
Machine Learning HW4

Recurrent Neural Networks

MLTAs

ntueemlta2021fall@gmail.com

Outline

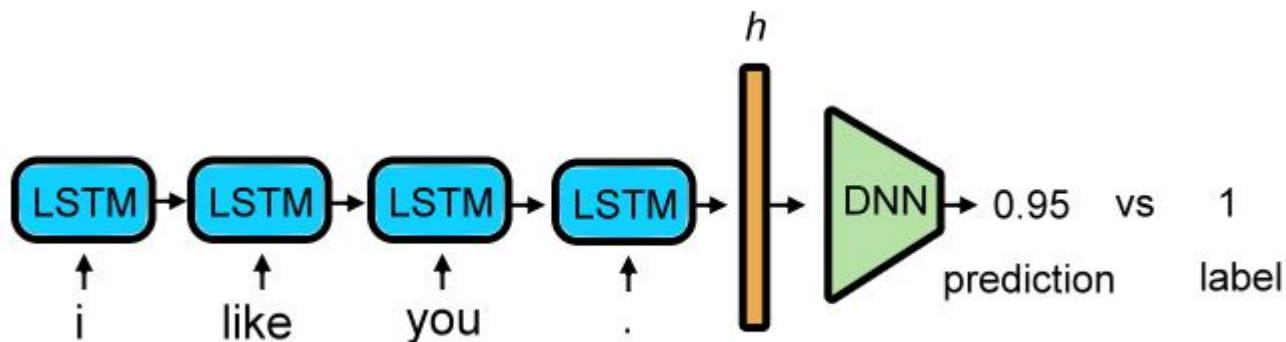
1. Task Introduction
2. Data Format
3. Kaggle
4. Rules, Deadline, Policy, Score
5. FAQ

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 :17 萬
- unlabeled training data :120萬
- testing data :2萬(10000 public, 10000 private)

Task and Dataset

- Task : **Text Sentiment Classification**

- Build your own model(ex: RNN/LSTM)
- Sample code:

- https://drive.google.com/file/d/1dJIB6Sbd_T_S7HsP0pDfLQKYaxnARG1g/view?usp=sharing

- Dataset :

- <https://drive.google.com/file/d/1dcc7RKlpzaOfHd3jVYznTa1fubvb4JO3/view?usp=sharing>

Kaggle Info & Deadline

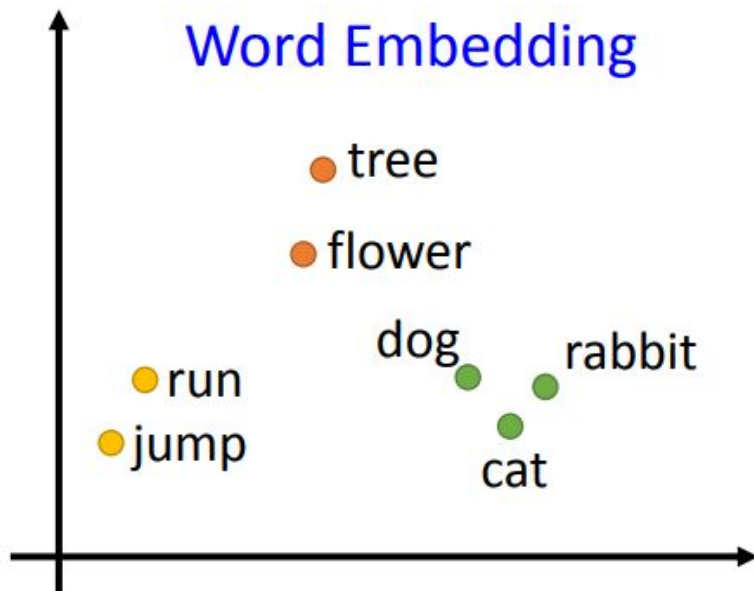
- Link: <http://www.kaggle.com/c/ml-2021fall-hw4>
- 個人進行、不須組隊
- Team Name:
 - 修課學生: 學號_任意名稱(ex: b09901666_name)
- Maximum Daily Submission: 5 times
- Kaggle Deadline: 12/9/2021 23:59:59 (GMT+8)
- Ceiba Deadline: 12/11/2021 23:59:59 (GMT+8)
- test set的20000筆資料將被分為兩份, 10000筆public, 10000筆private
- Leaderboard上所顯示為public score, 在Kaggle Deadline前可以選擇2份submission作為private score的評分依據。

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

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

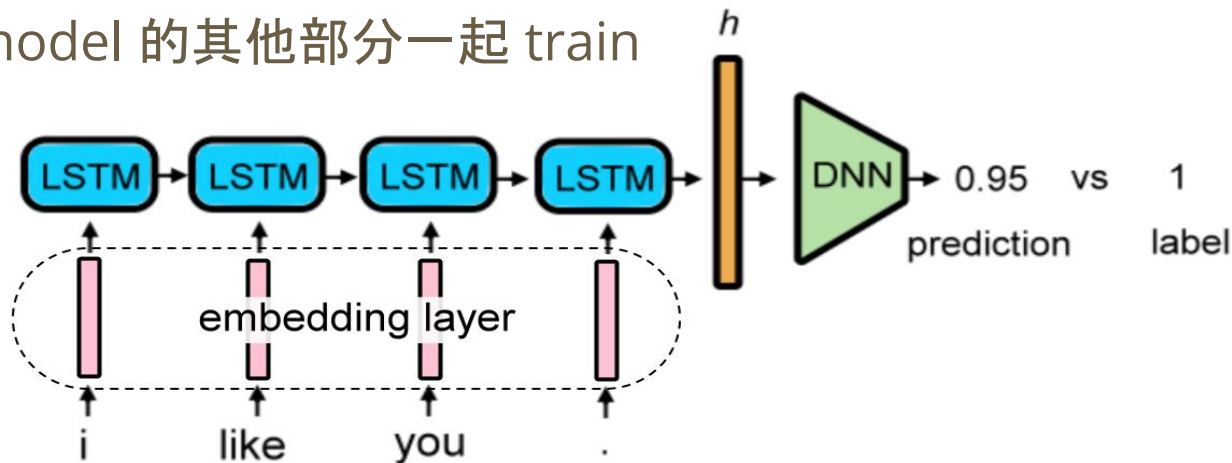
$$200000(\text{data}) * 30(\text{length}) * 20000(\text{vocab size}) * 4(\text{Byte}) = 4.8 * 10^{11} = \mathbf{480 \text{ GB}}$$

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 !

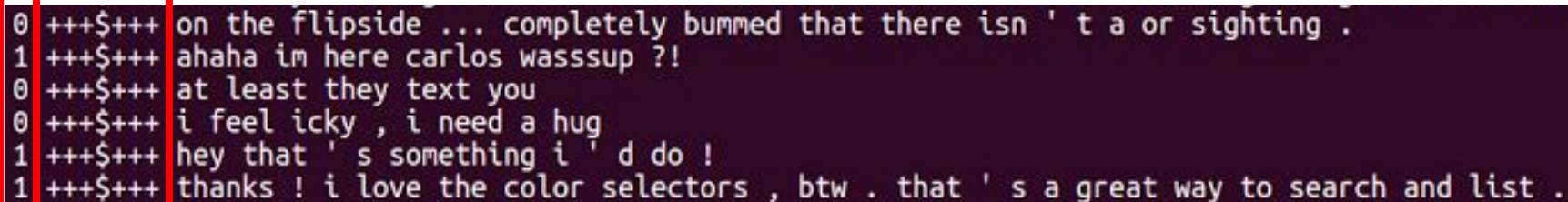
- 然後跟 model 的其他部分一起 train



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

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

Kaggle

Kaggle submission format

Kaggle link: <https://www.kaggle.com/c/ml-2021fall-hw4/leaderboard>

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

```
1 id,label
2 0,0
3 1,0
4 2,0
5 3,0
6 4,0
7 5,0
8 6,0
9 7,0
10 8,0
11 9,0
12 10,0
13 11,0
14 12,0
15 13,0
16 14,0
17 15,0
18 16,0
19 17,0
20 18,0
21 19,0
```


Rules, Deadline, Policy, Score

Ceiba Submissions

你的ceiba上請至少包含：

1. **report.pdf** : Please refer to report template and **show the checkpoint link in it**
2. your python (or ipynb) files
3. 請將參數連結附在report中

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

Report 格式

- 限制
 - 檔名必須為 report.pdf !!!
 - 檔名必須為 report.pdf !!!
 - 檔名必須為 report.pdf !!!
 - 請標明系級、學號、姓名，並按照report模板回答問題，切勿隨意更動題號順序
 - 若有和其他修課同學討論，請務必於題號前標明 collaborator (含姓名、學號)
- Report模板連結
 - 連結：
<https://docs.google.com/document/d/1mjawi2jtHhBrnxluXZ-Q2pNbh8YWuAm4HK3H6Khgoc8/edit?usp=sharing>
- 截止日期同 Ceiba Deadline: 12/11/2021 23:59:59 (GMT+8)

其他規定 Other Policy

- Lateness
 - Ceiba 每遲交一天(不足一天以一天計算) hw3 所得總分將 $\times 0.7$
 - 不接受程式 or 報告單獨遲交
 - 不得遲交超過一天, 若有特殊原因請儘速聯絡助教
- Runtime Error
 - 當 程式錯誤, 造成助教無法順利執行, 請在公告時間 內寄信向助教說明, 修好之後重新執行所得kaggle部分分數將 $\times 0.5$ 。
 - 可以更改的部分僅限 syntax 及 io 的部分, 不得改程式邏輯或是演算法, 至於其他部分由助教認定為主。

其他規定 Other Policy



- Cheating

- 抄 code、抄 report (含之前修課同學)
- 開設 kaggle 多重分身帳號註冊 competition
- 於訓練過程以任何不限定形式接觸到 testing data 的正確答案
- 不得上傳之前的 kaggle 競賽
- 教授與助教群保留請同學到辦公室解釋 coding 作業的權利, 請同學務必自愛

Score - Report.pdf

[Report link](#)

- (1%) 請以block diagram或是文字的方式說明這次表現最好的 model 使用哪些layer module(如 Conv/RNN/Linear 和各類 normalization layer) 及連接方式(如一般forward 或是使用 skip/residual connection), 並概念性逐項說明選用該 layer module 的理由。
- (1%) 請比較 word2vec embedding layer 初始設為 non-trainable/trainable 的差別, 列上兩者在 validation/public private testing 的結果, 並嘗試在訓練過程中設置一策略改變 non-trainable/trainable 設定, 描述自己判斷改變設定的機制以及該結果。
- (1%) 請敘述你如何對文字資料進行前處理, 並概念性的描述你在資料中觀察到什麼因此你決定採用這些處理, 並描述使用這些處理時作細節, 以及比較其實際結果, 該結果可以不用具備真正改進。如果你沒有作任何處理, 請給出一段具體描述來 說服我們為什麼不做處理可以得到好的結果, 這個理由不能是因為表現比較好。
- (1%) 請「自行設計」兩句具有相同單字但擺放位置不同的語句, 使得你表現最好的模型 產生出不同的預測結果, 例如 "Today is hot, but I am happy" 與 "I am happy, but today is hot", 並討論造成差異的原因。

Requirements

- 沒有特定限制model種類
 - RNN/LSTM
- 不能使用額外 data
- 如果你的code不只一個檔案(或有多個參數)請附上readme或shell script
- testing process要在10分鐘內跑完

Assignment Regulation

- Only Python 3.7 available !!!!
- 開放使用套件(或是你可以直接下載我們當初的[環境yaml檔案](#))
 - numpy == 1.19
 - pandas == 1.1.3
 - python standard library
 - pytorch == 1.10.0 (torchvision == 0.11.1)
 - tensorflow == 2.1.0
 - keras == 2.2.4
 - cv2
 - pillow >= 6.1.0
- 若需使用其他套件，請儘早寄信至助教信箱詢問，並請闡明原因。

配分 Grading Criteria-Kaggle(2%)

- Kaggle deadline : 12/9/2021 23:59:59 (GMT+8)
- Kaggle - 2%
 - ❑ 超過public leaderboard的simple baseline分數 : **0.5%**
 - ❑ 超過private leaderboard的simple baseline分數 : **0.5%**
 - ❑ 超過public leaderboard的strong baseline分數 : **0.5%**
 - ❑ 超過private leaderboard的strong baseline分數 : **0.5%**
- Bonus - 1%
 - (1.0%) private leaderboard 排名前五名, 並繳交投影片描述實作方法, 另外需錄製一份講解影片(少於三分鐘)作一個簡單的 presentation, 助教將公布給同學們參考

配分 Grading Criteria - report(8%)

- Programming Report - 4%
 - <https://docs.google.com/document/d/1mjawi2jtHhBrnxluXZ-Q2pNbh8YWuAm4HK3H6Khgoc8/edit?usp=sharing>
- Math Problem - 4%
 - <https://hackmd.io/@hAe95tLdTVqEePbZsJyqrw/BkWSTuqPF>
 - 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: ntueemlta2021fall@gmail.com

TA Hour

- TBD @ google meet
- ML2021 TA Hour
- 連結: <https://meet.google.com/zyi-gfgj-tdu>