Machine Learning HW5

MLTAs
mlta2020fall@gmail.com

Outline

- Task Description - Image Clustering
- Kaggle
- Requirements & Regulation
- Grading Policy
- Hand-by-hand
- FAQ

Outline

- Task Description - Image Clustering
- Kaggle
- Requirements & Regulation
- Grading Policy
- FAQ

Image clustering - outline 1/7

- 目標:分辨給定的兩張 images 是否為風景。
  - 除了 image 都是32*32*3的圖片，沒有任何 label
  - 只能用我們給的 data，不能使用額外的 dataset，也不能使用額外資料 train 的 model

V.S

Image clustering - data 2/7

- trainX.npy
  - 利用np.load()讀入資料。
  - 裡面總共有 9000 張 RGB圖片，大小都是32*32*3
- trainY.npy
  - 不能用於模型的訓練，只能用來繪製問題 c之圖。被發現使用trainY.npy於訓練者本次作業零分。
  - shape為(9000,)。

Image clustering - data 3/7

- sample_submission.csv
  - 第一行是 “id, label”
  - 之後每一行都會有 test case ID，以及對這個 test case 的 prediction
  - 如果 test case 的兩張 image 預測後是來自同一 dataset，Ans 的地方就是 1，反之是 0
    - 我們評分以Accuracy作為標準，前五個label皆為0。

Image clustering - methods 4/7

- 如果直接在原本的 image 上做 cluster，結果會很差 (有很多冗餘資訊)

  => 需要更好的方式來表示原本的image

- 為了找出這個更好的方式，可以先將原始 image 做 dimension reduction，用比較少的維度來描述一張 image

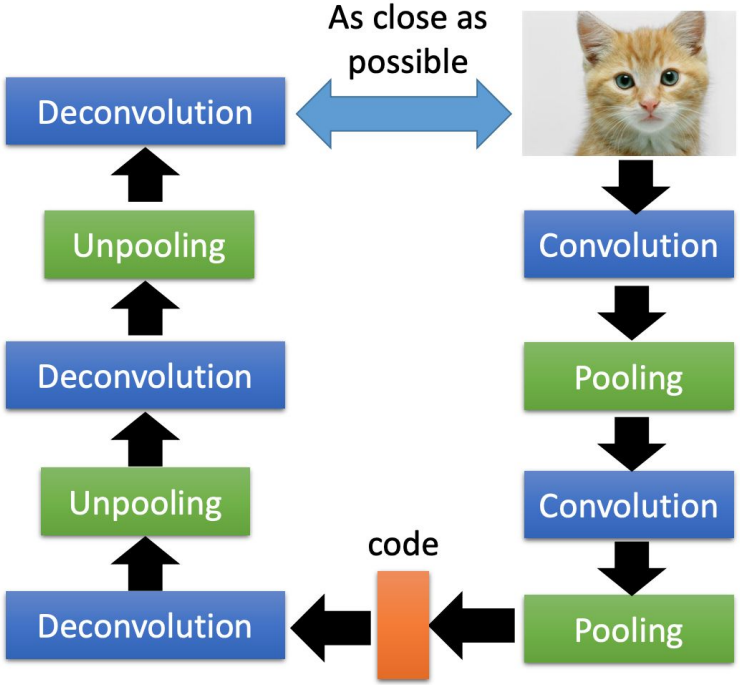
  e.g. autoencoder, PCA, SVD, t-SNE

Image clustering - requirements 5/7

1. 請實作用 **autoencoder** 將9000張圖片降維

2. 再利用降維過的latent code做分類

3. 預測9000筆測資是否來自相同的dataset



註: 同學實作的方法需含有autoencoder, 但還是可以將其他的降維方法一起搭配使用

Image clustering - methods (cont.) 6/7

- 接著對降維過後的數據做 cluster
  - cluster：可以試試 K-means

- 或者你可以衡量兩個降維過後的 images，他們之間的相似度 (similarity)。如果相似度大於一個設定好的 threshold，就把這兩個 images 當成同一類別
  - 算 similarity 的方法：euclidean distance, cosine similarity……

Image clustering - methods (cont.) 7/7

- 其他可能有幫助的事:
  - 更改model sturcture (增長 或是 縮小model)
  - 必須找個方法來衡量方法的好壞，一個直覺的方法是利用降維過後的feature 去 reconstruct 成原本的 image。如果 reconstruct 的結果越接近原本的image，可以一定程度的代表你抽出來的feature 越好
  - 對原始 image 做 data augmentation
  - try different number of cluster
  - 看看老師 unsupervised learning 上課內容
  - 看看網路上的 unsupervised learning 內容

Outline

- Task Description - Image Clustering
- Kaggle
- Requirements & Regulation
- Grading Policy
- FAQ

Kaggle - Info 1/2

- Kaggle 連結 : https://www.kaggle.com/c/ml2020fall-hw5-1/overview
- 個人進行，不需組隊
- 隊名:
  - 修課學生：學號_任意名稱（ex: b05902127_一不小心做成1）
  - 旁聽：旁聽__任意名稱
- 每天上傳上限 20 次
- 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。
- test set的資料將被分為兩份，一半為public，另一半為private。
- 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。
- ★ kaggle名稱錯誤者的分數將x0.7。

Kaggle - format 2/2

- 預測 9000 筆 training data 是否為風景還是物體，將預測結果上傳至kaggle
  - Upload format : csv file
  - 第一行必須是 id,label
  - 第二行開始，每行分別為id值及預測結果 (binary)，以逗號隔開

  - 預測後是來自同一類別, label 的地方就是 1，反之是 0
  - Evaluation: Accuracy

- 範例格式如右

sample_submission.csv

```
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
20,0
```

Outline

- Task Description - Image Clustering
- Kaggle
- Requirements & Regulation
- Grading Policy
- FAQ

Requirements

- 將預測結果上傳kaggle
- 用autoencoder 實作降維
- 回答report問題
- 不能使用額外的data訓練，也不能使用pre-trained model
- 不能 call 其他線上 API

Regulation - GitHub 2/3

- 你的 github 上 ML2020FALL/hw5/ 中請包含：
  - report.pdf
  - cluster.sh (for image clustering 那題，限制至少要使用autoencoder)（限時20分鐘）
  - your python files
  - your model files (can be loaded by your python file)

- **請不要上傳 dataset，請不要上傳 dataset，請不要上傳 dataset。**
- 如果你的 model 超過 github 的最大容量，可以考慮把 model 放在其他地方 (http://slides.com/sunprinces/deck-16#/2%EF%BC%89)。
- model 可以是多個檔案，例如 pytorch model。

Regulation 1/3

- 套件的部份請參考連結
- **Conda file**
- **若需要其它套件，請及早來信詢問。**若 import 有發生錯誤，分數將x0.7
- 如果對此env有問題，可以在FB/寄信問。

Regulation - Script Usage 3/3

- 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

  bash cluster.sh <trainX.npy path> <prediction file path>

  e.g. bash cluster.sh trainX.npy ans.csv

- Script 所使用之模型，如 hdf5, pt, pickle 檔等，可以於程式內寫死路徑，助教會 cd 進 hw5 資料夾執行 reproduce 程序。

Outline

- Task Description - Image Clustering
- Kaggle
- Requirements & Regulation
- Grading Policy
- FAQ

Grading Policy - Deadline 1/6

- Kaggle Deadline: 2020/12/18 23:59:59 (GMT+8)
- Github Deadline: 2020/12/18 23:59:59 (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

若遲交請寄信給TA說你遲交，其內容需要包含你的學號以及repo url。

Grading Policy - Evaluation (4% + Bonus 1%) 2/6

- (2%) 超過public leaderboard的simple baseline分數
- (2%) 超過public leaderboard的strong baseline分數
-
-
- (BONUS 1%) private leaderboard 排名前五名且於助教時間上台分享的同學
    - 這個還請前五位強者做一下 slides。

Grading Policy - Report 3/6

- Programming Report - 3%
  - https://docs.google.com/document/d/1mts0RLtxMRiKscXSE0tFC_A3GOFcCEvn0yZ_mtO5eIU/edit?usp=sharingLtxMRiKscXSE0tFC_A3GOFcCEvn0yZ_mtO5eIU/
- Math Problem - 3%(共兩題，每題1.5分)
  - Link
  - Type in latex(preferable) or take pictures of your handwriting
- Write them in report.pdf

Grading Policy - Report 4/6

2. Image clustering:

a. (1%) 請使用不同的Autoencoder model, 以及不同的降維方式(降到不同維度), 討論其reconstruction loss & public accuracy。(因此模型需要兩種, 降維方法也需要兩種, 但clustrering不用兩種。)

b. (1%) 從dataset選出5張圖, 並貼上原圖以及經過autoencoder後reconstruct的圖片。

c. (1%) 我們會給你dataset的label (trainY.npy) (不能用於模型之訓練)。請在二維平面上視覺化label的分佈。

★ trainY.npy 會在kaggle請用train好的模型去預測

Grading Policy - Report 5/6

- 限制
  - 檔名必須為 report.pdf !!!
  - 檔名必須為 report.pdf !!!
  - 檔名必須為 report.pdf !!!
  - 請用中文撰寫 report（非中文母語者可用英文）
  - 保留各題標題
  - 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）
  - 違反以上規定，report不予計分。
- Report模板連結
  - 連結：Link

- 截止日期同 GitHub Deadline: **2020/12/18 23:59:59 (GMT+8)**

Grading Policy - Other Policy 6/6

- **Lateness**
  - Github 遲交一天(不足一天以一天計算)
  - 遲交一天 *0.8, 兩天*0.6, 遲交超過兩天零分
  - 不接受程式 or 報告單獨遲交
  - 有特殊原因請找助教。

- **Script Error**
  - 當 script 格式錯誤, 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x0.7。
  - 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。
  - 不接受任何 py 檔的 coding 錯誤更改

FAQ

- 若有其他問題，請寄信至助教信箱，**請勿直接私訊助教**。
- 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
- 助教信箱
  mlta2020fall@gmail.com

相關連結

- 程式範例:https://colab.research.google.com/drive/1selHC_KXRwEL_rLWTV9fW2m3yHWE47bQ?usp=sharing
- conda file 連結:https://reurl.cc/6lMz7d
- kaggle:https://www.kaggle.com/c/ml2020fall-hw5-1/overview
- report 模板:https://docs.google.com/document/d/1mts0RLtxMRiKscXSE0tFC_A3GOFcCEvn0yZ_mtO5eIU/edit?usp=sharing
- github:https://classroom.github.com/a/d0bIiREB
- Math problem: https://drive.google.com/file/d/1-rmlFaIj_6hEfJGOHLKUxInoKMsKLHLf/view?usp=sharing