RAFT: Learning the Distributions of Adversarial Examples for an Improved Black-Box Attack on Deep Neural Networks

Yandong Li*1  Lijun Li*1  Liqiang Wang1  Tong Zhang2  Boqing Gong3

*Equal Contribution

1University of Central Florida  2Hong Kong University of Science and Technology  3Google
Adversarial Examples

\[ x \quad 82\% \text{ puma} \quad \xrightarrow{\text{Adversarial Noise}} \quad + \delta \quad x' \quad 90\% \text{ book jacker} \]
Popular: Gradient-Based Adversarial Attack

\[ x_{t+1} = x_t + \eta \text{sign}(\nabla_x L(x_t, y)) \]

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.
\[ \pi_S(x' | \theta) \]
ATTACK

Learn the distributions of adversarial examples

Smothes the optimization

Higher attack success rate

Reduce the “attack dimension”

Less queries into the network $\dim(\theta) \ll \dim(x')$

Characterizes the risk of the input example

New defense methods
\text{\textsc{NAttack}}

Learn the distributions of adversarial examples

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

\[ \pi_S(x' | \theta_0) \quad \pi_S(x' | \theta_1) \quad \pi_S(x' | \theta_2) \quad \pi_S(x' | \theta_3) \]
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)

➢ N\text{ATTACK}: 100% success rate on \textit{six out of the 13 defenses} and \textit{more than 90% on five of the rest}.

➢ Competitive with white-box attack: BPDA (Athalye et al., 2018).
The black-box baselines hinges on the quality of the estimated gradient.

Fail to attack Non-smooth DNNs.
In a nutshell, $\mathcal{N}\text{ATTACK}$

- Is a **powerful** black-box attack, $\geq$ 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