Machine Learning HW4

Recurrent Neural Networks

MLTAs
mlta2020fall@gmail.com

Outline

1. Requirements
2. Task Introduction
3. Data Format
4. Kaggle
5. Rules, Deadline, Policy, Score
6. FAQ

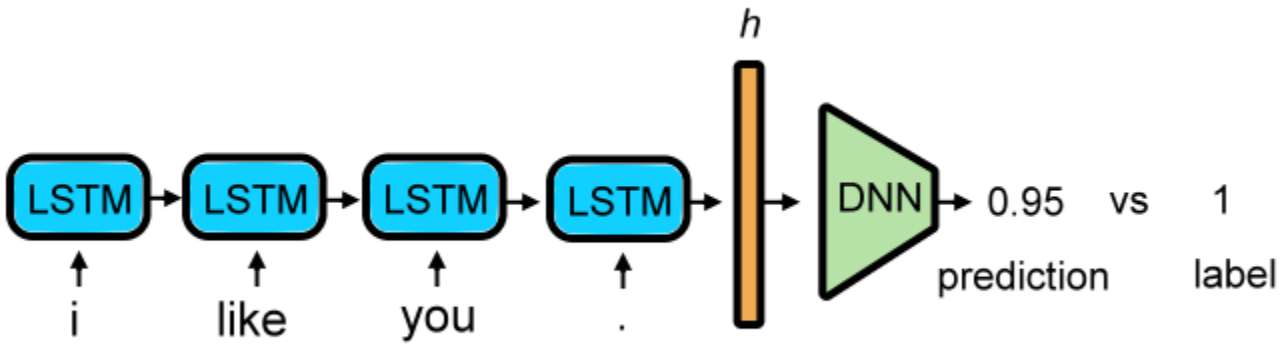
Requirements

- 請使用 RNN 實作
- 不能使用額外 data (禁止使用其他 corpus 或 pretrained model)
- 請附上訓練好的 best model (及其參數) 至 GitHub release 或 Dropbox，並於 hw4_test.sh 中寫下載的 command（可參照這裡的方法）
  - model 大小在 100MB 以內的可以直接上傳到 GitHub
- hw4_test.sh 要在 10 分鐘內跑完（model 下載時間不包含在此）
- 套件的部份請參考連結
- **Conda file**

Task introduction
(Text Sentiment Classification)

Task - Text Sentiment Classification

```
0 +++$+++ on the flipside ... completely bummed that there isn ' t a or sighting .
1 +++$+++ ahaha im here carlos wasssup ?!
0 +++$+++ at least they text you
0 +++$+++ i feel icky , i need a hug
1 +++$+++ hey that ' s something i ' d do !
1 +++$+++ thanks ! i love the color selectors , btw . that ' s a great way to search and list .
```



Text Sentiment Classification

本次作業為 Twitter 上收集到的推文，每則推文都會被標注為正面或負面，如：

```
1 +++$+++ thanks ! i love the color selectors , btw . that ' s a great way to search and list .
```

1：正面

```
0 +++$+++ i feel icky , i need a hug
```

0：負面

除了 labeled data 以外，我們還額外提供了 120 萬筆左右的 unlabeled data

- labeled training data　　：19萬
- unlabeled training data　：120萬
- testing data　　：1萬（ 5000 public，5000 private ）

Preprocessing the sentences

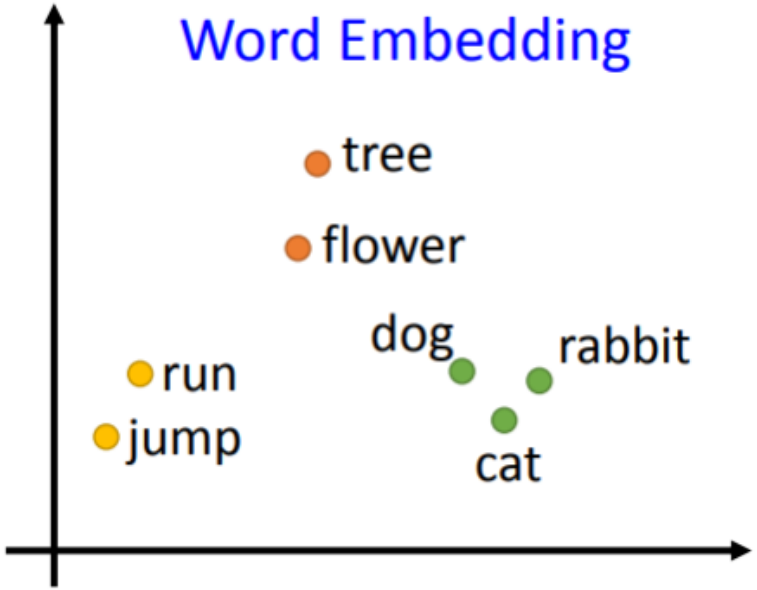
- 先建立字典，字典內含有每一個字所對應到的 index
  example:
  “I have a pen.” -> [1, 2, 3, 4]
  “I have an apple.” -> [1, 2, 5, 6]

- 利用 Word Embedding 來代表每一個**單字**，
  並藉由 RNN model 得到一個代表該句的 vector (這份投影片 p.5 的 h)

- 或可直接用 bag of words (BOW) 的方式獲得代表該句的 vector

What is Word Embedding

- 用一個向量 (vector) 表示字 (詞) 的意思



1-of-N encoding

- 假設有一個五個字的字典 [apple, bag, cat, dog, elephant]
  我們可以用不同的 one-hot vector 來代表這個字

  apple -> [1,0,0,0,0]
  bag -> [0,1,0,0,0]
  cat -> [0,0,1,0,0]
  dog -> [0,0,0,1,0]
  elephant -> [0,0,0,0,1]

- Issue :
  - 缺少字與字之間的關聯性 (當然你可以相信 NN 很強大他會自己想辦法)
  - 很吃記憶體

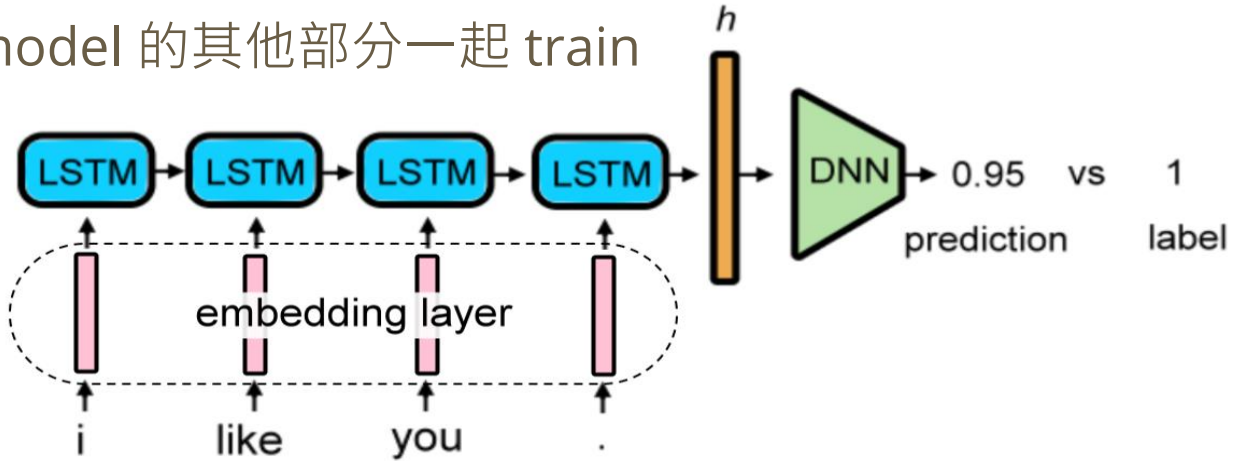
200000(data)*30(length)*20000(vocab size) *4(Byte) = 4.8*10^11 = **480 GB**

Word Embedding

- 用一些方法 pretrain 出 word embedding (e.g., skip-gram, CBOW.)
- Word2Vect 介紹

  小提醒：如果要實作這個方法，pretrain 的 data 也要是作業提供的！

- 然後跟 model 的其他部分一起 train



Bag of Words (BOW)

- BOW 的概念就是將**句子**裡的文字變成一個袋子裝著這些詞的方式表現，這種表現方式不考慮文法以及詞的順序。

例如：

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

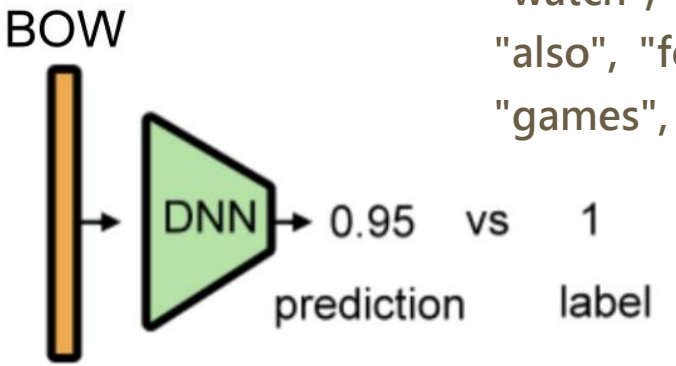
在 BOW 的表示方法下，會變

(1) -> [1, 2, 1, 1, 2, 0, 0, 0, 1, 1

(2) -> [1, 1, 1, 1, 0, 1, 1, 1, 0, 0

dictionary

["John", "likes", "to", "watch", "movies", "also", "football", "games", "Mary", "too"]



Data Format

Data Format (labeled data)

label +++$+++ text

```
0 +++$+++ on the flipside ... completely bummed that there isn ' t a or sighting .
1 +++$+++ ahaha im here carlos wasssup ?!
0 +++$+++ at least they text you
0 +++$+++ i feel icky , i need a hug
1 +++$+++ hey that ' s something i ' d do !
1 +++$+++ thanks ! i love the color selectors , btw . that ' s a great way to search and list .
```

Data Format (unlabeled data)

text

```
1 more day !
nursing celeste with a tummy ache .
hates being this burnt !! ouch
just couldn ' t sleep last night . working 7a 3p , than dinner with megan . happy bday jl !
i love slaves ! by david raccah , linkedin , rotfl
is being super organised and making up orders to post first thing tomorrow !
laying in the bed . it feels soooooo good . what a long day
finally , at the airport . currently chilling out at the citibank lounge . maaaan , the wi fi here doesn ' t work ! lameeee !
back and still feeling shattered . still no cockney ... i ' m ashamed to say .
so do i
```

Kaggle

Kaggle submission format

Kaggle link: https://www.kaggle.com/c/ml2020fall-hw4-3
請預測 testing set 中一萬筆資料並將結果上傳 Kaggle

1. 上傳格式為 csv 檔。
2. 第一行必須為 id, label，第二行開始為預測結果。
3. 每行分別為 id 以及預測的 label，請以逗號分隔。
4. Evaluation: accuracy

```
id,label
0,0
1,0
2,0
3,0
4,0
5,0
6,0
7,0
8,0
9,0
10,0
11,0
12,0
13,0
14,0
15,0
16,0
17,0
18,0
19,0
```

Rules, Deadline, Policy, Score

Deadline

Kaggle, Github deadline:

12/11/2020 23:59:59 (GMT+8)

遲交一天 *0.8，兩天*0.6，遲交超過兩天零分。

Policy

GitHub 上 hw4-<account> 裡面請至少包含：(1, 2, 3的檔名請務必**一模一樣**)

1. report.pdf
2. hw4_train.sh
3. hw4_test.sh
4. train/test Python files
5. model 參數 (Make sure it can be downloaded by your script.)
   - 請將 model 下載到與 script 相同的位置
   - 上傳的 model 總和大小建議在 600 MB 以內

請不要上傳 dataset，請不要上傳 dataset，請不要上傳 dataset

Policy

1. 以下的路徑，助教在跑的時候會另外指定，請保留可更改的彈性，不要寫死
2. Script usage:
   bash hw4_train.sh <training label data> <training unlabel data>
   training label data: training_label.txt 的路徑
   training unlabel data: training_nolabel.txt 的路徑
   bash hw4_test.sh <testing data> <prediction file>
   testing data: testing_data.txt 的路徑
   prediction file: 輸出結果的 csv 檔路徑
   (除非有狀況，不然原則上助教只會跑 testing，不會跑 training，因此請用讀取 model 參數的方式進行預測。)

Score - Report.pdf

Report link

- **(0.5%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線**

- **(0.5%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線。**

- **(0.5%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。**

- **(0.5%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy" 與 "I am happy, but today is hot" 這兩句話的分數（model output），並討論造成差異的原因。**

配分 Grading Criteria - kaggle (5% + Bonus 1%)

Kaggle and Github Deadline : 12/11/2020 23:59:59 (GMT+8)

Public simple baseline: 3%

Public strong baseline: 2%

Bonus - 1%

(1.0%) private leaderboard 排名前五名且於助教時間上台分享的同學

配分 Grading Criteria - report(5%)

Programming Report - 2%

https://reurl.cc/Q3Z9gp

Math Problem - 3%

https://drive.google.com/file/d/1fEu87banB4s6Yjku1dA5sMcnwCugEPBF/view?usp=sharing

Type in latex(preferable) or take pictures of your handwriting

Write them in report.pdf

FAQ

- 若有其他問題，請貼在 FB 社團裡或寄信至助教信箱，**請勿直接私訊助教**。
- 助教信箱：mlta2020fall@gmail.com

Links

- 雲端使用方法：http://slides.com/sunprinces/deck-16#/
- Kaggle：https://www.kaggle.com/c/ml2020fall-hw4-3
- Report template：https://reurl.cc/Q3Z9gp
- Sample code: https://reurl.cc/v1Kyp1
- Github classroom連結:https://classroom.github.com/a/KJF53BxF
- Conda file 連結：https://reurl.cc/EzdEN1