ADL HW2-Seq2Seq Report

b03705012 資管四 張晉華

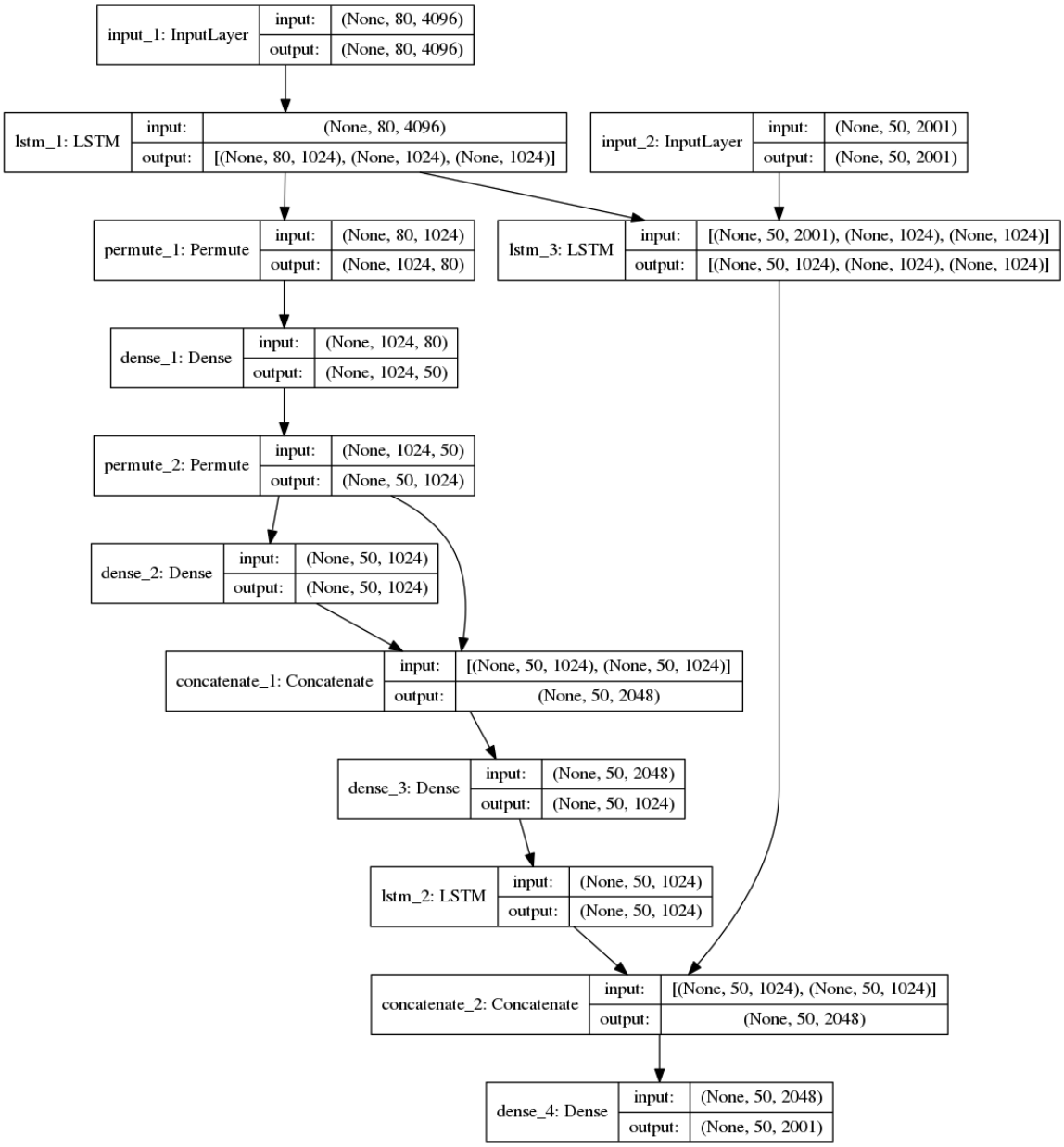
1.Model description

- 資料前處理

  將每個影片隨機取至多 5 個 captions，將 caption 轉成 one hot array，與影片的 feat data 一一組合成 training data，另外由於要做前後字之間關係的預測，因此將 caption 的 one hot array 複製成兩份，一份在開頭加入 begin of sequence(bos)，另一個在結尾加入 end of sequence(eos)，讓模型可以學習從前一字到下一字之間的關係，最後組成每個 sample 的資料為
  （X = [feat data,words with bos], Y = words with eos）

- 模型結構

  將 Video data 輸入進 encoder LSTM 內，將 return Sequence 通過 hidden layer 和 attention layer 傳到 decoder LSTM 內輸出成一個 dense，用來訓練 video 與 caption 之間的關係（下圖左側），而另外將 encoder 的 return state 作為另一個 LSTM 的 initial state，並輸入之前我們處理好的 words with bos，將 return Sequence 跟 Video data 的 decoder LSTM 輸出成的 dense 組合起來（下圖右側），再 softmax 成 one hot array，對應出新的 caption。



2.Attention mechanism

- 將 encoder 的 return sequences 用 Permute 和 Dense 從影片的 80 個 time step 轉換成我們預期的 caption time step(在 model 裡我假設為 50)作為 hidden layer，之後利用一層 softmax dense 對 hidden layer 做 attention，最後和 hidden layer 組合起來成為 decoder 的輸入，讓 decoder 除了考慮到每個 time step 的 encoder sequence 外也可以多考慮到每個 time step 的 encoder sequence 彼此的分佈和大小關係。
  - Compare(without attention → with attention)
    - Before
      - BLEU@1 – 0.2658
      - BLEU@new – 0.6406

        - after
            - BLEU@1 – 0.2693
            - BLEU@new – 0.5848
    - attention 會讓預測的 caption 傾向較常出現的字，但在 BLEU score 上效果沒有明顯提升。

3.How to improve your performance

- 降低 tokenizer 的辭彙數量

  train 時此選取最常出現的 2000 字做 one hot，過濾掉大部份只出現一兩次的字，可以讓 model 較容易收斂而且有較高的準確度。

- Beam search

  實作 beam search，讓預測時可以預測出更多的可能性，避免被侷限在常出現的字詞中。

    - Compare(以同一模型做比較)
        - Before
            - BLEU@1 – 0.2733
            - BLEU@new – 0.6126
        - after
            - BLEU@1 – 0.2838
            - BLEU@new – 0.6294

- 增加 LSTM 的 Units

  增加 encoder,decoder LSTM 的 Units 可以讓模型考慮到更多的 feature。

    - Compare(512 Units → 1024 Units)
        - Before
            - BLEU@1 – 0.2693
            - BLEU@new – 0.5848
        - after
            - BLEU@1 – 0.2733

        - BLEU@new – 0.6126

4.Experimental results and settings

- 2 LSTM stacks vs. Seq2Seq vs. 混合版
    - 2 LSTM stacks
        - video data → encoder LSTM → hidden layer → decoder LSTM → output
        - BLEU@1 – 0.1190

          BLEU@new – 0.4308
    - Seq2Seq
        - ref : https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html
        - video data → encoder LSTM -[return state, words with bos（每個預測結果的上一個字）]→ words with eos
        - BLEU@1 – 0.2807

          BLEU@new – 0.5933
    - 混合版
        - 同 Model description
            - BLEU@1 – 0.2733
            - BLEU@new – 0.6126
    - 就 caption 結果上混合板的在句子結構和影片內容上的綜合表現比較好。