HW2-1 Video caption generation

- Model description (3%)

架構為encoder-decoder，此處我們採單向的LSTM模型，其實與baseline model相去不遠，以下為參數：

LSTM dimension = 256

Learning rate = 0.001

Training epochs = 90

Batch size = 50

Word threshold = 3

將80*4096的video feature傳進encoder，encoder最後的final state則作為decoder的initial state，最後decoder會輸出256維的feature，再來output layer會將此feature轉成字典裡所有字數的維度，取argmax之後作為輸出。

在model中，我們加上了Attention model，採用tensorflow中的LuongAttention。

此外，還加上Schedule Sampling，使用tensorflow中的ScheduledEmbeddingTrainingHelper，sampling_probability =0.2。

而在文字的前處理部分，將每段台詞的標點符號移除，並加入了<bos>於句首、<eos>在句尾、<unk>取代出現頻率很低的單字、<pad>來補齊句子，將單字出現次數小於5次的移除等等。

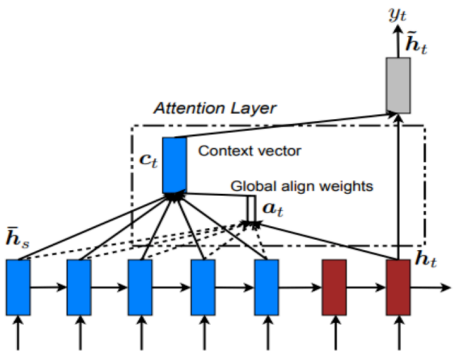
- How to improve your performance (3%)

(e.g. Attention, Schedule Sampling, Beamsearch...)

1.Write down the method that makes you outstanding (1%)

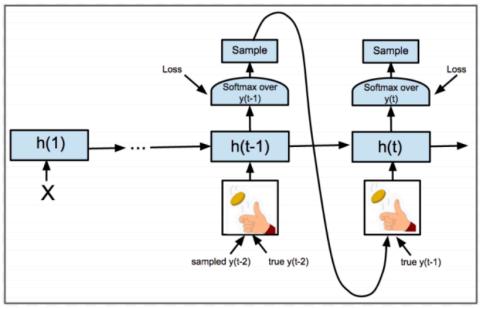
a.Attention

用tensorflow的LuongAttention，將所有encoder output作為memory傳進attention layer，對每個decoder output計算對每個encoder output的alignment score，所有score normalize後將每個encoder output與score相乘後加總成為context vector，最後同時考慮context vector及decoder outputs產生最後的結果。



b.Schedule Sampling

用tensorflow的ScheduledEmbeddingTrainingHelper，隨機讀model的output或直接讀reference的input，避免都看model的output會很難train，及直接讀reference的input會造成training跟testing不一致。



2.Why do you use it (1%)

因為實驗後，在BLEU@1 score有上升，表示這些tips有助於training的改善。

a.Attention

如果沒使用attention model，容易在訓練的過程中過度集中注意於某個frame上，而造成誤差。

b.Schedule Sampling

在訓練seq-to-seq時，還有個問題，training過程我們是以標籤的詞來當作input，但是在testing時，卻是拿上一個階段預測出的結果作為下一個階段的input，這會產生exposure bias問題。

而若是在training過程中直接以預測的結果當下一階段的input，則會有不穩定的現象出現。

因此需要使用Schedule Sampling。

3.Analysis and compare your model without the method. (1%)

a.Attention

	無Attention model	LuongAttention	BahdanauAttention
BLEU@1 score	0.5994	0.6059	0.5867

由結果可以看到，使用了LuongAttention後結果些微的上升，可見多多少少還是有影響。

b.Schedule Sampling

	無Schedule Sampling	加上Schedule Sampling
BLEU@1 score	0.5994	0.6478

由結果可以看到，有沒有做Schedule Sampling對結果影響很大，確實可以解決exposure bias問題，且使用sampling rate=0.2時效果最佳。

使用Attention mechanism及Schedule Sampling的BLEU@1 score為0.6510，又相較單做Schedule Sampling在更往上提升不少。

- Experimental results and settings (1%)

1.文字前處理

在資料前處理時，我們將所有標點符號去除，並在每句句子前加上<bos>、結束地方加上<eos>、句子長度以<pad>對齊，最重要的，我們發現出現過的文字有許多只出現一兩次，不具有代表性，反而會產生noise，最後實驗後決定調3為最佳。

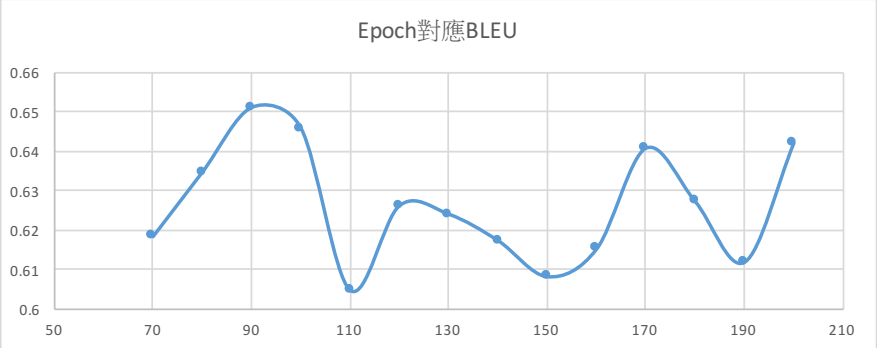
Threshold	1	2	3
BLEU@1 score	0.6252	0.6235	0.6420

2.Epoch

如果將training epochs調成300，則BLEU@1 score= 0.6213，所以如果太多epochs反而會過於overfitting，且出現的句子會過於複雜，如：ScdUht-pM6s_53_63.avi,a man is putting a piece of pizza out of a large。

Training epochs調成70，尚未收斂，BLEU@1 score= 0.6184，也不佳。

發現在最好的情況下，epoch應為90，BLEU@1 score= 0.6510。



3.維度

若dim_hidden設512，結果降為0.6245534372822313，表示如果太複雜的network也會造成結果不好。

4.Overall

word threshold : 3(剩2997個)

Attention model : LuongAttention

Schedule Sampling : sample rate = 0.2

Adam learning rate : 0.001

Epoch : 90

Loss : sparse_softmax_cross_entropy_with_logits

Accuracy : 0.6510