

---

# HW6 - Adversarial Attack

ML TAs

[ntu-ml-2020spring-ta@googlegroups.com](mailto:ntu-ml-2020spring-ta@googlegroups.com)

---

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# Task Description - Todo <sup>1/3</sup>

1. Fast Gradient Sign Method (FGSM)
  1. Choose any proxy network to attack the black box
  2. Implement non-targeted FGSM from scratch
  3. Tune your parameter  $\epsilon$
  4. Submit as `hw6_fgsm.sh`
2. Any methods you like to attack the model
  1. Implement any methods you prefer from scratch
  2. Beat the best performance in `hw6_fgsm.sh`
  3. Beat your classmates with lower L-inf. Norm and higher success rate
  4. Submit as `hw6_best.sh`

# Task Description - Fast Gradient Sign Method <sup>2/3</sup>

- Fast Gradient Sign Method (FGSM)

$$x^{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x, y_{true}))$$

where

$x$  is the input (clean) image,

$x^{adv}$  is the perturbed adversarial image,

$J$  is the classification loss function,

$y_{true}$  is true label for the input  $x$ .

Explaining and Harnessing Adversarial Examples: <https://arxiv.org/pdf/1412.6572.pdf>

Adversarial Machine Learning at Scale: <https://arxiv.org/pdf/1611.01236.pdf>

# Task Description - Evaluation Metrics <sup>3/3</sup>

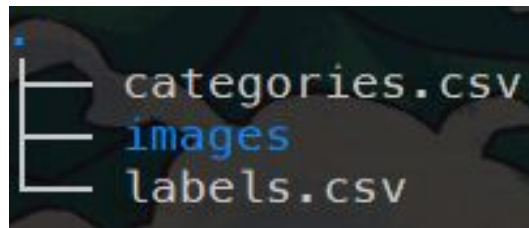
- Average L-inf. norm between all input images and adversarial images
- Success rate of your attack
- Priority: Success rate > Ave. L-inf. norm

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# Data Format <sup>1/2</sup>

- Download link: [link](#)
- Images:
  - 200 張 224 \* 224 RGB 影像
    - 000.png - 199.png
  - categories.csv: 總共 1000 categories (0 - 999)
  - labels.csv: 每張影像的 info



```
1 OriginId,ImgId,OriginImgUrl,TrueLabel,OriginalLandingURL,License,Author,AuthorProfileURL
2 0c7ac4a8c9dfa802,0,https://c1.staticflickr.com/9/8540/28821627444_0524012bdd_o.jpg,305,https://www.flickr.com/
  photos/gails_pictures/28821627444,https://creativecommons.org/licenses/
  by/2.0/,gailhampshire,https://www.flickr.com/people/gails_pictures/
3 f43fbfe8a9ea876c,1,https://c1.staticflickr.com/9/8066/28892033183_6f675dcc03_o.jpg,883,https://www.flickr.com/
  photos/barty/28892033183,https://creativecommons.org/licenses/by/2.0/,Barry Badcock,https://www.flickr.com/people/
  barty/
4 4fc263d35a3ad3ee,2,https://c1.staticflickr.com/8/7378/27465801596_a9dd11e5e2_o.jpg,243,https://www.flickr.com/
  photos/foxcroftacademy/27465801596,https://creativecommons.org/licenses/by/2.0/,Foxcroft
  Academy,https://www.flickr.com/people/foxcroftacademy/
5 cc13c2bc5cdd1f44,3,https://c1.staticflickr.com/9/8864/28546467522_56229f2bef_o.jpg,559,https://www.flickr.com/
  photos/o_0/28546467522/,https://creativecommons.org/licenses/by/2.0/,Guilhem Vellut,https://www.flickr.com/people/
  o_0/
```



# Data Format <sub>2/2</sub>

- 本次作業可以使用其他現成 pretrained 模型進行攻擊
- Black box 可能的模型如下：
  - VGG-16
  - VGG-19
  - ResNet-50
  - ResNet-101
  - DenseNet-121
  - DenseNet-169
- Model reference:
  - Keras: <https://keras.io/applications/>
  - PyTorch: <https://pytorch.org/docs/stable/torchvision/models.html>
  - Tensorflow: <https://github.com/tensorflow/models/tree/master/research/slim>

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# HW website - JudgeBoi <sup>1/2</sup>

- Link: [JudgeBoi](#)
- 個人進行, 不需組隊
- 以繳交作業的 GitHub 帳號登入, 嚴禁多重帳號
- 因為 Kaggle 沒有本作業要用的 evaluation metric, 所以本作業使用本課程自行研發的評估平台
- 霸脫不要亂搞QQ, 有任何問題請先回報給 TA

# HW website - JudgeBoi <sup>2/2</sup>

- 請將 200 張生成的 images 壓縮 .tgz 檔格式上傳
- Note: 解壓縮後不能包含資料夾
- E.g.,
  - `cd <your output image file>`
  - `tar -zcvf <compressed file> <all images>`
  - Ex. `tar -zcvf ../images.tgz *.png`
- 每日上傳上限 5 次 (更新時間為每天 00:00:00)
- 結束前請在 My submission 內選擇一個結果當作最後的結果, 若沒勾選會自動選擇最新上傳的

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# Submission Format - GitHub <sup>1/2</sup>

- GitHub 中 hw6-<account> 必須包含(注意格式):
  - report.pdf
  - hw6\_fgsm.sh
  - hw6\_best.sh
  - other files (e.g., attack.py)
  - 請不要上傳 dataset 和 output img
  - 如要上傳 model file, 請上傳至雲端(Dropbox, ...), 並在 script 中寫好下載的指令

# Submission Format - Bash Usage <sup>2/2</sup>

- 助教會以下指令執行程式，程式執行時間最多不能超過 **300 秒**
  - `timeout 300 bash hw6_fgsm.sh <input img dir> <output img dir>`
  - `timeout 300 bash hw6_best.sh <input img dir> <output img dir>`
  - input img directory: 為 200 張 original input img 之資料夾
  - output img directory: 為 200 張 adversarial output img 之資料夾
  - Ex. `bash hw6_fgsm.sh ./images ./output`
- Output file 中的 img 格式如同 input img
  - E.g., `./output/000.png`
- 路徑請勿寫死以免導致程式無法執行

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ



# Regulations

- Python 版本及套件規定請參考[期初公告](#)
  - 建議不要使用 Keras, 它的 pretrained model 在本次作業中不是好的 proxy model
  - **不得使用** cleverhans、deepfool、adversarial-robustness-toolbox **以及任何現成套件**。
- 若需使用其它套件, 請儘早寄信至助教信箱詢問, 並闡明原因。

# Outline

- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

# Grading Policy - Evaluation (4% + 1% bonus) <sup>2/7</sup>

- (2%)
  - success rate 高於 simple baseline
  - L-inf.norm 低於 simple baseline
- (2%)
  - success rate 高於 strong baseline
  - L-inf.norm 低於 strong baseline
- Simple baseline
  - Success rate: 0.310
  - L-inf. norm: 20.3450
- Strong baseline
  - Success rate: 0.915
  - L-inf. norm: 5.5
- (1% bonus) L-inf.norm 低於 strong baseline 的 submission 中的前五名

# Grading Policy - Reproduce <sup>3/7</sup>

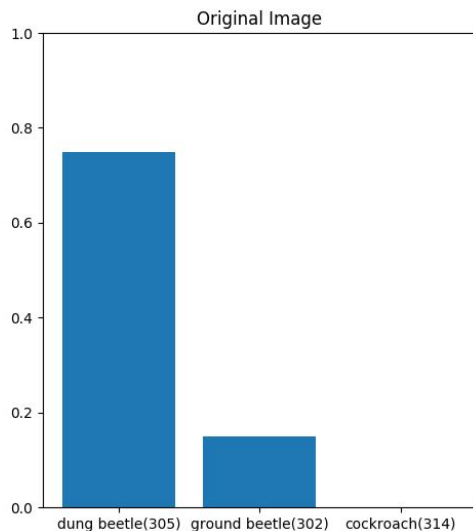
- 請務必隨時保留跑出最佳結果的 code 和結果
- hw6\_best.sh 執行後產生的 image, evaluation metric 需與 leaderboard 上一致, 否則 **evaluation 的成績將不予計分**

# Grading Policy - Report (6%) <sup>4/7</sup>

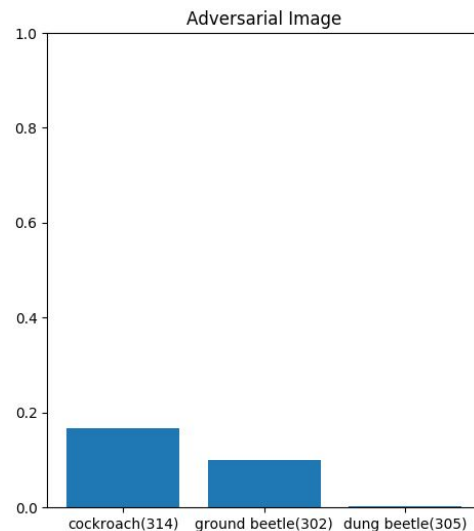
1. (2%) 試說明 hw6\_best.sh 攻擊的方法, 包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何? 如何影響你的結果? 請完整討論。(依內容完整度給分)
2. (1%) 請嘗試不同的 proxy model, 依照你的實作的結果來看, 背後的 black box 最有可能為哪一個模型? 請說明你的觀察和理由。

# Grading Policy - Report (6%) <sup>5/7</sup>

3. (1%) 請以 hw6\_best.sh 的方法, visualize 任意三張圖片攻擊前後的機率圖(分別取前三高的機率)。



Dung beetle 74.85%



Cockroach 16.65%

# Grading Policy - Report (6%) <sup>6/7</sup>

4. (2%) 請將你產生出來的 adversarial img, 以任一種 smoothing 的方式實作被動防禦 (passive defense), 觀察是否有效降低模型的誤判的比例。請說明你的方法, 附上你防禦前後的 success rate, 並簡要說明你的觀察。另外也請討論此防禦對原始圖片會有什麼影響。

Hint: some methods you may use:

- Gaussian filtering: [link](#)
- Median filter: [link](#)
- Bilateral filter: [link](#)
- Others: [link](#)

# Useful Links

- ❖ Data: <https://reurl.cc/vD3Yr1>
- ❖ Colab: <https://reurl.cc/Mv1pnn>
- ❖ 作業網站: <https://reurl.cc/exvR0R>
- ❖ Report: <https://reurl.cc/O17Zlr>
- ❖ 遲交表單: <https://bit.ly/39d2x2m>



# FAQ

- 若有其他問題，請在 FB 社團貼文或寄信至助教信箱，請勿直接私訊助教。
- 助教信箱：[ntu-ml-2020spring-ta@googlegroups.com](mailto:ntu-ml-2020spring-ta@googlegroups.com)
- 並請記得於標題以 [hw6] 註明作業編號。