

Deep Learning HW2 問題討論

309706033 資管碩一 康佑誠

第一部分 RNN、Transformer

1. Text Preprocessing:

- 本次文字前處理採用 NLTK 套件，作為 Tokenizer 的工具；且針對空白字元，我們採取忽略不 tokenize 它們。
- 加入 special token 的原因是因為每個標題長度要一致，所以我們要用 PAD 來填充過短的標題；另外如果字詞沒在字典裡出現過，就用 UNK 的替代它。
- Text Preprocessing 的程序為：
讓標題轉為小寫 --> 利用 Regex 去除標點符號 --> 去除 white spaces --> 用 NLTK token 文字 --> 去除 stop word --> 用 NLTK stemming 文字

2. RNN (Recurrent Neural Network) :

【Word Embedding】

- 這次使用 Pre-trained 的 word embedding 是使用 Glove.6B.200d 檔
- 會使用 pre-trained 的 word embedding 的原因是因為，這些像是 glove、w2v、fasttext 都是研究機構或是企業已經將好幾百萬、千萬的文章、文字都看過，並編成的 word vector，所以有點像是站在巨人的肩膀上一樣，可以將這些已經學過、訓練過的文字向量作為我們預設的向量，而不是一開始隨機預設然後再慢慢學習。
- 將每個標題轉為 Keras 可讀的形式，在截長補短 padding 到長度為 20

【Model Design】

- 本次任務使用了 GRU 作為預測模型
- 模型架構為: 64 個神經元的 GRU -> 32 個神經元的隱藏層 採用 relu 作為 activation function -> 最後一層為 softmax 5 個神經元的輸出

3. Transformer

Hyperparameter of Transformer:

- Attention head 設為 3
- Transformer block 裡面的隱藏層的 layer size 設為 16
- Transformer block 後面接一個 64nodes 的隱藏層，activation function 經實驗過後設為 tanh 表現最好