Practical No-box Adversarial Attacks against DNNs

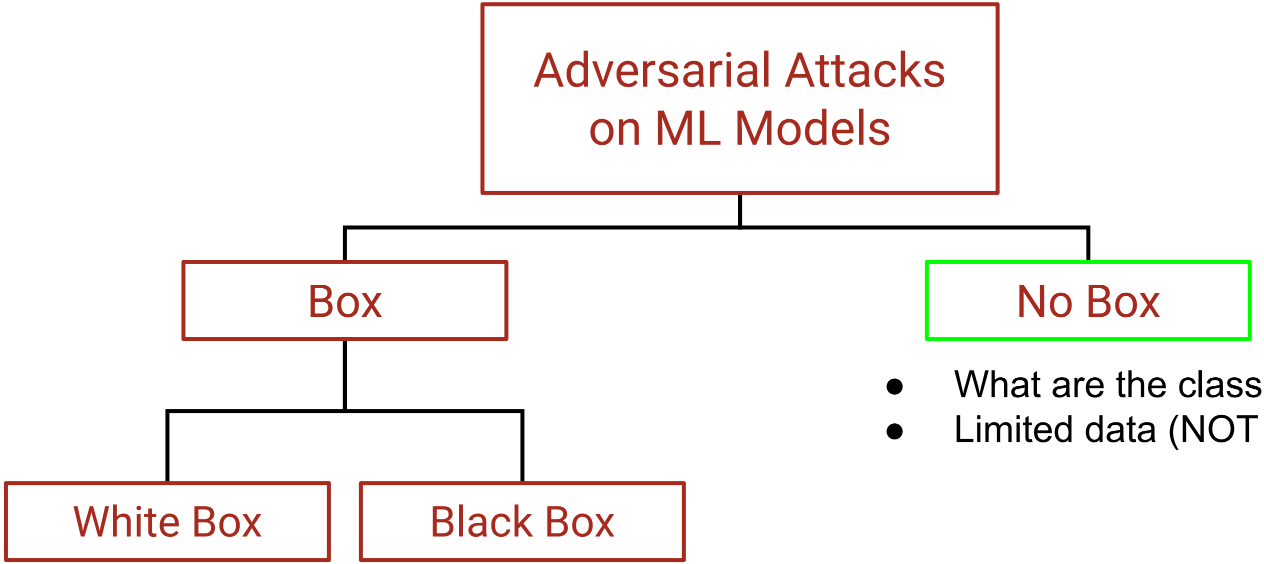
Qizhang Li [*]
ByteDance AI Lab
liqizhang@bytedance.com

Yiwen Guo [†]
ByteDance AI Lab
guoyiwen.ai@bytedance.com

Hao Chen
University of California, Davis
chen@ucdavis.edu

Talking points

- Setting up the Box Attacks - Definition, Meaning, Conventional methods
- Core Idea of the paper - 2 types of new attacks
    - Chaos Reconstruction Attack
    - Prototypical Reconstruction Attack
- Finally, how are attacks are conducted?
- NeurIPS Reviews of the paper - Conflicting ideas, future ideas, discussion



Adversarial Attacks on ML Models

Box

No Box

- What are the classes?
- Limited data (NOT training)

White Box

Black Box

- What are the classes?
- Run Queries on the model
- Model information
- Training data

- What are the classes?
- Run Queries on the model
- Training data

Problem Statement

AIM : Perturbate the sample (image) in some way such that it is misclassified by model X.

What Do we NOT have :

1) Access to model
2) Access to queried input/ouput
3) Access to training data

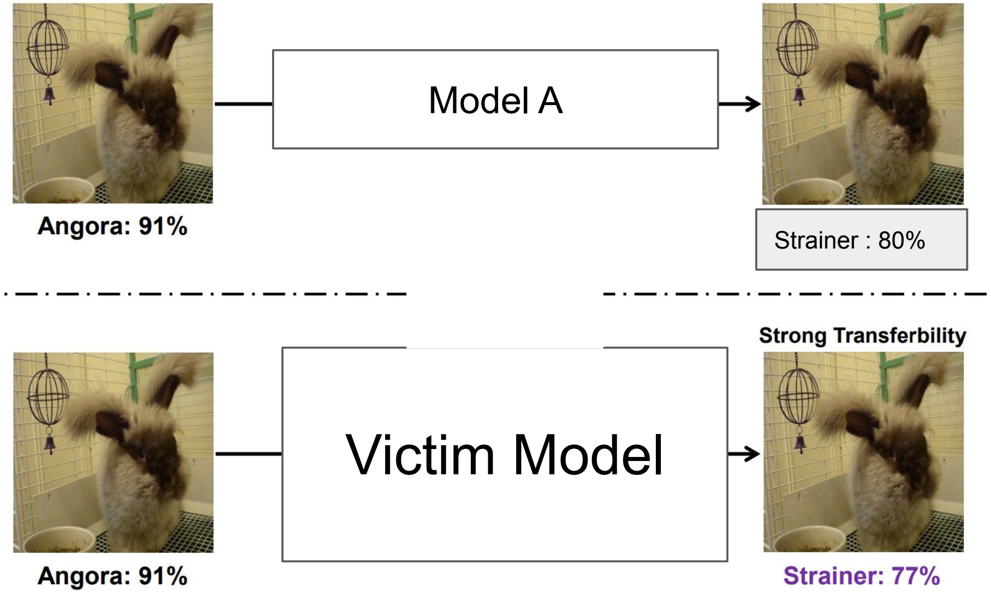
What do we have :

1) Limited samples (10-20)/class - NOT Training data
2) Limited number of classes (< 20)

Core Idea of the paper

Transferability

Adversarial examples crafted on one DNN may fool (i.e., transfer to) other DNNs.



Angora: 91%

Model A

Strainer : 80%

Strong Transferbility

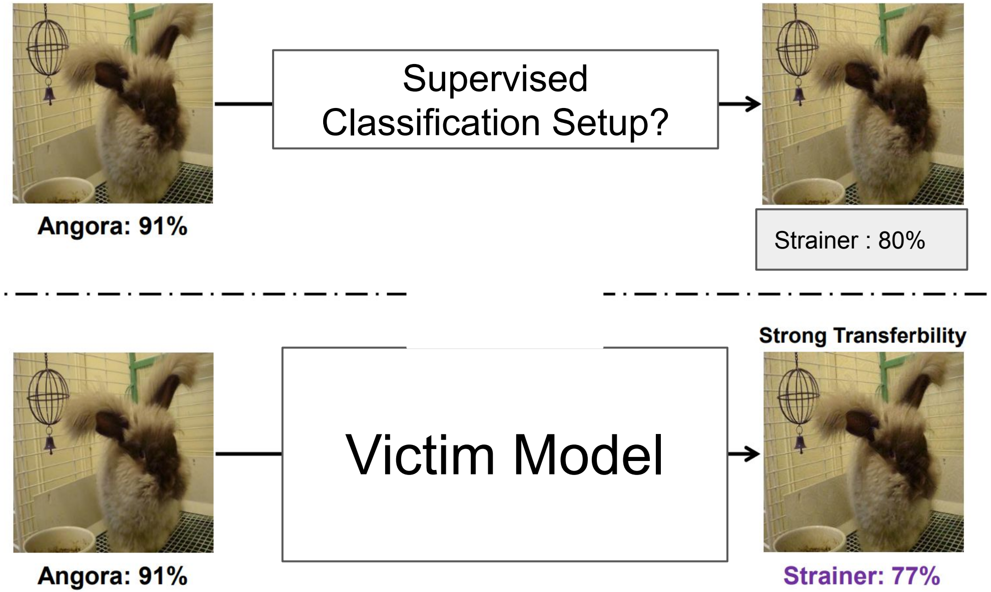
Angora: 91%

Victim Model

Strainer: 77%

Core Idea of the paper

- Little data
- Converges way too fast



Angora: 91%

Supervised Classification Setup?

Strainer : 80%

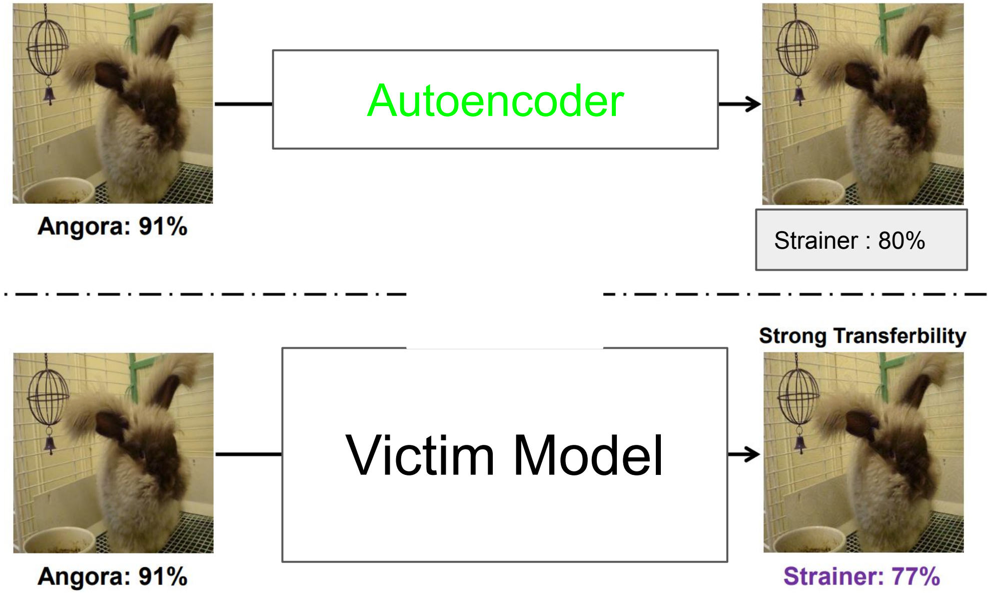
Strong Transferbility

Angora: 91%

Victim Model

Strainer: 77%

Core Idea of the paper



Autoencoder

Angora: 91%

Strainer : 80%

Victim Model

Angora: 91%

Strong Transferbility

Strainer: 77%

Encoder-Decoder Architecture

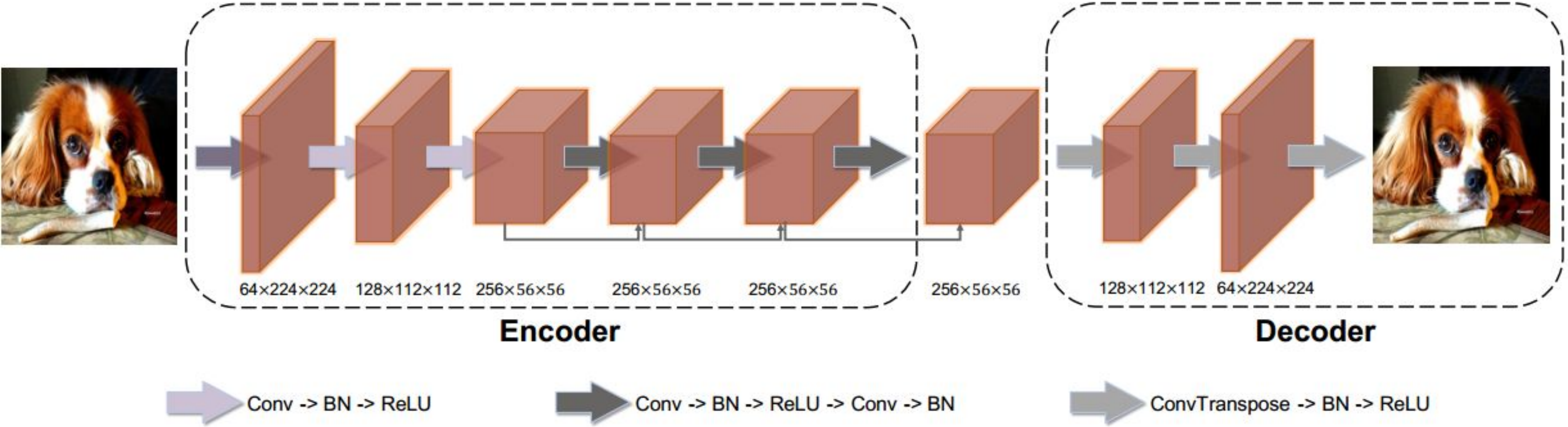
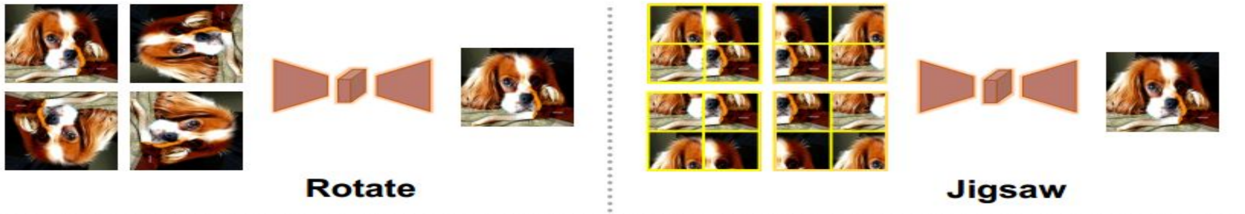


Figure 3: Illustration of the proposed training mechanisms for auto-encoding substitute models for no-box attacks, including two unsupervised mechanisms (*i.e.*, reconstruction from rotation/jigsaw) and a supervised mechanism (*i.e.*, prototypical image reconstruction).

Chaos Reconstruction Attack

- Unsupervised method which adds some pretext information to the model
- Perform tasks like image rotation, jigsaw puzzle configuration etc
- Lacks Discriminative ability

$$L_{\text{rotation/jigsaw}} = \frac{1}{n}\sum_{i=0}^{n-1} \|\mathrm{Dec}(\mathrm{Enc}(T(x_i))) - x_i\|^2$$

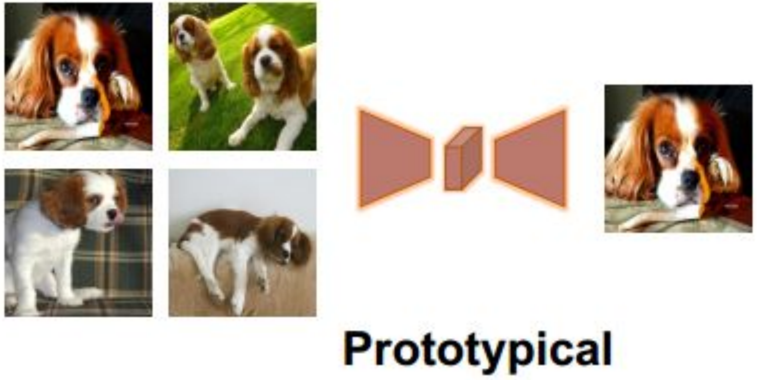


Prototypical image reconstruction

Supervised method

Reconstruct class-specific prototypes, such that direct supervision

Distinguish samples with different labels,



$$L_{\text{prototypical}} = \frac{1}{n}\sum_{i=0}^{n-1}\left((1-y_i)\left\|\text{Dec}(\text{Enc}(x_i)) - x^{(0)}\right\|^2 + y_i\left\|\text{Dec}(\text{Enc}(x_i)) - x^{(1)}\right\|^2\right)$$

How is an attack made?

Perform gradient based attacks like FGSM, I-FGSM using the above models and the below loss function on our auto encoder model.

$$L_{\text{adversarial}} = -\log p(y_i|x_i) \quad \text{where} \quad p(y_i|x_i) = \frac{\exp\left(-\lambda\|\text{Dec}(\text{Enc}(x_i)) - \tilde{x}_i\|^2\right)}{\sum_j \exp\left(-\lambda\|\text{Dec}(\text{Enc}(x_i)) - \tilde{x}_j\|^2\right)}$$

Get directional guides using gradient based methods then apply any intermediate level attack (like ILA).

Discussion

1) Implicit assumption that the model uses some kind of pre-trained convolutional neural network.
2) Real world models tend to be a mix of neural networks and downstream classifiers. Interesting to see how the results change.
3) Perturbation susceptible to human eyes.

https://papers.nips.cc/paper/2020/file/96e07156db854ca7b00b5df21716b0c6-Review.html