R06725048 資管碩一 陳信豪

MLDS HW2 Report

1. Model description
   - 文字處理
     - 濾掉標點符號、轉小寫、並以空白做 split 取出 word tokens
     - 將 captions 轉成 token sequences，並在前後加上起始結束的 tag
     - 設定 sequence 最大的長度 max_sent_len，不足的部分做 padding
       Ex: [<s>, a, man, is, walking, </s>, <pad>, <pad>, <pad>, <pad>]
     - 最後透過建立好的 dictionary，將 token sequence 轉成 index sequence 後再餵入 model

   - 由於一個 video 有多個 caption，我的處理方式是每次先選好 video 後，再隨機選擇其中的一個 caption 當 label，所以每次 epoch 跑的數量都會是一樣 1450 筆 data，但 video features 會在不同 epoch 中對應到不同 label (caption)。

   - S2VT Model

     為 Dynamic Conditional Generation。

     疊兩層 RNN (GRU/LSTM)，第一層作為 encoder，會拿 pre-trained CNN Outputs 進來做 training 取出影像的代表特徵；第二層作為 decoder，將代表特徵放入 input (condition) 做 training 得到結果。

     參數設置如下：

     batch size: 128

     layer_dim: 768 (for word_embedding, encoding rnn, decoding rnn)

     learning rate: 0.001

     max_sent_len: 15

     loss: cross-entropy （透過 masking 來濾掉 padding 的部分）

     optimizer: adam

2. Attention mechanism
3. How to improve your performance

   Scheduled sampling
4. Experimental results and settings