MLDS HW2 - Video Captioning

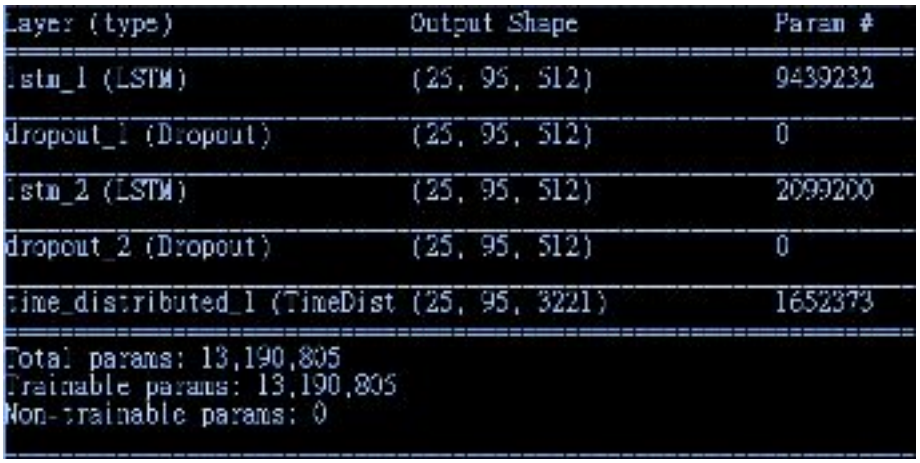
學號：B04901066 系級：電機三 姓名：洪國喨

1. Model description (2%)

使用keras實作，input=(1450, 80+15(輸出句子長度), 4096)，

兩層LSTM(512)，Dropout(0.25)

output=(1450, 80+15 , 3221(字典大小)) (activation='softmax')



因為觀察發現輸出句子通常很短，因此未實作將上個timestep output回授至下個LSTM輸入，影響為training時間變長。

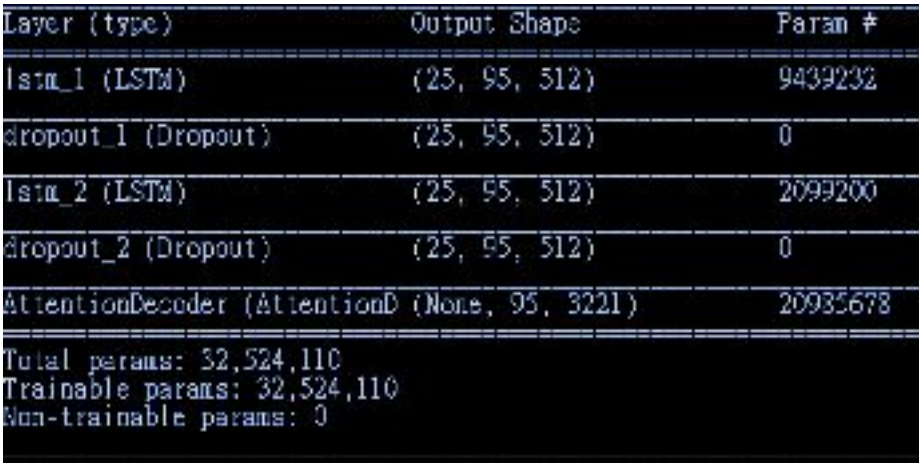
2. Attention mechanism(2%)

---How do you implement attention mechanism? (1%)

使用github code：

https://github.com/datalogue/keras-attention/blob/master/models/custom_recurrents.py

複製hidden state到序列，然後以 權重矩陣 乘上 重複的隱藏狀態，計算attention probabilities，再求出context vector。接著更新狀態：計算"r"、"z"gate，r-gate決定是否記憶狀態，z-gate決定step幅度，求得proposal hidden state，最後依此更新狀態即可。



---Compare and analyze the results of models with and without attention mechanism. (1%)

實際比對輸出字串，發現輸出內容之差異很小，新舊BLEU分數之差異也很小，主要明顯的影響是訓練的時間變長(參數2.5倍，時間約變為1.5~2.5倍)，而loss(cross entropy)也下降的較慢，推測是model變厚所致。猜測是因為LSTM本身就會做一些attention，而

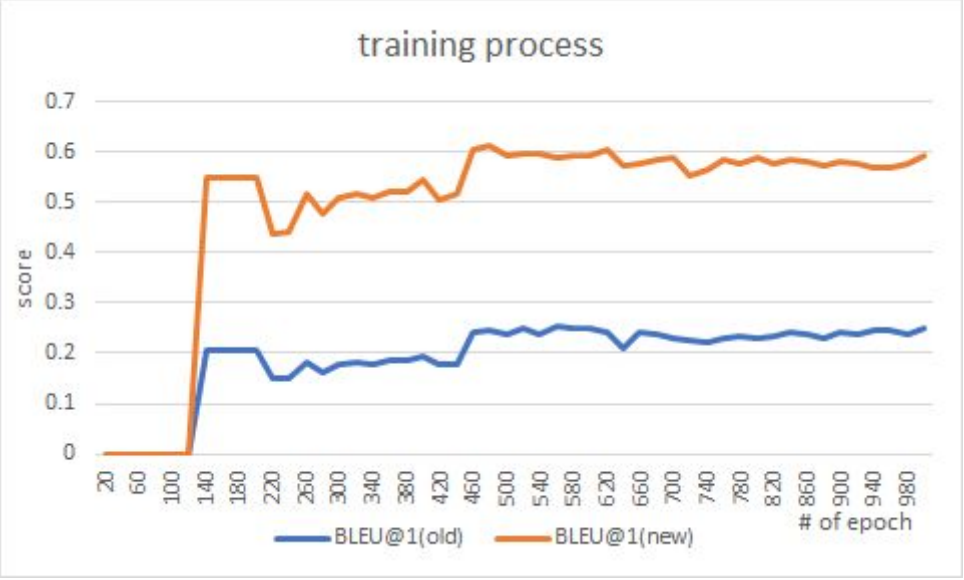
且我在字典有做特殊處理，所以才導致這種結果，若是使用最原始的simple seq2seq模型，使用attention mechanism之優化效果應該會較為明顯。

3. How to improve your performance (1%)
---Write down the method that makes you outstanding
---Describe the model or technique (0.5%)
---Why do you use it (0.5%)

使用One Hot Encoding。一開始字典直接使用單字(即將caption以空格分開)，訓練出來的模型輸出一直是 'A man is a is a a a.' , 'A woman is.'...之類的句子，因此我決定在字典動手腳，字典本來存單字，改成存詞組甚至片語，如此一來可強迫模型輸出較為豐富，採用以下演算法對caption做前處理：
取代逗號、句號為空白，取代" a "為" a_"，取代" an "為" an_"，取代" the "為" the_"，
取代" one "為" one_"，取代" two "為" two_"，取代" three "為" three_"，
取代" some "為" some_"，取代" there is "為" there_is "，取代" there are "為" there_are "，取代" is "為" is_"，取代" are "為" are_"，最後以空格切出詞組放入字典即可。
如此訓練出來的模型BLEU@1至少是2.0起跳。

因為未實作output回授，所以也未實作Schedule Sampling(導致增加訓練時間)和Beamsearch(導致降低BLEU分數，然而不一定表示輸出較不合理，請參考第四題後面之討論)

4. Experimental results and settings (1%)



由上圖(unit=128)可知，訓練愈久，BLEU分數會越高(雖然增加極緩)。
另外亦實測LSTM unit size對訓練及準確度之影響(以128、256、512測試，單個epoch時間分別為41秒、24秒、21秒)，發現三者訓練相同epoch數時，輸出之BLEU分數其實十分相近，然小model前100個epoch還不會有輸出，而大model就有，表示其收斂得較快(相同epoch數)。

除此之外，還發現新舊BLEU分數，在評價輸出句子上，其實都不是一個非常好的標準。像是'A man is a'在助教提供之testing_data分數分別為0.696和0.229，而我訓練出來之模型輸出一些符合影片內容且長度較長之句子，反而可能因為句子或文法架構與caption不同，使得分數反而較低。