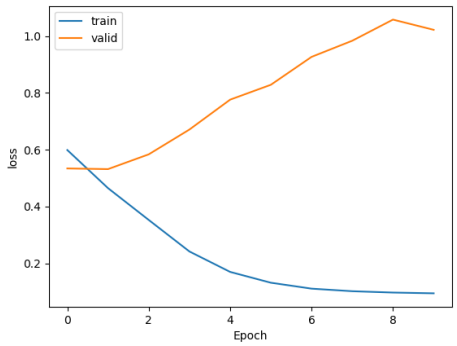


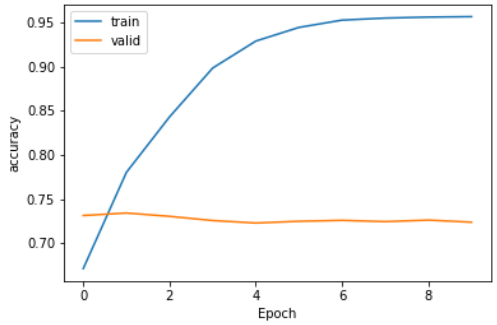
2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

   我先用 jieba 做斷詞，因為斷詞後的字太多，我取最常出現的 3 萬個字當 BOW 的 dictionary 後，將每個句子轉換成長度為 3 萬的 tensor，傳入 3 層的 fully connected 後得到結果，activation 為 RELU，dropout 和 batchnormalization 都有使用。

   kaggle public : 0.74380 kaggle private : 0.73650

左圖:BOW loss　　右圖:BOW accuracy





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。
   preprocess:我用 re 過濾標點符號、並調整 gensim Word2Vec 中 min_count 的數量，藉此捨去句了中不重要的資訊和太少出現的詞彙。
   embedding:用 pretrain Word2Vec 來初始化，並讓 embedding 繼續跟著 model 更新，來得到更好的 embedding 參數。除此之外，我的 embedding 還有多加 3 個詞，分別是 unknown、padding 和 EOF，並用 random uniform 的方法初使化這 3 個詞的參數，藉此提升 embedding 的解釋能力。
   架構:調整 GRU 的 hidden size，多加一層 GRU 和 bidirectional 都有助於 model 理解句子的前後文，加 dropout 可避免 overfitting，除此之外，因為這次很快就會知道 model 的好壞，因此 batch_size 調小一點也可使模型進步，還有我的 GRU 的參數用 orthogonal initialization。

4. (1%) 請比較不做斷詞 (e.g., 以字為單位) 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。
   無論是 RNN 或是 BOW，有做斷詞的正確率都比沒做斷詞好大約 0.01，我想是因為很多時候只看單一個字時，會無法理解它真實的意義，並影響 Word2Vec 的好壞，因為同一個字其實會有不同的意思。因此中文需要以詞為單位，這樣 model 才會得到比較好的結果。

5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "在說別人白痴之前，先想想自己 "與"在說別人之前先想想自己，白痴" 這兩句話的分數（model output），並討論造成差異的原因。
   在說別人白痴之前，先想想自己
   RNN: 0.4044　　BOW:0.9967
   在說別人之前先想想自己，白痴
   RNN: 0.4797　　BOW:0.9967

對 BOW 來說，兩句話的分數一樣高，因為 BOW 並不考慮詞的順序，因此兩句話對　BOW 來說基本上是一模一樣的輸入，所以得到一樣的分數。
RNN 因為有考慮詞彙順序的關係，因此兩句話的分數不同，即使第 2 句話的分數還是小於 0.5，還是可以看出 RNN 比起 BOW 更能理解一句話真正的意思。

P.S 我的 github 檔案中，word_matrix.pkl 只有跑 train 的時候才會用到，是 embedding 初始化的參數。word.pkl 是我將 gensim Word2Vec 用 KeyVectors 的方式存下來的，但因為限制只能用 Word2Vec，所以我另外存了一個 word_dict.pkl 的 dictionary 來處理。因此我的 test 只會用到 3 個 model 檔和 word_dict.pkl。