Machine Learning HW5

MLTAs
ntumlta2019@gmail.com

Outline

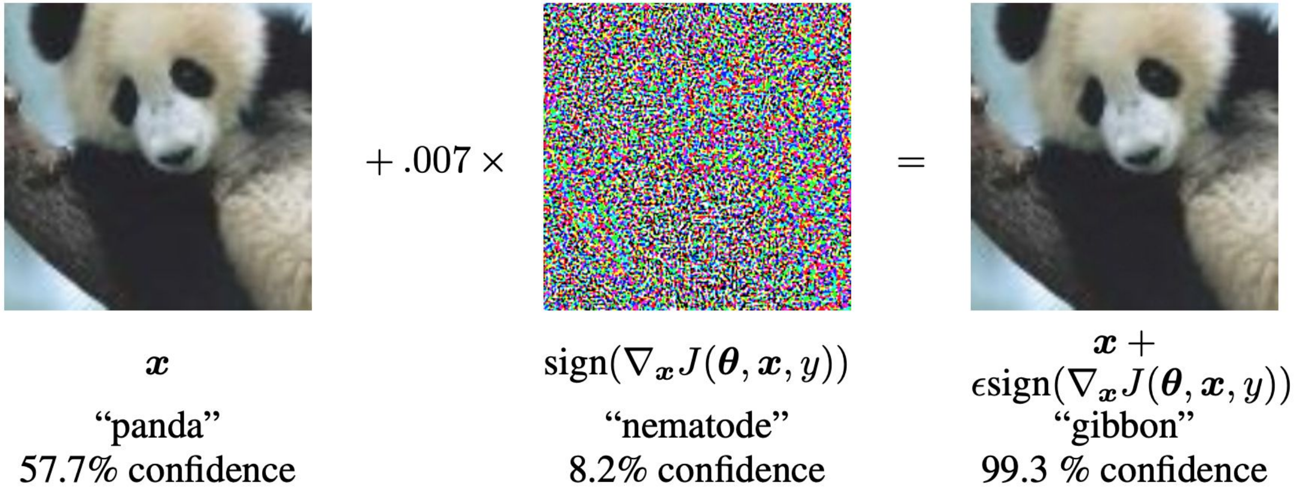
- Task Description
- Data Format
- HW website
- Submission Format (Code, Report)
- Regulations
- Grading Policy & Deadline
- FAQ

Outline

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

Task Description 1/5

- Goal：Non-targeted black box attack by using proxy network

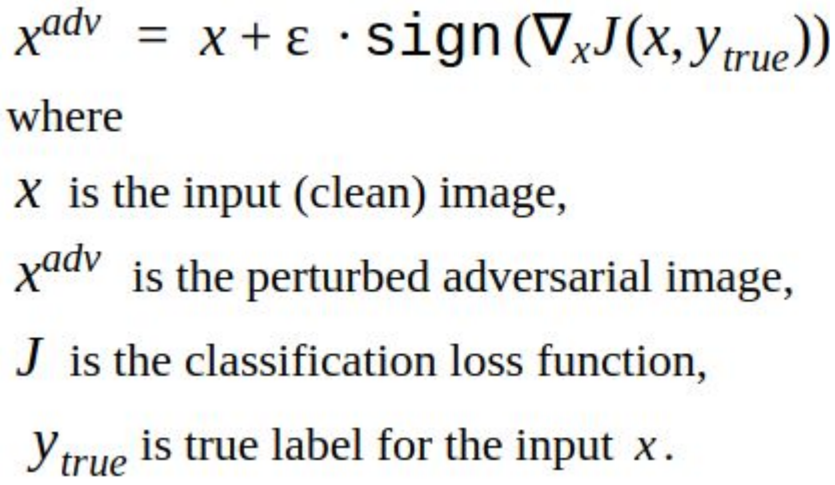

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Task Description - Todo 2/5

1. Fast Gradient Sign Method (FGSM)
    1.1. Choose any proxy network to attack the black box
    1.2. Implement non-targeted FGSM from scratch
    1.3. Tune your parameter ε
    1.4. Submit as hw5_fgsm.sh
2. Any methods you like to attack the model
    2.1. Implement any methods you prefer from scratch
    2.2. Beat the best performance in hw5_fgsm.sh
    2.3. Beat your classmates with lower L-inf. norm and higher success rate
    2.4. Submit as hw5_best.sh

Task Description - Fast Gradient Sign Method 3/5

- Fast Gradient Sign Method (FGSM)



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

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 4/5

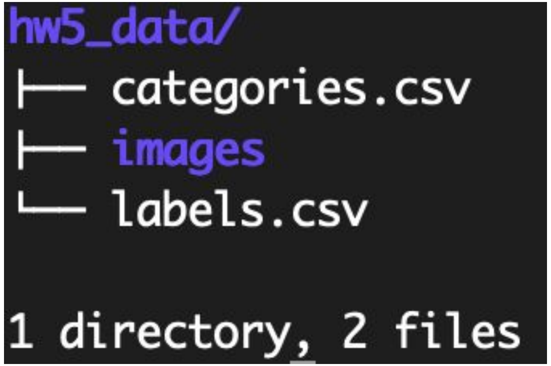
- Average L-inf. norm between all input images and adversarial images
- Success rate of your attack
- Priority: Success rate > Ave. L-inf. norm

Outline

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

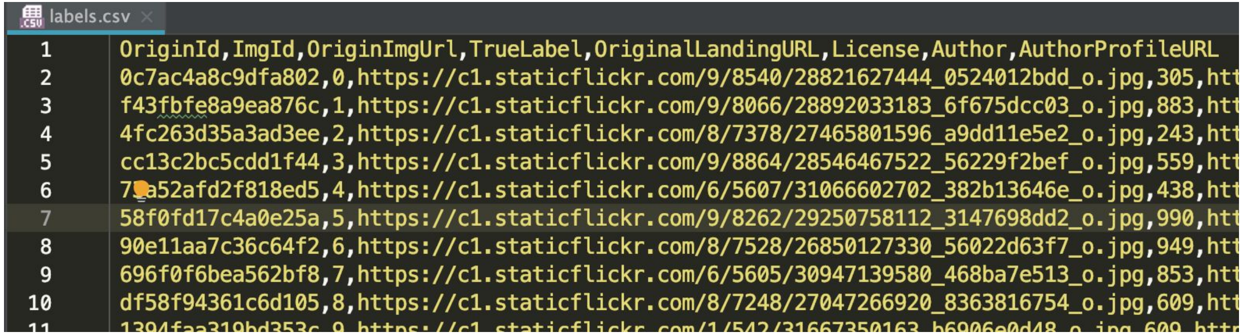
Data Format 1/2

- 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



```
hw5_data/
├── categories.csv
├── images
└── labels.csv

1 directory, 2 files
```



```
OriginId,ImgId,OriginImgUrl,TrueLabel,OriginalLandingURL,License,Author,AuthorProfileURL
0c7ac4a8c9dfa802,0,https://c1.staticflickr.com/9/8540/28821627444_0524012bdd_o.jpg,305,htt
f43fbfe8a9ea876c,1,https://c1.staticflickr.com/9/8066/28892033183_6f675dcc03_o.jpg,883,htt
4fc263d35a3ad3ee,2,https://c1.staticflickr.com/8/7378/27465801596_a9dd11e5e2_o.jpg,243,htt
cc13c2bc5cdd1f44,3,https://c1.staticflickr.com/9/8864/28546467522_56229f2bef_o.jpg,559,htt
7[illegible]a52afd2f818ed5,4,https://c1.staticflickr.com/6/5607/31066602702_382b13646e_o.jpg,438,htt
58f0fd17c4a0e25a,5,https://c1.staticflickr.com/9/8262/29250758112_3147698dd2_o.jpg,990,htt
90e11aa7c36c64f2,6,https://c1.staticflickr.com/8/7528/26850127330_56022d63f7_o.jpg,949,htt
696f0f6bea562bf8,7,https://c1.staticflickr.com/6/5605/30947139580_468ba7e513_o.jpg,853,htt
df58f94361c6d105,8,https://c1.staticflickr.com/8/7248/27047266920_8363816754_o.jpg,609,htt
1394faa319bd353c,9,https://c1.staticflickr.com/1/542/31667350163_b6906e0d48_o.jpg,609,http
```

Data Format 2/2

- 本次作業可以使用其他現成 pretrain 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 1/2

- Link: JudgeBoi beta 0.1.0
- 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
- 霸脫不要亂搞 TA 架設的網頁QQ，有任何問題請先回報給 TA

HW website - JudgeBoi 2/2

- 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
- Ex.
    - 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 1/2

- Github 中 ML2019SPRING/hw5 必須包含（注意格式）:
  - report.pdf
  - hw5_fgsm.sh
  - hw5_best.sh
  - other files (ex. 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 2/2

- TA 會以下指令執行程式
  - bash hw5_fgsm.sh <input img dir> <output img dir>
  - bash hw5_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 hw5_fgsm.sh ./images ./output
- Output file 中的 img 格式如同 input img
  - Ex. ./output/000.png, ./output/001.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 1/1

- Only Python3.6 is available!!!
- 開放使用的 Packages:
  - NumPy >= 1.14
  - Keras == 2.2.4 (Keras_Applications == 1.0.7)
  - PyTorch == 1.0.1
  - Tensorflow == 1.12.0
  - SciPy == 1.2.1, Pillow == 5.4.1, Scikit-Image == 0.14.2 (04/08 update)
  - Pandas >= 0.24.1
  - Scikit-learn == 0.20.0
  - python standard library (os, sys, …)
  - 不得使用之套件 : 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 - Deadline 1/9

- Early baseline deadline: 2019/04/11 11:59:59 (GMT+8)
- JudgeBoi deadline: 2019/04/25 11:59:59 (GMT+8)
- Github、Report deadline: 2019/04/26 23:59:59 (GMT+8)
- 助教會在 deadline 一到就 clone 所有的程式，並且不再重新 clone 任何檔案

Grading Policy - Evaluation (5% + Bonus 1%) 2/9

- (1%) hw5_fgsm.sh implementation
- (1%) Early Baseline: 2019/04/11 11:59:59 (GMT+8) 前皆通過 simple baseline
- (3%) Baseline 成績如下表
- (Bonus 1%) 綜合成績前五名(結束後由TA公佈)且於課堂時間上台分享

success rate L-inf. norm	低於 simple baseline	介於 simple baseline 和 strong baseline	高於 strong baseline
低於 simple baseline	0	0	0
介於 simple baseline 和 strong baseline	0	1	2
高於 strong baseline	0	2	3

Grading Policy - Evaluation (5% + Bonus 1%) 3/9

- 03/28 Simple baseline release
- 04/11 Strong baseline release
- 04/18 Strong baseline modified
- Simple baseline
  - Success rate: 0.305
  - L-inf. norm: 23.455
- Strong baseline
  - Success rate: 0.895
  - L-inf. norm: 5.675

Grading Policy - Reproduce 4/9

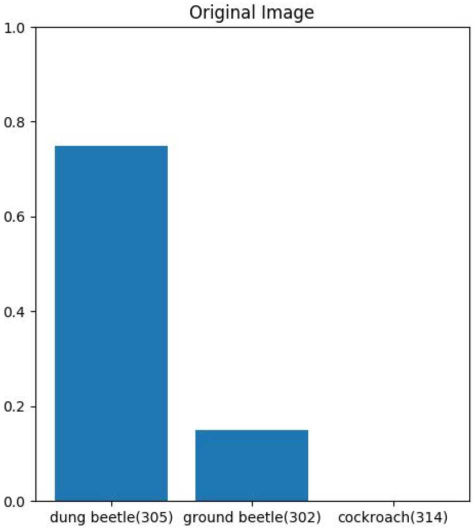
- 請務必隨時保留跑出最佳結果的 code 和結果
- hw5_best.sh 執行後產生的 img，evaluation metric 需與 leaderboard 上一致，否則 evaluation 的成績將不予計分

Grading Policy - Report (5%) 5/9

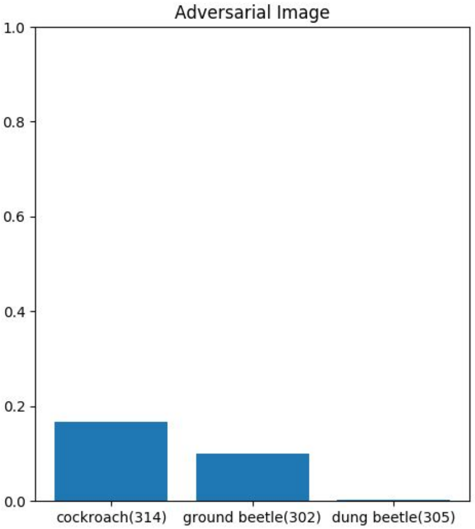
1. (1%) 試說明 hw5_best.sh 攻擊的方法，包括使用的 proxy model、方法、參數等。此方法和 FGSM 的差異為何？如何影響你的結果？請完整討論。(依內容完整度給分)
2. (1%) 請列出 hw5_fgsm.sh 和 hw5_best.sh 的結果 (使用的 proxy model、success rate、L-inf. norm)。
3. (1%) 請嘗試不同的 proxy model，依照你的實作的結果來看，背後的 black box 最有可能為哪一個模型？請說明你的觀察和理由。

Grading Policy - Report (5%) 6/9

4. (1%) 請以 hw5_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 (5%) 7/9

5. (1%) 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。**

   Some methods you can use:
   Gaussian filtering: link
   Median filter: link
   Bilateral filter: link
   Others: link

Grading Policy - Report (5%) 8/9

- Report template: link
- 請利用 template 撰寫 report，回答 report 的問題

Grading Policy - Other Policy 9/9

- 不接受 code 和 report 分開繳交
- Script 錯誤，作業以 0 分計
- 相關 format 錯誤，在助教公告的時間內修改程式，evaluation 部分成績 * 0.7，不予更改非 format 錯誤的程式碼
- Github 遲交，每遲交一天作業總成績 * 0.7，不得遲交超過一天，超過一天之後作業以 0 分計算，有特殊原因請先找助教。
- Github 遲交:
  - 遲交表單: link
  - 請先上傳好完整的作業至github 後再行填寫，助教會依填寫表單的時間手動clone下檔案

FAQ

- 若有其他相關問題，請留言在FB社團的討論或寄信至助教信箱，請勿直接私訊助教。
- 助教信箱：ntumlta2019@gmail.com