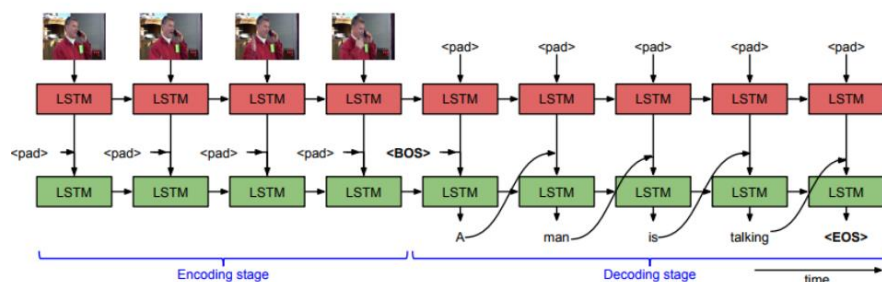


ADLxMLDS 2017 hw2

工科所碩三 葉峻孝 r04525061

Model description (2%)

我們參考助教上課說明建議使用的 Sequence to Sequence: Video to Text 這篇文章來實作 S2VT 的 model。



一開始我先將 feature 轉換跟 LSTM state size 一樣的維度。然後建置好兩個 LSTM 層叫做 LSTM1 跟 LSTM2。然後我使用 padding 的技巧把兩個 LSTM 層的輸入做 padding。然後按照 paper 實作 encoder and decoder。然後把 training 跟 predict 分開。等於說兩個情況會有不一樣的輸出。

Model description:

dim_hidden = 256

n_lstm_step = 80

n_epoch = 1000

batch_size = 50

learning_rate = 0.001

Attention mechanism(2%)

How do you implement attention mechanism? (1%)

我是根據課堂上所講解的 attention model 去實作，一開始使用 initial matching vector 對每個時間點 t 的 encoding stage 的輸出去計算 matching 的 score。然後再來是將 encoding stage LSTM 所輸出的 state 去做 weighting sum。然後得到一個 c 值。再來是將 c 輸入制 decoding stage LSTM2。然後 output 一個預測的字，最後將 LSTM output 取代一開始的 matching vector 成為新的數值。對這些步驟一直重複。

Compare and analyze the results of models with and without attention mechanism. (1%)

在我實作 attention 時，我發覺雖然有加入時都能比較好抓到主詞或者是這張圖的主體但是每次都會 predict 出來一些跟圖片無關的贅字，然後就會把 BLEU 的分數往下拉。當我使用 dim_hidden=256 然後 epoch=1000 時，我舉例一段影片 predict 出來的結果如下。

Model with attention, BLEU = 0.195	A pandas is a of a with to
Model without attention, BLEU = 0.276	A is pandas is climbing

雖然沒有 attention 的句子預測出來的字比較少但是比較精確，至少有動詞。

How to improve your performance (1%)

Write down the method that makes you outstanding

Describe the model or technique (0.5%)

Why do you use it (0.5%)

1. 我實做了 schedule sampling 來解決訓練端跟測試端不匹配的問題，然後我實測的結果發現效果有比沒有加之前的好很多，如果沒有加入 schedule sampling 的話其時通常都會出現一些不重要的字在句首，如果讓他一直 training 下去的話，那個字後面的所有字都會變成不匹配的字。實作下去之後發現 BLEU 分數至少提升 0.05 之多。
2. 然後再來有聽到助教給的建議說，我們應該可以把 vocabulary 的字數從一開始的 6000 多降成 3000，方法是看 training caption 裡面那些單字被使用過的次數，我是設定阈值直接取前 3000 個被使用比較多的字來當作 vocab。這樣會讓我 predict 出來的字都比較精確。減少不常用的字出現的機率。

Experimental results and settings (1%)

parameter tuning, schedual sampling ... etc

我一開始比較 training epoch 的數量來看最後的 performance 好壞。

Epoch = 500	BLEU score = 0.1861
Epoch = 700	BLEU score = 0.2316
Epoch = 1000	BLEU score = 0.2502
Epoch = 1200	BLEU score = 0.2211

我發現 training 大概 1000 個 epoch 來說效果會是最好，我覺的訓練到 1200 之後可能已經 overfitting 了所以最後的 performance 來看是比 1000 來的差。所以我最後選擇 epoch = 1000，擁有最好的 BLEU score = 0.2502。剛剛好壓線過 baseline。

我實測當 epoch = 1000 時有加入 schedule sampling 跟沒有加入的差別
最後時測的結果發現：

Epoch = 1000, without Schedule Sampling: BLEU score = 0.2502

Epoch = 1000, with Schedule Sampling: BLEU score = 0.2584

結論為有使用 SS 結果通常都會比較好，但是在 epoch 很高時卻沒什麼影響。等於說效果並不是非常的顯著。