Machine Learning HW10
Adversarial Attack

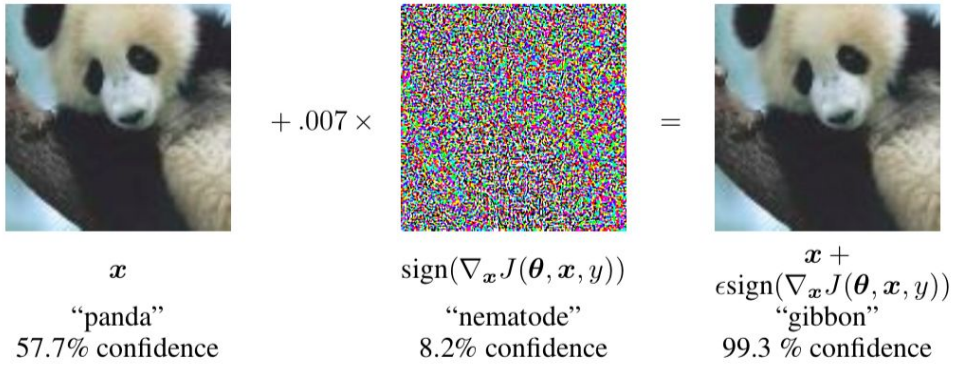
ML TAs
ntu-ml-2021spring-ta@googlegroups.com

Outline

- Task Description
- Data Format
- Grading
- Submission
- Regulations
- Contact

Task Description - Prerequisite 1/6

- Those are **methodologies** which you should be familiar with first
  - Attack objective: Non-targeted attack
  - Attack constraint: L-infinity norm and Parameter ε
  - Attack algorithm: FGSM attack
  - Attack schema: Black box attack (perform attack on proxy network)
  - Benign images vs Adversarial images



Task Description - TODO 2/6

1. Fast Gradient Sign Method (FGSM)
    1. Choose any proxy network to attack the **black box**
    2. Implement **non-targeted FGSM** from scratch
2. Any methods you like to attack the model
    1. Implement any methods you prefer from scratch
    2. Iterative Fast Gradient Sign Method (I-FGSM) --- **medium baseline**
    3. Model ensemble attack --- **strong/boss baseline**

Task Description - FGSM 3/6

- Fast Gradient Sign Method (FGSM)

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y; \theta))$$

x: benign image

x': (perturbed) adversarial image

L: classification loss function

y: ground truth of x

Task Description - I-FGSM 4/6

- Iterative Fast Gradient Sign Method (I-FGSM)

$$x'_0 = x$$

$$x'_{n+1} = \mathrm{Clip}_x^{\epsilon}(x'_n + \alpha \cdot \mathrm{sign}(\nabla_x L(x, y; \theta))$$

α: step size

n: iteration number

$\mathrm{Clip}_x^{\epsilon}$: adversarial images should be within the ϵ-ball (ϵ-limitation) to benign images

Task Description - Ensemble Attack 5/6

- Choose a list of proxy models
- Choose an attack algorithm (FGSM, I-FGSM, and so on)
- Attack multiple proxy models at the same time
- Delving into Transferable Adversarial Examples and Black-box Attacks
- Query-Free Adversarial Transfer via Undertrained Surrogates

Task Description - Evaluation Metrics 6/6

- Parameter ε is fixed as 8
- Distance measurement: L-inf. norm
- Model Accuracy is the only evaluation metrics

benign

adversarial (\eps = 8)

adversarial (\eps = 16)

Data Format 1/2

- Download link: link
- Images:
  - CIFAR-10 images
  - (32 * 32 RGB images) * **200**
    - airplane/airplane1.png, ..., airplane/airplane20.png
    - ...
    - truck/truck1.png, ..., truck/truck20.png
  - 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck)
  - 20 images for each class

Data Format 2/2

- In this homework, we can perform attack on pretrained models
- Pytorchcv provides multiple models pretrained on CIFAR-10
- A model list is provided here

Grading - Baseline Guide 1/3

- **Execution time: about 10 minutes**
- Simple baseline (public: 0.650)
  - Hints: FGSM (sample code)
- Medium baseline (public: 0.380)
  - Hints: Iterative-FGSM
- Strong baseline (public: 0.180)
  - Hints: Ensemble Attack, paper
  - TODO: build ensemble network and perform attack
- Boss baseline (public: 0.050)
  - Hints: Ensemble Attack with some techniques or luck, paper
  - TODO: trial-and-error to ensemble attack on different sets of models

Grading - Baselines 2/3

Simple baseline (public)	+1 pt (sample code)
Simple baseline (private)	+1 pt (sample code)
Medium baseline (public)	+1 pt
Medium baseline (private)	+1 pt
Strong baseline (public)	+0.5 pt
Strong baseline (private)	+0.5 pt
Boss baseline (public)	+0.5 pt
Boss baseline (private)	+0.5 pt
Upload code to NTU COOL	+4 pts

Total: **10** pts

Grading - Bonus 3/3

- **If you got 10 points**, we make your code **public** to the whole class.

- In this case, if you also submit **a PDF report briefly describing your methods** (<100 words in English), you get a bonus of **0.5 pt**.
  (your report will also be available to all students)

- Report template

Submission - Deadlines 1/6

- JudgeBoi

2021/05/28 23:59 (UTC+8)

- Code Submission (NTU COOL)

2021/05/30 23:59 (UTC+8)

No late submission!
Submit early!

Submission - JudgeBoi 2/6

- Parameter ε is fixed as 8, **any submissions exceeding this constraint will cause a submission error**
- The compressing code is provided in the sample code
- To create such a compressed file by yourself, follow steps below
  - Generate 200 adversarial images
  - Name each image **<class><id>.png**
  - Put each image in corresponding **<class> directory**
  - Use tar to **compress the <class> directories** with .tgz as extension
  - E.g.,
    - cd <output directory> (cd fgsm)
    - tar zcvf <compressed file> <the <class> directories> (tar zcvf ../fgsm.tgz *)

Submission - JudgeBoi 3/6

- **5 submission quota** per day, reset at midnight
- Please **select the final submission** before deadline, or we will use the private score of the **submission with the highest public score**
- Users not in whitelist will have no quota
- Only ***.tgz** file is allowed, file size should be **smaller than 2MB**
- The countdown timer on the homepage is for reference only
- If you cannot access the website temporarily, please wait patiently
- Please do not attempt to attack JudgeBoi, thank you
- Every Wednesday and Saturday from 0:00 to 3:00 is our system maintenance time

Submission - JudgeBoi 4/6

- The JudgeBoi server cannot serve too many submissions at the same time
- Under normal circumstances, JudgeBoi will complete the evaluation **within one minute**
- If pending conditions are encountered, it may be longer
- Please wait patiently after you submit
- However, if you have **waited more than two minutes** for the progress bar to finish, please **refresh the page and try to upload again**
- Please **DO NOT** upload at the last minute; no one knows if you can upload successfully

Submission - NTU COOL 5/6

- **NTU COOL** (4pts)
  - Compress your code and report into

    <student ID>_hwX.zip

    *** e.g. b06901020_hw10.zip**

    *** X is the homework number**

  - We can only see your last submission.
  - Do not submit your model or dataset.
  - If your code is not reasonable, your semester grade x 0.9.

Submission - NTU COOL 6/6

- Your .zip file should include only
  - **Code**: either .py or .ipynb
  - **Report**: .pdf (only for those who got 10 points)

- Example:

Regulations 1/2

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference. (*)
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- **Do NOT search or use additional data.**
- **You are allowed to use pre-trained models on any image datasets.**
- Your **final grade x 0.9** if you violate any of the above rules.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

(*) Academic Ethics Guidelines for Researchers by the Ministry of Science and Technology

Regulations 2/2

- Do NOT share your **ensemble model lists** or **attack algorithms** with your classmates.
- TAs will check the adversarial images you generate.

(*) Academic Ethics Guidelines for Researchers by the Ministry of Science and Technology

If any questions, you can ask us via...

- NTU COOL (recommended)
  - https://cool.ntu.edu.tw/courses/4793
- Email
  - ntu-ml-2021spring-ta@googlegroups.com
  - The title should begin with “[hwX]” (X is the homework number)
- TA hour
  - Each Monday 19:00~21:00 @Room 101, EE2 (電機二館101)
  - Each Friday 13:30~14:20 Before Class @Lecture Hall (綜合大講堂)
  - Each Friday during class