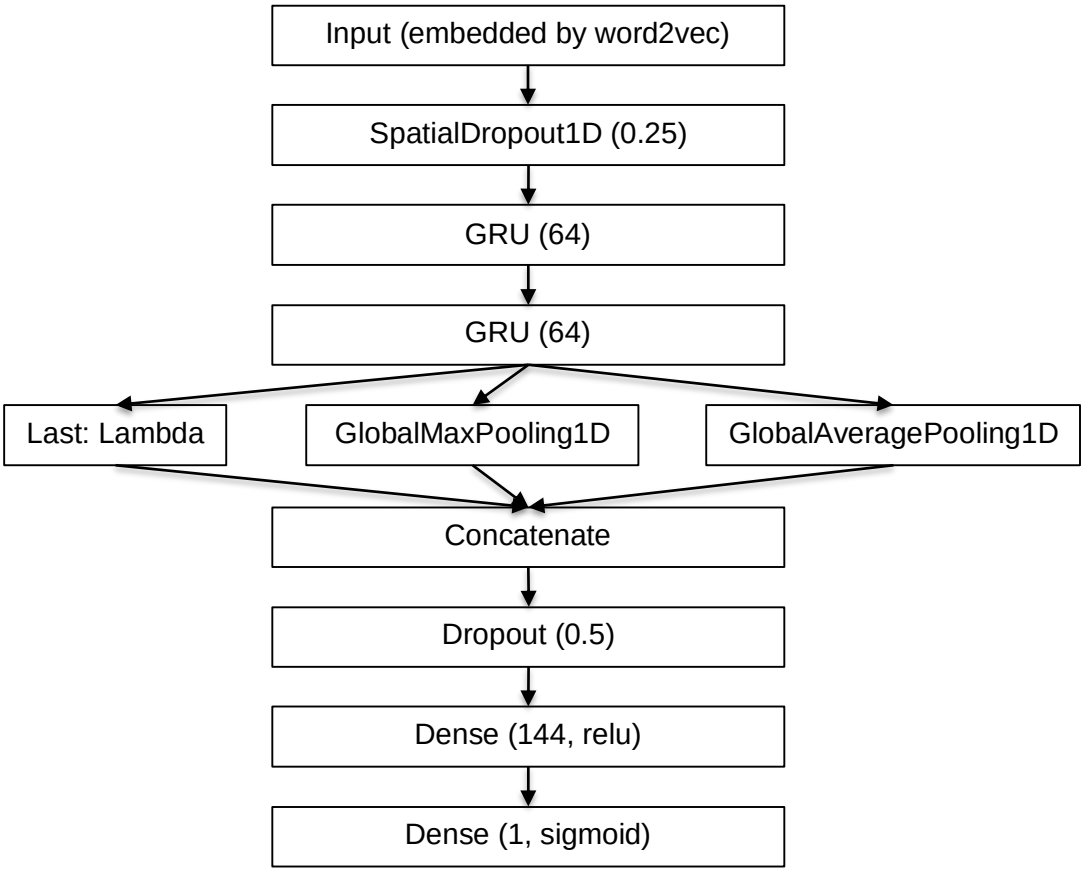
Machine Learning HW6 Report

學號：B05203050　　系級：化學三　　姓名：陳品翰

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*

   我參考 https://github.com/zake7749/DeepToxic 的模型架構，實作了以下的 RNN 模型：



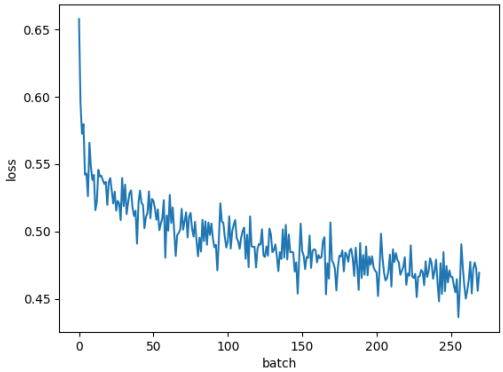
   Word embedding 使用 jieba 斷句後再做 word2vec，並將 train_x 及 test_x 都來 train，參數如下：

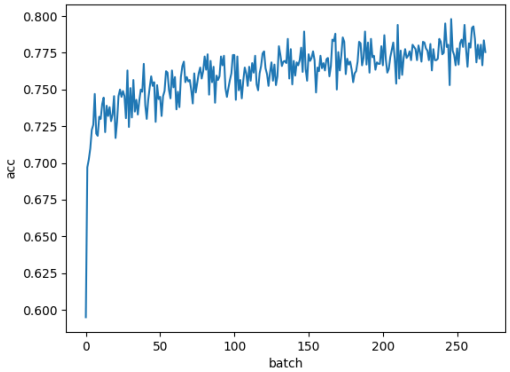
   size=300,window=5,min_count=2,workers=8,iter=80,negative=10

   正確率：

	private	public
Word2vec + RNN	0.75720	0.76090

   訓練曲線：

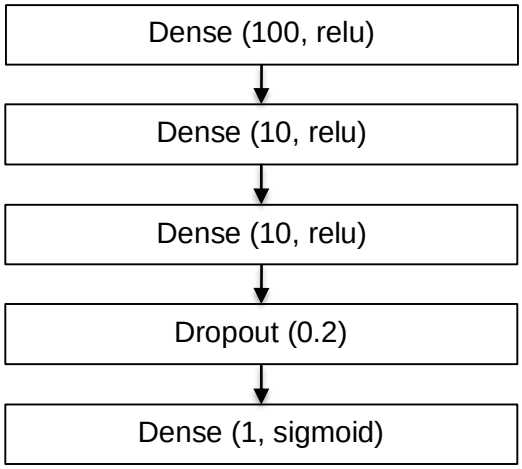




↑ loss to batch　　↑ acc to batch

以上為訓練 5 個 epoch (54 bathces per epoch) 的訓練曲線，可見在第一個 epoch 的第三個 batch，acc 即大幅提升至 0.7 以上。為節省運算時間，我的 val acc 是每個 epoch 結束後才算，因此沒有繪製曲線。其在三個 epoch 後即收斂於 0.76 附近。

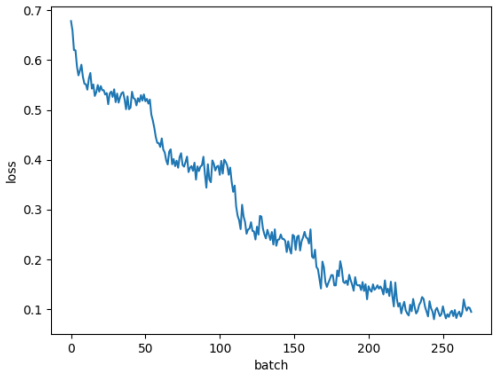
2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。



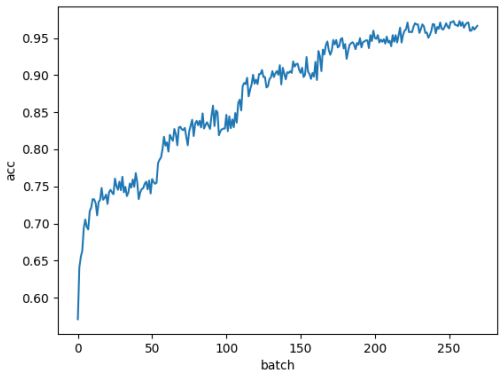
正確率：

	private	public
BOW + DNN	0.74160	0.73750

訓練曲線：



↑ loss to batch



↑ acc to batch

BOW 的 acc 很快速地上升，且每個 batch 後都大幅躍進，但 val acc 卻持續下降，五個 epoch 的 val acc 分別為 0.7543, 0.75, 0.7337, 0.725, 0.7519，可見 BOW 容易 overfit。

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:

我曾試著將句子最前面可能出現的 B1、B2，以及空格當作 stopword，在斷詞時不將其放進 input，但後來發現這樣做會使模型的表現變差，因此我沒有做任何 preprocess。

Embedding:

首先我在 word2vec 的參數中設置了負採樣，如此在訓練時，只會取幾個與輸入無關的輸出來更新 weight，減少運算的負擔。另外我在 embedding 後加入了 Spatial dropout，根據文獻 (https://arxiv.org/pdf/1512.05287.pdf) 指出，對輸入句中的幾個字進行 dropout，可以減少對特定字詞的依賴。實際上我在加入了 Spatial dropout 後，kaggle 的正確率平均由 0.7531 提升至 0.7591。

Model 架構:

我使用 GRU 而非 LSTM，其實兩者的表現差不多，但由於 GRU 較為簡化，因此訓練速度較 LSTM 快。

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。

正確率：

	private	public
With segmentation	0.75720	0.76090
Without segmentation	0.69480	0.69360

以字為單位時，模型需要再另外學習字與字之間是否有連成一詞的關聯性。以「當掉活該」為例：有做斷詞可分成「當掉」、「活該」，並從「活該」得知此句

為負面。但不做斷詞時，由於「當」、「掉」、「活」、「該」皆非負面用詞，因此需先從前後的字發現「活」、「該」連在一起是負面的，使問題更複雜，判斷錯誤的機率也提升。

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。

	第一句	第二句
RNN	0.57037306	0.73657346
BOW	0.679519	0.679519

對 BOW 而言，兩句有的詞皆相同，且皆有「白癡」這個負面詞，因此兩句分數相同且 > 0.5。而對 RNN 而言，第一句讀到「白癡」前面有「別人」，會使分數降低。第二句的「白癡」是在逗號後獨立斷開，因此分數較高。