MLDS HW2-1 Report: Video Caption Generation

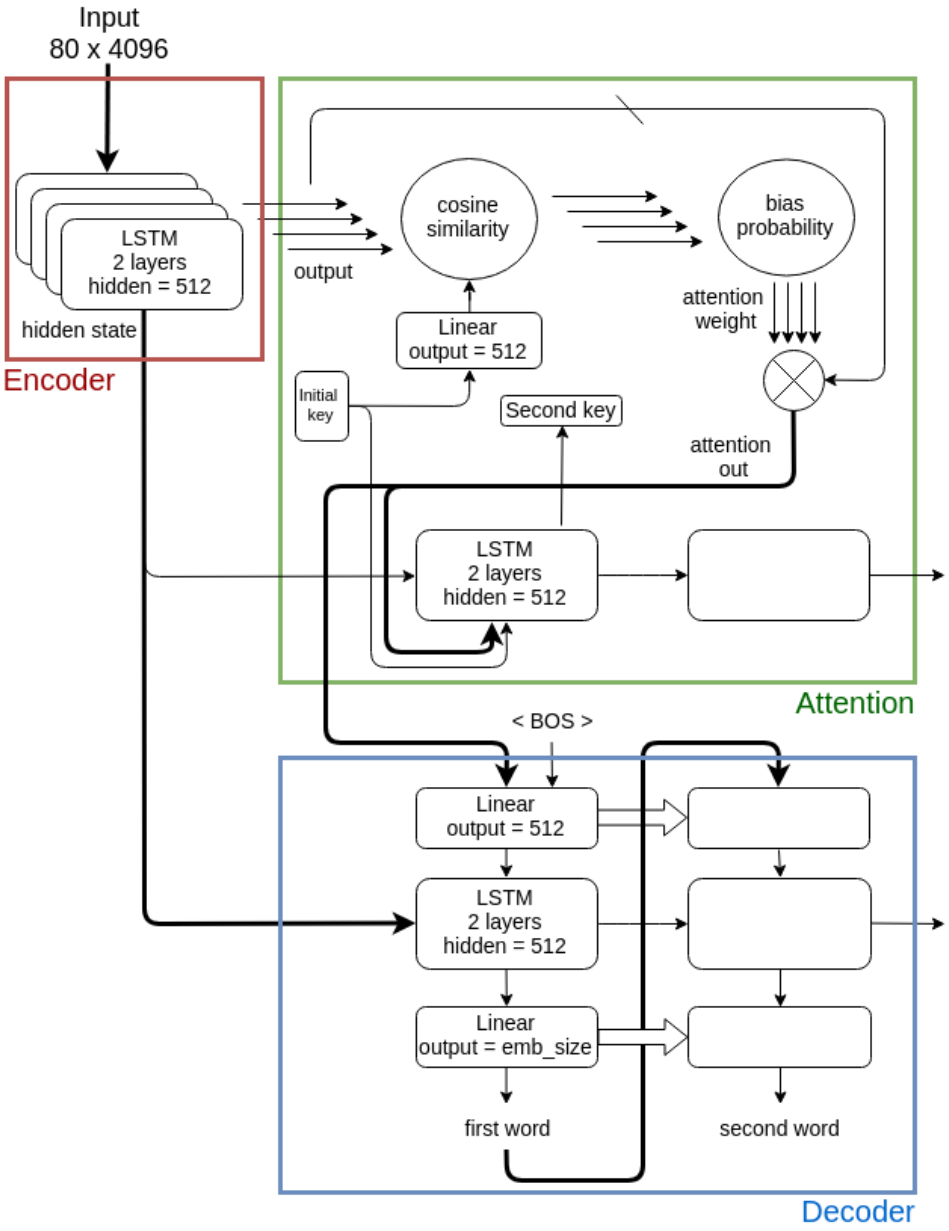
組員：b04901060 電機三 黃文璁　分工：code: [2-1, 2-2], report: [2-2-1]
b04901003 電機三 許傑盛　code: [2-1, 2-2], report: [2-1-1, 2-1-2, 2-1-3]
b04901096 電機三 蔡昕宇　code: [2-2], report: [2-2-2, 2-2-3]

2-1-1 Model Description

本次作業 2-1 實作的 seq2seq 改進方法為 attention。

我們的 model 主要分成三個部份：Encoder、Decoder、KeyRNN。Encoder 為兩層 LSTM，hidden size = 512，Decoder 由一層 linear 接上兩層 LSTM 再接上一層 linear 做為 output。KeyRNN 的部份由兩層 LSTM 加上一層 linear 所組成，另外還有一個 trainable 的 initial key 用來作為一開始拿來 attention 的 key。另外我們的 KeyRNN 和 Decoder 的 RNN 皆使用 Encoder 在最後一個 timestamp 的 hidden state 作為初始的 hidden state，attention 的部份我們使用 cosine similarity 來作為 attention 的權重，最後我們並沒有直接過 softmax，在後續的問題中我們會再做討論。

使用的模型架構如下圖：

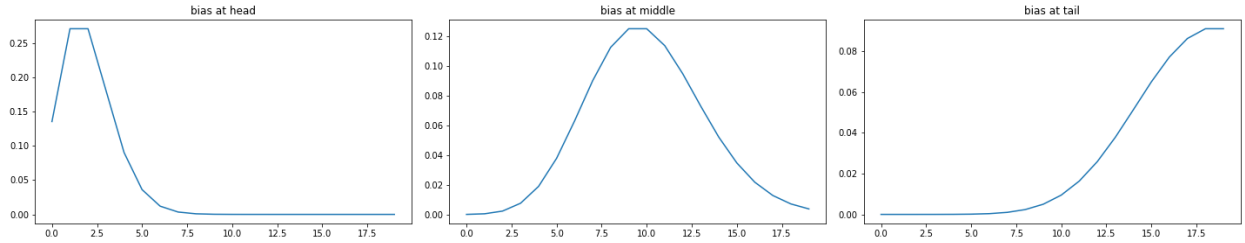


2-1-2 How to improve your performance

1. Methods that makes you outstanding

在這次的實驗中，我們主要針對 attention 的架構去做設計。首先我們的 attention 機制並不是直接使用 Decoder 的 hidden state 拿來跟 Encoder 的 output 計算 attention weight，而是像上面圖中所表示那樣，由另外的 RNN 來訓練 key 去計算 attention weight，這樣的方法有個好處是我們讓 attention 跟 Decoder 能各自分工，會這樣設計的原因是因為上課理解錯誤 XD，後來是經由組員提出疑問好像跟一般的 attention model 不太一樣，才發現這個特別的 attention 機制。

另外我們算完 attention 的 score (cosine similarity) 後並不是直接計算 softmax 得到 attention weight 而是會經過一層由不同 caption 位置所決定的 Poisson distribution，比如說如果我們對同一個影片取前端或中間或尾端的 caption 各會經過以下圖的 bias distribution 之後再過 softmax 得到 attention weight：



2. Why do you use it?

我們之所以會這樣設計主要有兩個原因，第一針對同影片我們的訓練資料有很多可能的 caption，但在訓練時輸入的影片是同一串 vector，因此在還沒採取此方法前我們在嘗試讓 model 對同個 input 產生不同的 ouptut，可能會讓 model 在訓練時有點無所適從。第二我們觀察幾個影片的 caption，發現 caption 對應時間軸有一定相關性。因此我們決定使用這樣的方法來計算我們的 attention，我們強迫模型在特定時間會有較高的 attention 特性，算是另類的教導？

在設計的過程中，我們嘗試了幾個不同的 attention 機制：

(1) cosine similarity -> softmax
(2) cosine similarty -> None
(3) cosine similarity -> bias distribution
(4) cosine similarity -> bias distribution -> softmax

我們會在後續的問題中再做探討

3. Analysis and compare your model without the method

使用 attention 與不使用最大的差異是使用 attention 會額外找出重要的關鍵字，例如

TZ860P4iTaM_15_28.avi：（影片中為一隻貓拍打琴鍵和用頭磨蹭琴鍵）

basic *a cat is playing the piano*
attention *a cat is playing a piano and rubbing its head*

UXs3eq68ZjE_250_255.avi：（影片中有人拿著木製杓子準備攪拌鍋中米飯）

basic *a person is stirring rice a wooden*
attention *someone is stirring a pot of rice with a wooden*

8HB7ywgJuTg_131_142.avi：（影片中有人鍋產翻炒著平底鍋中食物）

basic *a woman is mixing ingredients in a frying*
attention *a woman is cooking something a skillet meat and stirs it in a wooden*

4PcL6-mjRNk_11_18.avi：（影片中狗叼著球跑向鏡頭）

basic *a dog is running into something*
attention *a dog is taking a ball*

加上 attention 中的 model 中對 decoder 來說它得到額外關於某些特定時間點物件的資訊，因此有助於它在產生長一點的句子以最小化 cross entropy，不過對於沒有 attention 的 model 來說，傳進來的資訊只有 encoder 的 hidden state，不足以讓它產生更長的句子，因此有所侷限。

不過在 attention 的 model 中，由於它會有傾向根據所得的資訊把句子拉得很長，其中有會有一些是重複的字句，或是讓文法變得非常差的情況，但這些情況都能透過 post processing 把文法調得好看一點。還有一種情況是，可能訓練資料中並沒有出現的物件，兩種 model 都沒辦法正確描述影片中的東西，不過在 attention model 中會描述的比沒有 attention 的還要更貼近實際影片中的東西：

7HcYJKMxpcg_20_28.avi：（影片中為一隻獅子走過草皮）

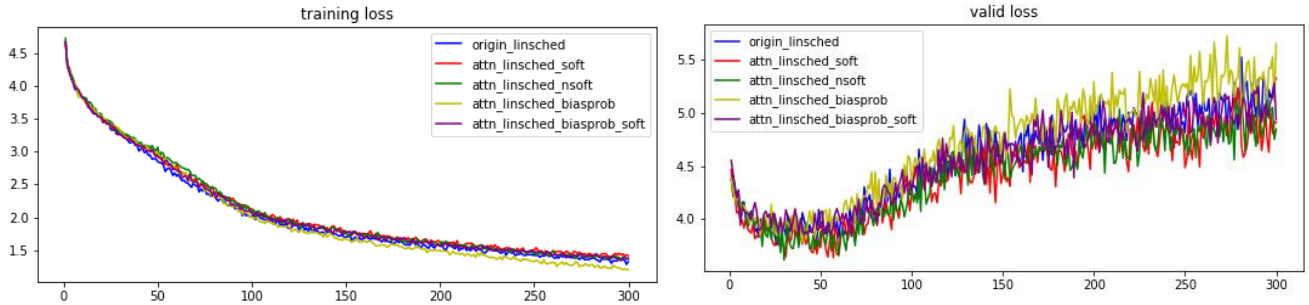
basic	*a man is walking in a huge barrel*
attention	*a dog is running through a field*

2-1-3 Experimental Results and Settings

在不同的 attention model 中，我們總共嘗試了四種不一樣的 attention 方法：

(1) cosine similarity -> softmax
(2) cosine similarty -> None
(3) cosine similarity -> bias distribution
(4) cosine similarity -> bias distribution -> softmax

第一種為比較一般的 attention 方式，第二種試著不讓 attention 過度集中在某個 frame 所以把 softmax 拔掉，第三種嘗試 bias distribution 的效果，第四種再讓 bias distribution 通過 softmax。



比較以上幾種不同的 attention 方法，訓練模型的過程的如上圖。其實在訓練過程中 training loss 下降的情況並沒有差太多，值得注意的地方是使用 bias distribution 而不加 softmax 的模型在後期的 training 中，loss 下降的速度比較快，原因有可能是我們所使用的 bias distribution 有正確達到我們所想要的指導效果，對於 overfit training data 效果卓越，不過 valid 的 loss 在這邊就不太具有意義，因為讓模型正確比對上 valid 的 caption 好像本來就不是很合理，在這邊只作為做訓練時的指標之一。在 testing 的部份，首先其實我們在未加上 attention 的 model 就能夠有相當不錯的 output 了。在加上不同的 attention 之後，其實每筆資料都會有著不同的表現，不過可大致上推論出下面幾點：

1. 加上 attention 之後能夠額外抓出與影片相符合的關鍵字
2. attention 有經過 softmax 的 model 會比較侷限產生長句子的能力
3. 有經過 bias distribution 會產生更能精緻描述場景的句子

範例：

8MVo7fje_oE_125_130.avi：（影片中一男子將塑膠盒子蓋上並做傾倒的動作）

basic	*a man is draining pasta of a plate pasta*
attention	*a man is pouring water from a plastic container*
attention + softmax	*a man is pouring a plastic container*
attention + bias	*a man is draining pasta a container*
attention + bias + softmax	*a man is pouring pasta out of a plastic container*

在上面的例子中，我們能看到有加上 attention 的模型不約而同的取出了“container”這個字，當 attention 有經過 softmax 時，它會侷限住產生長句子的能力，而選擇比較有把握的字來產生句子，而在最後的 bias + softmax model 中更另外表現出“out of”這樣的狀態。

J_evFB7RIKA_104_120.avi：（有人在用刀子切青椒）

basic	*a person is slicing a pepper*
attention	*a man is cutting a loaf of bread*
attention + softmax	*a man is slicing a pepper*
attention + bias	*a man is cutting something pieces of the and then cuts*
attention + bias + softmax	*a man is cutting a large something into pieces*

在這個例子中，attention 並沒有發揮比較大的功效，不過有幾點現象可以從這個例子中看出。首先 bias distribution 會想要表達出它切成塊狀（pieces），然後 softmax 會抑制 model 產生過長的句子。