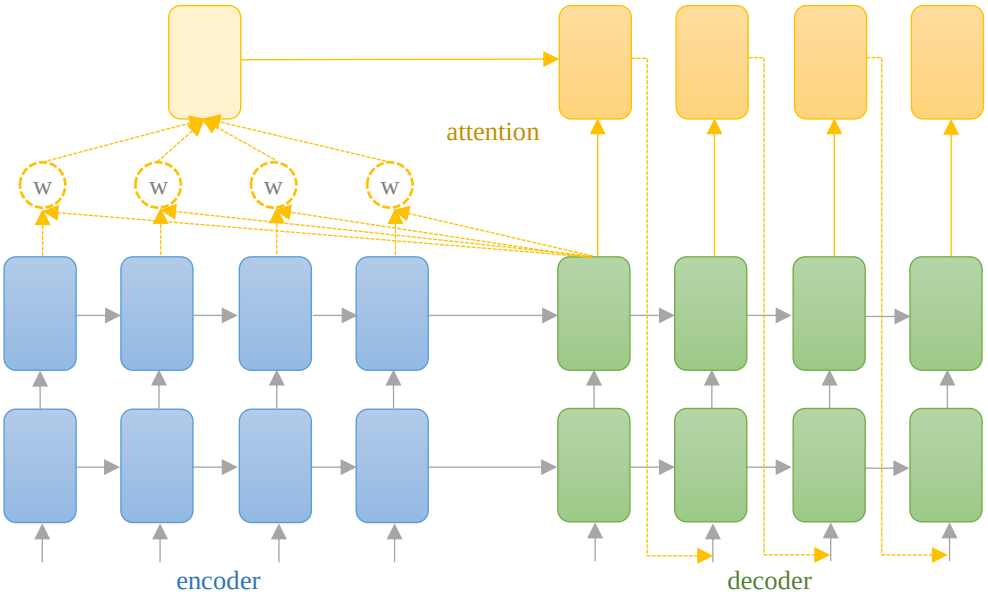
MLDS HW2-2 Report

B03901065 林宣竹、B03901101 楊其昇、B03901145 郭恆成

1. Model Description

Sequence-to-sequence model 的架構如下圖：



Encoder 與 decoder 分別由兩層 LSTM cell 建構而成，每個 LSTM cell 含 1024 hidden units，並且設定 dropout (keeping probability=0.75)，防止 overfitting 的狀況產生，然後加上 attention 的機制。

另外，embedding layer 的初始值依照 uniform distribution 產生。訓練資料在輸入 encoder 與 decoder 之前，會先經過此 embedding layer，將各個詞彙在詞典中的序號轉換成相對應的向量，詞典的詞彙量設定在五萬左右，向量長度則設為 250。詞典中除了一般的詞彙之外，還有<BOS>、<PAD>、<UNK>與<EOS>方便訓練。

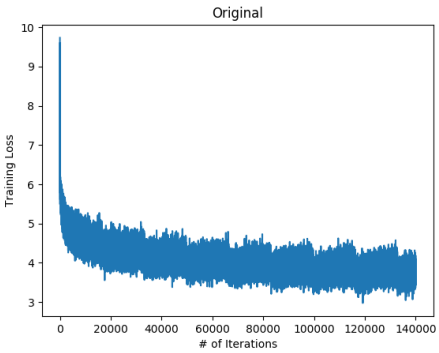
最後，由 decoder 輸出的資料會經過一層 hidden units 為詞彙量的 dense layer，且 kernel initializer 設定為 truncated normal initializer (mean=0, standard deviation=0.1)，然後得到最終結果。

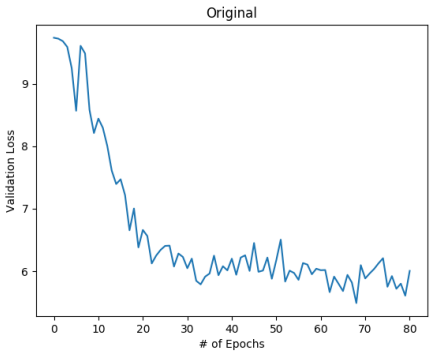
除了模型架構外，模型參數的設定如下：batch size 定為 50，optimizer 選定為 Adam optimizer，並將 learning rate 初始值設在 0.001。同時，將 gradient norm 限制在 5 之內，完成我們的訓練模型。

2. How to improve your performance

如果沒有加 schedule sampling 和 attention，training 結果如上，training loss

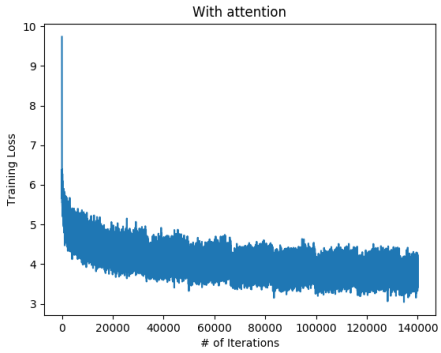
能掉到 3~4 之間，但是 validation loss 一直卡在 6 左右，所以我們嘗試使用其他技巧 improve 我們的 model。(因為運算資源有限，這裡只用 512 units，training data 取 80 萬筆)

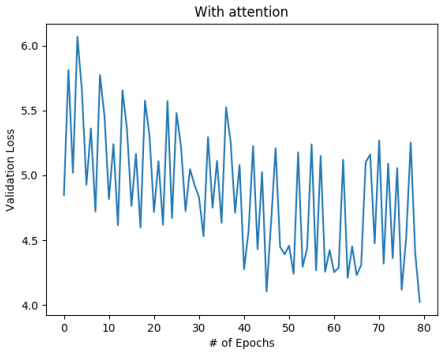




(1) Attention

我們先嘗試加了 Bahdanau attention，使得 decoder 在各個 time step 能過著重在 encoder output 不同的區段。從下圖可以發現加上 attention 之後，training loss 更快的降到 4~5 的區間，雖然之後持續在 3~4 之間擺盪，但是他的 validation loss 明顯表現更好，經過第一個 epoch 的 validation loss 就已經降到 6 以下。





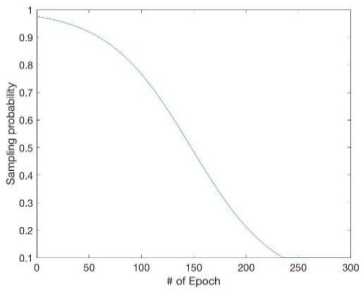

```
perplexity      : 9.102897 (baseline: < 100)
correlation score : 0.27774 (baseline: > 0.45)
```

至於評分的部分，卻過不了 correlation 的部分。雖然在訓練的過程中，我們能看到 output 與 input 蠻相關的，但仔細看了一下 testing output 的結果，發現句子常常會出現「我知道」、「我覺得」、「我是說」等比較籠統的詞，所以我們認為可能是因為在 training 時送進 decoder 的詞是 groundtruth input，並非上一個 step 輸出的詞，所以回話的表現會比較好，到 testing 時，儘管文法是合理的，但沒有 groundtruth input 當作 input 反而結果會爛掉，因此我們決定加上 schedule sampling 改善。

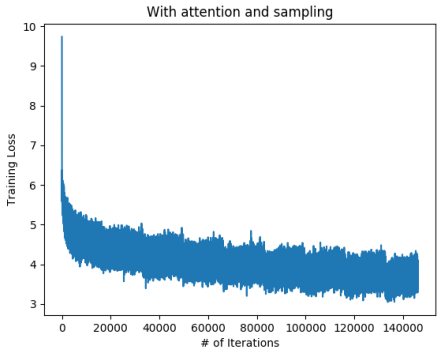
(2) Attention + Schedule sampling

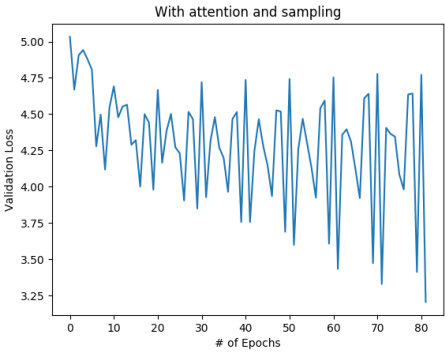
使用 schedule sampling 的時候，我們將其中的 sampling probability 設定

為隨 epoch 數量下降的值，希望模型在訓練初期，能由 groundtruth input 訓練 decoder，接著逐步降低 sampling probability，讓模型能夠在不需要 groundtruth input 的情況下，輸出有意義的句子。



從下圖可以看出加上 schedule sampling 後，validation loss 和 correlation score 都大幅進步，所以我們最後的模型是採用 attention 加上 schedule sampling。






```
perplexity       : 9.303208 (baseline: < 100)
correlation score : 0.56195 (baseline: > 0.45)
```

3. Experimental results and settings

(1) Data processing:

在訓練模型之前，我們對 training data 做了一些處理。選取常出現的五萬字，其他則為 unknown。為了能讓模型訓練的更快更好，我們將太長或是 unknown 字太多的句子刪掉，只有句子詞彙量在 2~15 之間，而且 unknown 的詞不超過 2 個的才能作為 training data。

(2) Model parameters:

Batch size = 50 (經我們測試 batch size 小一點，效果會比較好)。

Hidden units = 1024 (LSTM cell)。

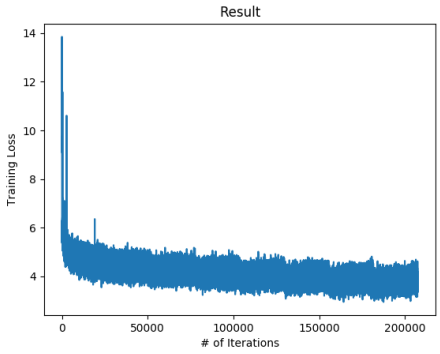
Dropout = 0.75 (keeping probability)

Schedule sampling probability 如 2-(2)所示。

Max gradient norm = 5.0

(3) Result:

訓練過程中的 training loss 趨勢如下：




```
perplexity      : 9.371748 (baseline: < 100)
correlation score : 0.52282 (baseline: > 0.45)
```

Perplexity 和 correlation score 都過 baseline。實際輸出文字也大致與 input 相關。

Input:

```
29 是你老婆嗎?
30 把槍放下,哈維
31 哈維,放下槍
```

Output:

```
29 前妻!前妻!前妻!
30 把槍放下
31 我求你了,哈維
```

4. README

Python3

TensorFlow r1.6

Numpy1.14.3

Python Standard Library

5. 分工表

B03901065 林宣竹	模型架構、調整參數
B03901145 郭恆成	資料處理、script