

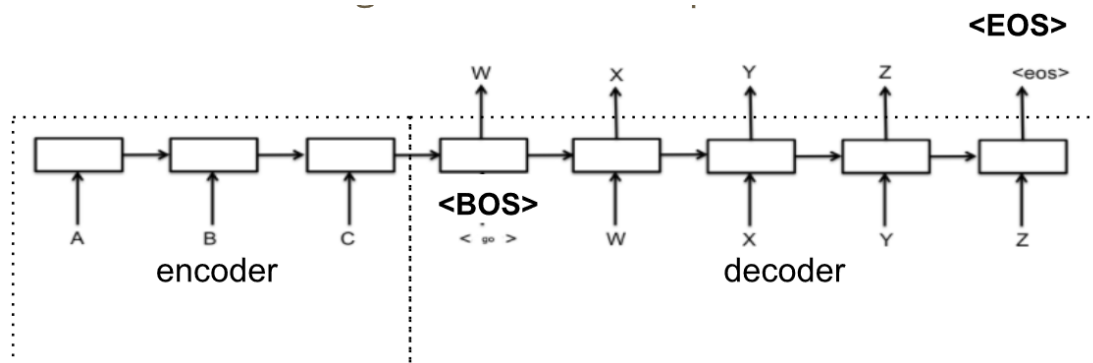
MLDS HW2

王垣尹 許哲瑋 李中原

2-2

Model Description(2%):

我們採用了簡單的 Seq2Seq Model，也就是一個 Encoder GRU 還有一個 Decoder GRU cell，Encode 結束之後馬上進行 Decode 的動作，不管是 encoder 或是 decoder，我們都重複疊了兩層，並且加上 inter-layer 的 Dropout(DropoutWrapper)，這題我們有使用 Attention 機制，採用的是 Luong Attention (2015)，另外值得一提的是由於跟 2-1 的影片輸入長度固定不同，這題的輸入是長度各自不同的句子，因此我們使用了 dynamic rnn 的 encode 跟 decode 來進行加速。



Perplexity:

首先瞭解到 perplexity 是 2 的 entropy 次方，要 minimize 這個值告訴我們的是我們希望找到一個 model 能夠讓預測下一個句子的確定性很低，以 one-hot encoding 來說，我們就是希望 Model 能夠 predict 接近 one-hot 的結果，而不是有很多機率接近的選項找最大的那一項。

Correlation: 0.296

試了各種方法，一直沒辦法超過 baseline 0.45。

Testing Perplexity and Correlation: ~46.7 and ~0.296

Write down the method that makes you outstanding (1%)

Analysis and compare your model without the method. (1%)

Why do you use it (1%)

How to improve the performance

由於有 Wrapper 的關係，我們三種方式都有嘗試，其中 Luong Attention 效果最好，Beam Search 效果不明顯但還是 train 的起來，Schedule Sampling 不知道為什麼都

train 不起來。

Luong Attention :

就是最經典的 Attention 機制，我們認為 Decode 時加上 Focus 的位置對於輸出會有更好的結果，原因都跟 paper 一樣。

Beam Search: Beam Width = 5

我們選擇在 decode 時保留機率最高的 10 個 beam，不過這個數字我們沒有多做太多的嘗試。不過就理論而言，Beam 的大小應該會對 Perplexity 很有影響，畢竟在一定意義下他們是在做一樣的事情；原因是由於用 Viterbi 去 decode 成本太大，我們不可能遍歷所有的 output 可能，因此我們為了減輕每一個 decoding time step 使用 greedy 所造成的結果差異。

Schedule Sampling:

完全做不起來，故略。

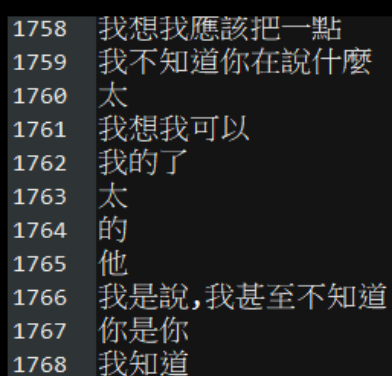
Experimental Result and Settings (1%):

```
rnn_size=1024,  
num_layers=2,  
embedding_size=1024,  
learning_rate=0.5,  
    learning_rate_decay_factor=0.99,  
batch_size=100,  
numEpochs=5000,
```

With Luong Attention: Perplexity 46.7 Correlation: 0.296

Without Luong Attention: Perplexity ~155.6up Correlation: 0.13

Sample Output:



```
1758 我想我應該把一點  
1759 我不知道你在說什麼  
1760 太  
1761 我想我可以  
1762 我的了  
1763 太  
1764 的  
1765 他  
1766 我是說,我甚至不知道  
1767 你是你  
1768 我知道
```

分工表

	王垣尹	許哲瑋	李中原
2-1	Model(S2VT) and Code And report	Technical Support	Model Fine Tuning and Report/ proper format
2-2	Report Support	Model(Seq2Seq) and Code And report	