ADL hw2 Report

B04902004 王佑安

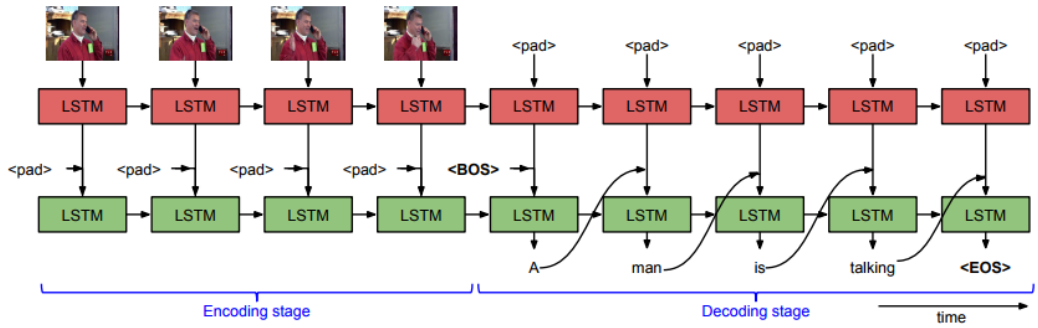
Used Library:

numpy==1.13.3

torch==0.2.0.post3

Model Description:

S2VT



Encoder：1 layer LSTM, hidden size=512

Decoder：1 layer LSTM, hidden size=512

Vocab size：4795

Optimizer：RMSprop, lr=1e-3, alpha=0.9

Attention Mechanism

類似（http://www.aclweb.org/anthology/D15-1166）Early attempt 的做法。

$$\boldsymbol{a}_t = \text{softmax}(\boldsymbol{W}_{\boldsymbol{a}}\boldsymbol{h}_t)$$

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{\boldsymbol{c}}[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

差別在於我直接將算出來的 attention vector 當作 decoder LSTM 的 hidden state 而非 input。

實驗結果：

Model	BLEU
Without Attention	0.303
With Attention	0.312

加了 attention 後直接看輸出沒有什麼明顯的特徵差異，但 BLEU 稍微提升了一些，推測是因為 attention 能選出哪些 frame 是比較重要的，所以對單的詞的預測結果較好，BLEU 較高。

How to Improve Performance

1. 移除 caption 太長的 training data。觀察 training data 的 captions 可以發現大部分的 caption 長度都在 5-10 之間，大於 10 的很少，model 可能會很難 fit 那些樣本數太少且文法較複雜的 caption，因此直接捨棄長的 caption。實驗結果也發現移除長的 caption 在 training 上 fit 的效率跟結果提升不少。
2. schedule sampling。在 training 的時候會發現 model 常常預測出文法不對的句子，因此用 schedule sampling 的方式讓 model 在 training 初期可以先從正確的 caption 學習正確的文法形式，到後面才能預測出正確的句子。

   schedule sampling 我使用（https://arxiv.org/pdf/1506.03099.pdf）提到的式子來調整使用 ground truth 的機率：

   - Inverse sigmoid decay: $\epsilon_i = k/(k+\exp(i/k))$ where $k \geq 1$ depends on the expected speed of convergence.

Experimental Results

Hidden Size：

Model	BLEU
LSTM(256)	0.298
LSTM(512)	0.312
LSTM(1024)	0.306

提高 hidden size 可以有效地讓預測結果變準，但 hidden size 太大到 1024 時 fit 不太起來，loss 降很慢，最後出來結果也稍微比 512 差一點。

Model With Different Technique：

Model	BLEU
base model (512)	0.279
base model + attention	0.281
base model + schedule sampling	0.303
base model + attention + schedule sampling	0.313

可以發現 schedule sampling 的 improvement 比 attention 明顯許多。尤其直接觀察預測結果，發現沒有 schedule sampling 的 model 預測出來的句子雖然有些單詞會正確但經常整句文法不通。而 attention 則只能提升對單詞的預測準確度。