Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation

Zeyu Qin*[1],Yanbo Fan*[2], Yi Liu[1], Li Shen[3], Yong Zhang[2], Jue Wang[2], Baoyuan Wu[1]

[1] The Chinese University of Hong Kong, Shenzhen

[2]Tencent AI Lab

[3] JD Explore Academy

NeurIPS 2022

Outline

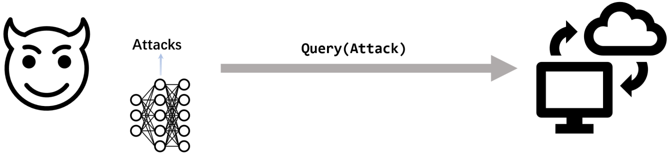
Introduction

Reverse Adversarial Perturbation

Experimental Evaluation

Severe Threats from Black-box Attacks

- **Transfer attacks: using $\boldsymbol{x}^{adv}$ from $\mathcal{M}^S$ to attack $\mathcal{M}^T$**
  - The attackers can **utilize same dataset to train the surrogate model** $\mathcal{M}^S$
  - Generating $\boldsymbol{x}^{adv}$ (white-box attacks) on $\mathcal{M}^S$, then attacking $\mathcal{M}^T$.
  - Don't need to iteratively query but it is not practical and performs poor attack performance.



Transfer attacks overfits to $\mathcal{M}^S$

- Taking the target attack as example, the general formulation of many existing transfer attack methods can be written as follows:

$$\min_{\boldsymbol{x}^{adv} \in \mathcal{B}_{\epsilon}(\boldsymbol{x})} \mathcal{L}(\mathcal{M}^s(\boldsymbol{x}^{adv}; \boldsymbol{\theta}), y_t). \tag{1}$$

  where $\mathcal{L}$ is the adversarial loss function, $\boldsymbol{y}_t$ is target label.
- The existing transfer attack methods exhibit poor transferability on $\mathcal{M}^T$ (not successfully attacking $\mathcal{M}^T$)
- $\boldsymbol{x}^{adv}$ severely depends on (overfits to) the decision boundaries of $\mathcal{M}^S$ and there are huge difference of decision boundaries between $\mathcal{M}^S$ and $\mathcal{M}^T$. [1,2,3]

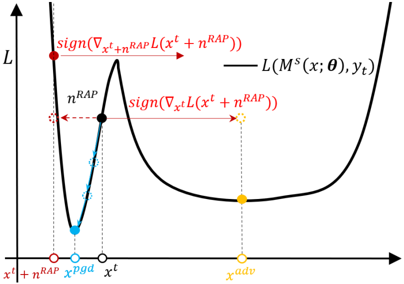
[1] Tramer et al., Ensemble Adversarial Training: Attacks and Defenses, ICLR 2018.

[2] Demontis et al., Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks, ACM CCS 2019.

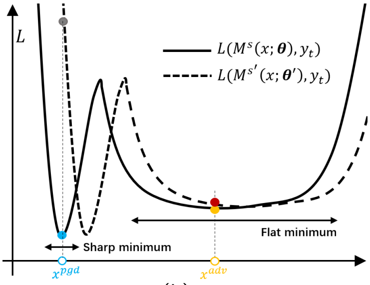
[3] Dong et al., Boosting adversarial attacks with momentum, CVPR 2018.

New Perspective to Interpret Adversarial Transferability

- We propose a new perspective to interpret the adversarial transferability, the flatness of (adversarial) loss landscape of $\boldsymbol{x}^{adv}$ on $\mathcal{M}^S$.
- The $\boldsymbol{x}^{adv}$ located at the flat local minimum is less sensitive to the changes of decision boundary (the difference of $\mathcal{M}^S$ and $\mathcal{M}^T$). Therefore, it could have the better adversarial transferability.



(a)



(b)

Outline

Finding x^{adv} located at a local flat region

- We encourage that **not only $\boldsymbol{x}^{adv}$ itself has low loss value, but also the points in the vicinity of $\boldsymbol{x}^{adv}$ have similarly low loss values**.
- We propose to **minimize the maximal loss value within a local neighborhood region** around $\boldsymbol{x}^{adv}$.
- The maximal loss is implemented by perturbing $\boldsymbol{x}^{adv}$ (adding perturbation $\boldsymbol{n}^{adv}$) to maximize the adversarial loss, named **Reverse Adversarial Perturbation (RAP)**. So, we aim to solve this problem,

$$\min_{\boldsymbol{x}^{adv} \in \mathcal{B}_\epsilon(\boldsymbol{x})} \mathcal{L}\left(\mathcal{M}^S\left(\boldsymbol{x}^{adv} + \boldsymbol{n}^{adv}; \boldsymbol{\theta}\right), y_t\right)$$

Where,

$$\boldsymbol{n}^{adv} = \underset{\|\boldsymbol{n}\|_\infty \le \epsilon_n}{\arg\max} \mathcal{L}\left(\mathcal{M}^S\left(\boldsymbol{x}^{adv} + \boldsymbol{n}; \boldsymbol{\theta}\right), y_t\right)$$

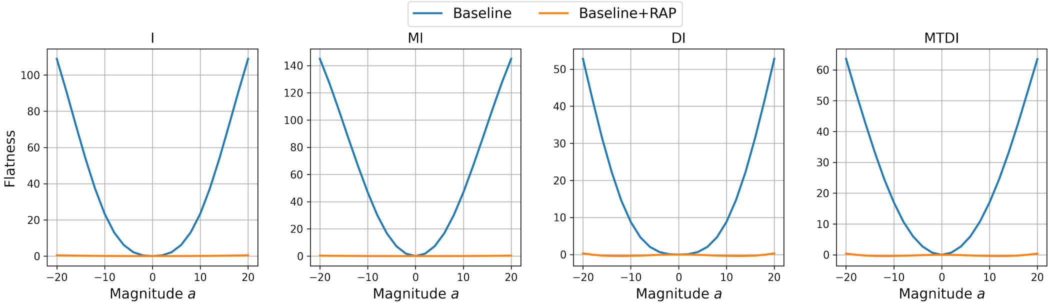
RAP with Late-Start

- In our preliminary experiments, we find that RAP requires more iterations to converge and the performance is slightly lower during the initial iterations.
- Hence, we further propose a better initialization, **late-start** (LS in following content) which first only runs the outer loop for several iterations then conducts the min-max loop.

Outline

A Closer Look at RAP

- First we **visualize the loss landscape around x^{adv} on $\mathcal{M}^S$** by plotting the loss variations. We can observe that RAP could help find x^{adv} located at the flat region.



RAP achieves the better attack performance

- Combined with existing attacks, RAP **further boosts their transferability for both untargeted and targeted attacks.**
  The below tables show the transfer targeted attack performance ($\mathcal{M}^S \Longrightarrow \mathcal{M}^T$).

Attack	**ResNet-50** $\Longrightarrow$			**DenseNet-121** $\Longrightarrow$		
	Dense-121	VGG-16	Inc-v3	Res-50	VGG-16	Inc-v3
MTDI / +RAP / +RAP-LS	74.9 / 78.2 / **88.5**	62.8 / 72.9 / **81.5**	10.9 / 28.3 / **33.2**	44.9 / 64.3 / **74.5**	38.5 / 55.0 / **65.5**	7.7 / 23.0 / **26.5**
MTDSI / +RAP / +RAP-LS	86.3 / 88.4 / **93.3**	70.1 / 77.7 / **84.7**	38.1 / 51.8 / **58.0**	55.0 / 71.2 / **75.8**	42.0 / 58.4 / **62.3**	19.8 / 39.0 / **39.2**
MTDAI / +RAP / +RAP-LS	91.4 / 89.4 / **93.6**	79.9 / 79.0 / **86.3**	50.8 / 57.1 / **64.1**	69.1 / 74.2 / **82.1**	54.7 / 63.1 / **69.3**	32.0 / 43.5 / **49.3**
Attack	**VGG-16** $\Longrightarrow$			**Inc-v3** $\Longrightarrow$		
	Res-50	Dense-121	Inc-v3	Res-50	Dense-121	VGG-16
MTDI / +RAP / +RAP-LS	11.8 / 16.7 / **22.9**	13.7 / 19.4 / **27.4**	0.7 / 3.4 / **4.6**	1.8 / **8.3** / 7.5	4.1 / **14.8** / 13.4	2.9 / 8.0 / **9.8**
MTDSI / +RAP / +RAP-LS	31.0 / 35.3 / **38.7**	41.7 / 44.4 / **49.6**	9.6 / **15.2** / 13.7	5.6 / **11.9** / 10.7	10.4 / **21.2** / 20.9	4.2 / **8.9** / 8.6
MTDAI / +RAP / +RAP-LS	36.2 / 39.0 / **43.1**	48.0 / 45.1 / **55.2**	11.6 / 17.1 / **17.6**	9.6 / 13.6 / **16.7**	17.9 / 27.5 / **31.6**	8.4 / 12.0 / **12.1**

RAP achieves the better attack performance on diverse architectures

- We also conduct experiments on diverse network architectures, ViT and ensemble models. Our RAP-LS achieves the better attack performance.

Attack	**Untarged**			**Targeted**			**Untarged**		**Targeted**	
	IncRes-v2	NASNet-L	ViT-B/16	IncRes-v2	NASNet-L	ViT-B/16	Inc-v3$_{adv}$	IncRes-v2$_{ens}$	Inc-v3$_{adv}$	IncRes-v2$_{ens}$
MTDI	83.4	89.0	27.9	14.8	32.1	0.4	68.1	50.9	0.8	0.0
MTDI+RAP-LS	**95.6**	**97.5**	**42.7**	**43.0**	**62.5**	**1.7**	**86.5**	**72.3**	**9.7**	**4.1**
MTDSI	95.7	98.0	43.0	45.5	67.9	2.6	90.0	79.6	12.7	6.7
MTDSI+RAP-LS	**98.6**	**99.7**	**57.4**	**64.0**	**80.4**	**5.3**	**96.5**	**91.5**	**31.0**	**22.0**
MTDAI	97.3	98.8	45.5	58.4	75.3	3.3	92.1	82.7	17.2	12.2
MTDAI+RAP-LS	**99.2**	**99.8**	**60.2**	**70.4**	**82.6**	**7.4**	**96.7**	**91.6**	**34.4**	**26.0**

RAP achieves the SOTA attack performance stronger defense models

- We also take a comparison on stronger defense models. Our methods also chieve the SOTA performance on defense models. Our RAP-LS achieves the better attack performance.

Attack	**Untarged**		
	Res-50 AT (ℓ_2)	Res-50 AT (ℓ_∞)	Feature Denoising
MTDI	42.5	32.4	44.1
MTDI+RAP-LS	**59.5**	**34.4**	**44.4**
MTDSI	56.6	35.8	45.0
MTDSI+RAP-LS	**70.3**	**36.6**	**45.7**
MTDAI	62.1	35.6	44.2
MTDAI+RAP-LS	**73.7**	**37.7**	**45.2**

Thank Y’all!