Lab5\_Conditional Sequence-to-Sequence VAE

機器人碩 309605014 王麒睿

1Introduction

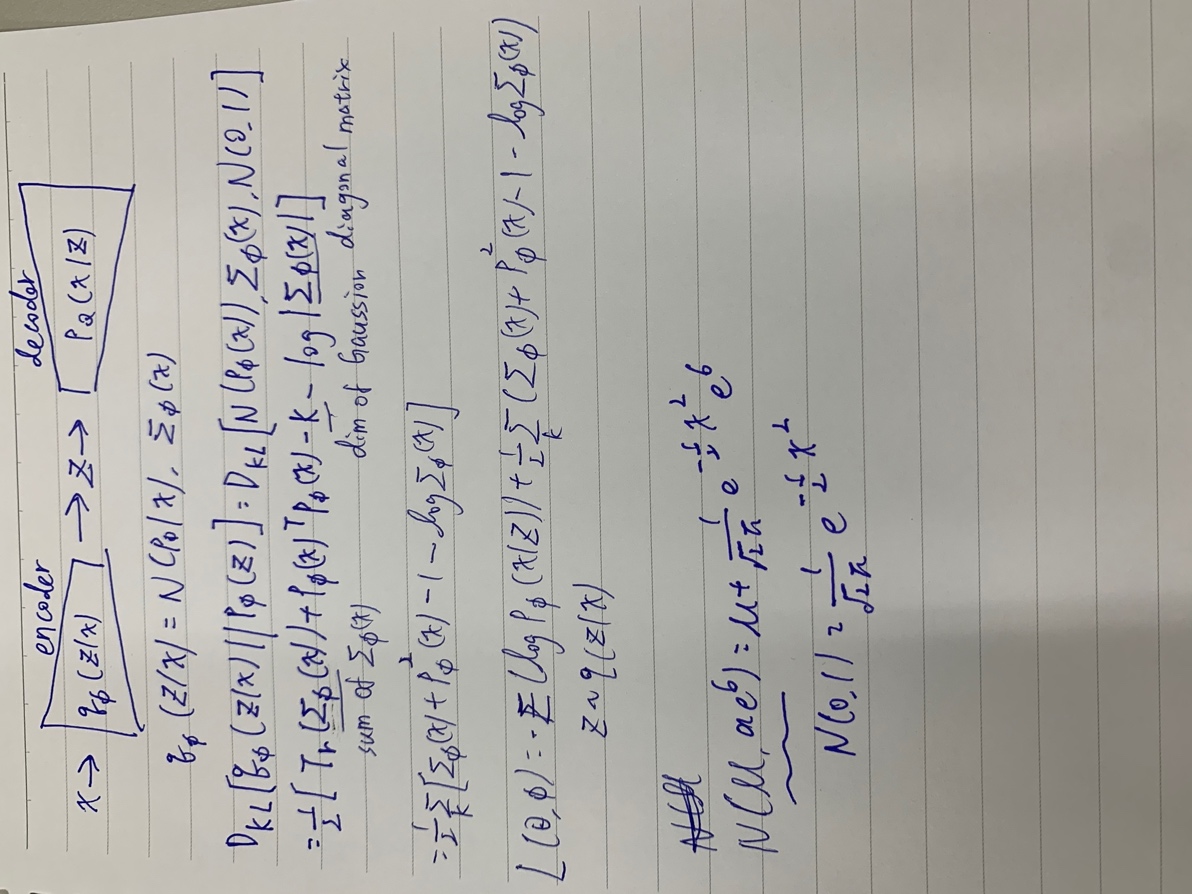
Every English verb has its tenses like simple、simple past etc. To convert different tense between input words and target words, this lab use tense as condition and English words as input and target.

Some requirements as follows:

• Implement a C seq2seq VAE model.  
• Adopt two method which teacher-forcing and kl loss annealing to train model.  
• Plot cross-entropy loss and kl loss curves during training.  
• Plot the BLEU-4 score curve of the test data during training  
• Show the results generated words with 4 tenses by Gaussian normal

distribution.

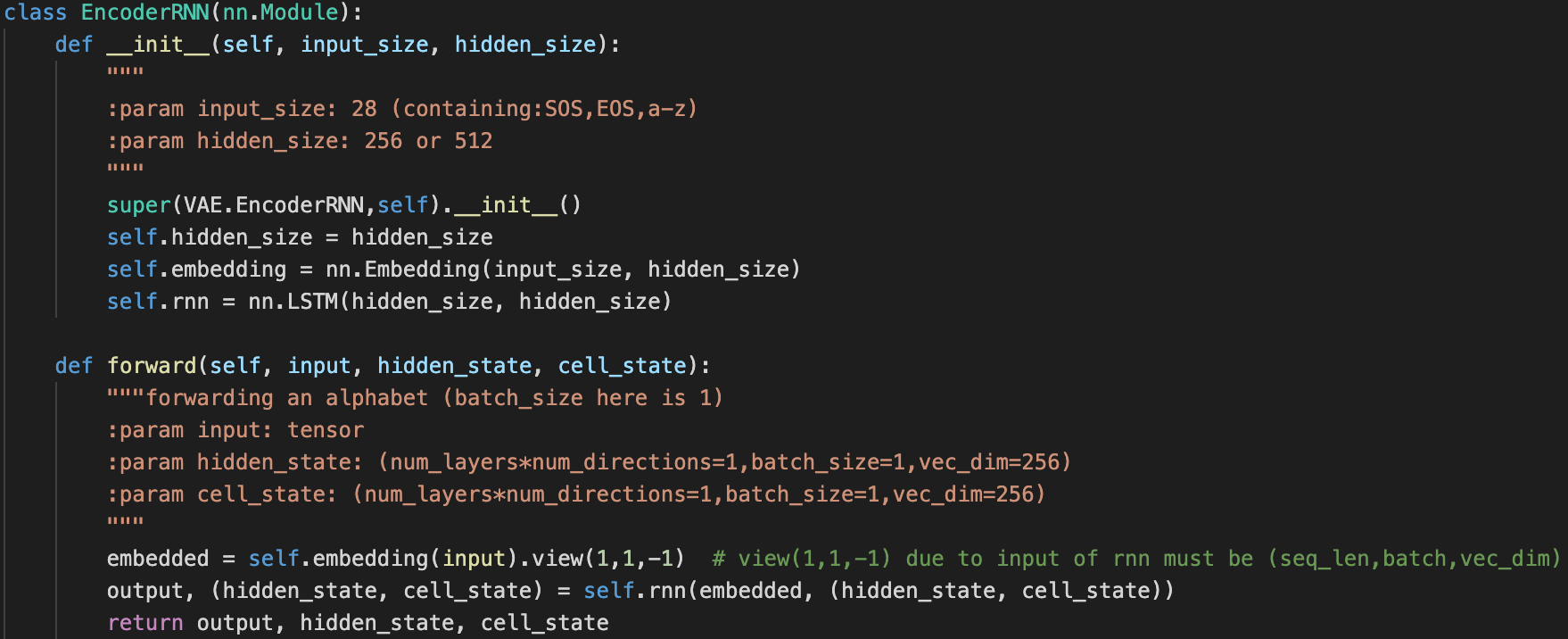
2.Derivation of CVAE



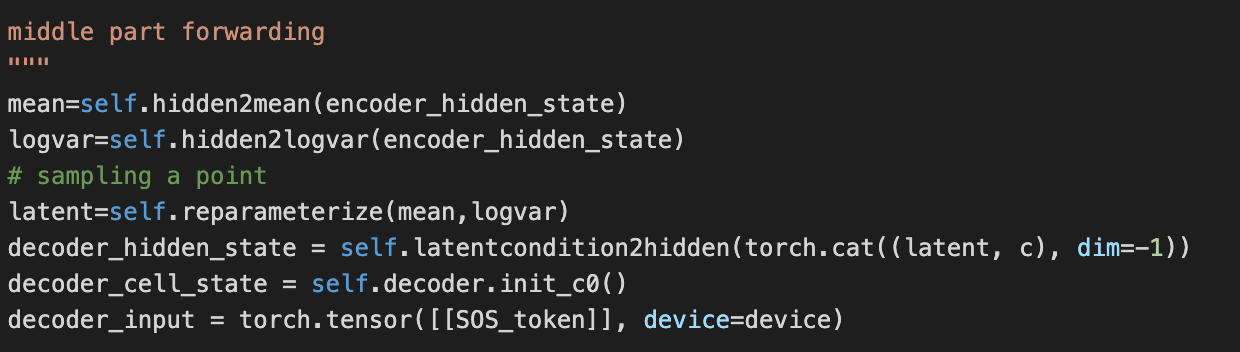
3.Experimental Setup

CVAE是由三個部份所組成:Encoder+中間sample part+Decoder。

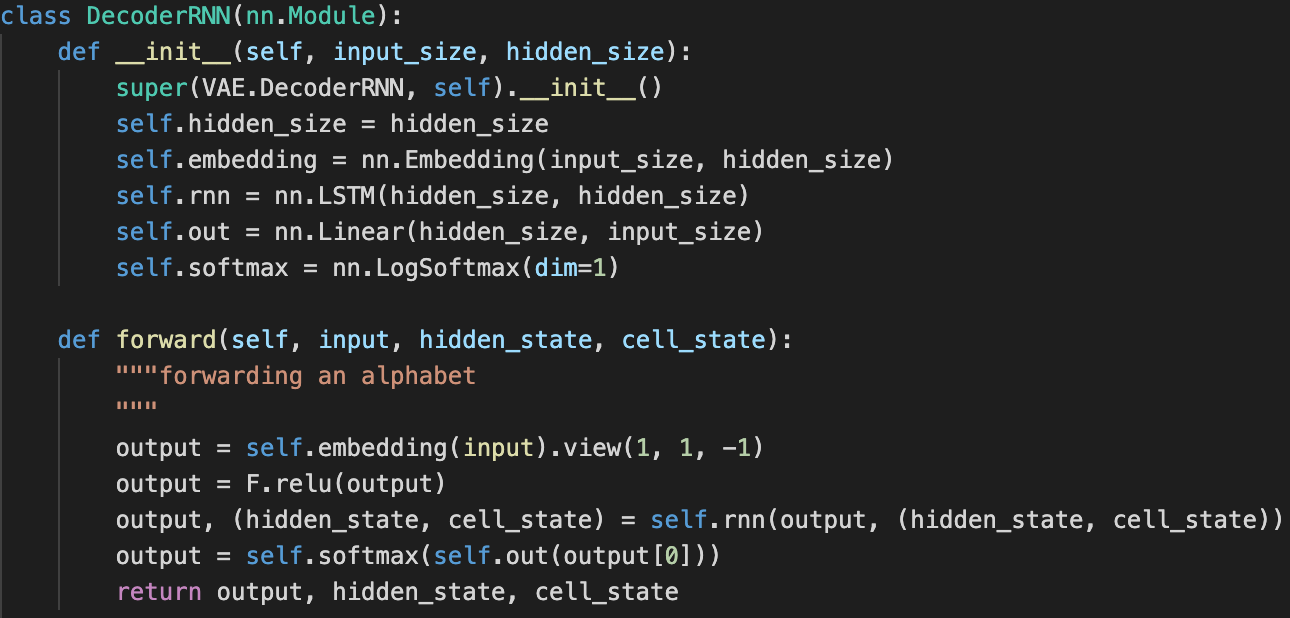
Encoder:  
hidden\_state加入會先把alphabet embedding成一個向量，再丟進LSTM去跑，最後輸出 output,hidden\_state,cell\_state



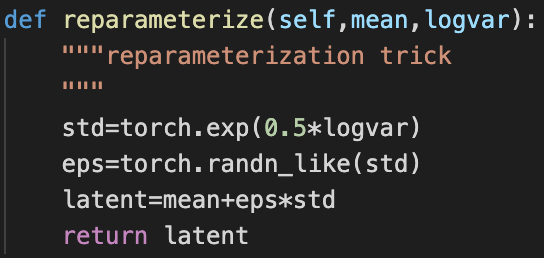
中間sample part:  
我們把encoder輸出的hidden\_state透過fully connected layer變為32-dim的  
mean與log variance，知所以用log variance是因為variance皆為正值，但fully connected layer 可能會輸出負數。  
有了mean與log variance後，我們就可以透過reparameterization trick sampling一個32-dim的 latent，32-dim latent與8-dim condition concate後，在透過一個fully connected layer轉換為 hidden\_state的維度。



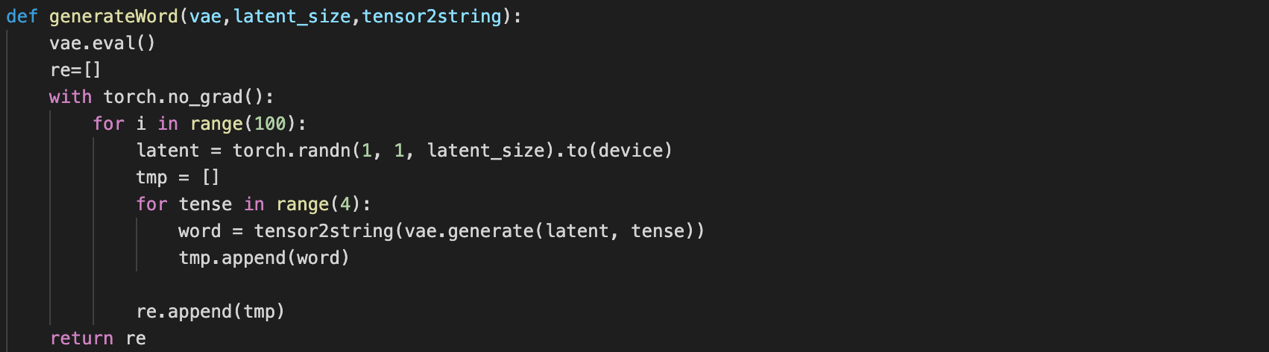
Decoder:  
輸入的hidden\_state為剛剛”中間sample part”的輸出，cell\_state則初始化為0 tensor。

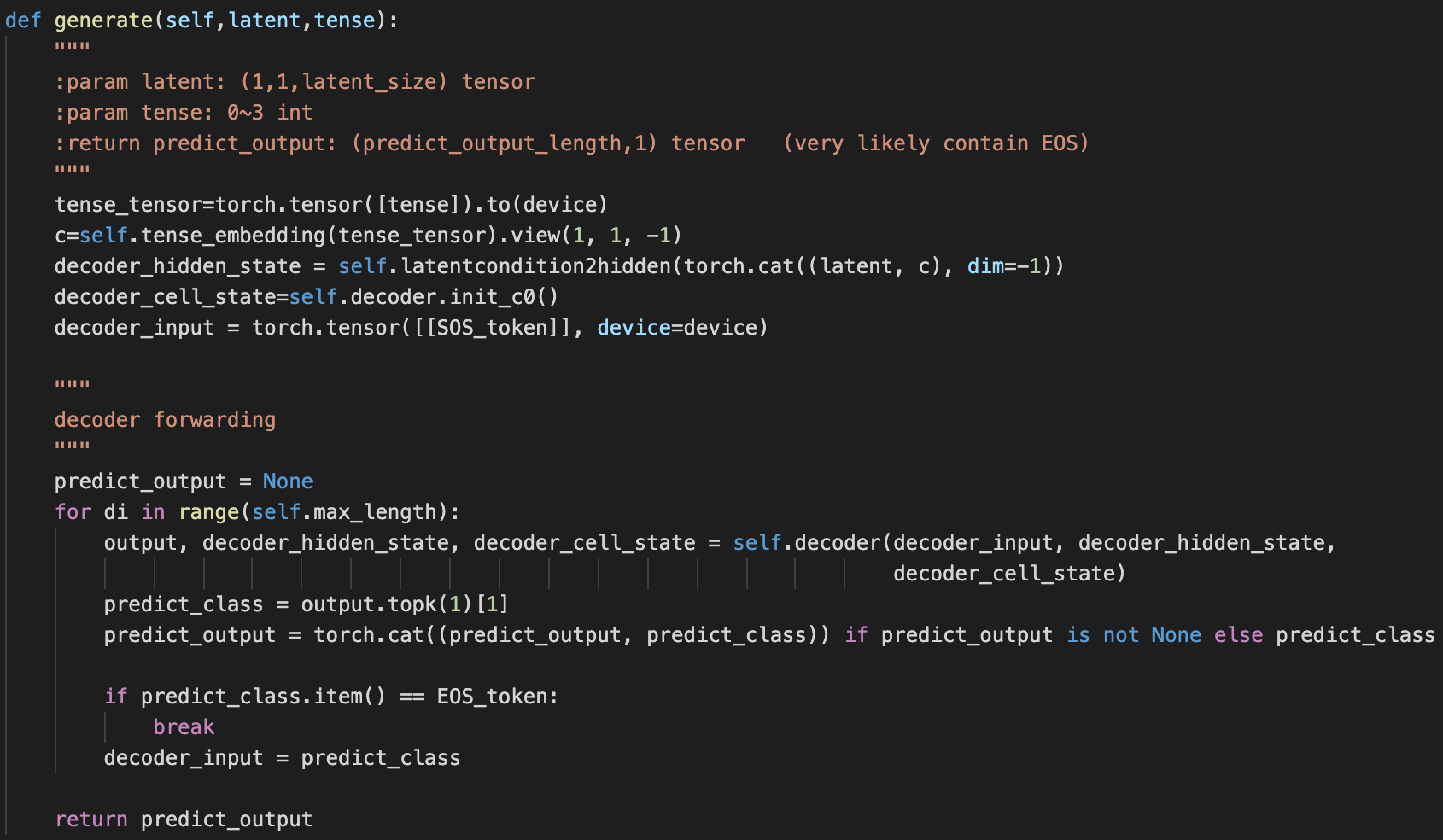


Reparameterization trick:  
從N(mean,exp(log variane))中sample一個點



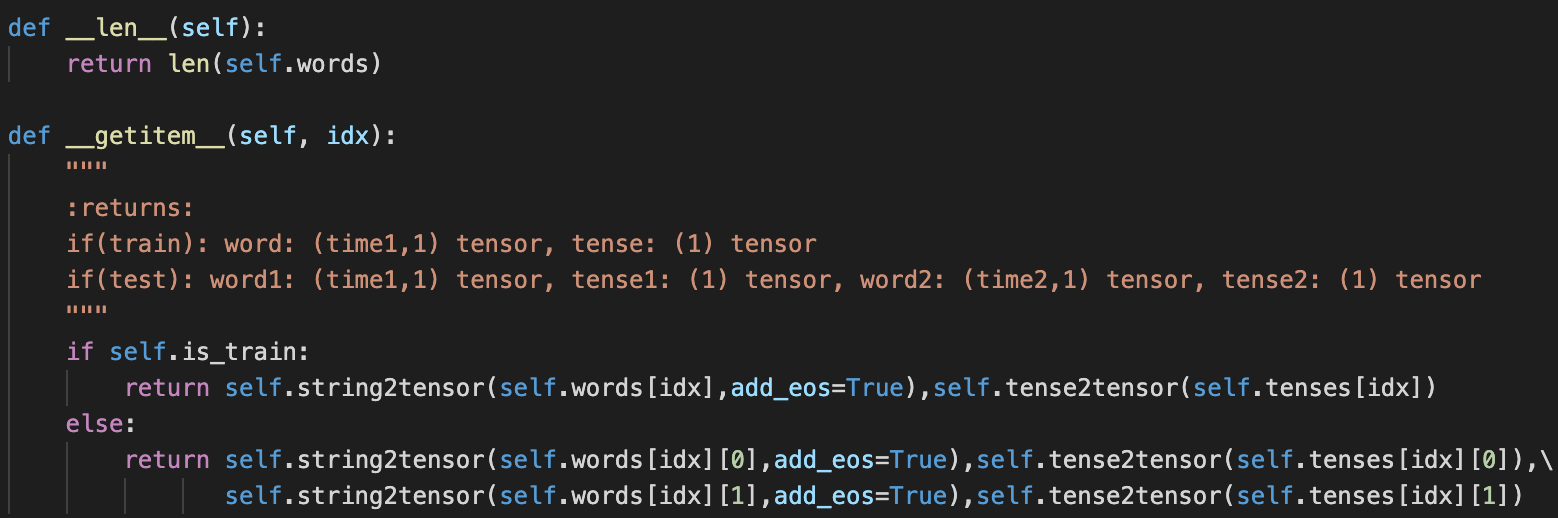
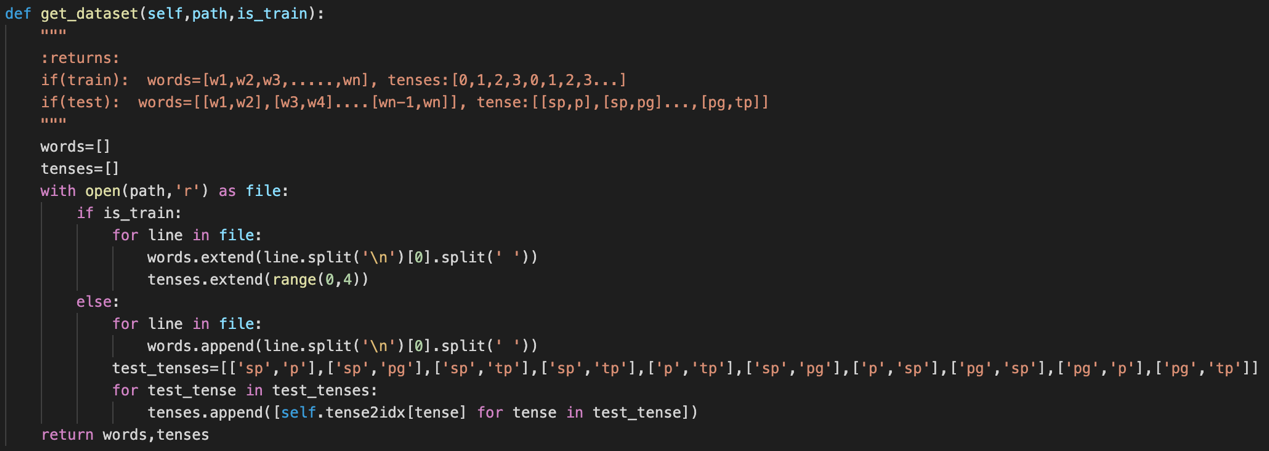
Text generation by Gaussian noise:  
用trorch.randn()隨機生成一個32-dim的latent tensor，再把這latent tensor與tense tensor concate，並作為decoder的hidden\_state





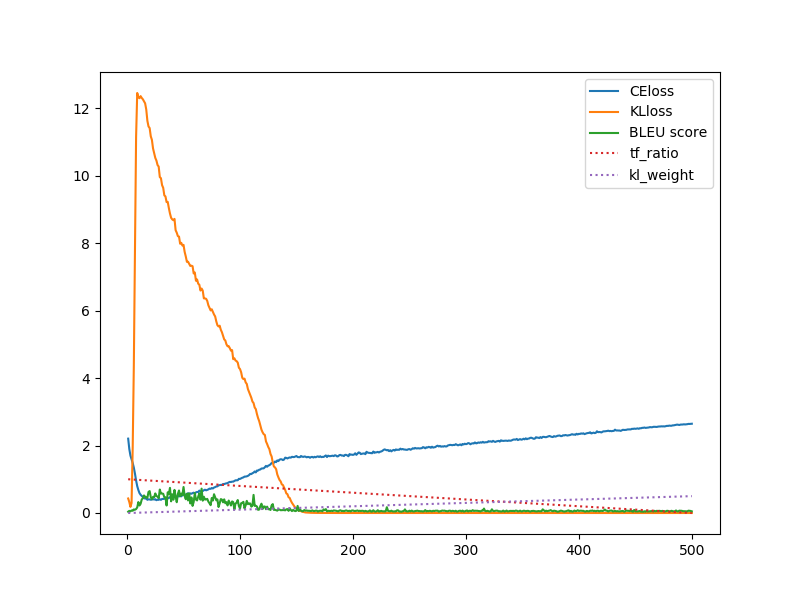
dataloader:

把SOS,EOS,a,b,c,....,z分別對應到0~27，以利將來torch.nn.embedding()



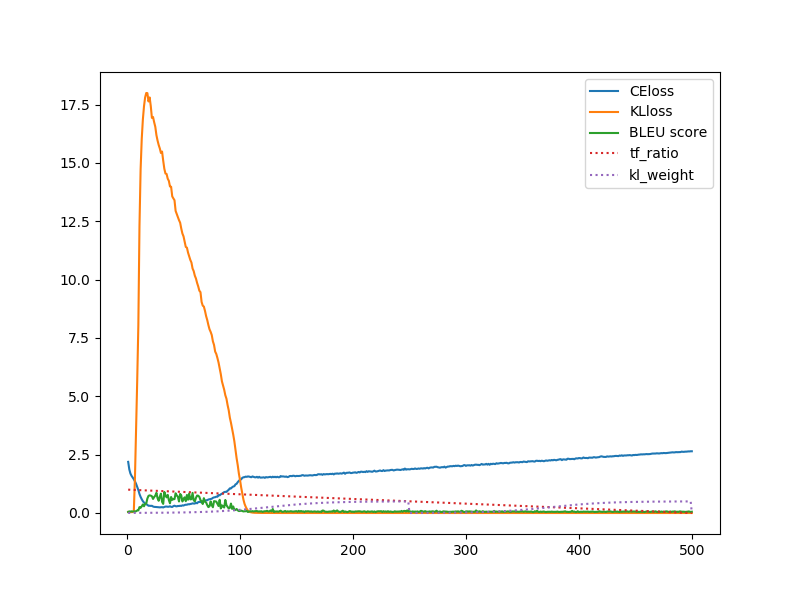
4.Experimental results

monotonic shedule:

500個epoch，teacher\_forcing\_ratio從1.0線性降至0.0，kl\_weight從0.0線性升至1.0 

cyclical schedule:

500個epoch，teacher\_forcing\_ratio從1.0線性降至0.0， 第1到250個epoch間與第251到500個epoch間kl\_weight從0.0線性升至0.5



Best score:



Bleu:0.9449

5.Discussion  
一開始CE loss大於KL loss，model還沒學到任何東西，BLEU分數也很小。 在大約第6個epoch，時CE loss逐漸下降，代表word的reconstuction成功了，因此BLEU提昇; 但latent distribution與prior:N(0,1)長得越來越不像，因此KL loss上升。

大約到第20個epoch時，由於kl\_weight變大，kl\_wight\*KL loss開始dominate整個loss function ，所以KL loss會被迫往下降，CE loss因而上升連帶影響BLEU下降。 在更後面的epoch，由於kl\_weigth 一直提高，迫使KL loss一直都很低，BLEU也連帶無法升高，也發現KL loss升高時，BLEU也會升高。  
在cycle shedule中，第250個epoch開始KL looss與BLEU都應該要上升才對，但並沒有，可能是KL weight的cycle頻率設的不對，或有其他原因。