DL LAB3 report

0412237

陳逸豐

1. Introduction (15%, 15%)

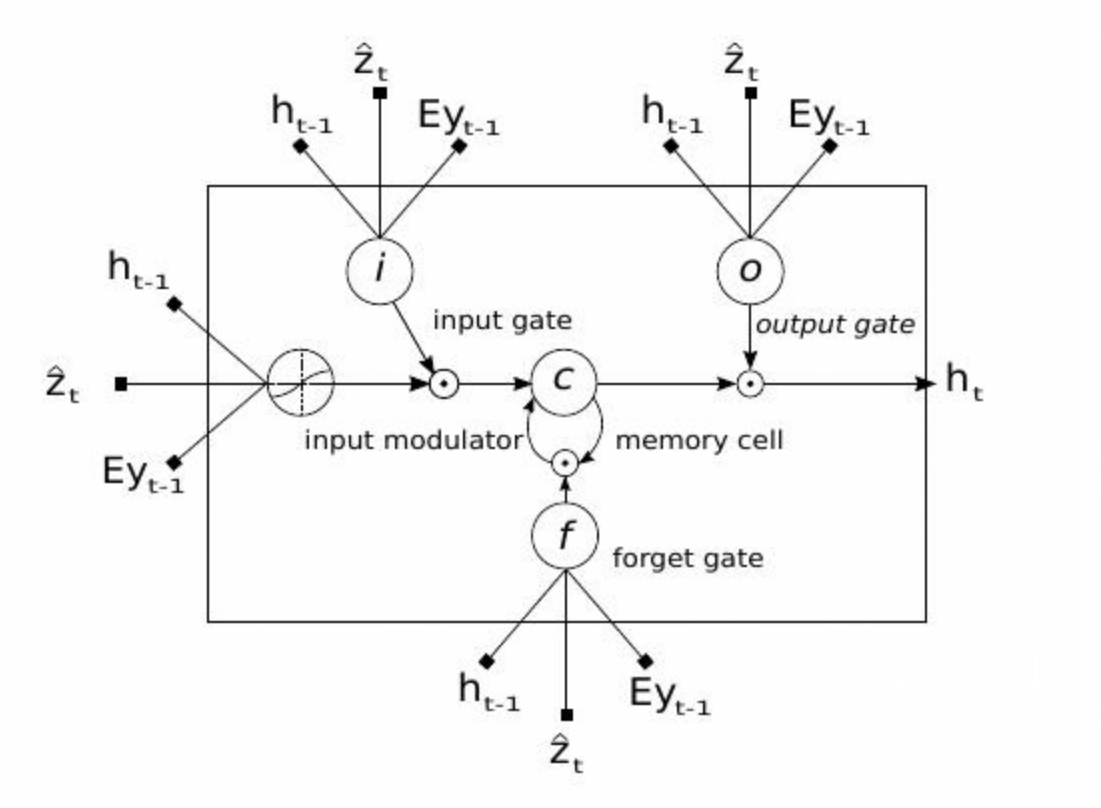
In this lab, we are going to implement the paper, “Show, attend and tell: Neural image caption generation with visual attention”, which will generate a caption according to the input image.

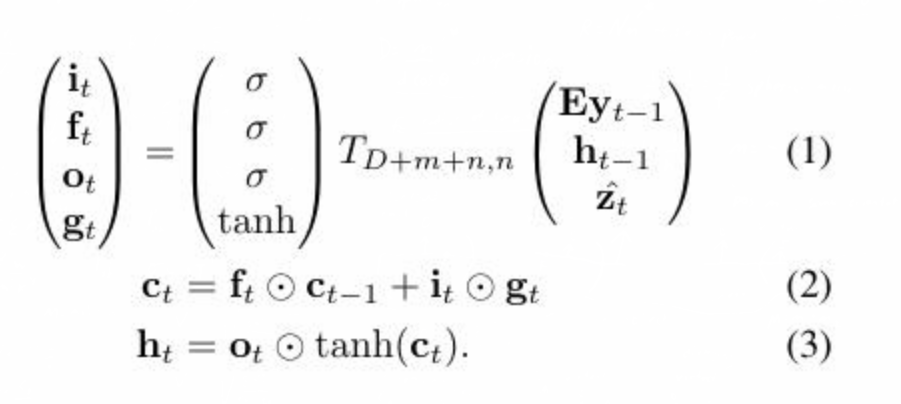
2. Experiment setup (15%, 15%)

 The detail of your model

My model is an auto-encoder. The encoder part is a pre-trained CNN model which is Resnet-101 and is able to extract the feature of the image and the decoder part is a LSTM which will decode the feature to the caption.

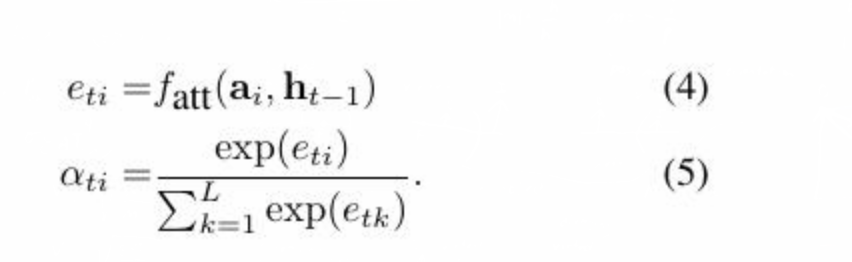
The pre-trained CNN model will generate an annotation per pixel.

The LSTM model is like the below. 



By the formula, we can see once we got Zt, we can get all other state in LSTM, such as it, ft, ot,gt.





And Zt can get from the formula above.

 Report all your training hyperparameters

batch size : 10

input encoding size : 512

rnn size : 512

att hid size: 512

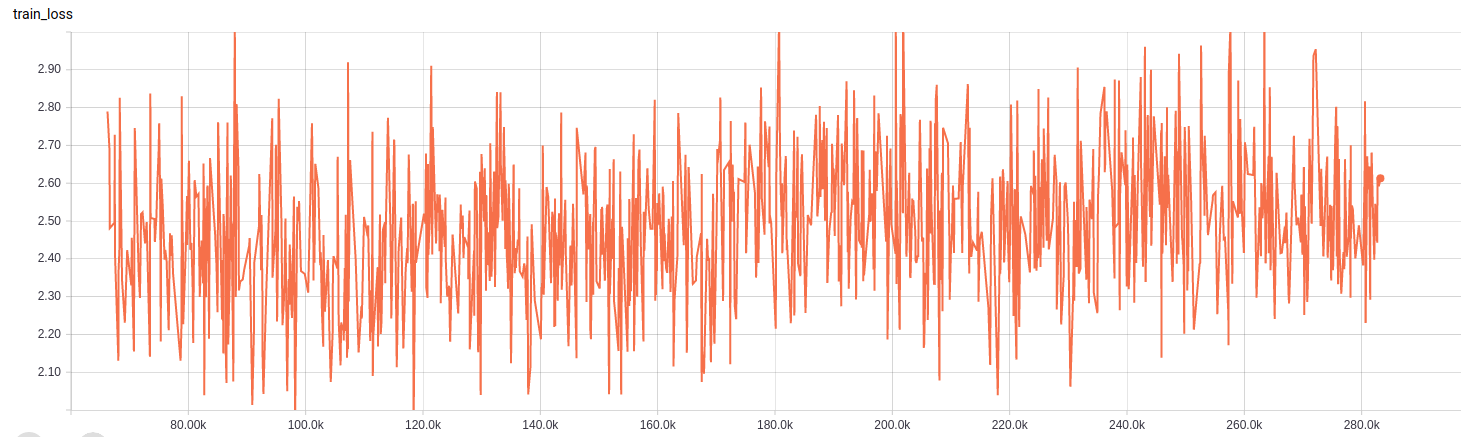
fc feat size: 2048

att feat size: 2048

rnn type: LSTM

3. Result (30%, 40%)

 Training loss of attention models.



 Caption of models.







