學號：b05901070　系級：電機四　　姓名：蔡昌廷

1. (20%) Teacher Forcing:
   a. 請嘗試移除 Teacher Forcing，並分析結果。
   以下為實測是否有 teacher forcing 下，其在 testing data 的表現

	Teacher Forcing	No Teacher
Testing loss	6.8086	3.9994
Bleu score	0.4966	0.4684

   從以上結果可以看出在有 teacher forcing 的情況下，單就 bleu score 的表現來看是比沒有 teacher forcing 來的好，原因可能是因為在有標準答案的訓練下，model predict 出來的 answer 會比較符合正確答案想要預測的。

2. (30%) Attention Mechanism:
   a. 請詳細說明實做 attention mechanism 的計算方式，並分析結果。
   在此處 attention 的方式，我先將 decoder 最後一層的 hidden layer 取出來，並跟 encoder output 做內積，形成 attention 的 weight (大小為 (seq_len, 1))，再將其通過 softmax 之後，把它跟 encoder 的 output 做內積形成 context vector，最後將 context vector 與 decoder 的 input 接在一起傳入 decoder。
   以下是我實測在 testing data 的結果

	Attention	No Attention
Testing loss	3.9148	3.9182
Bleu score	0.4780	0.4777

   從以上結果會發現時是否加 attention 對結果似乎沒有太大的變化，自己個人推測的原因是此處 attention 處理的方式不適合用在這次的實作，在處理 attention 時會將 attention weight 做 softmax 的處理，然而在這次的實作上因為做 softmax 的維度有 50 維，因此處理後的數值就很容易出現非 1 則 0 的情況，很可能就導致此處 attention 的效果不如預期。

3. (30%) Beam Search:
   a. 請詳細說明實做 beam search 的方法及參數設定，並分析結果。
   此處的 beam search 我選擇的 beam_size 是 4，針對每一個預測出來的結果取其前 4 高的機率的詞加入 candidate 的 sequence，此處機率最大的部份我將預測出來的值通過 log softmax 之後再相加去選擇數值最大的

sequence，例如：第一個詞取前四高機率的詞加入 candidate sequence，因此現在有四個 sequence 分別只有一個詞，再來預測第二個詞，分別對每個 sequence 取前四高機率的第二詞，現在便有 16 個 candidate sequence，接著對這 16 個 sequence 取最高的 4 個 sequence 當作 candidate，再繼續選擇第三個詞，以下以此類推。

以下是我實測在 testing data 的結果

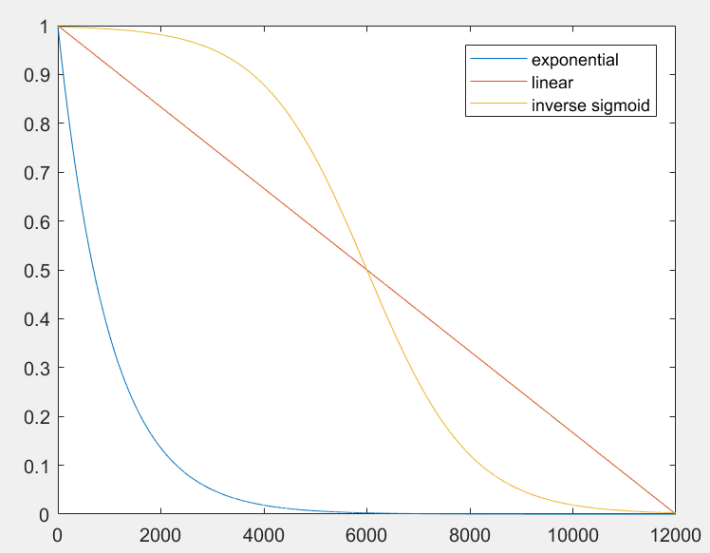
	Beam search	No beam search
Testing loss	3.9148	3.8899
Bleu score	0.4780	0.4740

根據上表會發現加了 beam search 效果似乎不是如此顯著，但是在表現上確實有略為提升，自己分析可能原因是 beam_size 不夠大，導致更好的解並沒有在 beam search 被找到，因此在這次實作上 beam search 的用途並不是特別明顯。

4. (20%) Schedule Sampling:

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

此處我實作三種不同的 sampling 函數，分別為：linear decay、exponential decay、inverse sigmoid decay，下圖為三種函數的作圖。



以下是我實測在 testing data 的結果

	linear	exponential	Inverse sigmoid
Testing loss	3.9148	4.0124	3.9192
Bleu score	0.4780	0.4656	0.4858

由上面的結果發現 exponential decay 表現的最差，這是可以預期的，從上面的分布圖可以看出其 decay 的速度最快，代表其在整個訓練過程大部分是沒有使用正解來訓練的，導致其在 testing 時的表現不如其他兩個 sampling function，至於另外兩個 inverse sigmoid 以及 linear 我認為兩者的表現不會差太多，最主要的差別在於訓練前期 inverse sigmoid 有更大的機率讓 decoder 去使用正解來學習，因此最後 testing 的結果也會使得 inverse sigmoid 較 linear 佳。