

# MLDS HW1 Report

## Members

D04944007 高瑞宗  
R05922139 林子芃  
R04921106 陳彥谷  
B02902105 廖瑋中

## Environment

我們用自己實驗室的Azure雲端計算資源:

Name	Cores	Memory	OS	GPU
mslabgpu2	6	56G	Ubuntu 16.04	Tesla K80, 12G Ram

## Library We Use

Natural Language Toolkit (NLTK)

## Model Description

### 1. SimpleRNN Language Model

這個方法使用單純的一層RNN架構，輸入的時候將one-hot encoding的單字轉為word embedding，再將embedding輸入到RNN內。Embedding方面我們用隨機產生的word embedding，而沒有使用已經train好的結果。利用tf.contrib.rnn.MultiRNNCell可以將RNN的層數調整為多層。

### 2. GloVe Word Embeddings + Cosine Similarity

這個方法使用GloVe中的一種open source pre-trained word embedding，得到不同word的word vector後，取需要填空的字前後n個window，計算那些2n個word embedding的sum vector，接著取用來填空的word vector選擇中和sum vector的cosine similarity最大的那個為predicition。我們在這個方法中嘗試了不同的window size和word embedding，以得到最好的效果。

# How to improve performance

## 1. Baseline RNN

- a. 我們使用基本的設定使我們的model達到baseline。
- b. 我們使用MultiRNNCell使得我們的RNN層數可以調整為多層。

## 2. GloVe Word Embeddings + Cosine Similarity

- a. 給予window中的每個字一個固定的weight，或是對weight進行training。
- b. 尋找更好的pre-trained word embedding，或者自己train一個。
- c. 將embedding的結果，再放入RNN去訓練。

# Experiment Settings and Results

## 1. Baseline RNN

Model	prediction acc. (public)	prediction acc. (private)
Basic RNN	0.32500	0.34615

Hyperparameters	Value
Epoch	2
Training time	3-4小時
Learning rate	1.0
Number of layers	2
Vocabulary size	12000
Hidden size of LSTM	200
Number of steps of RNN	20

## 2. GloVe Word Embeddings + Cosine Similarity

word embedding	window	prediction acc. (public)	prediction acc. (private)
glove.6B.300d (800MB)	5	0.32885	0.30769
glove.6B.300d (800MB)	10	0.32692	0.31154
glove.840B.300d (2GB)	5	0.31731	0.31346

## Some Issues

我們現在有出現一個情況是，我們的model在非當初train的主機上，可能會出現restore不出來的問題。發生這種情況時，model沒有讀出來，就會是initial的值，導致prediction很糟。目前發現的原因是當使用tensorflow的Supervisor做訓練，存起來的checkpoint換一個主機沒有辦法被讀出來。

## Team Division

Member	Work
D04944007 高瑞宗	Data preprocessing + baseline RNN + report
R05922139 林子芃	Cosine similarity + report
R04921106 陳彥谷	report
B02902105 廖瑋中	Data preprocessing + basic RNN + report