2025 Machine Learning Fall - HW4 Report Template

Student ID: r14921A13 Name:鄭皓中

1. (1.5%) Briefly describe the method(s) you used to achieve higher accuracy in your model.

我在Dataset & embedding的部分使用了
a. gensim.models.FastText：相較於單純的w2v，他使用了Sub-word infromation來拆解字根，可以比較容易判斷coool、hapyy這種打錯字的twitter訊息。
b. gensim.models.Phrases：使用Phrases將多次重複出現的句型視為片語，讓否定詞等可能要兩個字連在一起表示意思的句型更容易被判讀。
c. parsing_text：使用parsing_text過濾掉網址、UserID等無意義訊息，並將常見的emoji(包含 :)、:-)、:(等)改為情緒表達詞smile、sad……，讓模型將其視為訊息而非標點符號。

在LSTM的部分則是使用
a. Spatial Dropout：在訓練一定epoch後有發現overfitting，使用Spatial Dropout直接丟棄整條Embedding可以有效降低overfitting的狀況。
b. Bidirectional LSTM：雙向LSTM讓模型更了解語意。
c. Multihead self-attention header：計算attention並使用multihead讓模型可以從不同角度判讀句子。

最後Training的部分使用
a. R-Drop：將一份input丟進model兩次，由於有dropout結果會不同，最後則是讓模型愈靠近愈好。
b. Binary Focal Loss：取代標準的cross entropy，專注於較難判斷成功的樣本。
c. ReduceLROnPlateau：動態調整學習率，讓模型不會卡死在同一個點上。

2. (1.5%) Compare traditional RNN/LSTM architectures with modern decoder-only Transformer language models. Then, intuitively explain how the self-attention mechanism works and how it improves performance.

RNN/LSTM：

a. 運作方式：Sequential，需要一個一個接著讀，透過第t-1個字解讀第t個字。
b. 優點：理論上可以處理無限長度序列，參數量相對較少，適合小數據的資料集。
c. 缺點：因為需要一個一個接著讀，幾乎沒有辦法平行化，訓練速度較慢。距離太遠的資訊也有機會稀釋或丟失。

Decoder-only Transformer：

a. 運作方式：parallel，訓練時可以一次看到所有的單字，透過attention機制可以讓每一個字都看到前面一個狀態的所有單字。
b. 優點：高度平行化，可以讓GPU同時計算所有Token，訓練效率較高，且不會有長距離遺忘的問題，因為每個字都可以看到前面的所有單字。
c. 缺點：記憶體消耗量大，計算複雜度的話是O(N^2)

self-attention mechanism：

1. 運作原理：將輸入的字拆成Query、Key、Value，
   a. Query：查詢，目前要找的字是什麼。
   b. Key：索引，目前要找的字具有什麼特徵。
   c. Value：內容，實際上有什麼資訊。
2. 運作過程：
   a. 匹配Q、K：拿目前的Q去對所有的K做內積，內積愈大代表關聯性愈強。
   b. Softmax權重：將上面的內積拿去做Softmax轉成機率分布，得到每個K的Attention score。
   c. Weighted sum：用算出來的權重對所有的Value進行加權，得到新的向量。
3. 怎麼improve performance：
   a. 動態的文字學習：如果一個字同時具有兩種不同的意思，在使用上LSTM的輸入向量是一樣的，只能慢慢修正，而Transformer可以利用attention讓他快速地跟與其相似的字彙進行連結。
   b. 全部vs局部：Transformer可以一次看整個文字的情境分布，LSTM則是慢慢摸索文字之間的連結，所以Transformer可以解決長距離依賴而LSTM不行。