

# Sequence-to-sequence Model

B02902039 資工四 李奕皓

## Download the Checkpoint File

這次的 model (Machine Translation) 要 train 的 variable 非常多，所以 checkpoint 檔案很大，無法上傳到 GitHub。

Google Drive 下載連結：<https://drive.google.com/open?id=0B04Nu6u2W5BKS1d2OWJyNUdybjA>

請下載後放到 `hw4/translate/checkpoints/codalab.ckpt`。

## About the Model

這次作業的 model 是由 RNN encoder 和 RNN decoder 組成，decoder 的部份還有用到 attention mechanism，RNN cell 是用 GRU。

Encoder 的 input 是反著餵進去的。

實作上是用 `tf.nn.seq2seq.embedding_attention_seq2seq` 及 `tf.nn.seq2seq.model_with_buckets`。

## Hyperparameters

Hyperparameter	Machine Translation	NLG
Learning rate	0.001	0.001
Batch size	64	64
Size of RNN unit	512	256
Number of RNN layers (Deep RNN)	2	2
Source vocabulary size	40000	1000
Target vocabulary size	40000	1000

## My Improvements

本來用傳統的 Gradient Descent 收斂的速度非常慢，後來改成用 `AdamOptimizer` 收斂的速度就快非常多。

## Comment on Code

這次我是從 Tensorflow 官方 tutorial 的 code 修改而來。

<https://github.com/tensorflow/tensorflow/tree/v0.10.0/tensorflow/models/rnn/translate>

它會先載入 training data 建好 vocabulary，然後把 training data 和 validation data 裡面全部的字都轉換成代表每個字的整數。這份 code 只有按照標點符號以及空白把字切開，換一個好一點的 tokenizer 應該會得到更好的結果（例如可以把 `don't` 切成 `do` 和 `n't` 等等）。

它還有按照不同的 source, target sentence length 把 data 歸類到不同的 **bucket**，train 的時候一個 batch 裡面的 sentence 都出自同個 bucket，這樣一來可以確保 batch 內句子長度相近，不會有太多 padding。

另外還有 learning rate decay 的機制，要是 validation loss 一段時間沒有下降的話，就會降低 learning rate。