

---

---

# Machine Learning HW3

MLTAs

ntueemlta2023@gmail.com

---

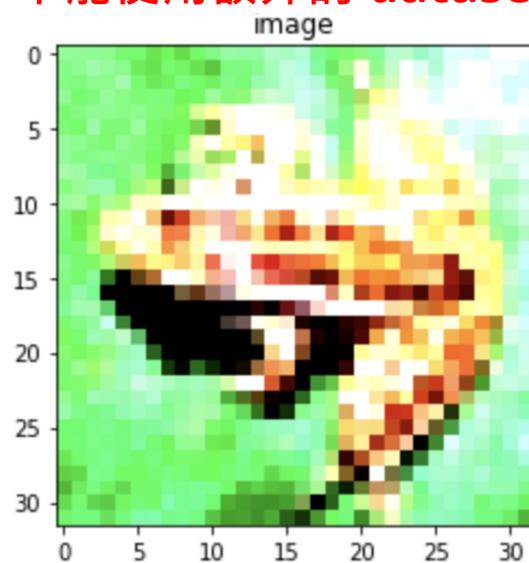
---

# Outline

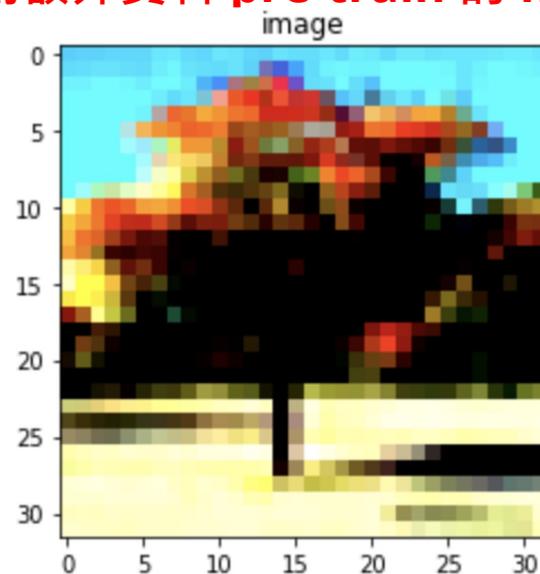
- Task Description - Embedding
    - Task: Image clustering
  - Kaggle
  - Requirements & Regulation
  - Grading Policy
  - FAQ
-

# 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
  - **不能使用額外的 dataset，也不能使用額外資料 pre-train 的 model**



V.S

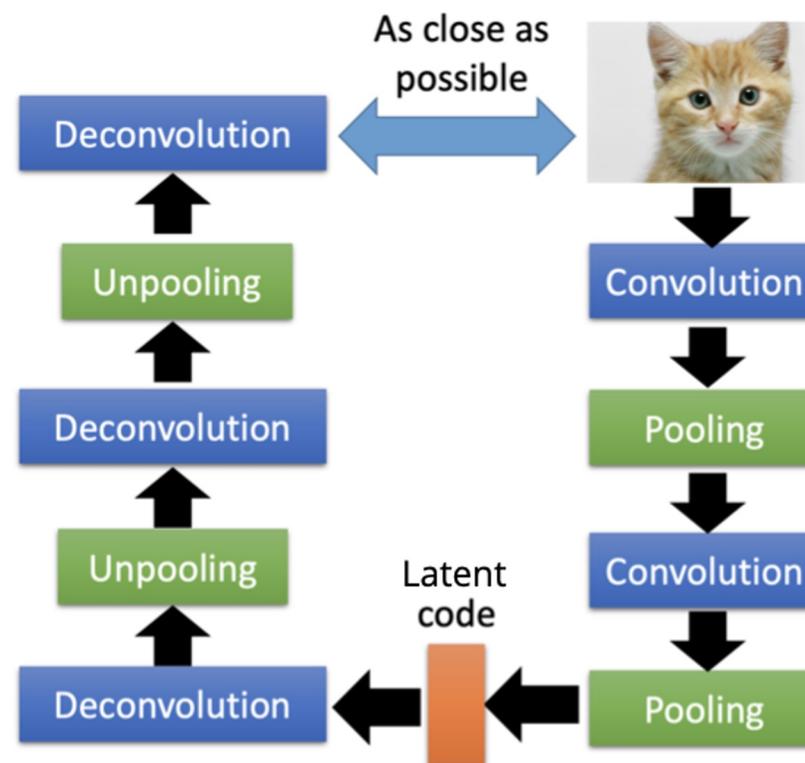


# Image clustering - methods <sup>2/7</sup>

- 如果直接在原本的 image 上做 cluster，結果會很差 (有很多冗餘資訊)
  - 更好的方式：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, or other embedding algorithms.
-

# Image clustering - requirements 3/7

1. 請實作用 **autoencoder** 將9000張圖片降維
2. 再利用降維過的 latent code 對這9000個 vector 去分類是否為風景



# Image clustering - data 4/7

- 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
- visualization.npy
  - 裡面總共有 5000 張 RGB圖片，大小都是32\*32\*3
  - 前一半 label 為 0，後一半 label 為 1
  - 該資料是寫 Report 用，**不能用於模型的訓練**。使用該資料訓練者，本次作業0分計算。

# Image clustering - data 5/7

- sample\_submission
  - 第一行是 "id, label"
  - 之後每一行都會有 image ID，以及對這個 image 的 prediction
  - 評分標準：Accuracy
    - 如果 test case 的兩張 image 預測後是來自同一類圖片，Ans 的地方就是 1，反之是 0
    - 若發現 accuracy 非常低，有可能是分群時 0,1 互換了，同學可以自行換回來。
      - 前五個 label 為 0, 1, 0, 1, 0

# Image clustering - Suggestion 6/7

- 對降維過後過後的數據做 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 - Suggestion 7/7

- 其他可能有幫助的事：
    - 對原始 image 做 data augmentation
    - try different number of cluster
    - 看看老師 unsupervised learning 上課內容
    - 衡量好壞：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，代表你抽出來的 feature 越好
    - 但同時 model 不能太複雜，生成能力太好會讓 latent code 不容易分群
-

# Outline

- Task Description - Embedding
    - Task: mage clustering
  - Kaggle
  - Requirements
  - Grading Policy
  - FAQ
-

# Kaggle - Info <sup>1/2</sup>

- Kaggle 連結：[ML2023 FALL HW3 | Kaggle](#)
  - 個人進行，不需組隊
  - 隊名：
    - 修課學生：學號\_任意名稱 ( ex: r10942198\_abc123 )
  - 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。
-

# Kaggle - format 2/2

- 預測 trainX.npy 的 data 來自哪個 dataset，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
  - 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 - Embedding
    - Task: mage clustering
  - Kaggle
  - Requirements
  - Grading Policy
  - FAQ
-

# Requirements

- Image clustering
    - 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
-

# 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>\_hw3.zip**
  - 該 zip 檔案內請包含：
    - **report.pdf**
    - **hw3.ipynb (or hw3.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).....)

# Outline

- Task Description - Embedding
    - Task: mage clustering
  - Kaggle
  - Requirements
  - Grading Policy
  - FAQ
-

# Grading Policy - Deadline

- Kaggle Deadline: 2023/11/11 01:00:00 (GMT+8)(請當成11/10截止 多一個小時給大家上傳調整)
  - Cool Deadline: 2022/11/12 23:59:59 (GMT+8)
-

# Grading Policy - Evaluation (2%)

- (0.5%) 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
-

# Grading Policy - Report (10%)

- Template Report - 4%
    - <https://docs.google.com/document/d/1n4RGlxXpLrTakT1TkypLvLs1dpuBJkaq/edit?usp=sharing&oid=112465961449455869485&rtpof=true&sd=true>
  - Math Problem - 6%
    - <https://ntueemlta2023.github.io/homeworks/hw3/ml-2023fall-hw3-math.pdf>
    - Type in latex(preferable) or take pictures of your handwriting
  - Write them in report.pdf
-

# Grading Policy - Report

- 限制
  - 檔名必須為 report.pdf !!!
  - 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
  - <https://docs.google.com/document/d/1n4RGlxXpLrTakT1TkypLvLs1dpuBJkaq/edit?usp=sharing&oid=112465961449455869485&rtpof=true&sd=true>

# Grading Policy - Other Policy

- **Lateness**

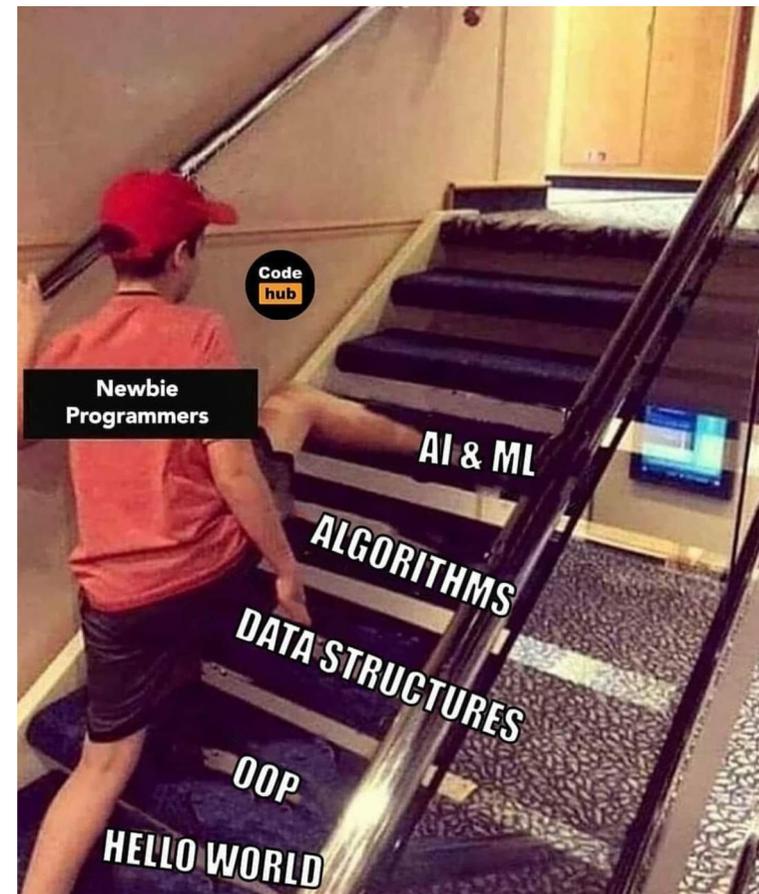
- Cool 遲交
  - 以最後一次繳交之時間為準
  - 一天: 以小時為單位，線性遞減至七折
  - 兩天: 以小時為單位，從七折線性遞減零分
- 不接受程式 or 報告單獨遲交
- 有特殊原因，請找助教說明。

- **Runtime Error**

- 若助教無法順利執行程式，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。
  - 若有大幅更動程式邏輯，請務必和助教說明清楚。
-

# 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\_hw3\_code]{your name}\_{title}  
[ML23\_hw3\_math]{your name}\_{title}



# TA

- [ntueemta2023@gmail.com](mailto:ntueemta2023@gmail.com) or [b08209023@ntu.edu.tw](mailto: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)
- TA hour:
  - 週二早上1000~1200 @ (待定)
  - 跟助教另外約時間

# 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



# 相關連結

- Kaggle:<https://www.kaggle.com/competitions/ml2023-fall-hw3>
  - Colab:[https://colab.research.google.com/drive/1\\_YkFH7Ajkd6zlee3feHcmOnwW0f98yDx?usp=sharing](https://colab.research.google.com/drive/1_YkFH7Ajkd6zlee3feHcmOnwW0f98yDx?usp=sharing)
  - Report Template  
<https://docs.google.com/document/d/1n4RGlxXpLrTakT1TkypLvLs1dpuBJkaq/edit?usp=sharing&oid=112465961449455869485&rtpof=true&sd=true>
  - Math problem:
    - <https://ntueemlta2023.github.io/homeworks/hw3/ml-2023fall-hw3-math.pdf>
-