Machine Learning HW10
Adversarial Attack

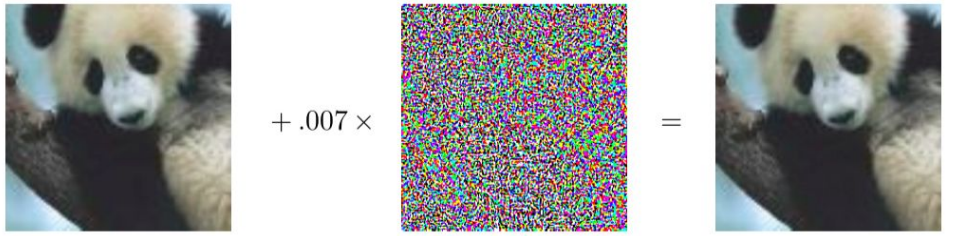
ML TAs
mlta-2023-spring@googlegroups.com

Outline

- Prerequisites
- Task Description
- Data Format
- Report & Grading
- Submission
- Regulations
- Contact

Prerequisites

- These 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)

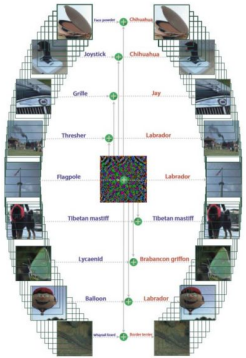


Prerequisites

*The images and videos are all from prof. Lee's previous lectures

If you are not familiar with them, you are strongly recommended to watch:

1. 【機器學習2021】來自人類的惡意攻擊 (Adversarial Attack) (上) – 基本概念
2. 【機器學習2021】來自人類的惡意攻擊 (Adversarial Attack) (下) – 類神經網路能否躲過人類深不見底的惡意？



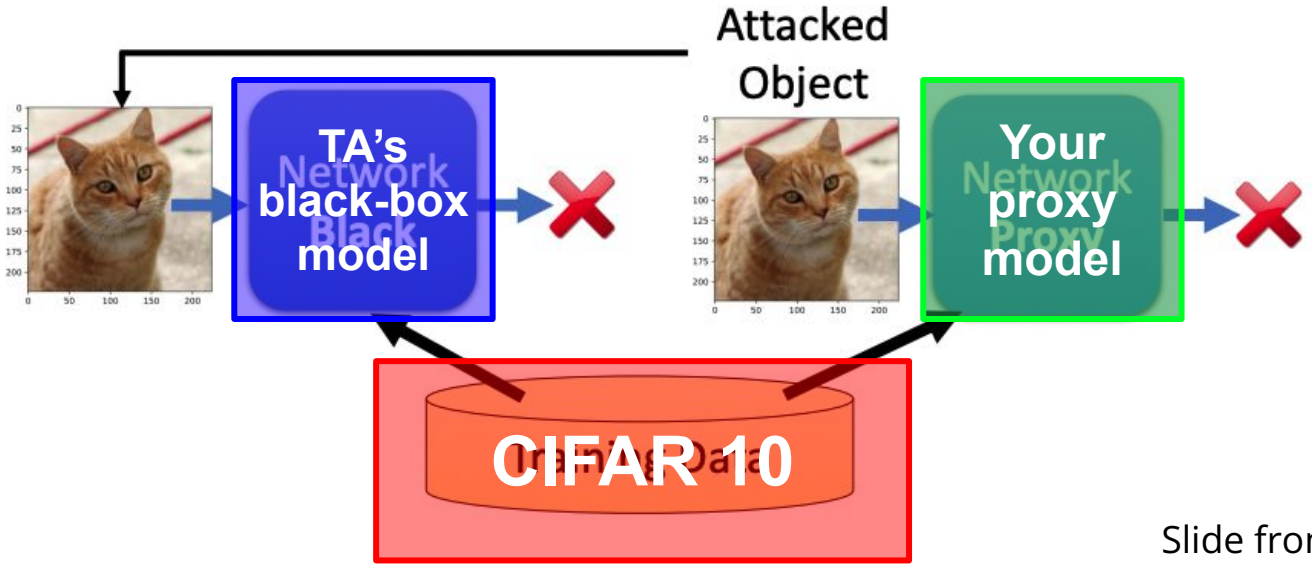
Black Box Attack is also possible!

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
- You can only use the models in the model list with suffix cifar10
- **[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

(Hint) Ensemble Attack - Example

```
'nin_cifar100': nin_cifar100,
'nin_svhn': nin_svhn,

'resnet20_cifar10': resnet20_cifar10,
'resnet20_cifar100': resnet20_cifar100,
'resnet20_svhn': resnet20_svhn,
'resnet56_cifar10': resnet56_cifar10,
'resnet56_cifar100': resnet56_cifar100,
'resnet56_svhn': resnet56_svhn,
'resnet110_cifar10': resnet110_cifar10,
'resnet110_cifar100': resnet110_cifar100,
'resnet110_svhn': resnet110_svhn,
```

Pretrained on cifar10

Pretrained on cifar100 (do not use this)

Evaluation Metrics

- Parameter **ε is fixed as 8**
- Distance measurement: **L-inf. norm**
- Model Accuracy is the only evaluation metrics (the lower, the better!)

benign

adversarial (**ε** = 8)

adversarial (**ε** = 16)

Data Format

- 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

NOTE:
- 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 several different proxy models, **but there's no performance guarantee**.

- Simple baseline (acc <= 0.70)
  - Hints: FGSM
  - Expected running time: 1.5 mins on T4
- Medium baseline (acc <= 0.50)
  - Hints: Ensemble Attack + ensemble random few model + IFGSM
  - Expected running time: 5 mins on T4
- Strong baseline (acc <= 0.25)
  - Hints: Ensemble Attack + ensemble many models + MIFGSM
  - Expected running time: 5.5 mins on T4
- Boss baseline (acc <= 0.10)
  - Hints: Ensemble Attack + ensemble many models + DIM-MIFGSM
  - Expected running time: 8 mins on T4

Report - Part 1: Attack in NLP (text) (3 pts)

In class, we've talked about attacks on "images". The goal of this problem set is to acquaint you with the attack methods in "text" domain. Let's begin with:

1. **(0.5 pt) Please imagine and describe a scenario of adversarial attacks on texts. Why and how this could be adverse and harmful for people?**

While the details of how the attack can be realized is totally optional, your answer should at least include 2 particular points: **(1) A concrete scenario** and **(2) The aftermath of it, the reason it's dangerous, and in what way it will turn out to be harmful.** The grading will based on the rationality of your answer.

Report - Part 1: Attack in NLP (text) (3 pts)

Now, please watch the 2 recommended videos (video1 and video2), and answer the following problems according to the content of the videos:

- **2. (0.5 pt) Why attacks in NLP are more difficult than those in CV?**
- **3. (0.5 pt) From video1, what's the four ingredients of evasion attacks?**
- **4. (1.5 pt) Among TextFooler, PWWS and BERT-Attack, choose an attack method you like and identify the components in each ingredient of the attack you choose and briefly summarize how they work.**

Here's something you need to notice:

1. There are **no partial credits** for problem 2 and 3.
2. For the four ingredients, please use the **same terminologies** as the videos

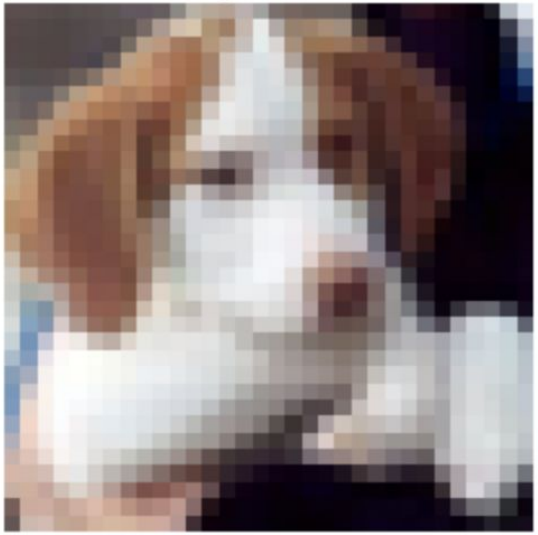
Report - Part 2: Defense (1 pt)

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

1. (0.2 pt) Is the predicted class wrong after fgsm attack? If so, change to which class? If not, simply answer no.
2. (0.3 pt) Implement the pre-processing method jpeg compression (**compression rate=70%**). Is the predicted class wrong after defense? Answer the question in the same manner as the first question.
3. (0.5 pt) Why jpeg compression method can defend the adversarial attack, improving the model accuracy?
    a. JPEG compression makes images more colorful.
    b. JPEGcompression 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: ?

Grading

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
- References if any
- Please upload to NTU COOL’s discussion forum of HW10 before **6/9 23:59**

Link

- Colab
- Judgeboi
- Gradescope

Submission - Deadlines 1/6

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

2023 6/2 23:59 (UTC+8)

No late submission is allowed!

Submission - JudgeBoi General Rules

- 5 submission quota per day, reset at **midnight**.
  - Guest users have no quota.
- 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 submission opportunities.
- Please do not attempt to attack JudgeBoi.
- Every **Saturday** 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/24108/discussion_topics/182915

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>_hw10.zip

    *** e.g. b08202033_hw10.zip**

  - 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.
  - It’s okay if NTU COOL automatically adds an index behind the original file name when you submit the homework more than once.

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 with suffix cifar10 in the provided model list only. It will be considered to be cheating if it's violated.
- 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)
  - HW10 discussion forum
- Email
  - mlta-2023-spring@googlegroups.com
  - The title should begin with “[hw10]”
- TA hour
  - Online (Mandarin & English): 5/19, 5/26, 6/2 19:00~21:00
  - Physical: During and after the classes