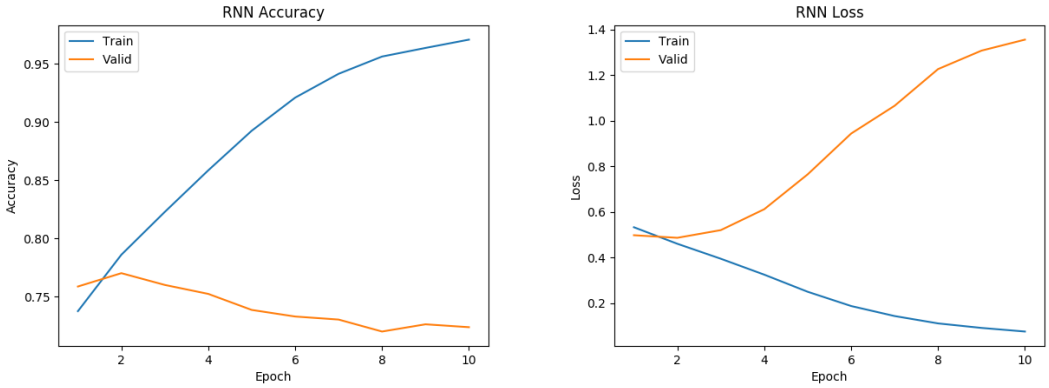
Machine Learning HW6 Report

學號：B04901147　系級：電機四　姓名：黃健祐

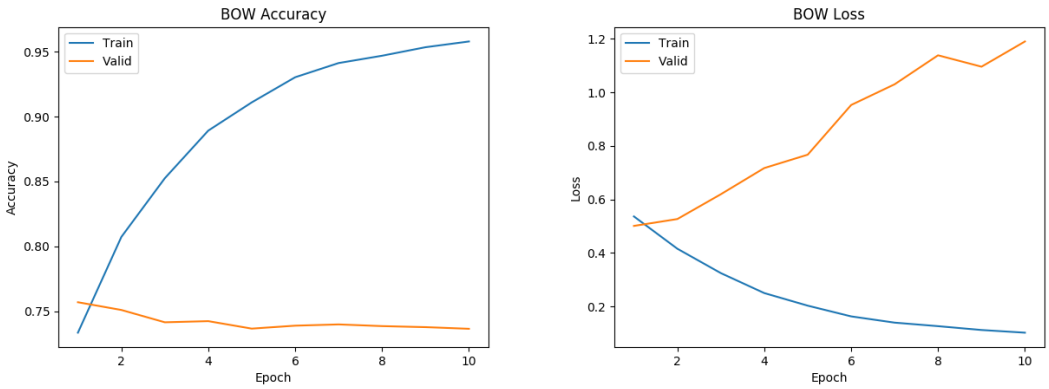
1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線。

   本次實作的 RNN 模型由一個 layer=2 的雙向 LSTM 以及 4 層 fully-connected layer 組成（以 LSTM 完整的 output 作為輸入）。Word embedding 是先由 gensim 產生之後，再和 model 的其他部分一起繼續訓練。單一 model 在 public 及 private 的 accuracy 分別為 0.76690 及 0.76300，而 ensemble 3 個 model 後的結果則是 0.76920 和 0.76610。



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。

   本次實作的 BOW 的字典包含大約 30000 個字詞，而 DNN 的部分則是由 4 層 fully-connected layer 組成。BOW model 在 public 及 private 上的 accuracy 分別為 0.73580 和 0.73400，比 RNN model 遜色一些。



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。

   資料前處理的部分，連續重複的字詞都被刪減到一個（例如：哈哈哈→哈）以避免大量無意義文字干擾判斷；embedding 則是由 gensim 產生後再繼續和 model 一起訓練，這樣可以讓 model 學習並調整字詞間的關係，並在大約 2-3 個 epoch 左右就達到 strong baseline 的水準。另外，fully-connected layer 是以 LSTM 完整的 output 而非只有最後的 hidden state 作為輸入，使得更完整的資訊可以被保留下來並使用到。

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。

	Public	Private
斷詞	0.76690	0.76300
無斷詞	0.75350	0.74740

   沒有做斷詞處理的 model，在 public 及 private 上表現都比有做斷詞的 model 遜色。因為詞才帶有語意，以詞為單位做 embedding 時，model 更能掌握詞與詞之間語意的關係，而字與字之間的關係就相對薄弱許多，導致 model 表現較差。

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。

   下表為 model 判斷語句為惡意留言的機率：

	在說別人白痴之前...	在說別人之前先想想...
RNN	0.30058	0.48822
BOW	0.21528	0.21528

   由表格可以觀察到，RNN 因為考慮了語句的順序，能夠分辨出同樣的字詞組成的兩句話有不同的語意（因此有不同的分數）；另一方面，對 BOW 而言，兩個句子的字詞組成是一樣的，因此儘管兩句話語意不同，分數也會相同。