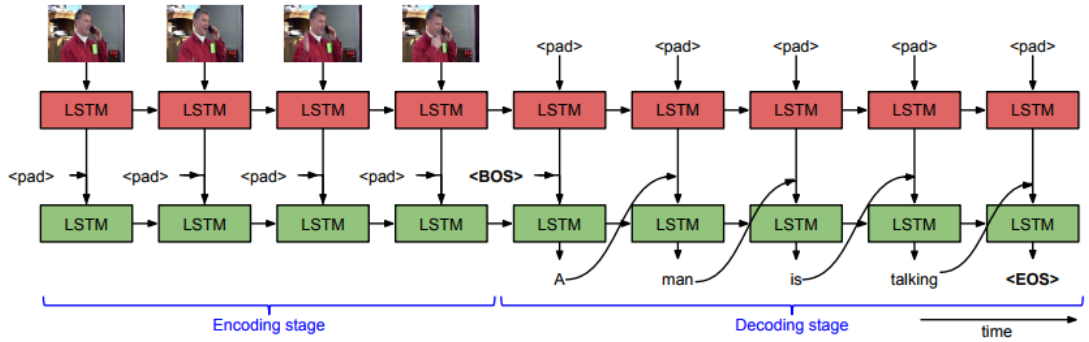
MLDS HW2 report

R04943151 梁可擎

Model Description

Sequence to sequence model

這次的作業模型，依照投影片建議實作 S2VT [1] 的 model，如下圖：



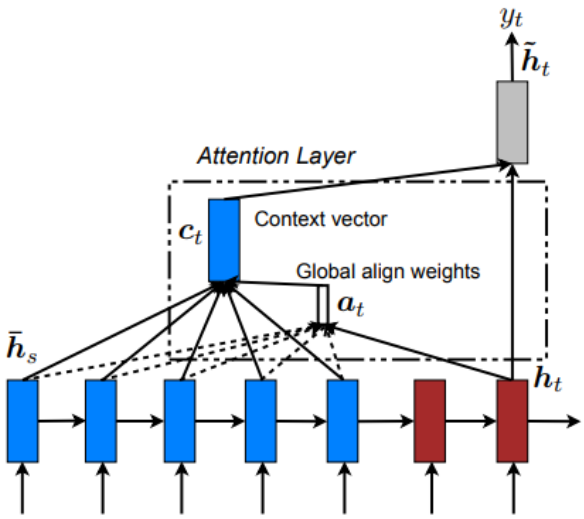
使用兩層 LSTM，Ecoder 長度為影片的 frames，Decoder 則為自己設定，依照 caption sequence 調整。第一層的 input 為影片 features，必須把 decoding stage 的 input 輸入 padding。第二層的 input 則為 caption sequences，前面 80 個 time steps 的輸入為 paddings。

Implementation Details

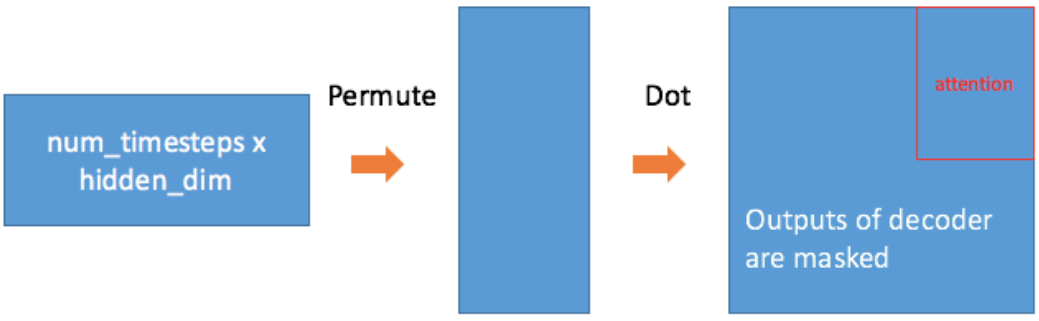
因為一開始選用 Keras，後來發現直接加兩層 LSTM 會有 Encoder 的 loss 算到 categorical_cross_entropy 的問題。因此在過了 Dense 之後做 soft max 前需要用一個 input tensor 將 encoder time steps 的輸出強制歸零。

Attention Mechanism

My Implementation of Attention Mechanism



使用參考論文 [2] 裡面描述的 dot 的方法，比較方便在 Keras 上實作。在 LSTM2 之後，將 output (None, 121, hidden_dim) 轉置然後做內積，得到一個 (121, 121) 的 tensor，為了只保留我們要的 weights，會用一個 input mask tensor 將只有 encoder output 內積的地方跟 decoder outputs 去掉。將 weight 與 LSTM2 outputs 內積後再 concatenate 回原本的 LSTM2 output 得到 (121, 512) 的 tensor 再去做 Dense 和 Activation。如下圖：



Results between with and without Attention

ID	S2VT	S2VT + Attention	Ground Truth
XOAgUVVwKEA_8_20.avi	a girl is eating a baby	a woman is sitting	A baby girl is eating spaghetti.
sZf3VDsdDPM_107_114.avi	a man is applying makeup	a man is speaking	The man smiled as he took a bite of food.
ufFT2BWh3BQ_0_8.avi	a baby panda is climbing	a panda is climbing a duck	Two pandas are playing with each other.

由上表可以看出，attention 能得到比較好的影格與影格中的理解，例如 a baby girl is eating 被看成 a girl is eating a baby，經過 attention 看出坐著的狀態。Two pandas are playing 則能看出有兩隻動物，雖然沒有看對動物種類，但經過人眼辨識發現被爬的那隻熊貓不太明顯，所以仍然可以說是有進步。

How to Improve Performance

在訓練的過程中發現多選用不同的 captions 組合相當有效。即是同一組 captions（一個 batch）不用訓練太多個 epochs，隨機多使用幾組不同的 captions。例如一開始用 10 組 captions 訓練了的 bleu score 剛好過 baseline，使用 20 組之後有提升到 0.27。可能是多看不同的字詞組合會增加學到的東西。0.27 的參數是 hidden_dim = 256、20 iterations，model.fit 的 batch_size = 32、epoch = 5。

Experimental Results

Environment Setting

OS	Ubuntu 16.04.3
CPU	Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz
GPU	NVIDIA GeForce GTX 750 Ti
RAM	16G
Library Version	tensorflow-gpu (1.4.0), numpy (1.13.3), Keras (2.0.9)

Results of S2VT model

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 121, 4096)	0	
lstm_1 (LSTM)	(None, 121, 128)	2163200	input_1[0][0]
input_2 (InputLayer)	(None, 121, 3726)	0	
concatenate_1 (Concatenate)	(None, 121, 3854)	0	lstm_1[0][0] input_2[0][0]
lstm_2 (LSTM)	(None, 121, 128)	2039296	concatenate_1[0][0]
time_distributed_1 (TimeDistrib	(None, 121, 3726)	480654	lstm_2[0][0]
input_3 (InputLayer)	(None, 121, 3726)	0	
multiply_1 (Multiply)	(None, 121, 3726)	0	time_distributed_1[0][0] input_3[0][0]
activation_1 (Activation)	(None, 121, 3726)	0	multiply_1[0][0]

Total params: 4,683,150
Trainable params: 4,683,150
Non-trainable params: 0

None

上圖為原本的 S2VT 模型，LSTM units 設為 256，訓練 20 個 iterations，每個 iteration，經過 20 iterations 後得到 Average bleu score 0.272592008029。經過實驗發現 LSTM units 越大 overfit 的情形會越嚴重，後來就維持 256。

Reference

[1] S. Venugopalan, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence video to text. In Proc. ICCV, 2015.

[2] Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In Conference on Empirical Methods in Natural Language Processing (2015).