Machine Learning HW10
Adversarial Attack

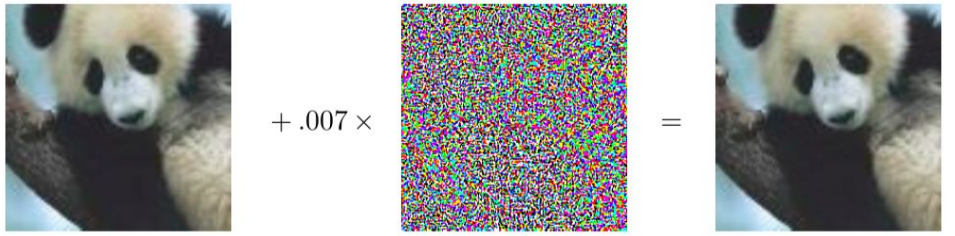
ML TAs
mlta-2022-spring@googlegroups.com

Outline

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

Task Description - Prerequisite

- 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/I-FGSM
  - Attack schema: Black box attack (perform attack on proxy network)

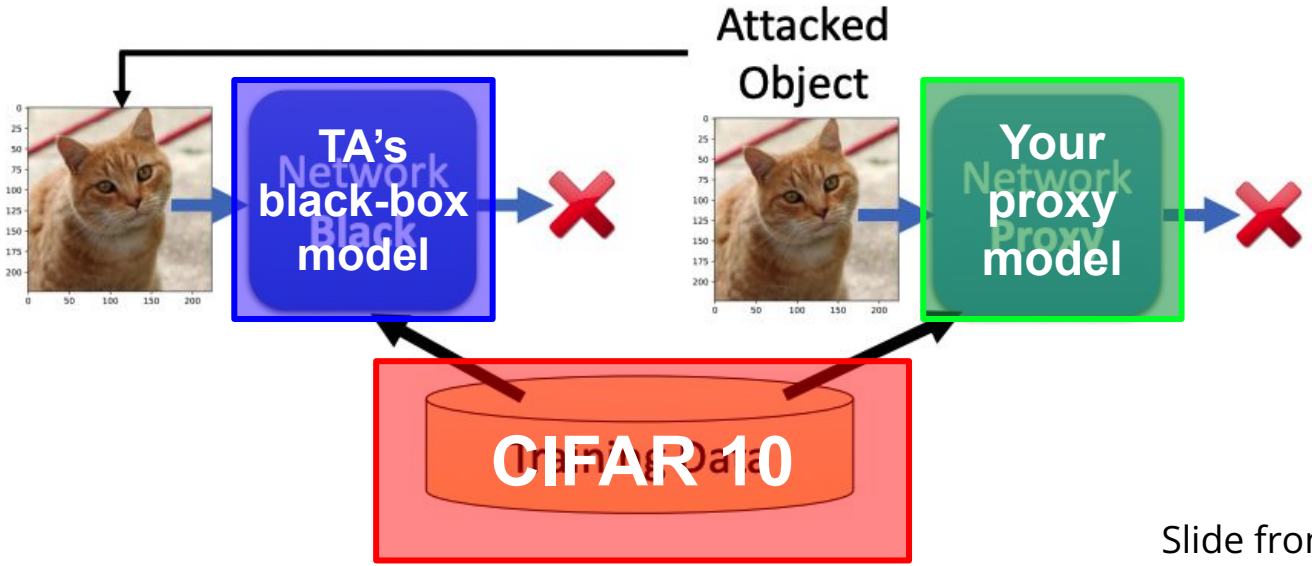


Task Description - Black-box attack

If you have the training data of the target network

Train a proxy network yourself

Using the proxy network to generate attacked objects



Slide from ML2021 attack

Task Description - TODO

1. Choose any proxy network to attack the **black box model from TA**
2. Implement **non-targeted adversarial attack method**
    a. FGSM
    b. I-FGSM
    c. MI-FGSM
3. Increase attack transferability by Diverse input (DIM)
4. Attack more than one proxy model - **Ensemble attack**

FGSM

- Fast Gradient Sign Method (FGSM)

$$\arg\max_{\boldsymbol{x}^{adv}} J(\boldsymbol{x}^{adv}, y), \quad \text{s.t. } \|\boldsymbol{x}^{adv} - \boldsymbol{x}^{real}\|_\infty \leq \epsilon.$$

$$\boldsymbol{x}^{adv} = \boldsymbol{x}^{real} + \epsilon \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^{real}, y))$$

I-FGSM

- Iterative Fast Gradient Sign Method (I-FGSM)

$$\boldsymbol{x}_0^{adv} = \boldsymbol{x}^{real}$$

for t = 1 to num_iter:

step size

$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y))$$

$$\mathrm{clip}\ \boldsymbol{x}_t^{adv}$$

you can define num_iter & step size by your own

(Hint) MI-FGSM

[paper] Boosting Adversarial Attacks with Momentum

Use **momentum** to stabilize update directions and escape from poor local maxima

for t = 1 to num_iter:

$$\boldsymbol{g}_{t+1} = \mu \cdot \boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_t^{adv}, y)\|_1},$$

decay factor μ

$$\boldsymbol{x}_{t+1}^{adv} = \boldsymbol{x}_t^{adv} + \alpha \cdot \text{sign}(\boldsymbol{g}_{t+1}),$$

clip $\boldsymbol{x}_t^{adv}$

Overfitting happens in adversarial attack too ...

- IFGSM greedily perturb the images in the direction of the sign of the loss gradient easily fall into the poor local maxima and overfit to the specific network parameters
- These overfitted adversarial examples rarely transfer to black-box models

How to prevent overfitting on proxy models, increasing the transferability of black-box attack?

Data augmentation!

(Hint) Diverse Input (DIM)

[paper] Improving Transferability of Adversarial Examples with Input Diversity

1. **Random resizing** (resizes the input images to a random size)
2. **Random padding** (pads zeros around the input images in a random manner)

$$T(X_n^{adv}; p) = \begin{cases} T(X_n^{adv}) & \text{with probability } p \\ X_n^{adv} & \text{with probability } 1 - p \end{cases}$$

e.g. DIM + MI-FGSM

$$g_{n+1} = \mu \cdot g_n + \frac{\nabla_X L(T(X_n^{adv}; p), y^{\text{true}}; \theta)}{||\nabla_X L(T(X_n^{adv}; p), y^{\text{true}}; \theta)||_1}$$

(Hint) Ensemble Attack

- Choose a list of proxy models
- Choose an attack algorithm (FGSM, I-FGSM, and so on)
- Attack multiple proxy models at the same time
- **[paper A] Ensemble adversarial attack:**
  Delving into Transferable Adversarial Examples and Black-box Attacks
- **[paper B] How to choose suitable proxy models for black-box attack:**
  Query-Free Adversarial Transfer via Undertrained Surrogates

Evaluation Metrics

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

benign

adversarial (**ε** = 8)

adversarial (**ε** = 16)

Data Format

- 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

Pre-trained model

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

TA's model

- We will use defense method, may include:
  1. Ensemble vanilla models
  2. Some passive defenses
- **Simply guess the exact models that TA used won't give better attack results**

Grading - Baseline Guide 1/3

NOTE:
- All the baselines need **below 20 mins** runtime on colab.
- You can pass all the baselines by simply choosing proxy models from **Pytorchcv**, so choosing the right models is important.
- We encourage you to try other proxy models, **but no performance guarantee**.

- Simple baseline (acc <= 0.70)
  - Hints: FGSM
- Medium baseline (acc <= 0.50)
  - Hints: Ensemble Attack + random few model + IFGSM
- Strong baseline (acc <= 0.30)
  - Hints:
    (1) Ensemble Attack + paper B (pick right models) + IFGSM /
    (2) Ensemble Attack + many models + MIFGSM
- Boss baseline (acc <= 0.15)
  - Hints: Ensemble Attack + paper B (pick right models) + DIM-MIFGSM

Grading - Baselines 2/3

Simple baseline (public)	+0.5 pt
Simple baseline (private)	+0.5 pt
Medium baseline (public)	+0.5 pt
Medium baseline (private)	+0.5 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
Report	+4 pts
Code submission	+2 pts

Total: **10** pts

Grading -- Bonus

If your **ranking in private set is top 3**, you can choose to share a report to NTU COOL and get extra 0.5 pts.

About the report

- Your name and student_ID
- Methods you used in code
- Reference
- in 200 words
- Deadline is same as code submission
- Please upload to NTU COOL's discussion of HW10

Report template

Report questions (4%)

Part 1: Attack

- **[Zh]** 根據你最好的實驗結果，簡述你是如何產生transferable noises, Judge Boi上Accuracy降到多少 (1pt)
- **[En]** Depending on your best experimental results, briefly explain how you generate the transferable noises, and the resulting accuracy on the Judge Boi. (1pt)

Part 2: Defense

[Zh] 當source model為**resnet110_cifar10**(from Pytorchcv), 使用最原始的**fgsm**攻擊在**dog2.png**的圖片。

1. 請問被攻擊後的預測的class是錯誤的嗎？(1pt)
   有個話:變成哪個class?
   沒有的話:則不用作答
2. 實作jpeg compression (**compression rate=70%**) 前處理圖片, 請問prediction class是錯誤的嗎？同第一題作答 (1pt)
3. Jpeg compression為什麼可以抵擋adversarial attack, 讓模型維持高正確率？(1pt)
   a. 圖片壓縮時讓色彩更鮮豔 b. 圖片壓縮時把雜訊減少
   c. 圖片壓縮讓圖片品質下降 d. 圖片壓縮時雜訊反而變大

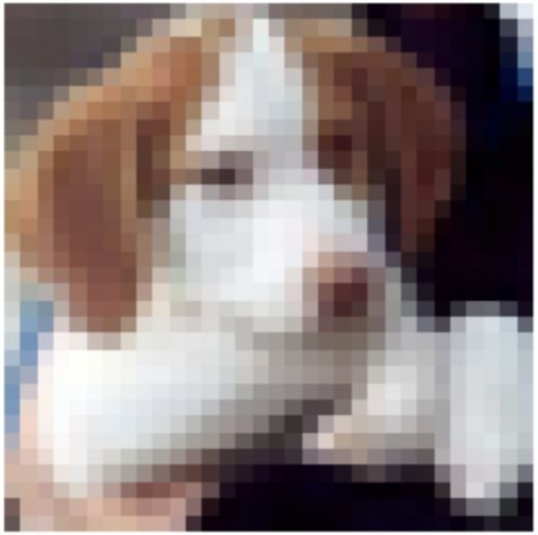
Part 2: Defense

[En] When the source model is **resnet110_cifar10** (from Pytorchcv), adopt the vanilla **fgsm** attack on image "**dog/dog2.png**" in data.zip.

1. Is the predicted class wrong after fgsm attack? If so, change to which class? If not, simply answer no. (1pt)
2. Implement the pre-processing method jpeg compression (**compression rate=70%**). Is the predicted class wrong after defense? Answer the question as the same manner as the first question. (1pt)
3. Why jpeg compression method can defend the adversarial attack, improving the model accuracy? (1pt)
    a. jpeg compression makes images more colorful
    b. jpeg compression reduces the noise level
    c. jpeg compression degrades the image qualities
    d. jpeg compression enlarges the noise level

Example

benign: dog2.png
dog: 99.64%



adversarial image
class: ? probability: ?

JPEG defense
class: ? probability: ?

Link

- Colab
- JudgeBoi
- Report (On Gradescope)

Submission - Deadlines 1/6

- JudgeBoi, Report (GradeScope), Code Submission (NTU COOL)

2022 5/27 23:59 (UTC+8)

No late submission!
Submit early!

Submission - JudgeBoi General Rules

- 5 submission quota per day, reset at midnight.
  - Users not in the whitelist will have no quota.
- The countdown timer on the homepage is for reference only.
- We do limit the number of connections and request rate for each IP.
  - If you cannot access the website temporarily, please wait a moment.
- The system can be very busy as the deadline approaches
  - If this prevents uploads, we do not offer additional opportunities for remediation
- Please do not attempt to attack JudgeBoi.
- Every Friday from 6:00 to 9:00 is our system maintenance time.
- For any JudgeBoi issues, please post on NTUCOOL discussion
  - Discussion Link: https://cool.ntu.edu.tw/courses/11666/discussion_topics/91777

Submission - JudgeBoi HW10-Specific Rules (1/2)

- 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 the following steps
  - 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
  - Steps:
    - cd <output directory> (cd fgsm)
    - tar zcvf <compressed file> <the <class> directories> (tar zcvf ../fgsm.tgz *)

Submission - JudgeBoi HW10-Specific Rules (2/2)

- Only *.tgz file is allowed, file size should be smaller than **2MB**.
- JudgeBoi should complete the evaluation within one minute.
  - You do not need to wait for the progress bar to finish

Submission - NTU COOL 5/6

- **NTU COOL**
  - Compress your code 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.

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/11666
- Email
  - mlta-2022-spring@googlegroups.com
  - The title should begin with “[hw10]”
- TA hour
  - Mandairn: Tuesday, 20:00~21:00
  - English: Friday, 22:00 ~ 23:00