Machine Learning - Homework 8

資工四 B05902023 李澤諺

May 21, 2020

在本次作業中，首先，我將所有句子的開頭與結尾分別加入 <BOS> 和 <EOS>，並在所有句子之後加入 <PAD> 使其長度爲 50，再將所有句子中的 token 轉爲 index，以此進行 data preprocessing。接著，以下爲我使用 PyTorch 所實作的 Seq2Seq 架構：

Encoder	
embedding	Embedding(num_embeddings , embedding_dim = 256)
dropout	Dropout(0.5)
recurrent	GRU(input_size = 256 , hidden_size = 512 , num_layers = 3 , bias = True , batch_first = True , dropout = 0.5 , bidirectional = True)

Decoder	
embedding	Embedding(num_embeddings , embedding_dim = 256)
dropout	Dropout(0.5)
recurrent	GRU(input_size = 1280 , hidden_size = 1024 , num_layers = 3 , bias = True , batch_first = True , dropout = 0.5 , bidirectional = True)
linear	Linear(2304 , 2048 , bias = True)
	Linear(2048 , 4096 , bias = True)
	Linear(4096 , num_embeddings , bias = True)

其中，我在 decoder 之中使用了 attention mechanism (其實作方式將於第 2 題之中說明)。最後，我使用了 Adam 訓練 Seq2Seq，其中 batch size 爲 64，learning rate 爲 0.0001，使用了 inverse sigmoid function 進行 scheduled sampling (其公式將於第 4 題之中說明)，以此訓練了 40 個 epoch，並且，在訓練過程之中，我每 5 個 epoch 就會進行 validation (其中 beam size 爲 4，而 beam search 的實作方式與其它大小的 beam size 所造成的影響將會於第 3 題之中說明)，並保存到目前爲止 BLEU score 最高的 model 之 parameter，以此得到最後的 model。

1. 請嘗試移除 Teacher Forcing，並分析結果。

我於 report 最剛開始所述的 Seq2Seq 其在 validation data 上所得到的 BLEU score 爲 0.53729，在將其 teacher forcing 移除之後，其在 validation data 上所得到的 BLEU score 下降爲 0.48403，由此可以看出 teacher forcing 對 Seq2Seq 訓練

過程的重要性，其原因可以詳見課程內容。

2. 請詳細說明實做 attention mechanism 的計算方式，並分析結果。

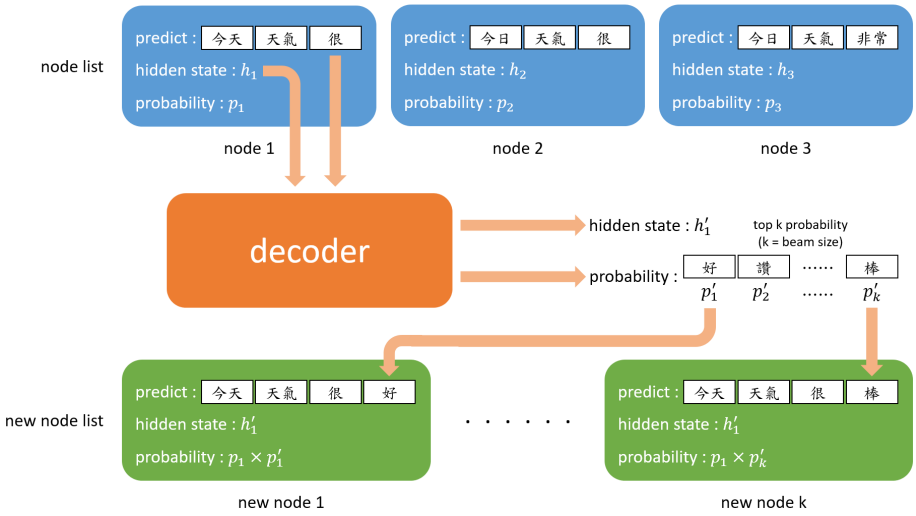
我所實作的 attention mechanism 參考自https://github.com/bentrevett/pytorch-seq2seq/blob/master/3%20-%20Neural%20Machine%20Translation%20by%20Jointly%20Learning%20to%20Align%20and%20Translate.ipynb。首先，我將 decoder 中最後一層 layer 的 hidden state 和 encoder 的所有 output 進行 concatenate，接著，我將所得到的所有 vector 輸入以下的 DNN 之中：

Linear(2048 , 1024 , bias = True)
Tanh()
Linear(1024 , 1 , bias = False)

我將該 DNN 所輸出的數值視爲各個 encoder output 的 attention score，並將所有 encoder output 的 attention score 進行 softmax，以得到各個 encoder output 的 attention weight，最後，將所有的 encoder output 以此進行 weighted sum，得到 attention vector。而在 attention vector 的使用上，首先，我會將 embedding vector 輸入 decoder 的 GRU layer 之中，接著將 embedding vector、GRU output、attention vector 進行 concatenate，再將其輸入 decoder 的 fully-connected layer 之中，以此得到 decoder 的 output。我於 report 最剛開始所述的 Seq2Seq 其在 validation data 上所得到的 BLEU score 爲 0.53729，在將其 attention mechanism 移除之後，其在 validation data 上所得到的 BLEU score 下降爲 0.48737，由此可以看出 attention mechanism 確實可以幫助 Seq2Seq 提高 performance。

3. 請詳細說明實做 beam search 的方法及參數設定，並分析結果。

我實作了一個 data structure，稱之爲 beam search node，用來記錄 decoder 到目前爲止所輸出的 output sequence、decoder 輸出該 output sequence 的 probability，以及 decoder 的 hidden state，並且，我使用了一個 list 來儲存 beam search node。首先，我將 list 中的各個 node 其紀錄的最後一個 decoder output 與其紀錄的 hidden state 輸入 decoder 之中，接著從 decoder 所輸出的 probability distribution 中找出 probability 最高的 k (k 爲 beam size) 個 output，並將這 k 個 output 分別加入原本的 node 的 output sequence 之中，以及將這 k 個 output 的 probability 分別乘以原本的 node 的 probability，以及紀錄 decoder 所輸出的 hidden state，並將所得到的新的 k 個 node 加入一個新的 node list 之中 (以上過程如下圖所示)，當原本的 node list 之中的所有 node 都經過以上的過程之後，便將新的 node list 中的 node 依照 probability 排序，並只保留其中 probability 最高的 k 個 node，再將該新的 node list 取代原本的 node list，如此不斷重複 t (t 爲 output sequence 的長度) 次，最後將 node list 之中 probability 最高的 node 其紀錄的 output sequence 作爲 prediction 輸出。



我將 report 剛開始所述的 Seq2Seq 使用不同大小的 beam size 進行 inference，其在 validation data 上所得到的 BLEU score 如下表所示：

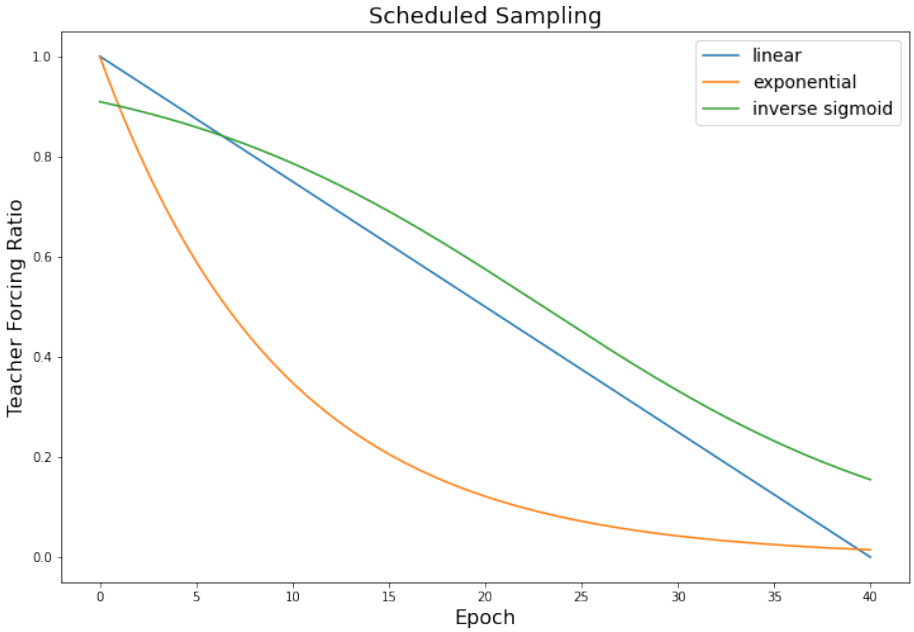
Beam Size	BLEU
1	0.54185
2	0.54312
4	0.53729
8	0.52326

由此可以看出，beam size 越大時，BLEU score 不一定也會越高，其可能是因爲 beam search 是在想辦法找出 probability 最高的 sequence，但其不一定會有最高的 BLEU score，兩者之間存在 bias 所造成。

4. 請至少實做 3 種 scheduled sampling 的函數，並分析結果。

我實作了以下 3 個 scheduled sampling function (其中 n 爲目前的 epoch 數目)：

$$\begin{aligned} \textit{linear} &: \max\left(1 - \frac{n}{40},\ 0\right) \\ \textit{exponential} &: 0.9^n \\ \textit{inverse sigmoid} &: \frac{10}{10 + e^{n/10}} \end{aligned}$$



我將以上 3 種 scheduled sampling function 使用在 report 剛開始所述的 Seq2Seq 之中，其在 validation data 上所得到的 BLEU score 如下表所示：

Function	BLEU
linear	0.53202
exponential	0.50335
inverse sigmoid	0.53729