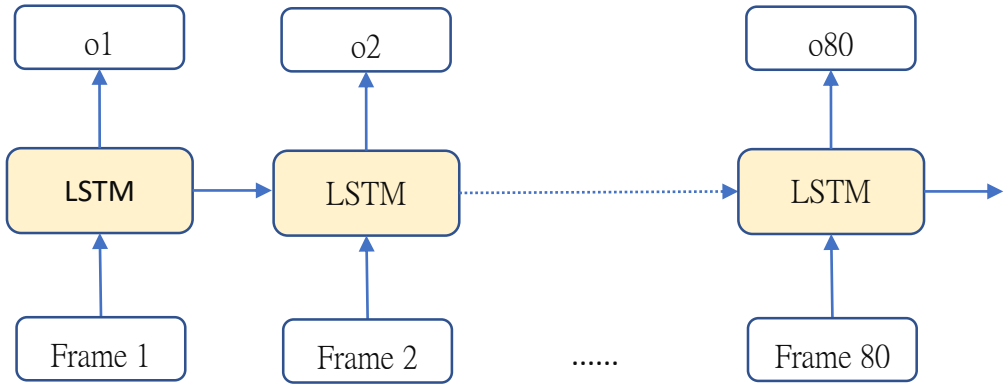
MLDS

Hw2 Report

R05725004 資管碩二 曾千蕙

1. Model description
   I. Encoder

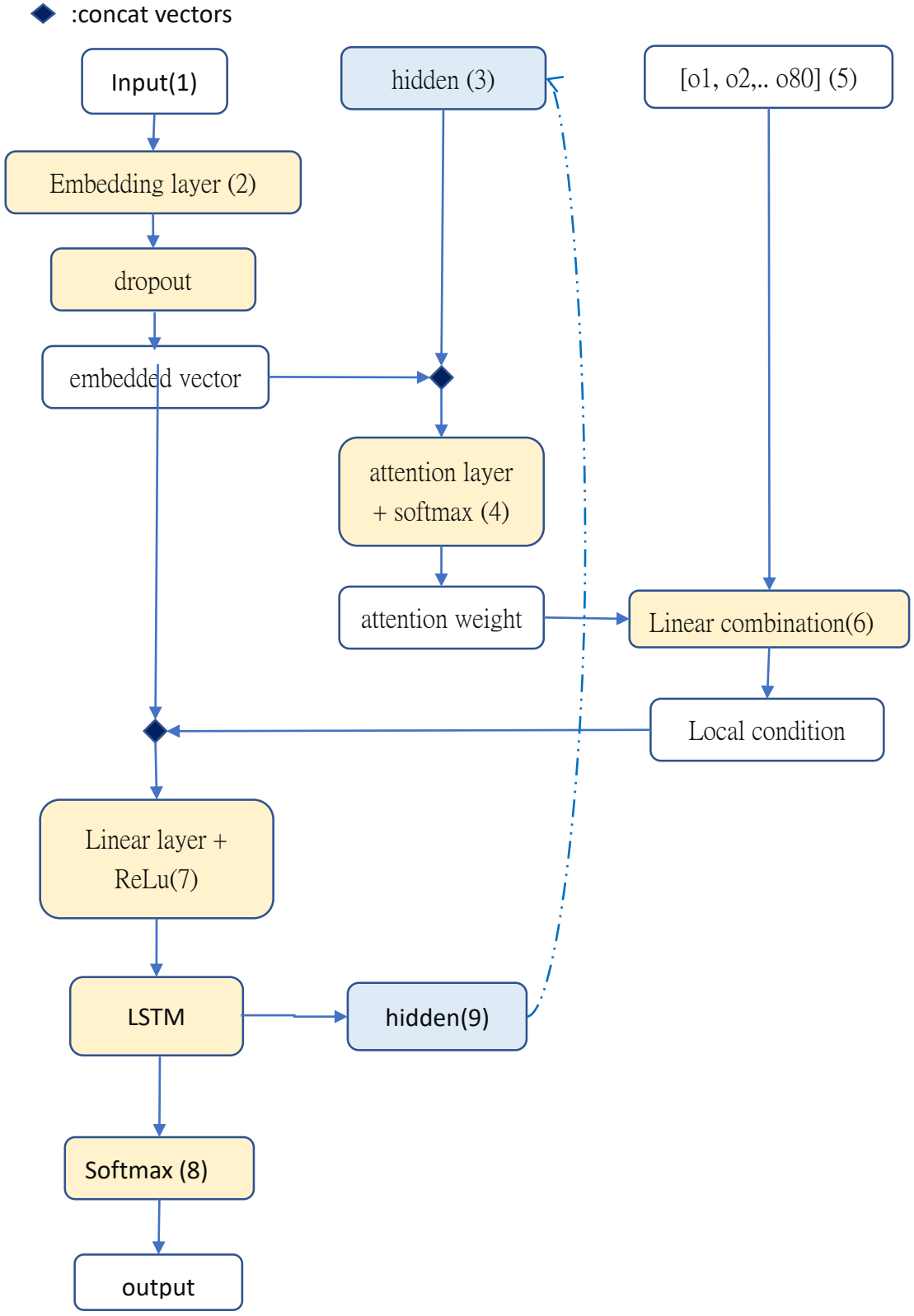
   將 input vectors(80 frames, 每個 frame 4096 個 features), 丟進 LSTM, hidden size=256



   II. Decoder (對應下頁的圖)

   將句子中的字一一轉為在字典中的 index, 再將這些 indexes 一一丟進 decoder 中

   (1) Input vector: 字所屬的 index
   (2) Embedding layer: one-hot encoding 對應到 embedding vector, 維度=256
   (3) hidden vector: decoder 的 LSTM 的 hidden vector, 用 encoder LSTM 最後一個狀態的 hidden vector 來做初始化
   (4) attention layer: 將 input 的 embedded vector 跟 hidden vector 串接之後丟進一 linear layer, output 的 units 數 = 80, 取 softmax 後每個值分別代表 encoder 各個時間下的 output vectors 的權重
   (5) [o1, o2,…,o80]代表 encoder 在每個時間點下的 output vectors
   (6) 將計算出的 attention 權重和 encoder 的這個時間下的 output vectors 做加權平均, 得出的結果可以視為帶有 local 資訊的 vector
   (7) 將 attention 過後的 vector 和 input 的 embedded vector 串接起來(維度=512)後丟進一 linear 層, output 維度 = 256
   (8) 將先前的結果丟進 LSTM, output 維度 = 字典大小, 最後經過 softmax 得出每個字做為最後結果可能的機率


:concat vectors
Input(1)
hidden (3)
[o1, o2,.. o80] (5)
Embedding layer (2)
dropout
embedded vector
attention layer
+ softmax (4)
attention weight
Linear combination(6)
Local condition
Linear layer +
ReLu(7)
LSTM
hidden(9)
Softmax (8)
output

2. Attention mechanism

(1) 如同上題所述的流程，將 input 的 embedded vector 跟 hidden vector(代表 encoder 看完整串 input 所得到的 global 的資訊)串接之後丟進一 linear layer, 計算出 encoder 各個時間下的 output vectors 的權重, 將 encoder 的 output vectors 依照此權重做加權平均, 得到的 vector 可以視為是強調 encoder 局部資訊的 vector, 將此 vector 做為 decoder LSTM 的 condition 使用。

(2)

隨機找 4 個影片比較有無 attention 機制之差異：

id	No attention	With attention	Labels
UXs3eq68ZjE_250_255.avi	A man is frying a potato	A man is adding water to a pot	Someone is pouring rice into a pot of water.
s1ZABV7AQdA_38_48.avi	A small girl is flying a basket	The men are getting up a street	The men are running down the street.
inzk2fTUe1w_1_15.avi	A man is a banana	A boy is the a piece of paper	Someone peels a banana.
qvg9eM4Hmzk_4_10.avi	A man is shooting a gun	A man is the the car by the rear bumper	A young man is lifting a pickup by the rear bumper.

發現有 attention 的 model 較容易判斷出正確的動作與主要的人事物，且 BLEU score 也較高(0.631 v.s. 0.596)

3. Methods to improve performance

I. Attention

同上題所述

II. Scheduled sampling:

如果訓練的時候都只有給 decoder 看標準答案，當測試的時候很容易出現標準答案中沒有出現過的結果，此時很容易造成後續的結果也都接連著錯，因此在訓練的時候也讓 decoder 有機會使用自己預測的字做為下一次的 input, 增加模型預測的穩定性，做法是每次隨機產生一個 0~1 的數值，若數值在 0.5 以下則使用標準答案，反之使用 decoder 自己的預測的字。而且由於 decoder 自己預測的結果容易變動，會造成模型很難收斂，因此選擇在模型 error 下降到一定程度的時候(觀察大約 negative log likelihood 低於 0.8 時)，才開始使用 decoder 自己預測的結果。

4. Experimental results and settings

I. Scheduled sampling：

	Blue score
全部使用標準答案	0.24
標準答案與 decoder outputs 機率各 0.5	0.268
全部使用 decoder outputs	0.261

雖然全部使用 decoder outputs 的結果看似沒有差太多，但訓練時收斂速度很慢，且相當多不完整的句子。

id	標準答案與 decoder outputs 機率各 0.5	全部使用標準答案	全部使用 decoder outputs	Labels
UXs3eq68ZjE_250_255.avi	A man is adding water to a pot.	A person is mixing ingredients in a bowl.	Someone man is some powder a pot.	Someone is pouring rice into a pot of water.
s1ZABV7AQdA_38_48.avi	The men are getting up a street.	The men are riding in the street	A are is a.	The men are running down the street.
inzk2fTUe1w_1_15.avi	A boy is the a piece of paper.	A woman is pouring a brown liquid from a glass	A woman is a a of	Someone peels a banana.
qvg9eM4Hmzk_4_10.avi	A man is the the car by the rear bumper.	A man is riding a bike.	A man is riding on skateboard.	A young man is lifting a pickup by the rear bumper.

II. Decoder 的初始化 hidden vector 使用 encoder 最後一個 output vector 或使用最後一個 hidden vector, bleu score 分別為 0.268 和 0.271, 結果差異不大。

III. Beam search

本來認為 decode 時考慮較多路徑應該可以得到較好的結果，但不知為何反而結果較差(bleu score = 0.123)。

id	beam search	No beam search
UXs3eq68ZjE_250_255.avi	The person poured water to pan.	A man is adding water to a pot
s1ZABV7AQdA_38_48.avi	A are riding the bicycle	The men are getting up a street
inzk2fTUe1w_1_15.avi	The boy showed a zucchini	A boy is the a piece of paper
qvg9eM4Hmzk_4_10.avi	A guy drove up car	A man is the the car by the rear bumper