

---

---

# Machine Learning HW4

MLTAs  
mlta2019fall@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 <sup>1/7</sup>

- 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
  - Kaggle Deadline後會公佈在Kaggle上, 這個檔案用於report。
  - 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

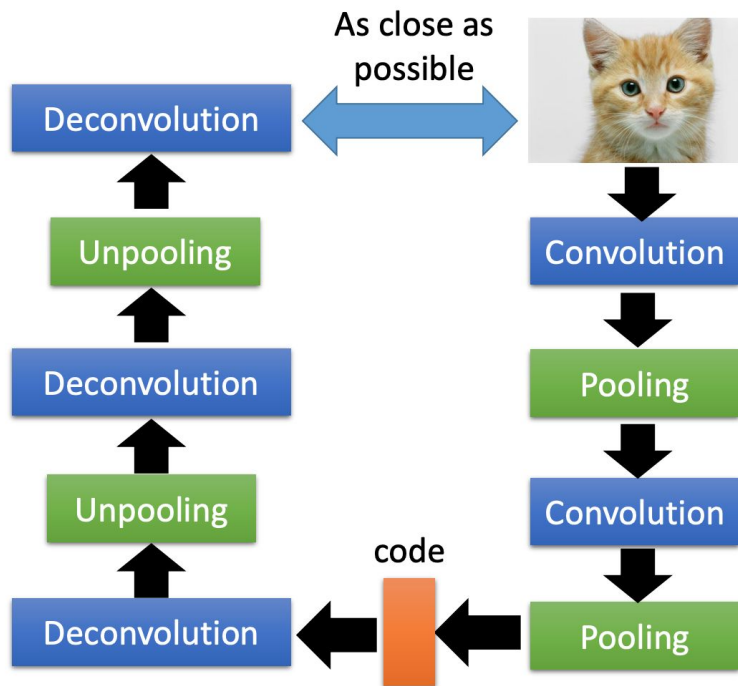
- 為了找出這個更好的方式, 可以先將原始 image 做 dimension reduction, 用比較少的維度來描述一張 image

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：
  - 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 <sub>1/2</sub>

- Kaggle 連結 : <https://www.kaggle.com/t/60de6a53d3074491a0daa3bad5591505>
- 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 5 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 x
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
22 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 <sup>1/3</sup>

- **Python Only**, 請使用 Python 3.6
- Python standard library are available
- **Image clustering:**
  - Numpy  $\geq 1.14$ , Pandas  $\geq 0.24.1$
  - Keras  $\geq 2.2.4$ , Tensorflow  $\geq 1.14.0$  , pytorch  $\geq 1.1.0$
  - Scipy  $\geq 1.2.1$
  - scikit-image  $\geq 0.15.0$
  - scikit-learn  $\geq 0.20.3$
  - Pillow  $\geq 6.0.0$
  - cv2
- **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
- [Conda File](#), 如果對此env有問題, 可以在FB/寄信問。



# Regulation - GitHub 2/3

- 你的 github 上 ML2019FALL/hw4/ 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%E4%B8%99>)。
- **model 可以是多個檔案, 例如 pytorch model.**

# Regulation - Script Usage <sup>3/3</sup>

- 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 hw4 資料夾執行 reproduce 程序。

# Outline

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

# Grading Policy - Deadline <sup>1/6</sup>

- Early Simple Deadline: 2019/11/15 23:59:59 (GMT+8)
- Kaggle Deadline: 2019/11/28 23:59:59 (GMT+8)
- Github Deadline: 2019/11/29 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%) <sup>2/6</sup>

- **(0%)** 超過public leaderboard的simple baseline分數
- (1%) 超過public leaderboard的strong baseline分數
- (1%) 超過private leaderboard的simple baseline分數
- (1%) 超過private leaderboard的strong baseline分數
- (1%) **2019/11/15 23:59:59 (GMT+8)**前超過public simple baseline
- (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 <sup>3/6</sup>

- Programming Report - 3%
  - [https://docs.google.com/document/d/1mts0RLtxMRiKscXSE0tFC\\_A3GOFcCEvn0yZ\\_mtO5eIU/edit?usp=sharingLtxMRiKscXSE0tFC\\_A3GOFcCEvn0yZ\\_mtO5eIU/](https://docs.google.com/document/d/1mts0RLtxMRiKscXSE0tFC_A3GOFcCEvn0yZ_mtO5eIU/edit?usp=sharingLtxMRiKscXSE0tFC_A3GOFcCEvn0yZ_mtO5eIU/)
- Math Problem - 3%
  - [Link](#)
  - Type in latex(preferable) or take pictures of your handwriting
- Write them in report.pdf

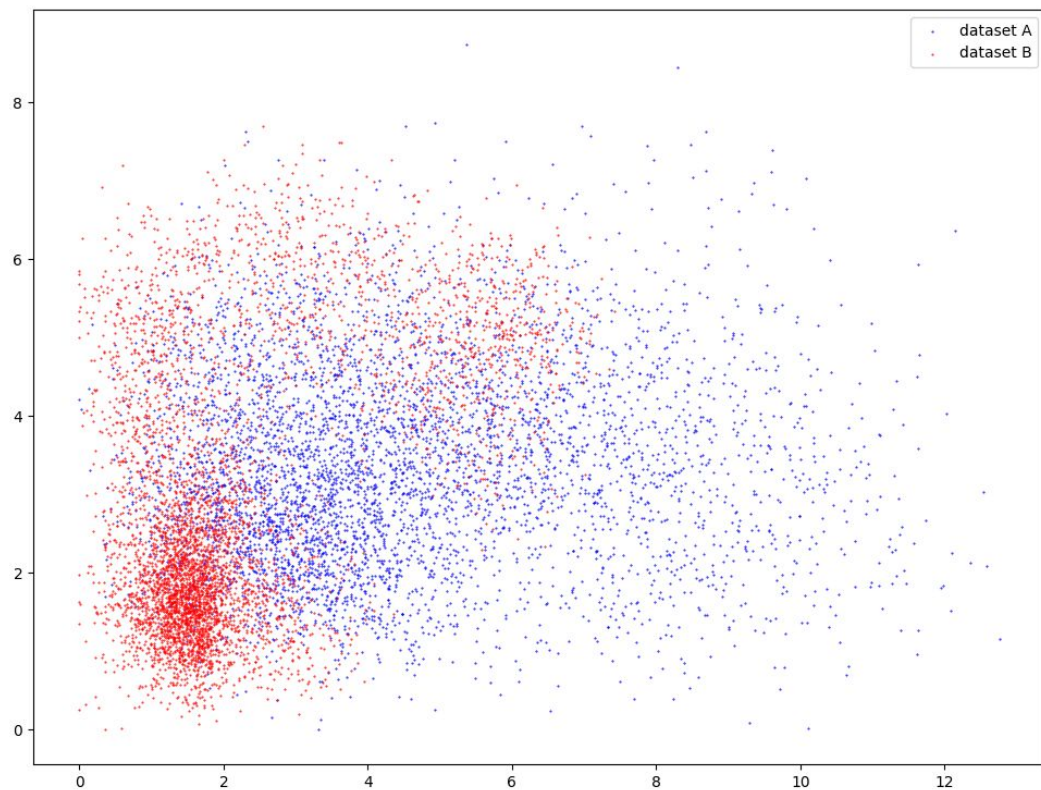
# Grading Policy - Report 4/6

## 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 / private 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2張圖, 並貼上原圖以及經過autoencoder後reconstruct的圖片。
- c. (1%) 之後我們會給你dataset的label (trainY.npy)。請在二維平面上視覺化label的分佈。

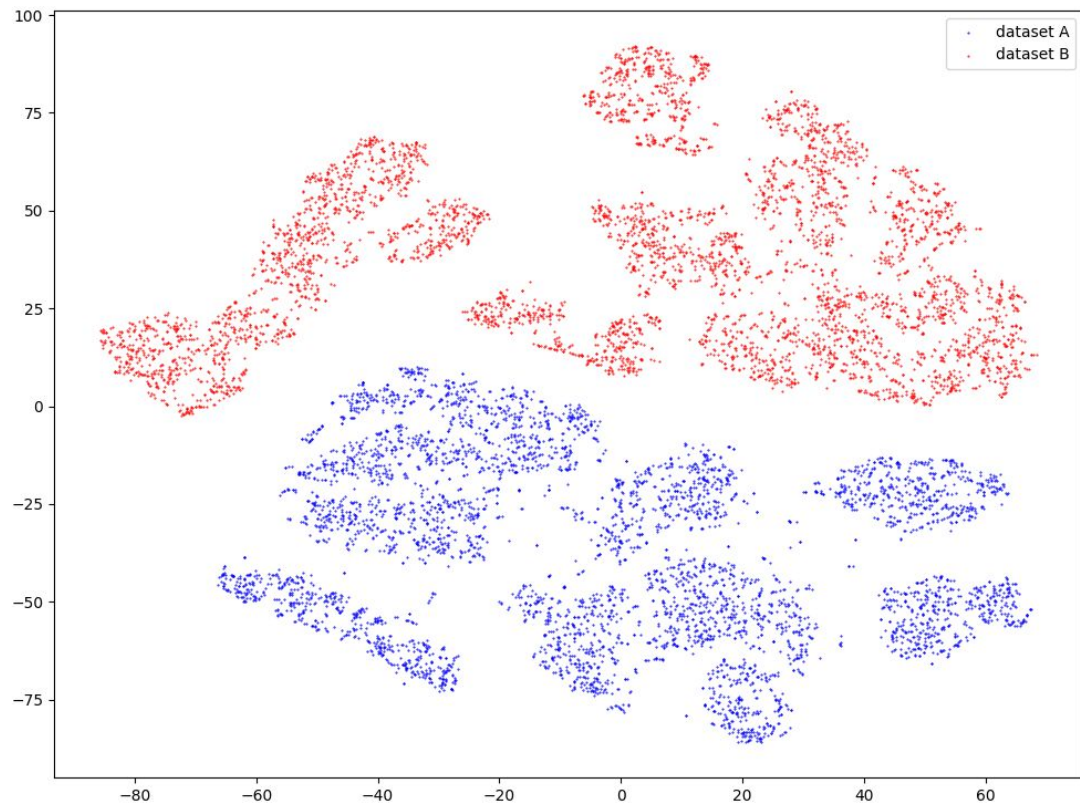
★ trainY.npy 會在kaggle請用train好的模型去預測

- 取降維過後的 feature 前兩個維度作圖





- 把降維過後的 feature 再用 t-SNE 投影到二維



# 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: 2019/11/22 23:59:59 (GMT+8)**

# Grading Policy - Other Policy <sup>6/6</sup>

- **Lateness**

- Github 遲交一天(不足一天以一天計算) hw4 所得總分將  $\times 0.7$
- **不接受程式 or 報告單獨遲交**
- 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: 遲交請先上傳遲交檔案至自己的github 後再填寫遲交表單, 助教群會以表單填寫時間作為繳交時間手動clone 檔案。

- **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 $\times 0.7$ 。
- 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

# Hand-by-hand 教學 <sup>1/1</sup>

- <https://slides.com/arvinliu/ml2019-fall-hw4-handbyhand>

# 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  
[mlta2019fall@gmail.com](mailto:mlta2019fall@gmail.com)

