MLDS HW2-2 Report

林宣竹 B03901065 楊其昇 B03901101 郭恆成 B03901145

1. Model Description

- Data preprocessing:

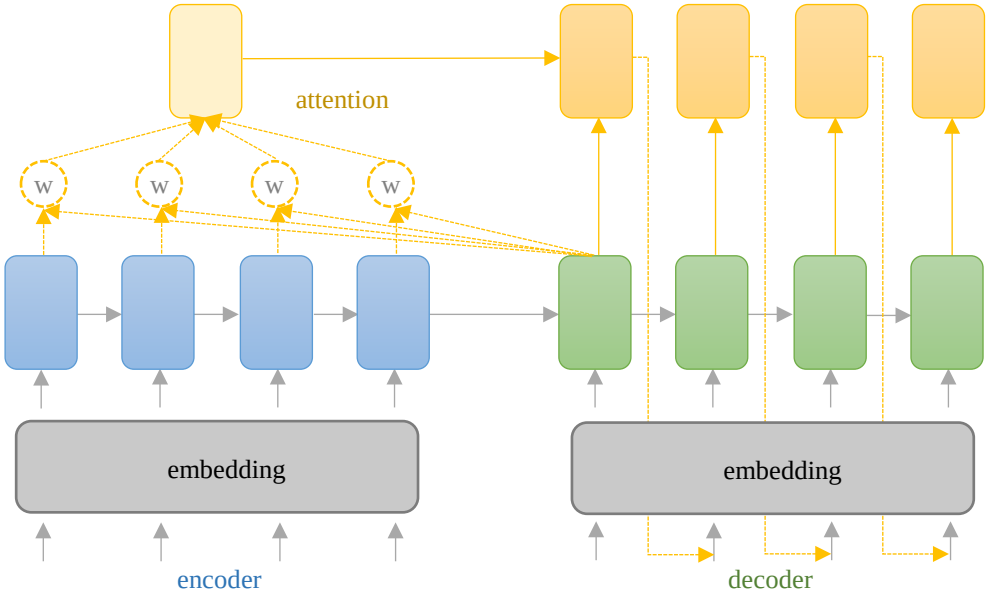
  將 training label 中所有的 caption 當作 corpus，設定 min count=3，得到 2430 個字的 vocabulary。然後加上<BOS>, <EOS>, <PAD>, <UNK>等特殊符號轉成 word-based 的 one-hot word embedding。

  為了避免同個 video feature 但卻有不同 caption 的問題，每個 epoch 會隨機對 video feature 加上 mean=0, std=0.05 的 Gaussian noise。

- Training parameters

  Adam Optimizer, learning rate=0.001, Early stop period=10

- Sequence-to-sequence model + Attention：

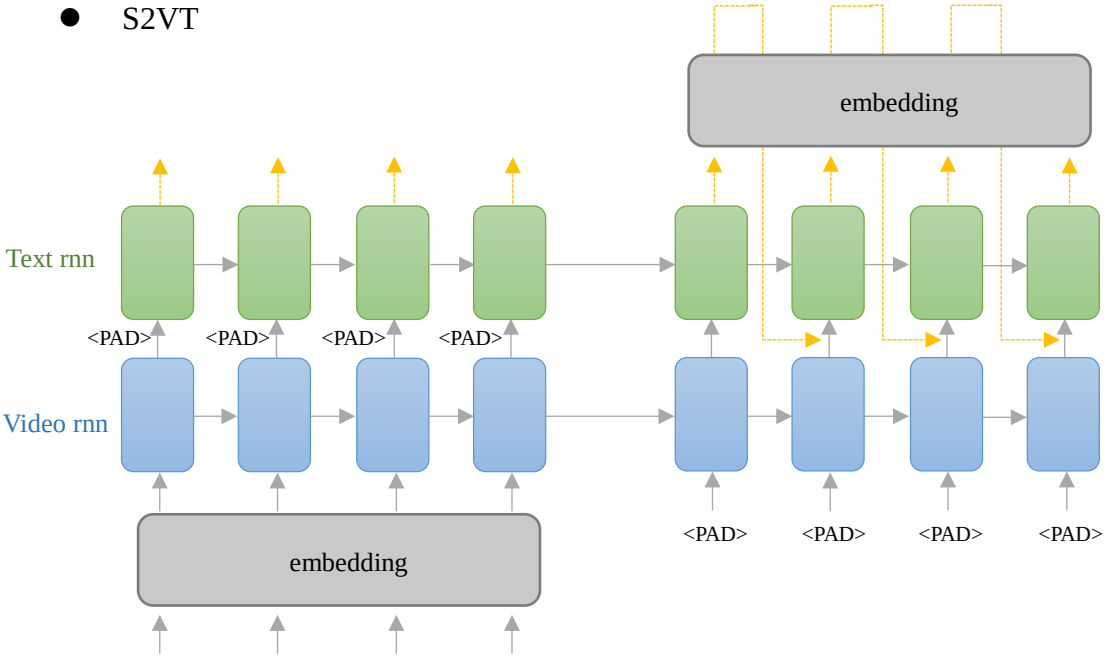


首先 video feature 跟 one-hot embedding 會分別各自經過 400 units 沒有 bias 的 DNN 組成的 embedding layer，而 embedding layer 為單純的 linear transform 映射到低微度的 feature。

而 encoder 跟 decoder 皆是由一層 512 hidden unit 的 LSTM 組成，並且設定 dropout(keeping probability=0.7)，防止 overfit 的狀況產生，然後加上 DNN 組成的 attention。

Decoder 則會把上一個 timestep 的 output 當作 input，直到達到設定好的 max length 或是<EOS>產生。

- S2VT



S2VT 由兩個 RNN stack 而成，第一層是 Video RNN 第二層是 Text RNN，而前半段的 timestep 是 Video 的 encode，後半段的 timestep 則是 Text 的 decode。Text RNN 的 input 是 Video RNN 的 output 串接前一個 timestep 的 prediction。當 Video encode 的階段 Text RNN 輸入 <PAD> 當作 input。Text decode 時 Video RNN 亦然。

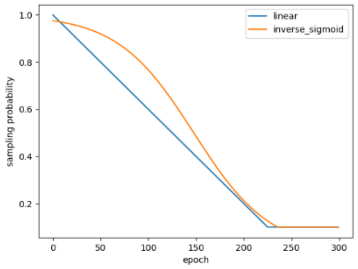
而 S2VT 的參數設定與 sequence-to-sequence 相同。Video feature 與 word embedding 皆會經過 400 unit DNN 的 embedding layer。Video RNN, Text RNN 皆是由一層 512 hidden unit 的 LSTM 組成，並且設定 dropout(keeping probability=0.7)，防止 overfit 的狀況產生。

2. How to improve your performance

- Schedule sampling

(1) Write down the method that makes you outstanding

一定的機率會將正確的 target 當作 input 輸入 decoder 其中取樣機率初始值為 1.0，然後隨著 epoch decay，最後 clip minimum sampling probability 為 0.1。而 decay function 則有 linear decay 與 inverse sigmoid decay。



(2) Why do you use it

希望藉由 schedule sampling 達到 teacher forcing 的效果，來加強 model 在 testing data 上的表現。

(3) Analysis and compare your model without the method

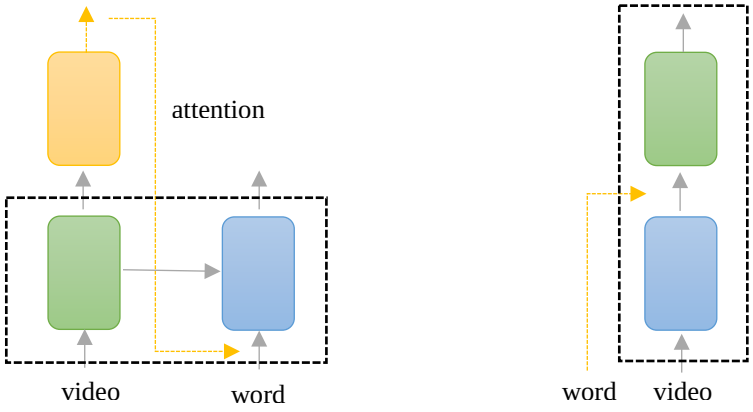
當使用 Schedule Sampling 的時候，model 在 training data 上的 loss 會比未使用 Schedule Sampling 還要高一點，但是在 testing data 上的 bleu score 表現卻會比較好。推測原因為 model 在未使用 Schedule Sampling 的情況下，往往可能是硬背常出現的句子組合，例如: a man is playing …，所以在 testing data 上的表現會比較糟。相反地，若使用 Schedule Sampling 的方法，機器在訓練過程中，會慢慢的學到用自己 predict 的字來產生 caption，雖然在 training loss 會在某個程度就飽和無法再下降(因為愈來愈多 predict 出來的詞會拿來當 decoder 的 input)

- S2VT

(1) Write down the method that makes you outstanding

本次作業中我們有實做第一部分介紹的 sequence to sequence model 以及 S2VT，架構如同上述所說，而 S2VT 也達到的較好的成果。

(2) Why do you use it



上左圖是 seq2seq + attention model 未根據 timestep 展開的結構，而上右圖則是 S2VT model 未根據 timestep 展開的結構。

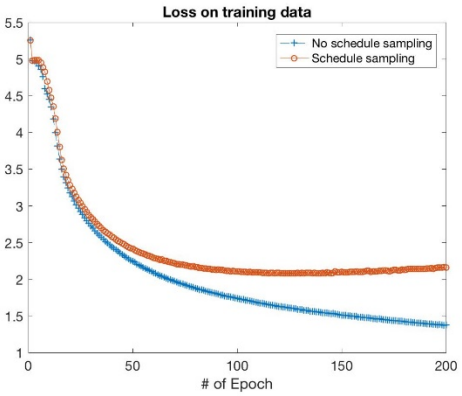
由上面兩張圖我們可以明顯看出 S2VT 有兩個很大的優點:

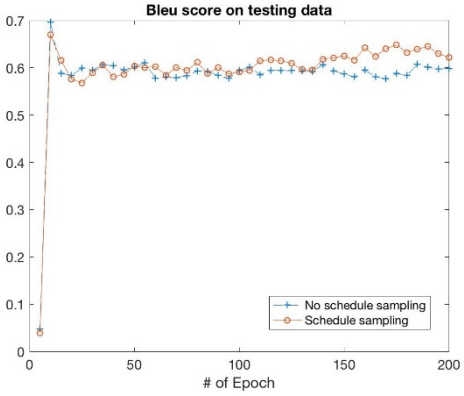
1. S2VT 將 encoder 跟 decoder stack 在一起，保持 decoder 及 encoder 的功能，同時用 deep 的結構去取代 shallow，根據 HW1.1 的實驗，可以預期 deep 的結構有較強的能力。
2. 利用 peephole 將 word embedding 輸入第二層 RNN，而第一層的 hidden 則可視為 dynamic condition 幫助 decode。

(3) Analysis and compare your model without the method

在 training 過程中可以發現 seq2seq model 的 bleu score 通常高於 S2VT model 的 bleu score。但是印出 testing data 的 prediction 可以發現，seq2seq model 的預測以不完整短句子例如 a man is a 占了大多數，導致較高的 bleu score，而 S2VT 則擁有較完整的文法，但是 train 過度的 S2VT 則會產生過長不斷接龍的句子。

3. Experimental results and settings





本題根據結果最好的 S2VT + schedule sampling 的 training processing 回答。由 epoch-loss 做圖可以發現如 題目 2. 所述 schedule sampling 的 loss 較高，但是得到的 bleu score 也較高。

Parameter Setting:

Word based one-hot word embedding

SGD, initial leaning rate=0.001

Initial schedule sampling probability =1.0

Linear sampling probability = 0.02

4. Readme

Numpy 1.14.3

TensorFlow r1.6

Keras 2.0.7

5. hw2_1分工表：
b03901145 郭恆成：seq2seq model
b03901101 楊其昇：s2vt model