HW2 report — Seq2seq & Attention

R05943135 杯屎賴—

R05943135江承恩 R05943011 沈恩禾

R05921016 傅鈞笙 R05942072 吳昭霆

環境

CPU	GPU	Memory	OS	Libraries
Intel(R) Xeon(R) CPU E5-2630 v2 @ 2.60GHz	GTX 980	HYNIX HMT42GR7BFR4C-RD MEMORY 16GB * 8	Ubuntu 15.04 Mint 17.1 Rebecca	Tensorflow 1.0

Model 參數

- ***LSTM cell*** **：** LSTM 組成單元，使用 tf.contrib.rnn.LSTMCell
- ***hidden size*** **：** LSTM 單元之 input , output 皆為 unit_num 大小的 float32 向量。
- ***num layers: LSTM 堆疊層數***

DEJ	Attention 1	Attention 2
2	1	

- ***Single Decoder –***

  Lstm 的 input dimension 為[batch_size * hidden_size] * 2，在 decoding stage 時，分別將 embedding 過後的 video 以及 zero padding(此時不須 decode 文字)當作 input；而在 encoding stage 時，改成將 zero padding(此時不須 encode 影片)以及 embedding 過後的 caption 當作 input。

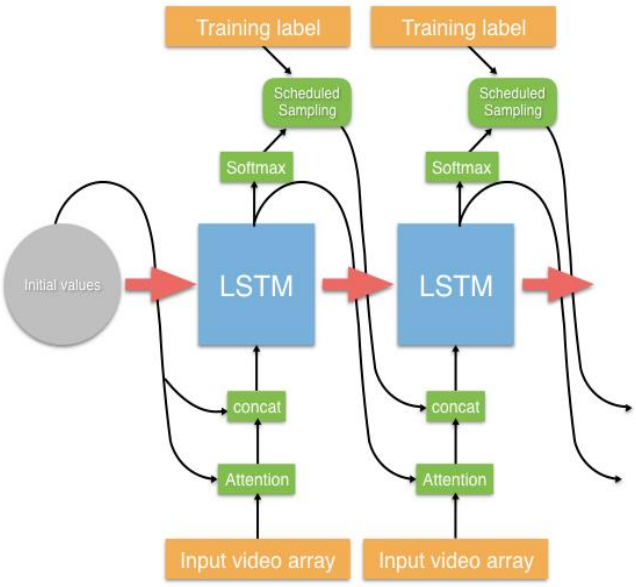
- ***Decoder + Encoder jointly trained –***

  Double-layer lstm 的架構，第一層的 decoder input dimension 為[batch_size * hidden_size]，主要負責吃進 video inputs，第二層的 encoder input dimension 為[batch_size*hidden_size] * 2，負責吃 caption inputs，和 single decoder 雷同但差別在第二層的 lstm 會多吃第一層的 lstm 的 output。

- **Attention 1–** *based on the reference paper: Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle and Aaron Courville(2015). Describing Videos by Exploiting Temporal Structure(ICCV 2015).*
  - **Key work –** $\mathbf{p}_t = \text{softmax}(\mathbf{U}_p \tanh(\mathbf{W}_p[\mathbf{h}_t, \varphi_t(V), \mathbf{E}[y_{t-1}]] + \mathbf{b}_p) + \mathbf{d}),$
    為了得到 decoder output p_t， 將前一個 time step 的 hidden state h_t ，前一個 time step 的 decoder output y_t， 以及在 encoding stage 每一個 time steps 的 outputs 通通 collect 起來當做所謂的 attention states V， 通過 linear 的 transform 以及 non-linear 的 activation 後產生而成。
- **Attention 2–** *based on the reference paper: Dzmitry Bahdanau, KyungHyun Cho and Yoshua Bengio(2015). Neural Machine Translation by Jointly Learning to Align and Translate(ICLR 2015).*
  - **Key work –** 主要修改 tensorflow 已經包好的 wrapper tf.contrib.legacy_seq2seq.embedding_attention_seq2seq 的 function, 將其原本的 EmbeddingWrapper 換成 InputProjectionWrapper，這樣替換 input 端才可以符合我們這次的需求。而演算法的部分 tf.contrib.legacy_seq2seq.embedding_attention_decoder 所使用的演算法即是這篇 paper 所用到的。

Model 描述

- **Model input:** Video sequence.
- **Model output:** A sentence of caption.
- **Attention based model:**
  使用上個 step 的 LSTM output 計算 attention.
- **Scheduling Sampling:**
  輸入 video array 經由 attention 後，會與另外一個 array S 串接，作爲 LSTM input，Array S 可以是來自上個 step 輸出的 most likely caption，或是 Training label，由 Scheduled sampling 決定。



Training Data 處理參數

- **video_size, video_step**： 4096*80，影片 feature 維度。
- **caption_size**： 擷取所有 training label 最多共 6193 個單字，並加上起始<bos>，結束<eos>，填補<pad> Token。
- **caption_steps**： 最長可能輸出句子長度，產生較短句子皆會補足到 caption_step 長度。
- **Sentence concatenation**：將多組句子組合為一組 train_num_steps 長度的句子，即跨句 training。
- **Training label choice**：分為 all, one。可選出所有句子當 label，或者選出一句長度接近 caption_step 者使用。皆使用 all。
- **Training bound**：可做為 label 最長之句子字數，使用 20。

Training 參數

- **batch_size**：DEJ model 使用 10，Attention 系列使用 100,128。
- **learning rate**：learning rate 起始值，使用 0.001。
- **sampling_choice**：傳入 linear, exponential, inverse sigmoid 之機率分布給 scheduled sampling 使用，僅嘗試 linear，至 500,1000 epoch 機率下降為零
- **使用 AdamOptimizer**。

Pre-trained Embedding

GloVe： Global Vectors for Word Representation by Jeffrey Pennington, Richard Socher, Christopher D. Manning - Standford NLP Group

- 使用 glove.42B.300d, glove.6B.100d, glove.6B.300d
- **42B/6B**:GloVe training 文本大小
- **300d/100d**: embedding 向量大小
- glove.42B 共有 1.9M 單字，glove.6B 則有 40K 單字。

實驗

DEJ :

Bleu	Num vocab	Hidden size	Epoch	Learning rate	embedding	meaningful	Replicated words
NaN	6043	300	363	0.001	non	Unrelated	No

Attention 1:

Bleu	Num vocab	Hidden size	Epoch	Learning rate	embedding	schedule	meaningful	Replicated words	embedding
28.11	6193	300	420	0.001	yes	non	主詞正確,動詞無關,無受詞	No	42B 300d
29.98	6193	300	380	0.001	Yes	yes	主詞正確,動詞接近	severe	42B 300d

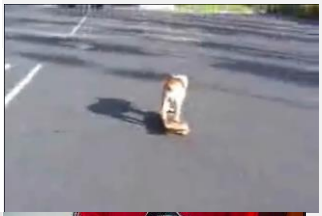
Compare different method：

Exp	Bleu	Captions
A	0.236	A man is cutting a box of food
B	0.342	A girl is cutting a
C	0.458	A woman is cutting an
D	0.366	a woman is slicing a woman

A	0.386	a baby is playing with a white ball
B	0.271	a cat is a a
C	0.32	a cat is on a
D	0.263	a man is cutting a small ball

A	0.322	A man is riding a bike
B	0.279	A man is riding a
C	0.279	A man is riding a
D	0.357	a man is riding a motorcycle

A	0.454	**A** man **is riding a** motorcycle
B	0.407	A girl is **riding** a **on** on a
C	0.41	A **dog** is on the road road speeding
D	0.454	a man is riding a motorcycle



A	0.454	A man is slicing a red onion
B	0.407	A man is slicing a red red
C	0.308	A man is slicing **an onion**
D	0.314	a man is slicing a man is cutting a knife

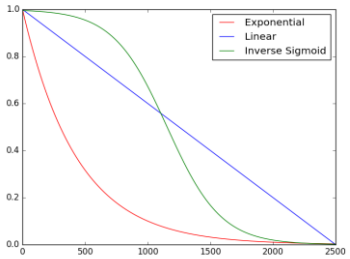
A	attention w/o scheduling	epochs:420
B	attention with scheduling	epochs:380
C	attention with scheduling	epochs:430
D	attention with scheduling&word embedding	epoch 100

嘗試改進方法

- **schedule：based on the reference paper: Samy Bengio, Oriol Vinyals, Navdeep Jaitly, Noam Shazeer, "Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks"：**

  實作三種 schedule — linear, exponential, and Inverse sigmoid，如圖：

  其中縱軸值代表選取 Training Label 的機率，



分析

- 在 attention 1 中，加上 schedule 之後，雖然 BLEU 分數提升不多，但是實際上看輸出句子的話，發現可輸出動詞、受詞等，可見對於結果是有正面的提升。
- 有時候只單看 BLEU 分數會不太準確，因為 A , man , is ,此類詞可輕鬆使 BLEU 分數提高，可見 BLEU 在 30 以下並不能代表 cation 的精準性。
- Scheduled sampling 使主詞有機會跳脫 a man 此類能輕易降低 training loss 但詞意不清確的 cation，動詞也容易產生較精確的選擇，但 training 時間不夠長的話無意義重複字詞很多，受詞也無法產生。
- Model 明顯偏好特定主詞與動作，可能表示 training data 較少。就算 training 時間不夠，整體比現不好，仍有些許影片能產生完整又正確的 caption，如在碗裡攪拌食物，切菜等等。

分工

江承恩	沈恩禾	傅鈞笙	吳昭霆
lstm decoder+attention	Pretrain + Tune model	lstm decoder+attention	Fix little bug