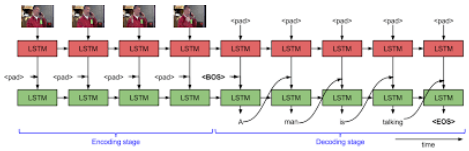
ADLxMLDS HW2 Report — Video Captioning

資工所 碩一 R06922055 吳均庭

- Model Description

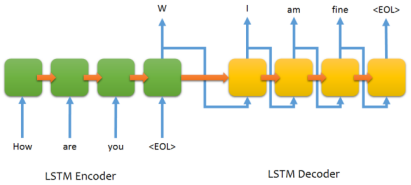
  S2VT

  Model 架構為 encoder-decoder 架構, 由兩層size = 256之LSTM 所組成, encoder decoder共享權重。在encoder 階段, 與第一次作業相同, 將training set的影片對80個frame抽出的4096維feature傳入第一層RNN, 並將輸出與padding concat傳入第二層RNN。在decoder 階段則在第一層RNN傳入padding, 並將輸出上一個time step的 embedding concat之後傳入第二層RNN, decoder每個time step 皆會輸出一個256維feature, 經過output layer 轉成 [word_count] 維向量, 取argmax之後作為輸出。

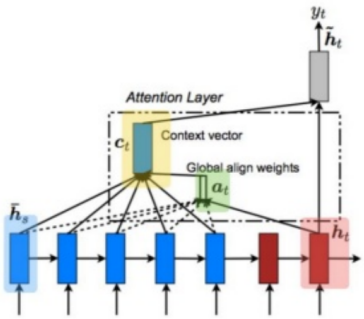


  Video-Caption translation

  參考Neural Machine Translation 之架構來做修改, 分為encoder decoder 兩個階段, 首先先將80個frame的4096維feature餵進encoder, 接著把encoder state 複製到decoder, 上一個time step的embedding concat之後傳入第二層RNN, decoder每個time step 皆會輸出一個256維feature, 經過output layer 轉成 [word_count] 維向量, 取argmax之後作為輸出。



- Attention mechanism
  - implementation

    這次attention我使用tensorflow中的 LuongAttention layer, 將所有的encoder output 作為memory 傳進Attention layer, 對每個decoder output 會透過一個content base function, 計算對每個encoder output 計算alignment score, 對所有score normalize 後將每個encoder output與對應score 相乘後加總(weight average)成為context vector, 最後同時考慮context vector與 decoder outputs 產生最後的結果。

  - Compare and analysis

    概念上加入attention可以讓decoder 同時看所有encoder output並找出最重要的部分, 原先未加入attention時, 在training epoch = 80時得到最佳之BLEU@1 score為 0.296 / 0.655, 加入attention之後為 0.309 / 0.678, 雖然不多, 但可以發現結果確實有所提升。

- How to improve your performance (1%)
  - 在special task中使用S2VT model, 並實作attention mechanism, 效果只有些微的進步, 之後嘗試用修改nmt 架構來實作, 發現BLEU@1 score結果比原先S2VT架構更好, 所以改使用nmt架構來做修改, 來作為本次作業model。
  - 原本使用的dictionary為所有training 與 testing label 中所出現的字, 發現有許多字在caption中只出現過一次, 不具有代表性, 反而會成為結果中的noise, 所以必須把這些字濾掉。最後model中, 使用的

threshold為2，將出現次數太少的字捨棄，剩下3874個字。

- 另外，model有嘗試使用schedule sampling 想要改善training 過程可能產生exposure bias的問題，嘗試過一些實驗之後，發現結果確實有所提升，最後使用 sampling rate = 0.2 時可以使結果上升最多。

- 在inference 過程中，也有嘗試使用過 beam search 來做輸出，在n條beam中再挑出score 最高的output 作為最後的輸出，但也沒有在BLEU@1 score 中看到太明顯的進步。

- Experimental results and settings (1%)

  - 實驗中發現，epoch不用太多約100 epoch 以內就已經收斂，再繼續train會發生overfit 造成BLEU@1 score 下降。最終設定為使用以下設定，得到最佳之兩種BLEU@1 為0.309 / 0.678 。

    - model: Video-Caption translation
    - word threshold = 2 (3874 words)
    - Luong Attention
    - schedule sampling = 0.2
    - Adam lr =0.002
    - epoch = 80
    - loss = sparse_softmax_cross_entropy_with_logits
    - Max accuracy： BLEU@1 = 0.309 / 0.678