MLDS HW2

R05921035 陳奕安
R04921055 劉叡聲
R05921043 林哲賢
R05548020 吳侑學

Model Description

Model以類似S2VT的架構為基礎，有兩層的RNN分別作為encoder和decoder，並分為encode和decode兩個階段。而RNN cell使用LSTM加上DropoutWrapper。

Encode

此階段的step次數依據影片的frame數量而定。在每一個RNN的time step，會將4096維的frame feature經過一個full conneted layer作為embedding後，作為第一層RNN－即encoder部分的input。最後保留此層RNN在每一個time step的output留作attention。

Decode

此階段的step次數則是根據caption的字數而定。在每一個time step，將real word經過一層embedding後作為第一層RNN的input，並取其output與Encode階段的output作attention。
得到attention vector後，將attention vector作為第二層RNN的input，再將output經過一層full connected layer得到one of n的機率分布vector。

Attention

Attention部分採用[1]中的Local Attention，將Decode階段第一層output(h)與兩需學習參數運算後乘上frame數量得到要做attention的中心位置p。並取Encode階段該中心前後一範圍內的第一層RNN output，與h和另一個需學習attention參數運算得到一個align score，最後將align score乘上一個以t為中心的gaussian distribution後作為最後的權重，將權重乘上對應vector後即是attention vector。

參數配置

one-hot word dimension = 3000
frame embedding = 300
word embedding = 300
RNN hidden layer neuron = 200

Encoding階段step = 20
Decoding階段step = 45
Local attention範圍 = 2

How do you improve your performance

Schedule Sampling

在training的過程中根據一機率選擇使用real word作為decoding階段第一層RNN的input或者是用上一個step預測的結果作為input。
而此一機率在剛開始training的過程會接近1，並在過程中逐漸遞減，在此我們的model中讓這個機率linear的遞減。

不過我們在implement的過程中因為一個意外的bug發現把scheduled sampling反過來做竟然能train出比較好的結果，所以我們最後的版本是在一開始decoder是feed_previous的在後面才開始拿gorund truth作為input。

Beam Search

在Testing時，若僅用greedy的方式，找到的句子很容易並非是真正機率最高的word sequence，因此在每一個step時分別找出到目前為止機率最大的3個path直到最後一個step，以此方法可以增加找到最高機率句子的機會。

在實作上需要將decoder每個step分拆開來再取每個one hot(去除padding unk id)前beam_size再重新feed進LSTM Cell，因此架構上需要進行大幅度的修改，所以在最後版本的model中來不及實作上。不過在github上的beamsearch branch有實做完成的版本。

Encoding Layer

在encoding時嘗試以兩層的RNN作為encode。
也有嘗試過用bidirectional rnn當作encoder再取forward state和cell往後傳作為decoder stage的RNN cell。
不過效果似乎不好，因此最後也沒有採取這個版本。

Experiment settings and results

Preprocessing

由於這次的作業是要產生對影片敘述的句子，因此在文字的前處理部分不需移除stop word，否則少了定冠詞與be動詞等字眼的句子並不像自然語言。因此直接對training data中的所有字作word count，得到共6085個字。將出現次數只有1次、2次字刪除－這些字或許是其他字的同義詞，或者是影片中較少人注意的細節，因此在caption中出現次數較少，對於簡述一個影片內容並不具有太大意義－之後得到3000個字作為我們的dictionary。

對於每個caption，我們利用每個字在dictionary的index將其轉換成對應的數字，加入<BOS>和<EOS>，並加入<PAD>一直到45個字。

每一個影片總共有80個frame，但影片的時間卻大概只有5秒上下，因此我們認為80個frame之中有太多重複的資訊，並不需要全部放進RNN中，因此我們每四個frame取一個feature，總共使用20個frame來encode。

Experiment Attempt

1. 有paper說將frame反過來丟入encoder，可以學得更好，但我們試出來的結果略低於正常的丟法

Team division

- 陳奕安（R05921035）：preprocessing、S2VT model
- 林哲賢（R05921043）：beam search 、S2VT model
- 劉叡聲（R05921043）：preprocessing、S2VT model
- 吳侑學（R05548020）：attention model inplement, try global and local attention

Reference

[1] Effective Approaches to Attention-based Neural Machine Translation
http://www.aclweb.org/anthology/D15-1166 (http://www.aclweb.org/anthology/D15-1166)