$\mathcal{N}$ATTACK: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks

Yandong Li*[1] Lijun Li*[1] Liqiang Wang[1] Tong Zhang[2] Boqing Gong[3]

*Equal Contribution

[1]University of Central Florida [2]Hong Kong University of Science and Technology [3]Google

Adversarial Examples

x

82% puma

Adversarial Noise

$+\delta$

x'

90% book jacker

Popular: Gradient-Based Adversarial Attack

$$x_{t+1} = x_t + \eta sign(\boxed{\nabla_x L(x_t, y)})$$

Gradient of classifier output according to x.

White-box:

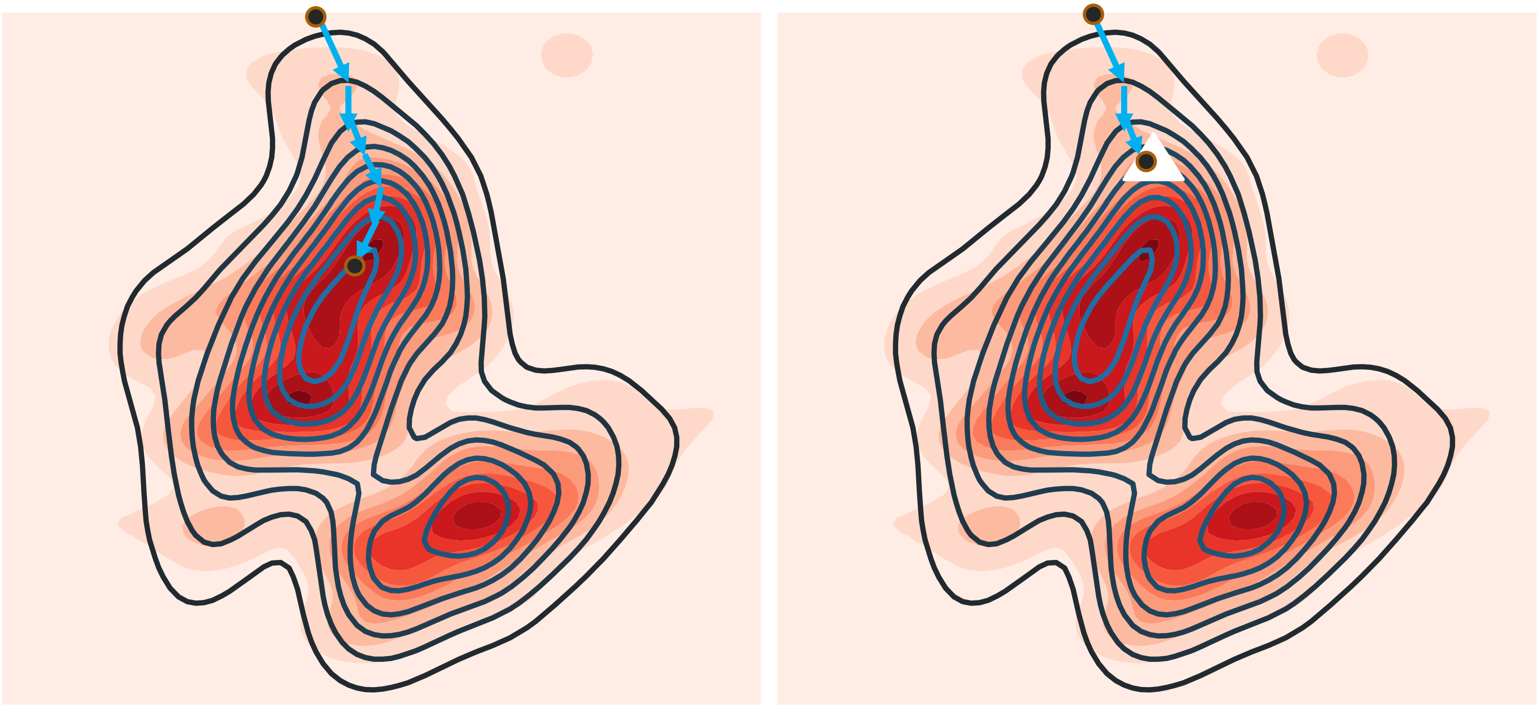
- FGS (Goodfellow et al. 2014)
- BPDA (Athalye et al., 2018).
- PGD (Madry et al., 2018)
- …

Black-box:

- ZOO (Chen et al. 2017)
- Query-Limited (Ilyas et al. 2018)
- …

One? Adversarial Perturbation (For an Input)

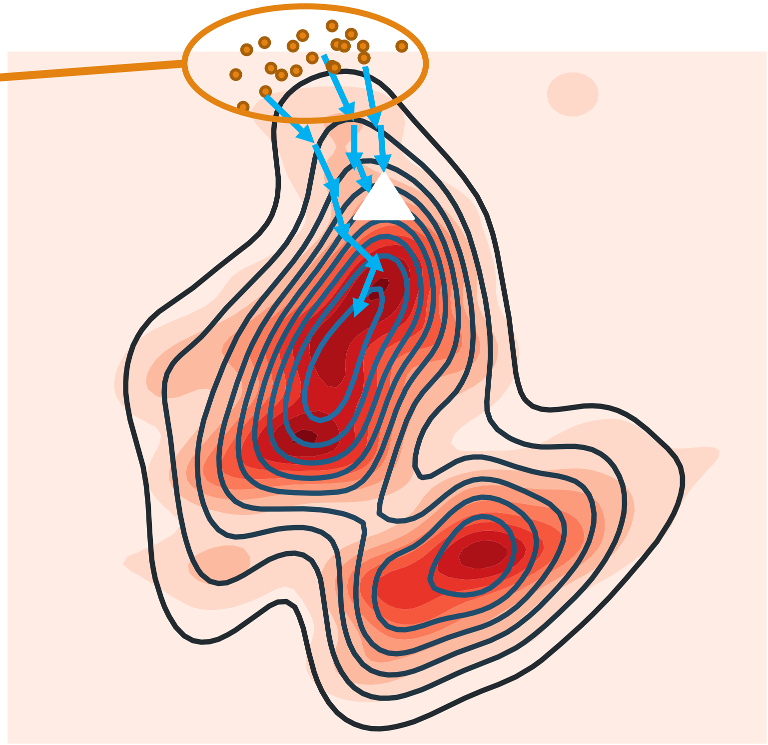
Bad local optimum, non-smooth optimization, curse of dimensionality, etc.



$\mathcal{N}$ATTACK

Learn the distributions of adversarial examples

$$\pi_S(x'|\theta)$$



$\mathcal{N}$ATTACK

Learn the distributions of adversarial examples

Smoothes the optimization

Higher attack success rate

Reduce the "attack dimension"

Less queries into the network

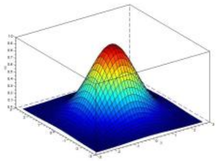
$$dim(\theta) << dim(x')$$

Characterizes the risk of the input example

New defense methods

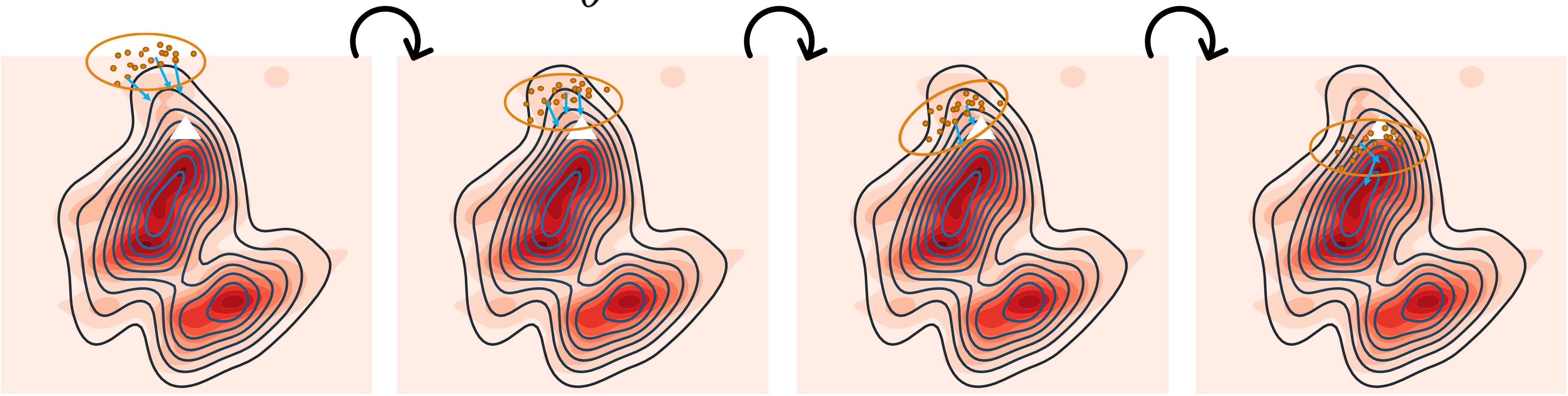
$\mathcal{N}$ATTACK

$\pi_S(x'|\theta) \longrightarrow x' \sim \mathcal{N}(x'|\mu, \sigma^2)$



Learn the distributions of adversarial examples

$$\max_{\theta} \ \mathbb{E}_{x' \sim \pi} L(x', y)$$



$\pi_S(x'|\theta_0)$ $\pi_S(x'|\theta_1)$ $\pi_S(x'|\theta_2)$ $\pi_S(x'|\theta_3)$

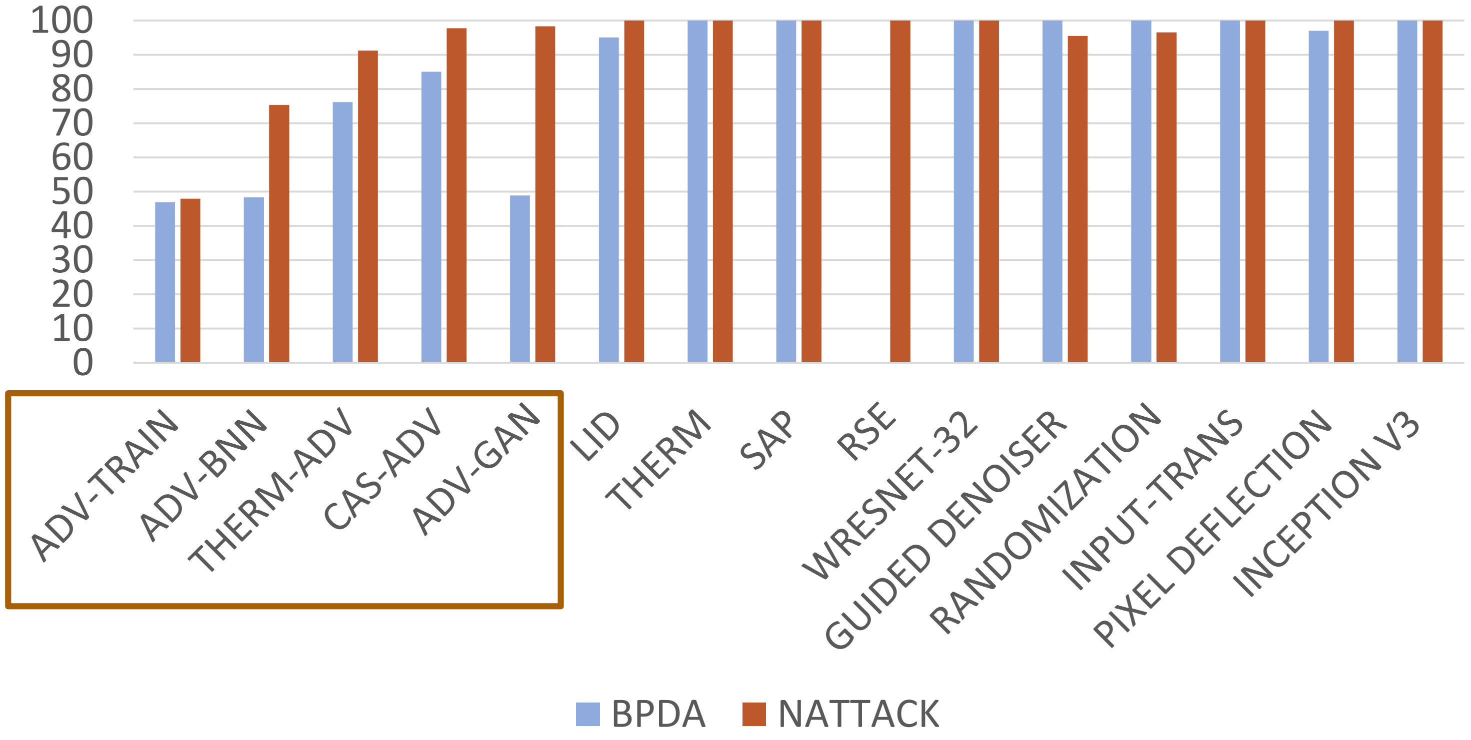
$\mathcal{N}$ATTACK

- *How to define the distributions* of adversarial examples?
- *Optimization*: *how to maximize* the objective function.

$$\max_{\theta} \; \mathbb{E}_{x' \sim \pi} L(x', y)$$

Poster session: Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #69

Experiments (Comparison with BPDA)

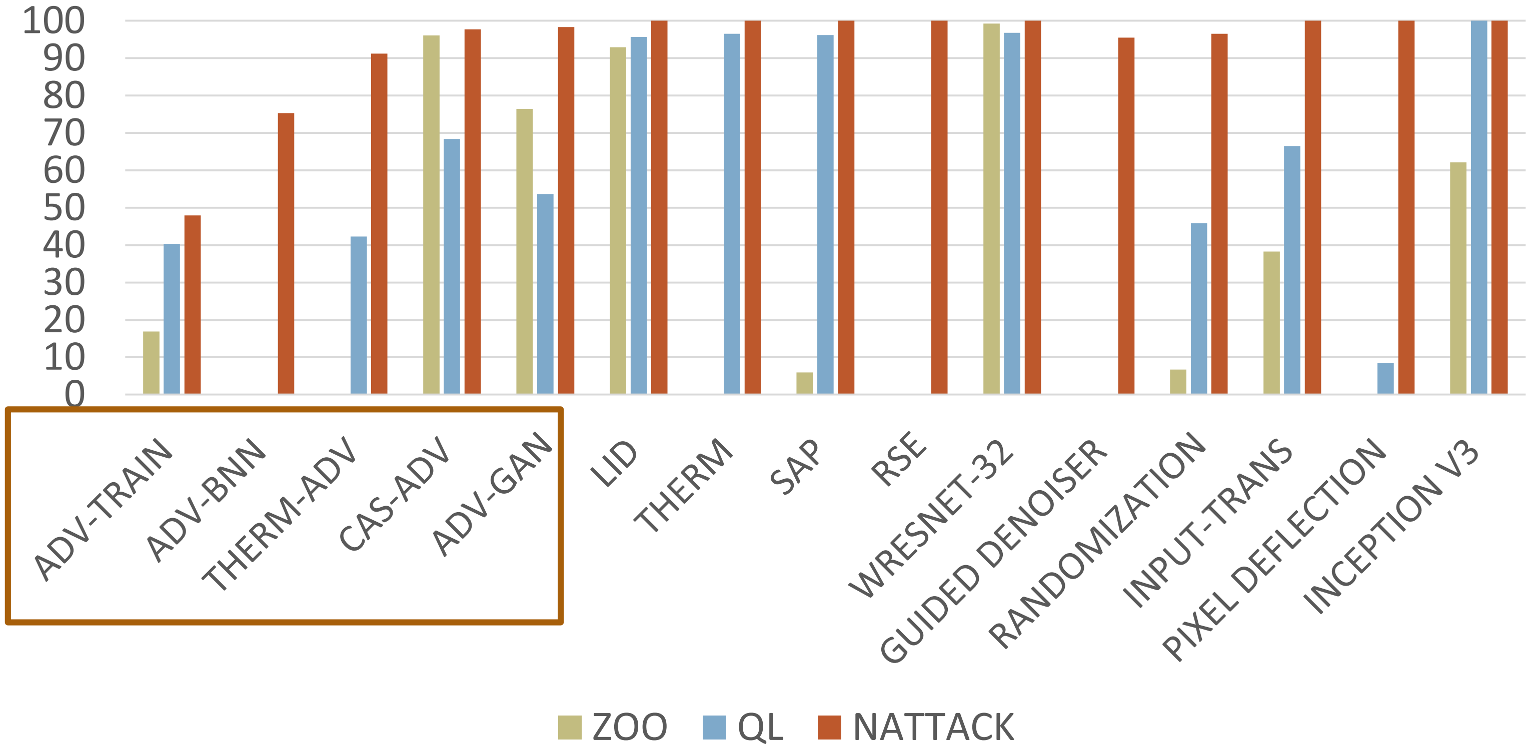


CIFAR10: ADV-TRAIN, ADV-BNN, THERM-ADV, CAS-ADV, ADV-GAN, LID, THERM, SAP, VANILLA WRESNET-32

IMAGENET: GUIDED DENOISER, RANDOMIZATION, INPUT-TRANS, PIXEL DEFLECTION, VANILLA INCEPTION V3

- $\mathcal{N}$ATTACK: 100% success rate on ***six out of the 13 defenses*** and ***more than 90% on five of the rest***.
- Competitive with white-box attack: BPDA (Athalye et al., 2018).

Experiments (Comparison with Black-box Approaches)



CIFAR10: ADV-TRAIN, THERM-ADV, CAS-ADV, ADV-GAN, LID, THERM, SAP, VANILLA WRESNET-32

IMAGENET: RANDOMIZATION, INPUT-TRANS, VANILLA INCEPTION V3

- The black-box baselines hinges on the quality of the ***estimated gradient.***
- Fail to attack ***Non-smooth DNNs.***

In a nutshell, $\mathcal{N}$ATTACK

- Is a ***powerful*** black-box attack, >= white-box attack.
- Is ***universal***: fooled different defenses by ***a single algorithm***.
- Characterize the distributions of adversarial examples.
- Reduce the “attack dimension”

Poster session: Wed Jun 12th 06:30 -- 09:00 PM @ Pacific Ballroom #69