ADL x MLDS assignment2 report

學號：B03901030 系級： 電機四 姓名：蕭晨豪

1. Model Description

我採用的是 S2VT 的架構，在 encode 和 decode 部分 share 兩個 LSTM，其中 LSTM1 處理圖像資訊，LSTM2 處理文字。

在 encoding stage: LSTM1 輸入我們抽取出來的 frame feature，將這個 feature 和一個 zero vector(因為此時沒有文字資訊) concatenate 起來後傳入 LSTM2。在 decoding stage: 由於沒有 frame 因此我們在 LSTM1 輸入一個 zero vector，將輸出和 word embedding concatenate 起來後輸入 LSTM2，其中這個 word embedding 是由上一個 time step 的 LSTM2 output 經過 word embedding 產生的。

(LSTM2 的 output 實際上會先通過一個 linear 將 output size 轉成 vocabulary size，通過 softmax 後選出機率最大的字進行 word embedding)

另外我也實作了 Attention, schedule sampling 和 beam search，將在回答以下的問題時逐一介紹。

2. Attention mechanism

我實作了 Luong Attention，不過是 global 的版本。一樣使用 S2VT 的架構，只有 decoding stage 的 LSTM2 input 有一些更動。我選用 encoding stage LSTM1 的 hidden 和 decoding stage LSTM2 的 hidden 去計算相似度(我使用內積)，以這個相似度作為 weight 乘上 LSTM1 的 hidden 計算出每個 time step 的一個 attention context(意即在該 decoding time step 中所「更注重」的 encoding feature)，將這個 context 和原本 S2VT 架構的 word embedding 和 LSTM1 output 三者 concatenate 起來後輸入 LSTM2，之後的 output 一樣通過 softmax 產生一個 word 的機率分布。

在使用 attention 之前我的句子有時會出現許多重複的 is 和 a ，使用 attention mechanism 後我的句子更注重在各個名詞與動詞上，減少了較不重要的 is 和 a 的數量，形成較完整的句子。

3. How to improve your performance

Methods:

Schedule sampling: 選定一個機率，根據那個機率決定 decoding stage 的 LSTM2 的 embedding 部分要傳上一個 time step 的 output 的 embedding 還是正確的 target word 的 embedding。使用 Schedule sampling 可以避免因為傳入上一個錯誤的字而造成後面繼續預測錯誤的情況，同時因為傳入正確的 target 可以加快收斂速度。

Beam search: 在 testing 時使用 beam search decoding，並非是每一個 time step output 機率最高的字，而是每個 time step 有一些候選的 sequence，將這些 sequence 和當前 time step 的 word probability 做組合，再次選出機率最高的前幾個候選 sequence，直到全部的候選 sequence 都達到 EOS 或是長度限制為止。這個方法可以避免只看到上個 time step 的字就決定下個字的問題，考慮了更多的情況，但同時 time complexity 也會隨之增高。(基於效率和一些記憶體的問題我上傳的 code 並沒有使用 beam search)

Clip gradient norm: 這是訓練 RNN 時常使用的技巧，由於 RNN 的 loss surface 在某些部分較為陡峭，即使 learning rate 隨著 epoch 下降也有可能在數個 epoch 之後碰到峭壁然後 loss 突然增高，使用 clip gradient norm 的話便可人為決定 gradient norm 的範圍，使 loss 更容易收斂

4. Experimental Results

首先我認為這次的 bleu score 在差距很小的情況下(0.02 左右)很難拿來衡量 model 的好壞，有時候形如 A man is + 動詞 的句子(明顯是未完成的句子) bleu score 就可能很高(只要動詞有 match)，但文法較通順完整的句子有可能因為幾個中間的名詞 mismatch 反而使 bleu score 變低(因此需要 peer review)。因此這個 experiment result 我選擇數個過 baseline (0.25)的 model，主要探討各個參數訓練時的收斂速度以及自己主觀衡量 testing output 的成果。

Schedule sampling ratio:

由於 schedule sampling 是依一定機率傳入正確的 target word，因此他可以加快收斂速度並且增加 training 的穩定性(不會一直用上一個錯誤的字往後傳)，當此機率越高時收斂速度越快，但主觀判斷下 schedule sampling 機率越高時在 testing data 上越容易不知所云，因為 testing data 無法傳 target word 只能傳上一個的 output word，因此若 training 時 schedule sampling ratio 設太高的話會有 exposure bias 的問題(其實 ratio 設越高就越接近 teacher forcing，所以會有一樣的問題是合理的)根據實驗結果 0.5~0.6 左右是這次的 task 最合適的 ratio

Hidden dimension:

我嘗試了 128,256,512,1024 作為 hidden dimension。很直覺的參數越多 loss 下降越快，但有時候會 overfit training data 而使 testing 結果不理想(句型正確但名詞錯誤)，最後選用 256

5.README

requirements:

PyTorch 0.2.0

Numpy 1.13.3