
Machine Learning HW4

Recurrent Neural Networks

MLTAs

ntueemlta2023@gmail.com

Outline

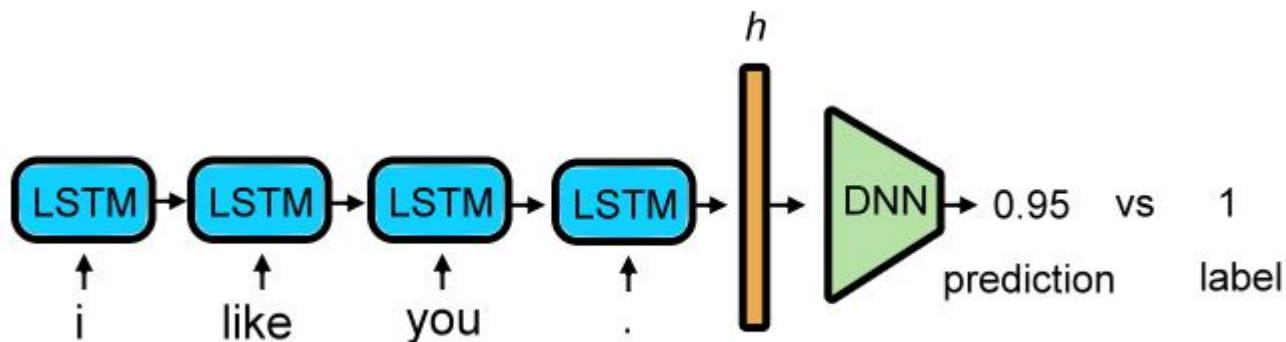
1. Task Introduction
2. Data Format
3. Kaggle
4. Rules, Deadline, Policy, Score
5. FAQ

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 (X) 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://colab.research.google.com/drive/16Pvhbl67Ork5c16C0XluRljaks9ru3-N0?usp=sharing>

- Dataset :

- <https://drive.google.com/drive/folders/1786AXjRAtgFvWMBeh-bLm4MtU21lQpBg?usp=sharing>

Kaggle Info & Deadline

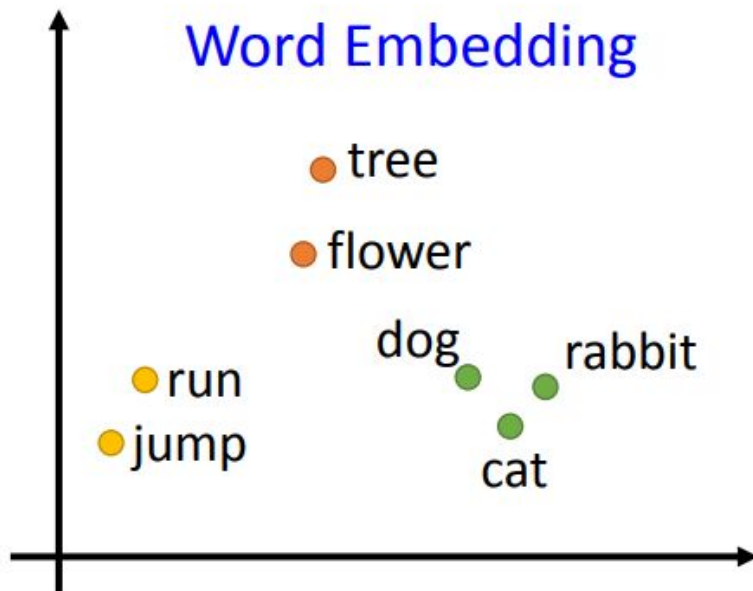
- Link: <https://kaggle.com/competitions/ml-2023-fall-hw3> -> 網站的url確實是hw3...
- 個人進行、不須組隊
- Team Name:
 - 修課學生: 學號_任意名稱 (ex: b09901666_name)
- Maximum Daily Submission: 5 times
- Kaggle Deadline: 11/25/2023 01:00:00 (GMT+8)
- Cool Deadline: 11/26/2023 23:59:59 (GMT+8)
- test set的20000筆資料將被分為兩份, 10000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

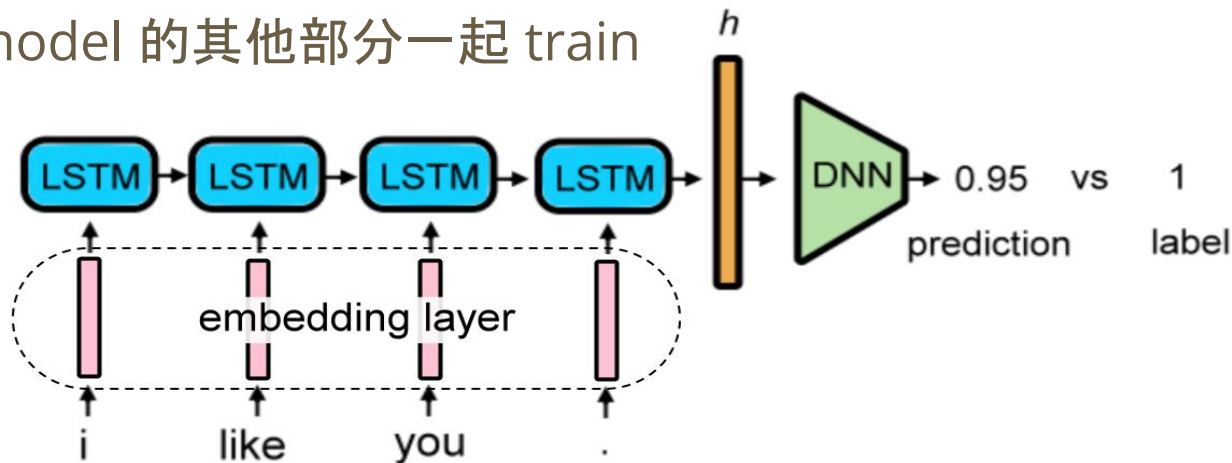
$$200000(\text{data}) * 30(\text{length}) * 20000(\text{vocab size}) * 4(\text{Byte}) = 4.8 * 10^{11} = \mathbf{480 \text{ GB}}$$

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)

id,label, text

```
73967,1,i am the rhtymn police and i will arrest anyone dancing off beat 2day
73968,1,"has finished her photog exam tomorrow and all week , blahhhh"
73969,1,"quick meeting , everything is in order ! now going to lay in the heat"
73970,0,"just woke up from grand sleep !! this is the life , sad its only till 1 . 7 . 09 then back to work"
73971,1 not yet sleepy now watching ang spoiled bubble gang haha ! it has a replay ! aww
```

Data Format (unlabeled data)

text

```
sorry cant make it staying home with sinus infection instead siliconbeach  
don ' t say such silly things ! twitter aint shit ! but you ... you be amazing  
setting up prices is difficult work  
"just logged in to for the first time in months and i can see from my feed or lack of it , everyone left"  
breakfast later we do a paper round  
very sunny and great environment . do come !  
"madcatz fightstick te has arrived !... but , i only managed to get 5 units and they have all been pre booked next batch  
shud be in next"
```

Kaggle

Kaggle submission format

Kaggle link: [ML 2023 FALL HW4 | Kaggle](#)

請預測 testing set 中2萬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/1cO0q5AA_QTyIokC_eOdJJCrqfZeGkDL094AuoiSoStU/edit?usp=sharing
- 截止日期同 Cool Deadline: 11/26/2021 23:59:59 (GMT+8)

其他規定 Other Policy

- Lateness
 - Cool 每遲交一天(不足一天以一天計算) hw4 所得總分將x0.7
 - 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x0.5。
 - 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

https://docs.google.com/document/d/1cO0q5AA_QTylokC_eOdJJCrqfZeGkDLo94AuoiSoStU/edit?usp=sharing

- (1%) 請以block diagram或是文字或者繪圖的方式 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 的理由以及設計想法。
- (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, 這個理由不能只是單純因為表現比較好。
- (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, 但請不要用範例句子。

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

Requirements - environment issue

- [environment.yaml](#)
- 若需要其它套件，請及早來信詢問。
 - 套件版本與python版本並沒有強制限制，以colab能跑為主

```
$pip list
```

```
$python -version
```

```
$nvidia-smi
```

如果助教跑你的code跑不動會寄信與同學確認
 - 強烈建議不要在WSL上嘗試裝設nvidia-driver
在Windows/Ubuntu/MacOS環境下直接跑反而會更加順利
繳交時記得注意一下資料夾結構跟檔名，推薦在.ipynb檔頭用註解寫上跑的系統

```
#@system[Ubuntu-22.04 LTS/Windows10/CentOS/MacOS/Colab...]
```


Requirements - file uploading

- 你的上繳至 cool 中的檔案請壓縮在同一個資料夾，並取名為<學號>_hw4.zip
 - 該 zip 檔案內請包含：
 - report.pdf
 - hw4.ipynb (or hw4.py)
 - 請將參數連結(最佳model, 或其他reproduce必須的檔案)附在report中
 - 也可以上傳自己的雲端, 在 code內用 gdown 指令。
 - 範例：

```
1 import gdown
2
3 url = <你的 model 壓縮檔 Google雲端連結>
4 output = "model.zip"
5 gdown.download(url=url, output=output, quiet=False, fuzzy=True)
6 !unzip -o model.zip
```

- Optional: 其他可以幫助說明你的 code 的文件
- 簡述一下使用到的套件名稱以及套件版本(ex:python3.10 & numpy(a.b.c).....)

配分 Grading Criteria-Kaggle(2%)

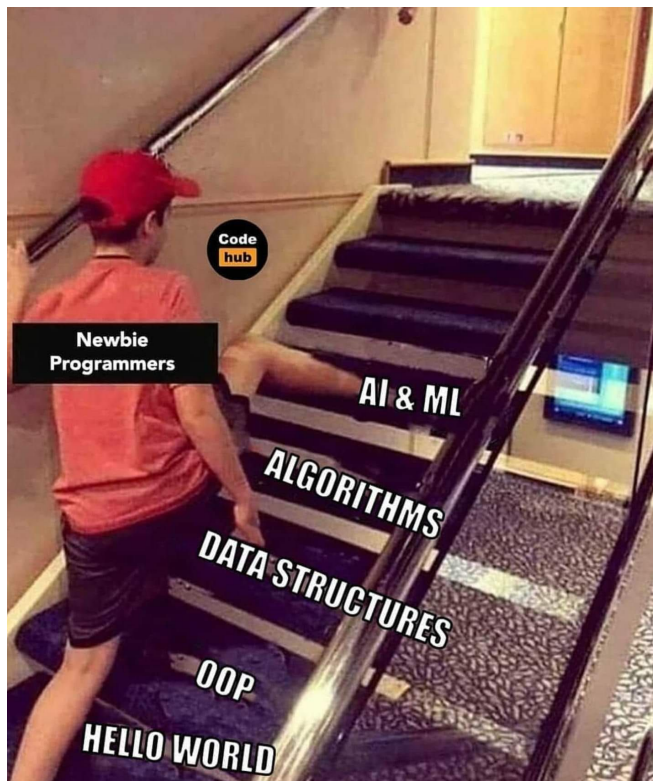
- Kaggle deadline : 11/25/2023 01:00:00 (GMT+8)
- Kaggle - 2%
 - ❑ 超過public leaderboard的simple baseline分數 : **0.5%**
 - ❑ 超過private leaderboard的simple baseline分數 : **0.5%**
 - ❑ 超過public leaderboard的strong baseline分數 : **0.5%**
 - ❑ 超過private leaderboard的strong baseline分數 : **0.5%**
- Bonus - 1%
 - (1.0%) 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
 - 請在12/2/2023之前(卅一周內)寄信繳交

配分 Grading Criteria - report(10%)

- Programming Report - 4%
 - https://docs.google.com/document/d/1cO0q5AA_QTyIokC_eOdJJCrfZeGkDLo94AuoiSoStU/edit?usp=sharing
- Math Problem - 6%
 - <https://ntueemlta2023.github.io/homeworks/hw4/ml-2023fall-hw4-math.pdf>
 - Type in latex(preferable) or take pictures of your handwriting
- Write them in report.pdf

FAQ

- 環境問題請善用 google。
 - `pip install xxx`
 - `apt-get install xxx`
- 有問題建議可以在 FB Group 裡面留言發問，可能很多人都有一樣的問題。
- 若有其他問題，請寄信至助教信箱，**請勿直接FB私訊助教**。
 - Mail title:
[ML23_hw4_code]{your name}_{title}
[ML23_hw4_math]{your name}_{title}



TA

- ntueemlta2023@gmail.com or b08209023@ntu.edu.tw

- Title : [ML23_hw3_code]{your name}_{title}

- 關於環境/運行時間/註解問題:

1. 環境部分基本上以 colab 可以順利執行為主, 推薦使用 python3.7, 但 python3.10 也可以, 倘若有套件衝突需要檢查自己電腦環境有沒有爛掉, 強烈不建議用 WSL 跑, 如果都是用助教的範例 code 為主並且在 colab 上可以順跑, 沒有奇怪的 import 的話則不需要特別註解跟提交。

2. 時間限制非強制, 但還是推薦壓在 10min 左右可以跑 10 個 epoch, 基本上助教檢驗 code 時不要讓助教的電腦跑太久即可

(助教顯卡約莫 1650 等級, 不要拿 RTX ada A6000 跑個十分鐘或者拿工作站的大量 GPU unit 去 train)

這次作業跑很快

- TA hour:

週二早上 1000~1200 @ IPCS 206, 但 **程式助教 11/17~21 不在台北**, 因此強烈推薦寄信問跟助教另外約時間/google meet

學術倫理

- 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

