Machine Learning HW4

TAs
ntu.mlta@gmail.com

Outline

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

Requirements

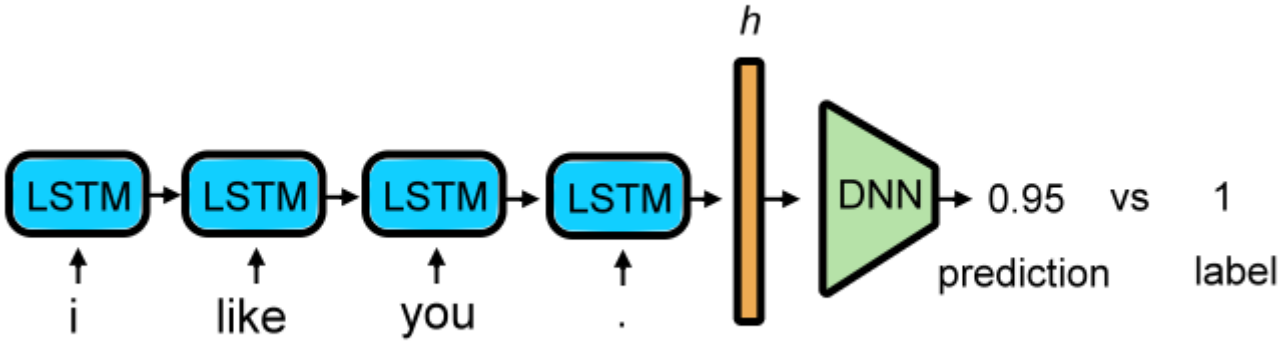
1. 請使用RNN實作model
2. 不能使用額外data (禁止使用其他corpus pretrain好的model)
3. 請附上訓練好的best model (及其參數)至github release或dropbox，並於hw4_test.sh中寫下載的command（請參照以下網站中方法：http://slides.com/sunprinces/deck-16#/2 ）
4. model大小在100mb以內的可以直接上傳到github
5. hw4_test.sh要在10分鐘內跑完（model下載時間不包含在此）
6. Available Toolkit Versions:
    a. Only Python3.5+
    b. 可使用numpy, pandas0.20+以及python standard library
    c. 可額外使用tensorflow1.3, keras2.0.8, pytorch0.2.0, gensim, GloVe
    d. 使用nltk需下載一些額外的data 所以禁止使用nltk

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

除了有label的data以外，我們還額外提供了120萬筆左右沒有label的data

- labeled training data　　：20萬
- unlabeled training data　：120萬
- testing data　　：20萬（10萬public，10萬private）

Preprocessing the sentences

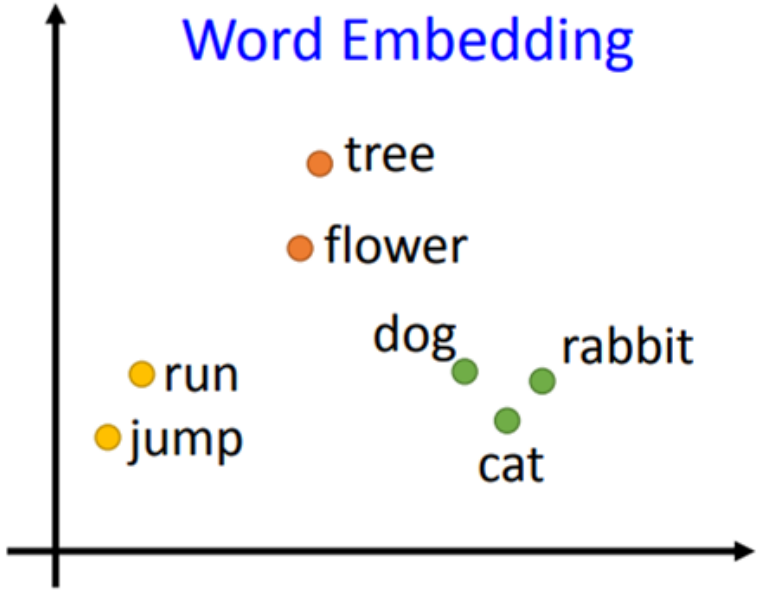
- 先建立字典，字典內含有每一個字所對應到的index
  example:
  “I have a pen.” -> [1, 2, 3, 4]
  “I have an apple.” -> [1, 2, 5, 6]

- 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(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 [1,2,3,4,5]
  我們可以用不同的one-hot vector來代表這個字
    1 -> [1,0,0,0,0]
    2 -> [0,1,0,0,0]
    3 -> [0,0,1,0,0]
    4 -> [0,0,0,1,0]
- Issue :
  a. 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)
  b. 很吃記憶體

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

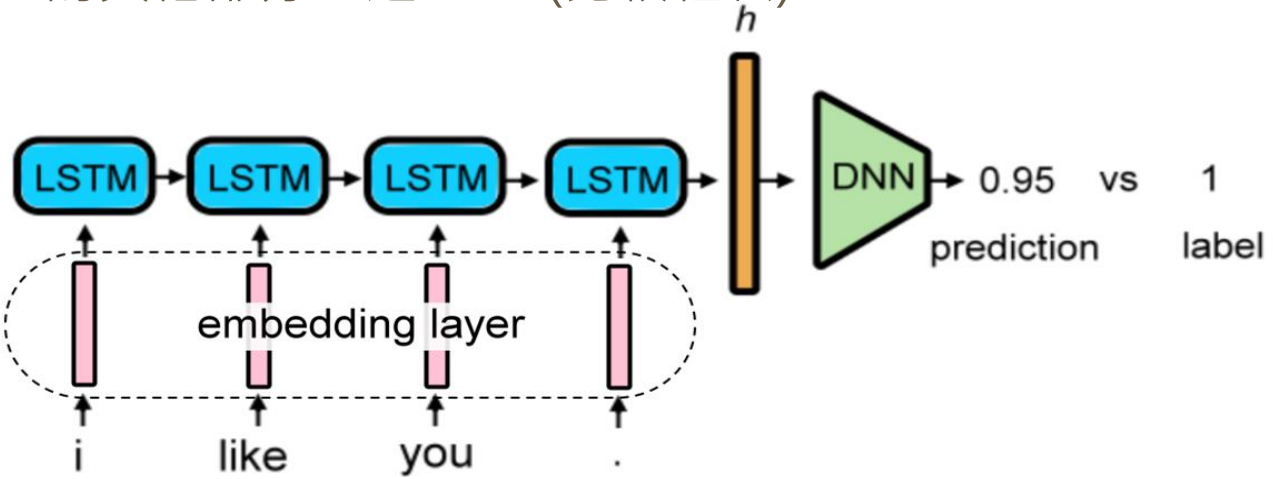
Word Embedding(*)

1. 用一些方法pretrain 出word embedding (ex：skip-gram、CBOW)

   reference : http://speech.ee.ntu.edu.tw/~tlkagk/courses/ML_2017/Lecture/word2vec%20(v2).pdf

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

1. 跟model的其他部分一起train (比較輕鬆)



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.

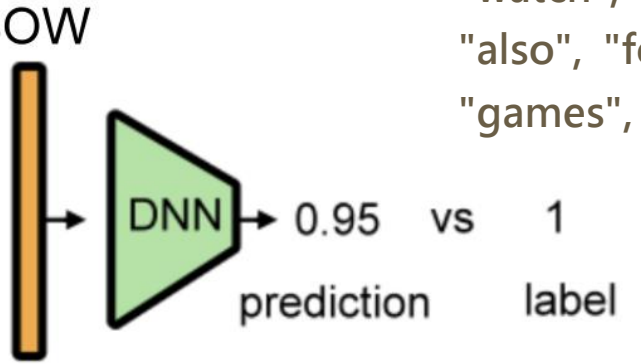
在BOW的表示方法下，會變成

(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"]



Semi-Supervised learning

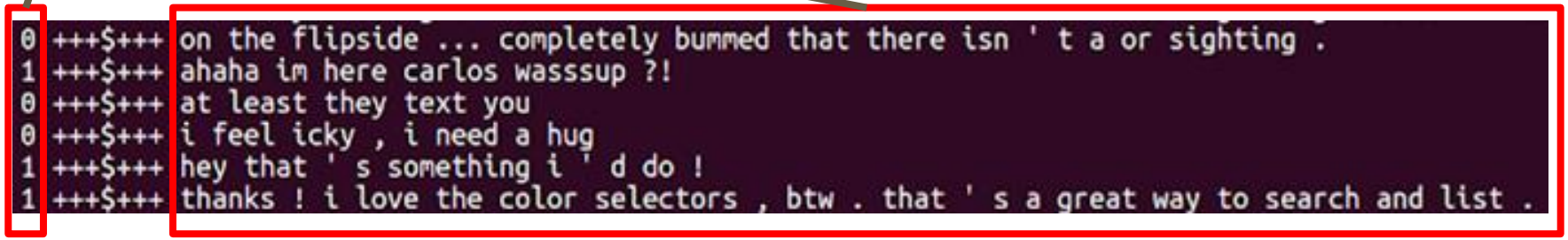
- semi-supervised 簡單來說就是讓機器自己從unlabel data中找出label，
  而方法有很多種，這邊簡單介紹其中一種比較好實作的方法 Self-Training。
- Self-Training：

  把train好的model對unlabel data做預測，並將這些預測後的值轉成該筆
  unlabel data的label，並加入這些新的data做training。
  你可以調整不同的threshold、或是多次取樣來得到比較有信心的data。

  ex：設定pos_threshold=0.8，只有在prediction>0.8的data才會被標上1的lab

Data Format (labeled data)

label +++$+++ text



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 . maaaan , 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

1. kaggle_url：https://www.kaggle.com/t/98e81d1542fe47e092d3e102dfe42360
2. 請使用先前使用的kaggle帳號登入。
3. 個人進行，不需組隊。
4. 隊名:學號_任意名稱(ex. r06666666_許哲瑋打球)，旁聽同學請避免學號開頭。
5. 每日上傳上限5次。
6. test set的資料將被分為兩份，一半為public，另一半為private。
7. 最後的計分排名將以2筆自行選擇的結果，測試在private set上的準確率為準。

★ kaggle名稱錯誤者將不會得到任何kaggle上分數。

★ 本次作業為期三週，strong baseline將在第二週結束時公布

Kaggle submission format

請預測test set中二十萬筆資料並將結果上傳Kaggle

1. 上傳格式為csv。
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
```

Deadline

- Kaggle: 2017/12/7 11:59 p.m. (GMT+8)
- report, github: 2017/12/8 11:59 p.m. (GMT+8)

助教會在deadline一到就clone所有程式，並且**不再重新clone任何檔案**

Policy

github上ML2017fall/hw4/裡面請至少包含：

1. Report.pdf
2. hw4_train.sh
3. hw4_test.sh
4. your python files
5. model參數 (Make sure it can be downloaded by your script.)
   (*請將model download到與script相同的位置)

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: 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predict。)

Score - Kaggle Rank

- (0.8%) 超過public leaderboard的simple baseline分數
- (0.8%) 超過public leaderboard的strong baseline分數
- (0.8%) 超過private leaderboard的simple baseline分數
- (0.8%) 超過private leaderboard的strong baseline分數
- (0.8%) 2017/11/30 23:59 (GMT+8)前超過public simple baseline
- (BONUS) kaggle排名前五名(且願意上台跟大家分享的同學)
- **其中，前五名排名以public平均為準，屆時助教會公布名單**

Score - Other Policy

- script 錯誤，直接0分。若是格式錯誤，請在公告時間內找助教修好，修完kaggle分數*0.7。
- Kaggle超過deadline直接shut down，可以繼續上傳但不計入成績。
- Github遲交一天(*0.7)，不足一天以一天計算，不得遲交超過兩天，有特殊原因請找助教。
- Github遲交表單：(遲交才必需填寫)
  - Github CODE 遲交表單：https://goo.gl/TzKypu
  - Github REPORT 遲交表單：https://goo.gl/9jVX9E
    遲交請「先上傳程式」Github再填表單，助教會根據表單填寫時間當作繳交時間。
- 上傳的model總和大小建議在500MB以內。

Score - Report.pdf

[注意] 這次報告建議頁數為三頁，請盡量精簡內容

- (1%) 請說明你實作的 RNN model，其模型架構、訓練過程和準確率為何？
- (1%) 請說明你實作的 BOW model，其模型架構、訓練過程和準確率為何？
- (1%) 請比較bag of word與RNN兩種不同model對於"today is a good day, but it is hot"與"today is hot, but it is a good day"這兩句的情緒分數，並討論造成差異的原因。
- (1%) 請比較"有無"包含標點符號兩種不同tokenize的方式，並討論兩者對準確率的影響。
- (1%) 請描述在你的semi-supervised方法是如何標記label，並比較有無semi-surpervised training對準確率的影響。

作業網址參考: https://ntumlta.github.io/2017fall-ml-hw4/

Report template:https://goo.gl/vb6Baq

Collaborators請附上學號與姓名

小老師制度（手把手教學）

- 在11/30以前超過simple baseline並願意在12/1在上課時間教導同學撰寫作業四程式，請填寫一下表單：https://goo.gl/forms/vi8RbHKATZNkwHS42
- 11/30將公布小老師名單在作業網頁，人數太多將以符合以下標準的同學為主:
  1. 沒有當過小老師
  2. Kaggle Public Leaderboard成績排名較高 (但請不要因此想overfit public set)
- 小老師當次成績+1%

FAQ

- 若有其他問題，請po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： ntu.mlta@gmail.com

Link

- 雲端使用方法：http://slides.com/sunprinces/deck-16#/2 ）
- Kaggle：https://www.kaggle.com/t/98e81d1542fe47e092d3e102dfe42360
- 作業網址：https://ntumlta.github.io/2017fall-ml-hw4/
- Report template: https://goo.gl/vb6Baq
- Github 遲交表單:
  - Github CODE 遲交表單：https://goo.gl/TzKypu
  - Github REPORT 遲交表單：https://goo.gl/9jVX9E
- 小老師報名表單：https://goo.gl/forms/vi8RbHKATZNkwHS42