HW4 Report

R04725040

黃柏睿

Environment:
Ubuntu 14.04
Tensorflow 0.11.0

本次作業參考了：https://www.tensorflow.org/tutorials/seq2seq/

1. 前處理：把句子照空格跟標點符號 split，然後看過全部 training data 後，挑出最常出現的字，建立一個 100000 字的字典(但 generation 並沒有那麼多字)，並把 training data 跟 valid data 轉乘 token id，沒出的字就用 _UNK 取代。Generation 的部分，直接 parse json 檔，把第一句當作要翻譯的句子，並把下面兩句當翻譯出來的句子，使用同一個 model 來 train。
2. bucket 的部分，translation 選擇用原本的 bucket set，而在 generation 的部分，我直接選擇開一個很大的 bucket，做出來的效果會比較好。
3. model 的參數，每一個 model layer 用 1024，總共三個 layer，使用 batch size 64，每隔 500 個 step，儲存一次，最後用的 model 是 42000 steps(translation) 跟 17500 steps(generation)
4. 由於 dictionary 跟 model 太大，所以我把他壓縮上傳到 dropbox，並在 shell script 裡面用 wget 抓取，並解壓縮。