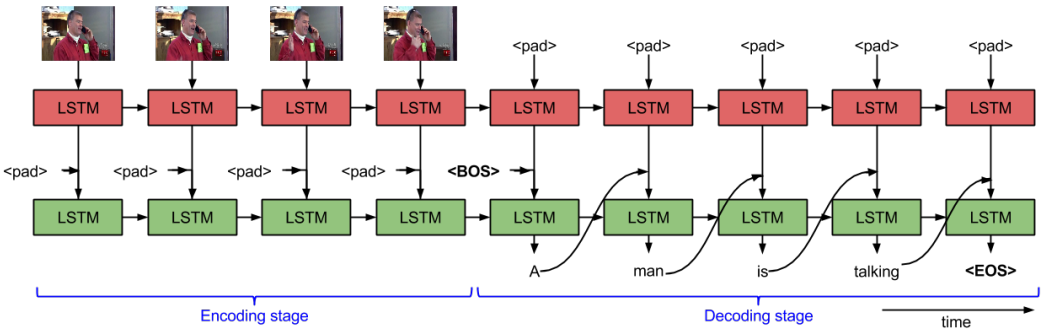
ADLxMLDS 2017 Fall
HW2 - Video Captioning

B05901189 吳祥叡

November 20, 2017

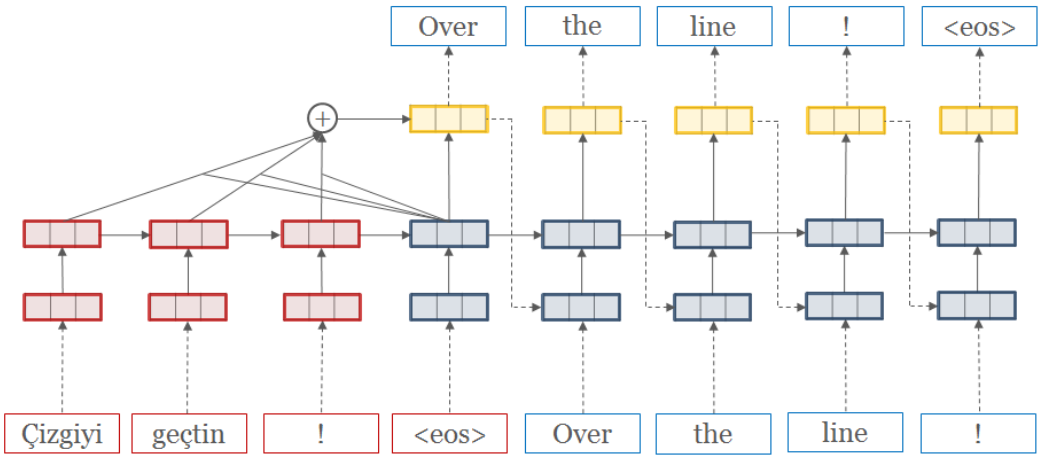
1 Model Description

1.1 S2VT



在 S2VT model 中有嘗試用 scheduled sampling.

1.2 seq2seq



圖為 Neural Machine Translation 的 seq2seq 模型, 將前面輸入換成 vgg 抽出的 feature 即可用在這次作業.
在這個 seq2seq model 中有嘗試用 attention, scheduled sampling, beam-search.

2 Attention Mechanism

2.1 如何實做

用 tensorflow.contrib 的 seq2seq library 可以選用 Bandanau 和 Luong Attention.

2.2 結果比較

用 seq2seq model 上實驗發現 luong attention 最穩定,bahdanau attention 變動較大但平均來看表現和 luong 差不多.
另外比較沒有 attention 發現 attention mechanism 使 BLEU 有顯著的提昇.

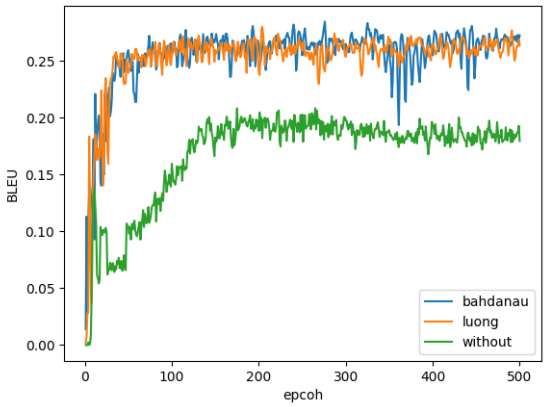


Figure 1: BLEU

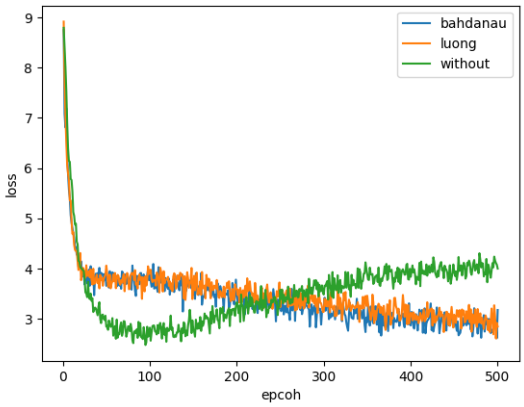


Figure 2: Loss

3 Experiment Settings and Results

3.1 實驗設置

1. 運算資源: 有兩個 K80 GPU 的 Azure NC6
2. 使用套件: Tensorflow
3. schedule sampling 函數: prob = $exp^{-n_{epochs}/200}$
4. word embedding dimension: 300
5. rnn cell type : GRUCell

3.2 實驗結果

3.2.1 不同 word embedding 初始化

比較使用 Glove 或零初始化 word embedding. 發現 Glove 可以讓 model 更穩定, 但是後期表現差不多.

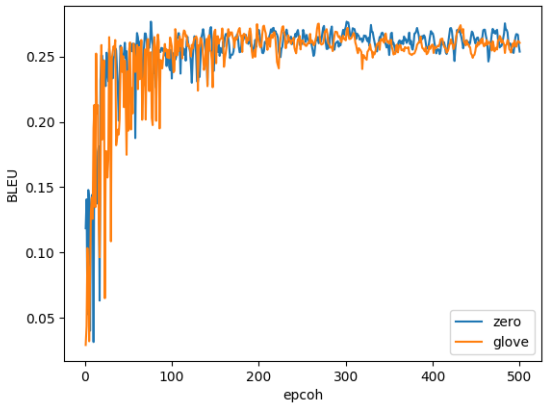


Figure 3: Glove vs Zero

3.2.2 S2VT vs seq2seq

比較參數量大致相同的 S2VT 和 seq2seq model(Luong Attention). 可以看到 seq2seq 比 s2vt 表現好一點點.

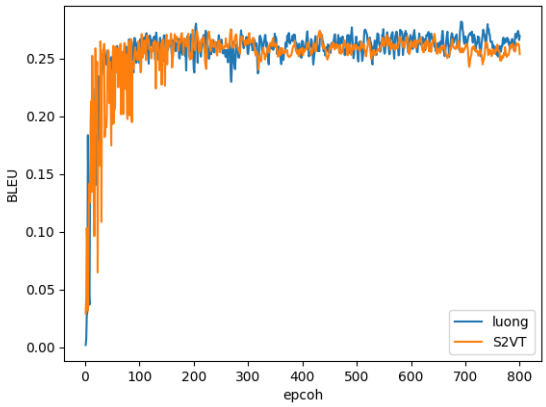


Figure 4: S2VT vs seq2seq

3.2.3 加入雜訊

因為每個影片有多個參考解答, 原先我是在產生 batch 的時候隨機挑出一個, 但是發現這樣不太合理, 因為 model 看過的影片總共只有一千多個, 很容易 overfit. 所以我才會想在影片裡加上一點雜訊, 想要達到類似 image augmentation 的效果. 但是我們拿到的是已經抽好 feature 的 4096 維的 vector, 不能學 image 做旋轉平移鏡射. 所以我採取的是將某些 frame 糊化的方法.
也就是 random 取其中約 20 個 frame, 將他們和前後兩個 frame 平均.

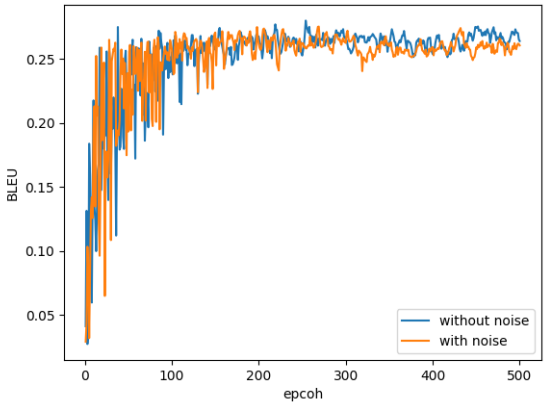


Figure 5: with or without noise