Why Does Little Robustness Help? A Further Step Towards Understanding Adversarial Transferability

Yechao Zhang[1], Shengshan Hu[1], Leo Yu Zhang[2], Junyu Shi[1],
Minghui Li[1], Xiaogeng Liu[1], Wan Wei[1] and Hai Jin[1]

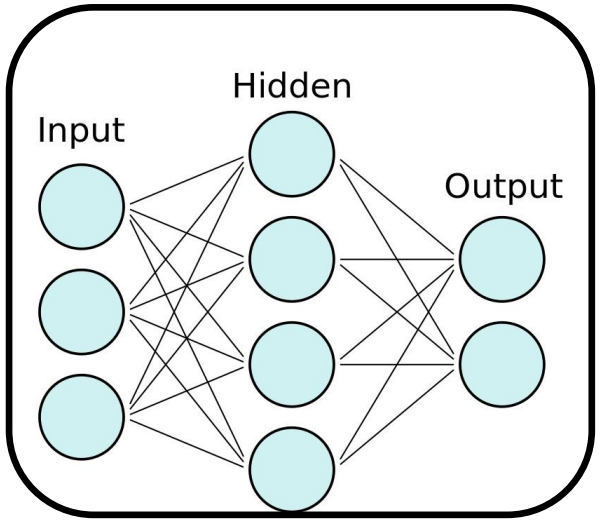
[1] Huazhong University of Science and Technology

[2] Griffith University

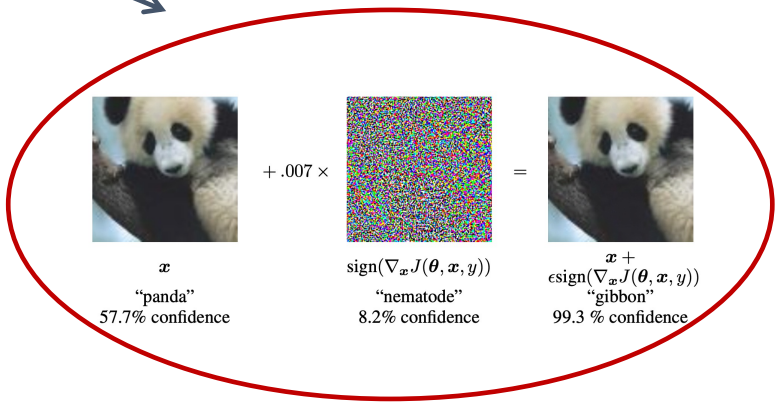
To Appear at IEEE Symposium on Security and Privacy 2024

Background: Transfer Attack

Adversary



White-box surrogate model



Optimizing adversarial examples

Different AE generation methods can be applied

Transfer to

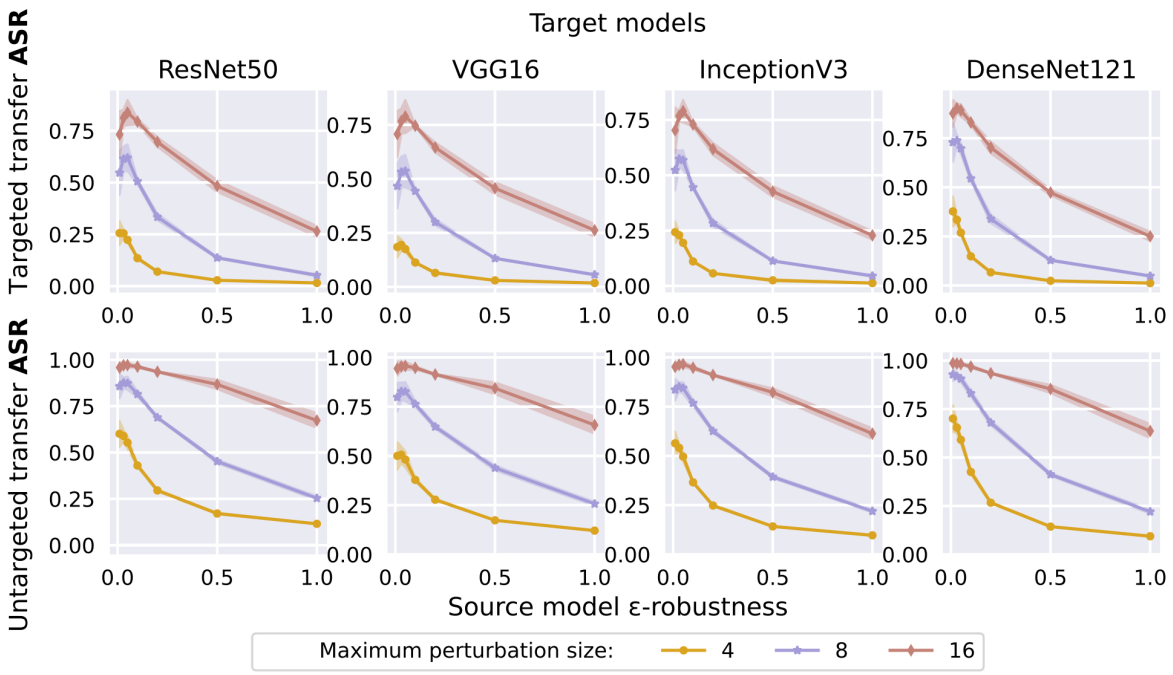
Adversarial transferability: the ability of an adversarial example to generalize across different models.

Remark: Most work focus on optimizing the adversarial examples, we investigate what kind of surrogates are more suitable for transfer attacks.

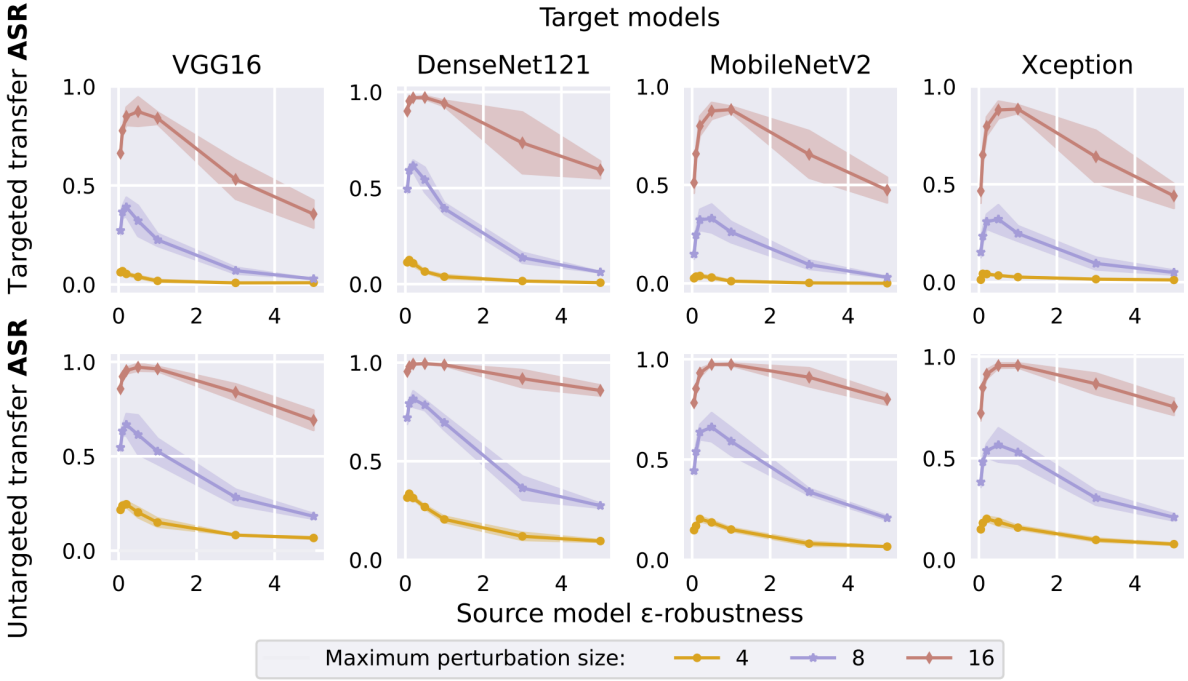
Let's start from an interesting observation

Adversarial training with small perturbation leads to better surrogates

Transfer attack using adversarial training surrogates



(a) CIFAR-10

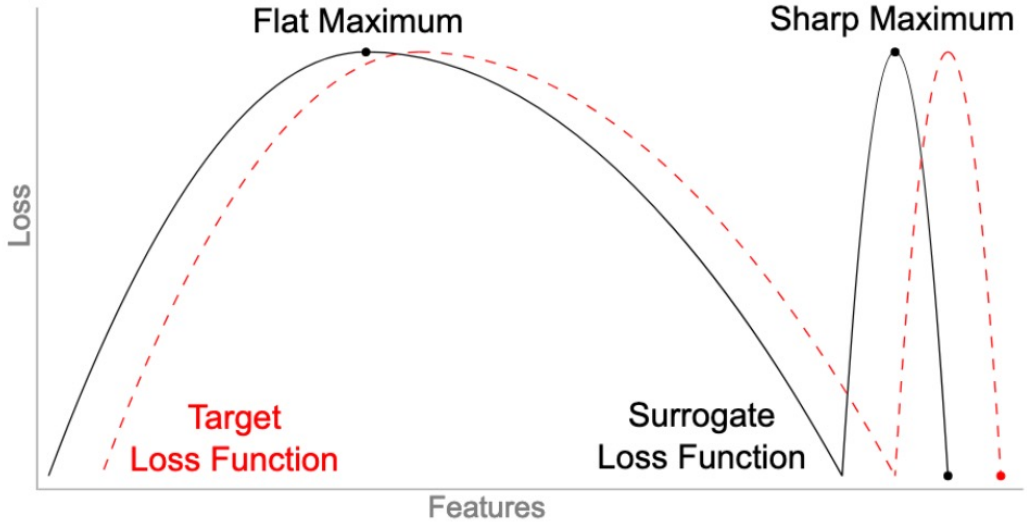


(b) ImageNette

Question: *Why does "little robustness" exhibit this benefit whereas "much robustness" does not?*

Intuitions and deciding factors behind adversarial transferability

Two factors that are believed to be essential to adversarial transferability

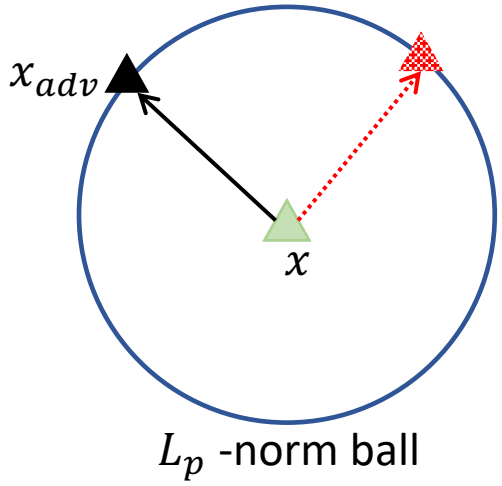


Flat optima is more stable, more likely to transfer to target model.

Model smoothness: how smoothness the input loss landscape in the model.

$$\sigma_{\mathcal{F},\mathcal{D}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\sigma(\nabla_x^2 \ell_{\mathcal{F}}(x,y))]$$

Second-order derivatives



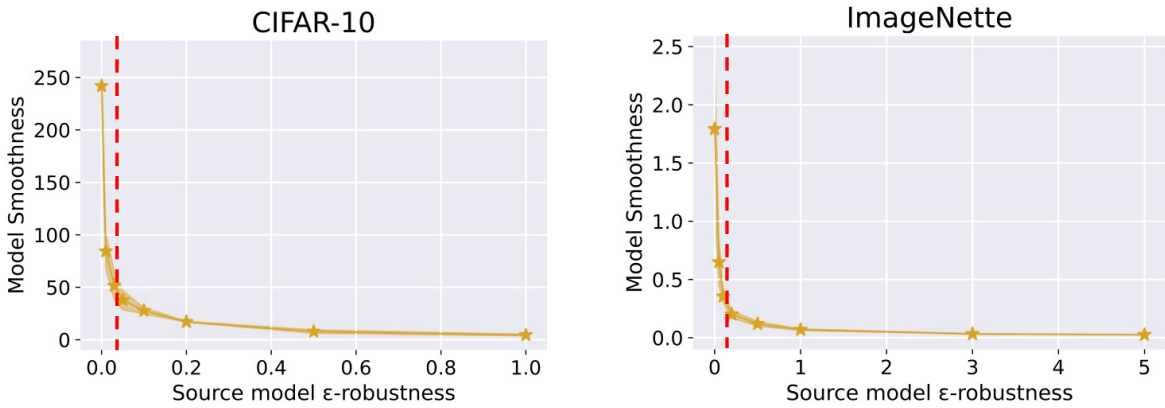
Gradient direction between surrogate and target models is more similar, more likely to transfer.

Gradient similarity: how smoothness the minimum found in the model.

$$\mathcal{S}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}}, x, y) = \frac{\nabla_x \ell_{\mathcal{F}}(x,y) \cdot \nabla_x \ell_{\mathcal{G}}(x,y)}{\|\nabla_x \ell_{\mathcal{F}}(x,y)\|_2 \cdot \|\nabla_x \ell_{\mathcal{G}}(x,y)\|_2}$$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{S}(\ell_{\mathcal{F}}, \ell_{\mathcal{G}}, x, y)]$$

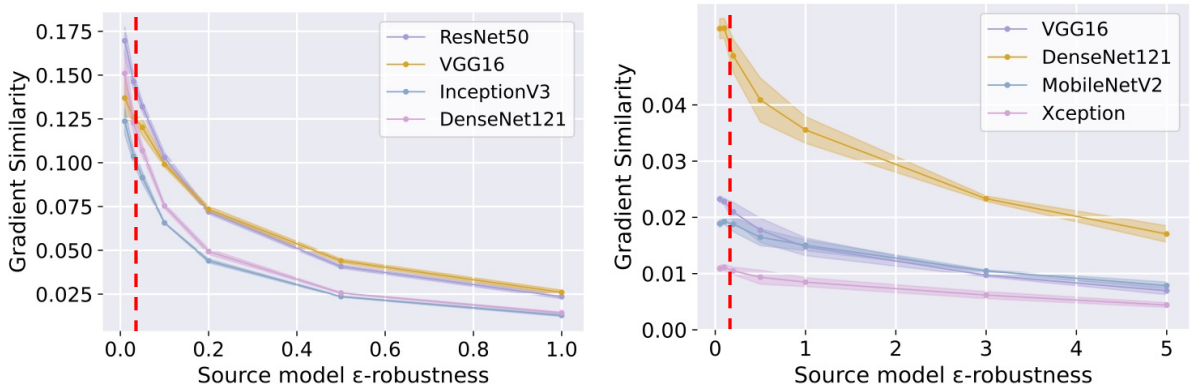
The trade-off between smoothness and similarity in adversarial training



(a) ResNet18 on CIFAR-10 (b) ResNet50 on ImageNette

Model smoothness in adversarial training

The bigger the adversarial budget, the smoother the model.



(a) CIFAR-10 (b) ImageNette

Gradient similarity in adversarial training

The bigger the adversarial budget, the more dissimilar between the gradient directions.

Conjectures:

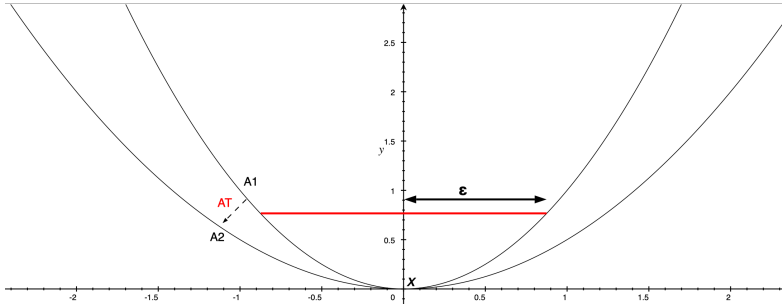
- The quick improvement in transferability for small ε occurs could be the cause of the rapid gains in smoothness and small decays in gradient similarity.
- The degradation in transferability for large ε occurs may because the smoothness gains have approached the limit while gradient similarity continues to decrease.

Why adversarial training benefits smoothness?

This can be intuitively explained and theoretically proved

$$\max_{\|\delta\|_2<\epsilon} \ell(f(x_i+\delta), y_i) = \ell(f(x_i), y_i) + \boxed{[\max_{\|\delta\|_2<\epsilon} \ell(f(x_i+\delta), y_i) - \ell(f(x_i), y_i)]}$$

non-smoothness



Applying Taylor expansion, If x is a local minimum, then

$$\max_{\|\delta\|_2\leq\epsilon} \ell(f(x_i+\delta), y_i) - \ell(f(x_i), y_i) = \frac{1}{2}\sigma\big(\nabla_x^2 \ell(f(x_i), y_i)\big) \cdot \|\delta\|_2^2 + O\big(\|\delta\|_2^3\big)$$

Thus, it also provably suppresses the second-order derivatives.

But why adversarial training degrade similarity?

This cannot be well-explained theoretically

We make an intuitive hypothesis:

Data distribution shift impairs gradient similarity*.

A long-held belief in the literature:

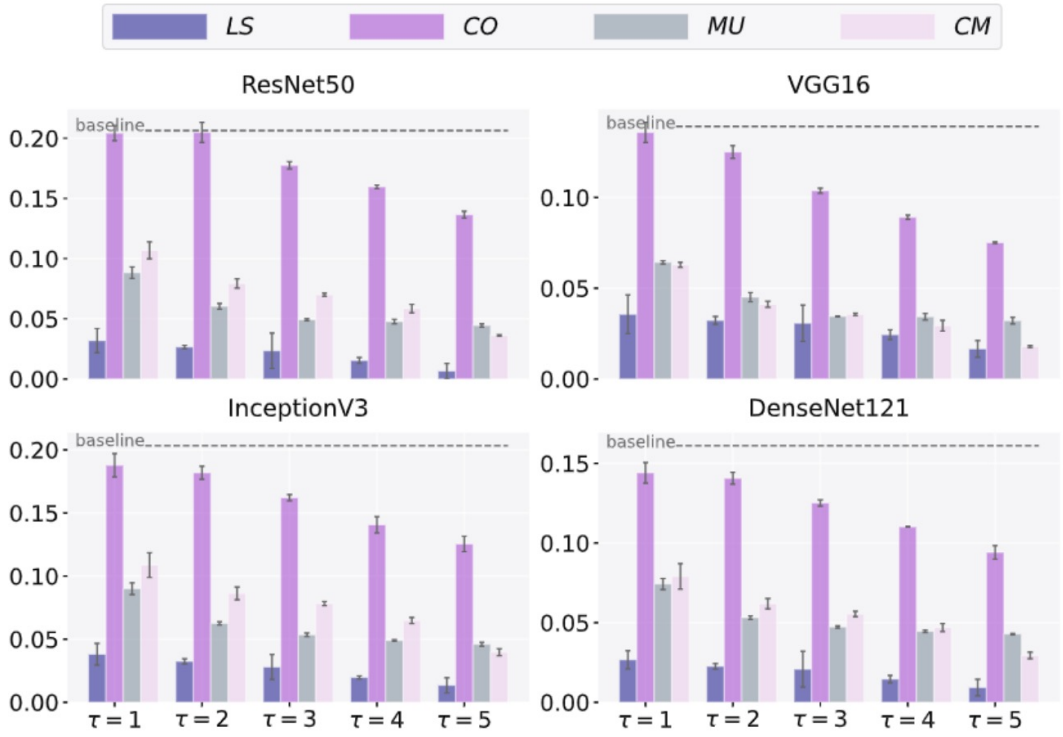
"Clean data lies in a low-dimensional manifold. Even though the adversarial examples are close to the clean data, they lie off the underlying data manifold."

We believe the off-manifold adversarial samples in the training data cause the gradient dissimilarity.

*Note that there is a premise that target model does not change the data distribution

Verifying the distribution shift hypothesis

Experimental evaluations on 4 data augmentations



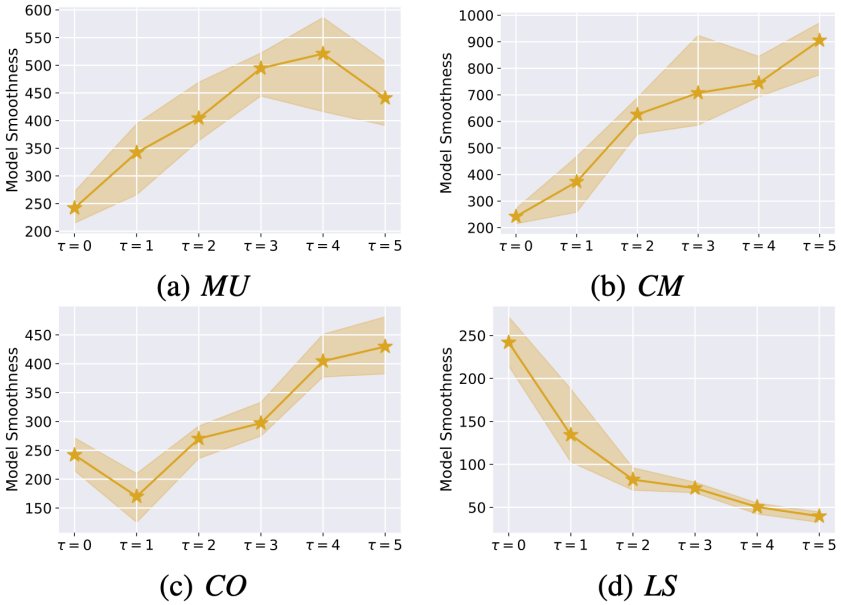
Similarity between augmented models and target models

- The results support the hypothesis that data distribution shift impairs gradient similarity.
- Similarity degradation:

The distribution shift in $\mathcal{P}_y$ may have a greater negative impact on similarity than that in $\mathcal{P}_x$

CO	*MU*	*CM*	*LS*
Cutout	< Mixup = Cutmix	<	Label smoothing
$\mathcal{P}_x$		$\mathcal{P}_{x \times y}$	$\mathcal{P}_y$

Data augmentation generally yields worse surrogates



(a) *MU* (b) *CM*

(c) *CO* (d) *LS*

Smoothness in data augmentations

- The smoothness in data augmentation does not exhibit a uniform tendency.
- However, the transfer attack success rate are uniformly worse than the baseline.

Trade-off under data augmentations is quite complex, and no single augmentation can produce good surrogates

CIFAR-10												
	4/255				8/255				16/255			
	ResNet50	VGG16	InceptionV3	DenseNet121	ResNet50	VGG16	InceptionV3	DenseNet121	ResNet50	VGG16	InceptionV3	DenseNet121
ST	$41.2_{\pm 5.6}$	$30.1_{\pm 3.5}$	$41.7_{\pm 5.3}$	$58.6_{\pm 7.1}$	$62.9_{\pm 7.9}$	$54.5_{\pm 6.5}$	$64.8_{\pm 6.0}$	$80.5_{\pm 6.8}$	$83.2_{\pm 5.2}$	$81.0_{\pm 5.5}$	$84.2_{\pm 3.8}$	$91.0_{\pm 3.3}$
MU, $\tau = 1$	$27.9_{\pm 5.1}$	$20.9_{\pm 2.5}$	$29.0_{\pm 4.8}$	$40.2_{\pm 7.0}$	$48.5_{\pm 6.9}$	$40.6_{\pm 3.7}$	$51.4_{\pm 5.9}$	$65.3_{\pm 7.5}$	$73.1_{\pm 6.4}$	$71.3_{\pm 4.5}$	$76.5_{\pm 4.7}$	$85.2_{\pm 5.1}$
MU, $\tau = 3$	$20.4_{\pm 0.6}$	$16.4_{\pm 0.5}$	$23.3_{\pm 1.2}$	$27.8_{\pm 1.6}$	$37.7_{\pm 0.9}$	$31.2_{\pm 1.1}$	$42.0_{\pm 1.4}$	$51.1_{\pm 1.3}$	$68.0_{\pm 1.9}$	$64.7_{\pm 1.7}$	$72.9_{\pm 1.2}$	$80.4_{\pm 1.2}$
MU, $\tau = 5$	$18.8_{\pm 0.3}$	$15.7_{\pm 0.3}$	$20.8_{\pm 1.0}$	$24.5_{\pm 0.8}$	$33.8_{\pm 1.7}$	$28.6_{\pm 0.9}$	$37.5_{\pm 1.0}$	$44.9_{\pm 1.6}$	$64.1_{\pm 1.5}$	$60.5_{\pm 1.9}$	$68.3_{\pm 0.4}$	$76.4_{\pm 1.5}$
CM, $\tau = 1$	$22.4_{\pm 0.9}$	$16.7_{\pm 1.3}$	$21.8_{\pm 0.6}$	$30.3_{\pm 1.3}$	$39.6_{\pm 1.7}$	$32.42_{\pm .5}$	$39.0_{\pm 1.0}$	$53.4_{\pm 2.5}$	$64.5_{\pm 2.0}$	$62.4_{\pm 3.2}$	$64.9_{\pm 1.7}$	$77.8_{\pm 3.2}$
CM, $\tau = 3$	$14.3_{\pm 1.2}$	$11.5_{\pm 0.4}$	$14.2_{\pm 1.1}$	$17.2_{\pm 2.0}$	$25.8_{\pm 3.0}$	$20.2_{\pm 1.8}$	$25.1_{\pm 2.6}$	$32.1_{\pm 4.4}$	$48.9_{\pm 4.1}$	$47.1_{\pm 2.7}$	$49.8_{\pm 3.0}$	$60.6_{\pm 4.6}$
CM, $\tau = 5$	$12.3_{\pm 0.5}$	$10.4_{\pm 0.2}$	$12.0_{\pm 1.1}$	$13.7_{\pm 1.2}$	$21.9_{\pm 1.3}$	$17.5_{\pm 0.6}$	$21.3_{\pm 0.9}$	$26.5_{\pm 1.8}$	$44.4_{\pm 1.6}$	$41.5_{\pm 1.6}$	$44.3_{\pm 1.9}$	$53.4_{\pm 1.8}$
CO, $\tau = 1$	$40.7_{\pm 7.0}$	$31.1_{\pm 7.0}$	$39.8_{\pm 5.9}$	$55.9_{\pm 8.1}$	$62.4_{\pm 6.9}$	$55.3_{\pm 8.1}$	$63.1_{\pm 6.2}$	$78.2_{\pm 5.9}$	$82.5_{\pm 5.6}$	$79.9_{\pm 6.3}$	$82.8_{\pm 5.3}$	$89.9_{\pm 3.7}$
CO, $\tau = 3$	$34.6_{\pm 5.2}$	$24.9_{\pm 2.8}$	$33.7_{\pm 5.1}$	$47.1_{\pm 8.0}$	$55.5_{\pm 5.8}$	$46.0_{\pm 4.7}$	$55.8_{\pm 6.7}$	$70.7_{\pm 8.0}$	$79.6_{\pm 5.4}$	$74.4_{\pm 5.1}$	$79.2_{\pm 5.6}$	$86.8_{\pm 5.0}$
CO, $\tau = 5$	$30.8_{\pm 2.5}$	$22.3_{\pm 2.0}$	$31.0_{\pm 3.3}$	$42.1_{\pm 6.7}$	$49.5_{\pm 5.0}$	$41.3_{\pm 4.8}$	$51.3_{\pm 5.8}$	$63.9_{\pm 7.8}$	$73.2_{\pm 5.6}$	$69.3_{\pm 6.3}$	$74.5_{\pm 6.4}$	$81.7_{\pm 7.0}$
LS, $\tau = 1$	$35.5_{\pm 4.8}$	$28.2_{\pm 5.7}$	$35.5_{\pm 3.3}$	$48.4_{\pm 5.7}$	$54.3_{\pm 9.4}$	$49.7_{\pm 10.7}$	$56.0_{\pm 6.4}$	$69.4_{\pm 8.4}$	$76.1_{\pm 7.8}$	$76.9_{\pm 8.3}$	$78.3_{\pm 6.0}$	$85.9_{\pm 5.1}$
LS, $\tau = 3$	$34.2_{\pm 7.5}$	$27.3_{\pm 4.6}$	$33.9_{\pm 5.6}$	$44.1_{\pm 2.8}$	$55.0_{\pm 13.6}$	$50.0_{\pm 10.0}$	$55.9_{\pm 11.0}$	$68.2_{\pm 5.9}$	$76.8_{\pm 12.1}$	$76.6_{\pm 10.0}$	$78.3_{\pm 9.3}$	$85.5_{\pm 5.3}$
LS, $\tau = 5$	$31.3_{\pm 3.5}$	$26.6_{\pm 3.5}$	$30.2_{\pm 3.1}$	$41.2_{\pm 2.9}$	$51.1_{\pm 8.2}$	$48.6_{\pm 8.0}$	$51.2_{\pm 6.9}$	$65.6_{\pm 7.0}$	$72.2_{\pm 7.7}$	$74.5_{\pm 8.2}$	$73.5_{\pm 6.2}$	$83.3_{\pm 5.4}$

Transfer attack success rate under data augmented surrogates

How to stably find better surrogates?

Resorting to smoothness-promoting methods that does not change data distribution

The rationale is:

- In the real scenario, similarity is a pair-wise metric referring target model, which we cannot access.
- We do not know what kind of approach will benefit similarity.
- We do know what will degrade similarity.
- Smoothness is a standalone concept, which can be independently regulated and measured.

Thus, we believe we can stably increase the transferability by promoting the smoothness of surrogate alone while do not change the data distribution shift.

Promoting smoothness through input gradient regularizations

- The most direct way to promote smoothness is to minimize the loss surface curvature, $\sigma(\nabla_x^2 \ell(x))$
- However, computing the second-order derivative is extremely expensive, let alone optimizing it.

Solution: Approximating the first-order derivatives through first-order derivative.

$$\sigma(\nabla_x^2 \ell(x)) \leq \|\nabla_x^2 \ell(x)\|_F \approx \|\nabla_x \ell(f(x))^T \nabla_x \ell(f(x))\|_F \leq \|\nabla_x \ell(f(x))\|^2$$

Input gradient regularization (*IR*):

$$L_{ir} = \frac{1}{\|\mathbf{S}\|} \sum_{i=1}^{\|\mathbf{S}\|} [\ell(f(x_i)) + \lambda_{ir} \|\nabla_x \ell(f(x_i))\|],$$

Input Jacobian regularization (*JR*):

$$L_{jr} = \frac{1}{\|\mathbf{S}\|} \sum_{i=1}^{\|\mathbf{S}\|} [\ell(f_\theta(x_i) + \lambda_{jr} \|\nabla_x f_\theta(x_i)\|_F]$$

Proposition 1. *Let a neural network parameterized by θ, and f_θ represents its logit network. Given a sample (x, y), if $\|\nabla_x f_\theta(x)\|_F \to 0$, $\|\nabla_x \ell(f_\theta(x), y)\| \to 0$, where ℓ denotes the cross-entropy loss function.*

Promoting smoothness through weight gradient regularizations

- Researches have proved that the gradient regularizing pressure on the weight space $\|\nabla_\theta f_\theta(x)\|_F$ can transfer to the input space $\|\nabla_x f_\theta(x)\|_F$.

Theorem A.1 (Transfer Theorem). *Consider a network $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ with l layers parameterized by $\theta = (w_1, b_1, \ldots, w_l, b_l)$, then we have the following inequality*

$$\|\nabla_x f_\theta(x)\|_F^2 \leq \frac{\|\nabla_\theta f_\theta(x)\|_F^2}{T_1^2(x,\theta) + \cdots + T_l^2(x,\theta)}, \tag{43}$$

- Thus, we also consider two weight space gradient regularizations as follows:

Explicit gradient regularization (*ER*):

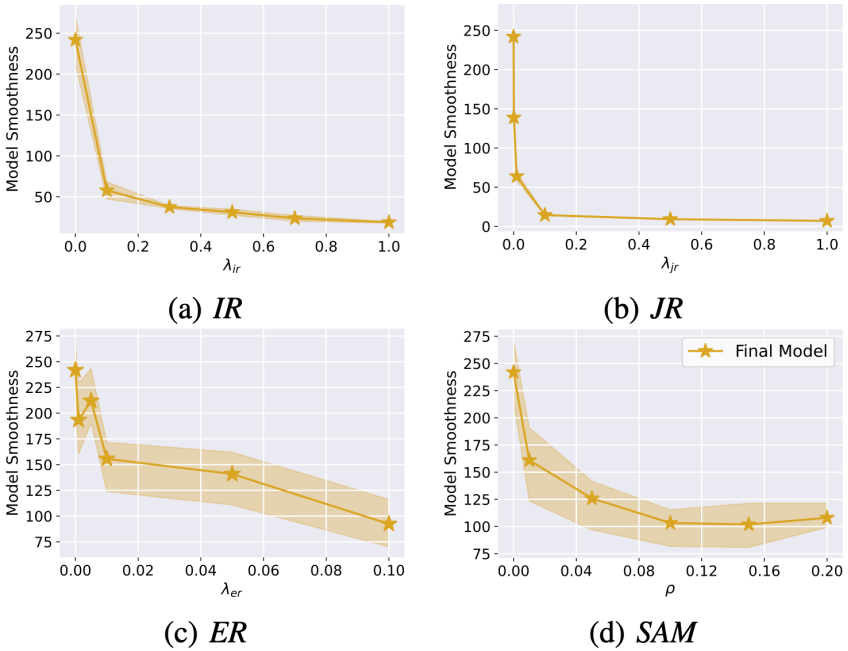
$$L_{er}(\theta) = L(\theta) + \frac{\lambda_{er}}{2}\|\nabla_\theta L(\theta)\|^2$$

Sharpness-aware minimization (*SAM):**

$$\nabla_\theta L_{sam}(\theta) \approx \nabla_\theta L(\theta + \rho \frac{\nabla_\theta L(\theta)}{||\nabla_\theta L(\theta)||}).$$

*Recent researches establish *SAM* as a special kind of gradient normalization.

Gradient regularizations yield better surrogates



(a) *IR* (b) *JR* (c) *ER* (d) *SAM*

These results indicate:

- All gradient regularizations generally improve the smoothness and transferability.
- *IR, JR* yield better smoothness than *ER, SAM.*
- Surprisingly, in terms of transferability, *SAM* is generally better than *IR* and *JR* in CIFAR-10, and worse in ImageNette.

It suggests an overall examination of these training mechanisms, as they may weigh differently on similarity.

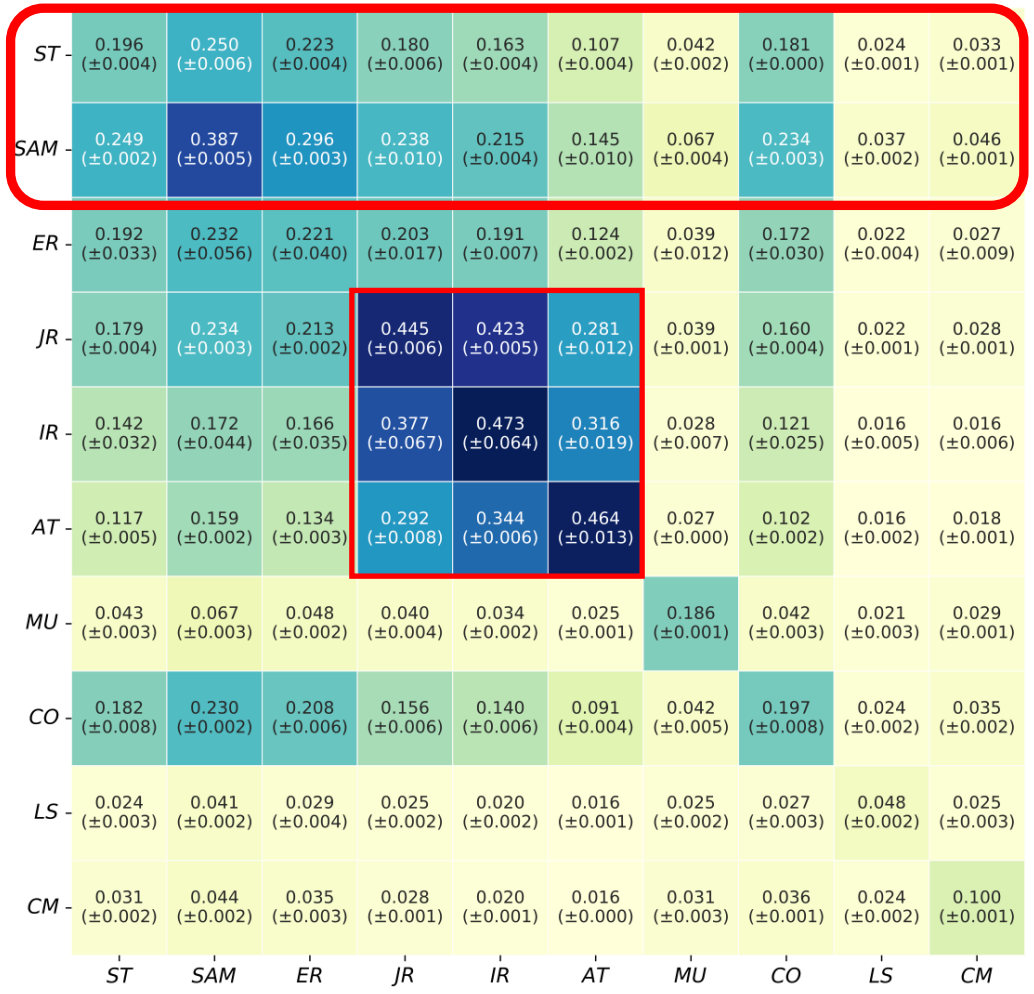
CIFAR-10

	4/255				8/255				16/255			
	ResNet50	VGG16	InceptionV3	DenseNet121	ResNet50	VGG16	InceptionV3	DenseNet121	ResNet50	VGG16	InceptionV3	DenseNet121
ST	$41.2_{\pm 5.6}$	$30.1_{\pm 3.5}$	$41.7_{\pm 5.3}$	$58.6_{\pm 7.1}$	$62.9_{\pm 7.9}$	$54.5_{\pm 6.5}$	$64.8_{\pm 6.0}$	$80.5_{\pm 6.8}$	$83.2_{\pm 5.2}$	$81.0_{\pm 5.5}$	$84.2_{\pm 3.8}$	$91.0_{\pm 3.3}$
IR	$51.9_{\pm 1.4}$	$45.3_{\pm 1.1}$	$48.0_{\pm 0.6}$	$54.4_{\pm 1.1}$	$85.9_{\pm 3.4}$	$82.4_{\pm 3.9}$	$83.8_{\pm 3.3}$	$87.4_{\pm 3.2}$	$92.4_{\pm 2.1}$	$92.2_{\pm 2.3}$	$91.8_{\pm 2.1}$	$92.0_{\pm 2.0}$
JR	$63.5_{\pm 8.7}$	$53.1_{\pm 7.1}$	$62.4_{\pm 5.7}$	$75.1_{\pm 4.5}$	$78.9_{\pm 6.8}$	$73.7_{\pm 6.1}$	$79.5_{\pm 4.1}$	$88.7_{\pm 2.6}$	$87.6_{\pm 4.3}$	$87.2_{\pm 3.9}$	$88.1_{\pm 2.4}$	$92.1_{\pm 2.1}$
ER	$55.0_{\pm 5.5}$	$42.9_{\pm 5.6}$	$46.2_{\pm 7.4}$	$55.2_{\pm 13.8}$	$82.3_{\pm 5.5}$	$74.1_{\pm 7.9}$	$76.1_{\pm 8.8}$	$81.1_{\pm 12.2}$	$90.6_{\pm 4.0}$	$89.3_{\pm 4.6}$	$88.9_{\pm 4.0}$	$89.8_{\pm 4.4}$
SAM	$66.1_{\pm 8.6}$	$53.4_{\pm 5.8}$	$66.0_{\pm 8.8}$	$81.1_{\pm 6.9}$	$88.4_{\pm 4.7}$	$83.4_{\pm 4.5}$	$88.7_{\pm 4.8}$	$94.1_{\pm 2.4}$	$94.3_{\pm 3.0}$	$94.0_{\pm 3.0}$	$94.2_{\pm 2.9}$	$94.8_{\pm 2.4}$

ImageNette

	4/255				8/255				16/255			
	VGG16	DenseNet121	MobileNetV2	Xception	VGG16	DenseNet121	MobileNetV2	Xception	VGG16	DenseNet121	MobileNetV2	Xception
ST	$10.1_{\pm 0.7}$	$16.0_{\pm 1.0}$	$7.1_{\pm 0.5}$	$6.7_{\pm 0.1}$	$27.4_{\pm 3.0}$	$41.0_{\pm 2.8}$	$17.7_{\pm 0.7}$	$16.9_{\pm 1.1}$	$61.5_{\pm 6.5}$	$83.3_{\pm 4.9}$	$51.5_{\pm 2.7}$	$44.8_{\pm 4.6}$
IR	$19.1_{\pm 6.1}$	$29.9_{\pm 6.3}$	$16.7_{\pm 5.1}$	$16.7_{\pm 4.7}$	$60.7_{\pm 14.7}$	$82.7_{\pm 9.5}$	$64.1_{\pm 16.7}$	$55.9_{\pm 13.5}$	$91.2_{\pm 7.6}$	$96.2_{\pm 0.8}$	$92.0_{\pm 5.6}$	$90.3_{\pm 6.5}$
JR	$23.4_{\pm 2.2}$	$37.3_{\pm 2.3}$	$20.0_{\pm 1.8}$	$19.1_{\pm 1.1}$	$67.3_{\pm 3.5}$	$88.2_{\pm 0.5}$	$66.6_{\pm 1.8}$	$57.7_{\pm 3.5}$	$93.1_{\pm 1.5}$	$97.4_{\pm 0.4}$	$93.5_{\pm 0.3}$	$91.9_{\pm 1.1}$
ER	$12.1_{\pm 2.7}$	$19.6_{\pm 2.4}$	$7.5_{\pm 0.5}$	$8.1_{\pm 1.5}$	$32.5_{\pm 9.3}$	$51.9_{\pm 9.5}$	$22.1_{\pm 4.5}$	$21.5_{\pm 5.5}$	$65.7_{\pm 17.9}$	$89.2_{\pm 5.2}$	$55.0_{\pm 8.4}$	$50.1_{\pm 11.9}$
SAM	$22.6_{\pm 2.6}$	$30.1_{\pm 0.5}$	$12.7_{\pm 0.5}$	$11.5_{\pm 0.1}$	$55.5_{\pm 1.9}$	$72.5_{\pm 3.1}$	$39.4_{\pm 3.4}$	$34.8_{\pm 3.0}$	$89.8_{\pm 1.0}$	$97.1_{\pm 0.7}$	$86.1_{\pm 10.3}$	$82.0_{\pm 13.8}$

An overall examination on all the training mechanisms



Examine the gradient similarity between all the training mechanisms:

- *SAM* improves gradient similarity towards every training solution, (compared to *ST*).
- Input regularizations (*IR*, *JR*) and adversarial training (*AT*) align with each other very well.

Observations on model smoothness:

- *IR*, *JR* yield better model smoothness.

Observations on transferability:

- *SAM perform* better than *IR* and *JR* in CIFAR-10, and worse in ImageNette.

SAM and input regularizations (*JR*, *IR*) are highly complementary!

Boosting adversarial transferability with *SAM&IR* and *SAM&JR*

Transfer attack against target models trained without distribution shift:

	Untargeted											
	4/255				8/255				16/255			
	VGG16	DenseNet121	MobileNetV2	Xception	VGG16	DenseNet121	MobileNetV2	Xception	VGG16	DenseNet121	MobileNetV2	Xception
ST	$11.2_{\pm 1.0}$	$17.5_{\pm 1.1}$	$7.6_{\pm 0.4}$	$8.7_{\pm 0.5}$	$28.7_{\pm 1.1}$	$42.7_{\pm 3.5}$	$18.3_{\pm 0.7}$	$18.5_{\pm 1.3}$	$61.7_{\pm 4.7}$	$81.4_{\pm 4.6}$	$49.3_{\pm 2.7}$	$44.1_{\pm 5.1}$
AT	$14.7_{\pm 2.1}$	$20.5_{\pm 1.5}$	$15.3_{\pm 1.3}$	$15.7_{\pm 0.9}$	$52.1_{\pm 6.3}$	$68.7_{\pm 2.9}$	$59.3_{\pm 7.1}$	$52.3_{\pm 5.1}$	$96.5_{\pm 1.5}$	$98.8_{\pm 0.4}$	$97.5_{\pm 0.5}$	$95.7_{\pm 0.9}$
IR	$21.0_{\pm 6.6}$	$31.7_{\pm 6.9}$	$18.5_{\pm 5.1}$	$19.5_{\pm 4.3}$	$62.5_{\pm 14.3}$	$85.6_{\pm 9.0}$	$65.2_{\pm 16.4}$	$58.4_{\pm 14.0}$	$94.7_{\pm 6.9}$	$99.5_{\pm 0.3}$	$95.7_{\pm 5.3}$	$94.6_{\pm 5.0}$
JR	$26.3_{\pm 2.1}$	$37.3_{\pm 0.7}$	$21.7_{\pm 1.3}$	$20.5_{\pm 1.3}$	$68.5_{\pm 2.1}$	$87.0_{\pm 0.2}$	$67.6_{\pm 3.6}$	$56.0_{\pm 5.6}$	$95.0_{\pm 1.4}$	$99.5_{\pm 0.1}$	$94.7_{\pm 1.1}$	$89.8_{\pm 3.0}$
ER	$13.2_{\pm 2.8}$	$20.9_{\pm 1.7}$	$8.9_{\pm 1.5}$	$9.6_{\pm 1.0}$	$33.3_{\pm 9.9}$	$51.1_{\pm 8.7}$	$20.7_{\pm 3.7}$	$20.5_{\pm 3.3}$	$67.5_{\pm 15.9}$	$89.3_{\pm 6.7}$	$54.5_{\pm 9.1}$	$50.1_{\pm 8.7}$
SAM	$22.8_{\pm 1.4}$	$29.7_{\pm 1.5}$	$12.8_{\pm 0.8}$	$12.8_{\pm 0.0}$	$58.7_{\pm 2.7}$	$72.8_{\pm 2.0}$	$42.3_{\pm 1.1}$	$34.7_{\pm 1.7}$	$91.4_{\pm 2.0}$	$97.1_{\pm 0.3}$	$81.0_{\pm 1.8}$	$75.3_{\pm 1.7}$
SAM&IR	$26.9_{\pm 5.5}$	$39.8_{\pm 4.8}$	$24.3_{\pm 4.7}$	$23.2_{\pm 4.0}$	$72.5_{\pm 10.9}$	$92.6_{\pm 2.8}$	$78.5_{\pm 9.5}$	$68.9_{\pm 10.3}$	$98.1_{\pm 2.1}$	$99.8_{\pm 0.2}$	$98.7_{\pm 1.5}$	$97.7_{\pm 2.3}$
SAM&JR	$32.7_{\pm 4.3}$	$46.3_{\pm 3.7}$	$28.1_{\pm 4.5}$	$24.9_{\pm 4.1}$	$76.6_{\pm 5.6}$	$93.6_{\pm 1.2}$	$73.1_{\pm 7.7}$	$65.9_{\pm 7.7}$	$96.7_{\pm 7.3}$	$99.8_{\pm 0.2}$	$96.6_{\pm 1.6}$	$93.9_{\pm 4.1}$

Transfer attack against target models trained with distribution shift:

	AT, $\epsilon = 0.01$		*AT*, $\epsilon = 0.05$		*CM*		*CO*		*LS*		*MU*	
	T	U	T	U	T	U	T	U	T	U	T	U
ST	20.18	51.33	4.83	22.49	42.42	79.49	38.80	76.33	34.62	72.34	37.46	69.25
★	43.01	76.96	35.54	69.56	14.20	54.57	32.91	79.11	11.90	46.66	15.81	54.43
IR	60.08	91.23	31.84	73.98	59.80	91.82	60.78	92.14	63.20	93.52	60.79	90.06
JR	57.24	83.54	14.57	41.77	73.44	92.47	74.77	93.03	72.27	92.30	68.34	86.90
ER	32.61	77.14	9.55	40.41	41.94	86.92	43.04	87.39	44.51	87.77	42.12	82.66
SAM	40.35	79.31	8.41	35.05	75.34	96.59	67.72	94.84	62.68	93.57	65.37	90.43
SAM&IR	**71.88**	**94.01**	**48.38**	**80.88**	70.81	93.75	73.05	94.31	74.02	95.47	70.81	92.68
SAM&JR	58.81	89.23	15.11	47.11	**83.56**	**97.57**	**80.19**	**97.02**	**77.95**	**96.92**	**75.06**	**94.49**

Transfer attack against 3 MLaaS commercial platforms:

Model	Untargeted						Targeted					
	AWS		Baidu		Aliyun		AWS		Baidu		Aliyun	
	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{8}{255}$	$\frac{16}{255}$	$\frac{8}{255}$	$\frac{16}{255}$
ST	9.4	13.4	30.0	53.8	18.6	52.2	23.0	23.8	11.6	16.6	2.6	6.4
MU	8.6	10.8	20.4	39.6	12.4	33.8	20.4	23.8	9.6	15.0	1.4	2.8
CM	8.0	10.8	19.4	35.0	12.4	29.2	22.2	23.0	10.0	11.2	1.2	3.0
CO	9.6	13.0	27.2	48.4	18.6	51.8	23.2	26.4	14.0	16.8	2.0	6.8
LS	9.6	12.0	23.4	43.0	12.2	33.6	22.2	23.0	12.6	14.4	1.4	2.0
AT, $\epsilon = 1$	9.6	24.0	38.2	60.4	35.0	82.6	24.8	33.2	14.8	27.6	3.6	13.8
AT, $\epsilon = 5$	7.8	16.0	22.2	53.4	15.6	60.0	22.0	27.6	11.0	17.0	1.2	5.0
IR	10.4	23.6	45.8	60.0	48.8	86.0	25.0	34.4	17.2	26.4	8.0	15.4
JR	9.8	22.6	46.4	60.6	48.2	79.6	26.8	32.0	17.2	25.0	7.8	12.4
ER	9.8	13.4	28.2	56.2	24.0	56.6	23.8	26.4	14.2	20.0	3.2	8.0
SAM	10.2	16.4	37.2	59.2	36.2	68.2	23.4	30.0	13.8	19.2	4.6	9.8
SAM&IR	11.0	28.4	54.2	63.8	55.8	93.4	29.2	36.2	16.8	28.0	9.8	18.0
SAM&JR	11.4	16.2	45.8	61.2	48.0	79.0	25.4	33.2	18.8	21.8	8.2	11.8

In all these scenarios, the best surrogate is either *SAM&IR* or *SAM&JR*.

A good surrogate is better than good AE generation methods

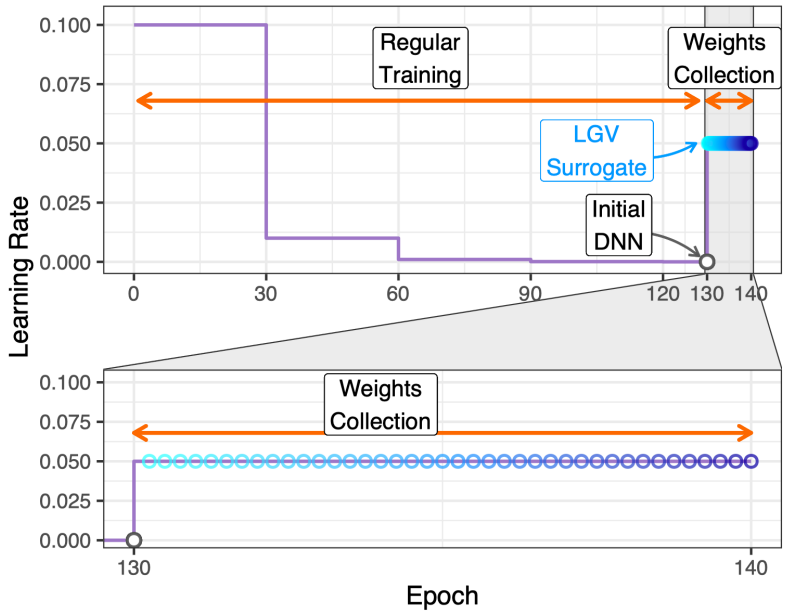
MI, DIM are two most representative transferable AE generation methods.

The results show that:

- *SAM +* MI > *SAM, SAM +* DMI > *SAM*; *SAM&JR* + MI > *SAM&JR*, *SAM&JR* + DMI > *SAM&JR*.
  Good surrogates perform better with better generation methods.
- *SAM, SAM&JR* > *ST*+MI, *ST*+DIM.
  Bad surrogates with better generation methods still underperform good surrogates .

	Untargeted								**Targeted**							
	4/255				8/255				4/255				8/255			
	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121
ST	$55.0_{\pm 2.4}$	$43.3_{\pm 2.1}$	$54.9_{\pm 2.5}$	$71.7_{\pm 1.2}$	$79.6_{\pm 2.6}$	$71.3_{\pm 3.2}$	$80.2_{\pm 3.8}$	$93.1_{\pm 1.3}$	$19.0_{\pm 1.2}$	$13.2_{\pm 1.0}$	$21.1_{\pm 1.5}$	$36.4_{\pm 1.8}$	$41.1_{\pm 4.7}$	$34.3_{\pm 3.7}$	$45.8_{\pm 4.2}$	$70.8_{\pm 4.6}$
SAM	$76.8_{\pm 4.7}$	$65.0_{\pm 5.3}$	$74.9_{\pm 0.6}$	$87.4_{\pm 1.2}$	$97.5_{\pm 1.2}$	$94.3_{\pm 2.2}$	$97.0_{\pm 0.2}$	$99.5_{\pm 0.1}$	$37.1_{\pm 6.4}$	$27.2_{\pm 5.0}$	$38.5_{\pm 2.0}$	$55.8_{\pm 3.7}$	$77.2_{\pm 9.2}$	$68.3_{\pm 8.4}$	$78.4_{\pm 2.7}$	$93.4_{\pm 0.9}$
SAM&JR	$\mathbf{81.2}_{\pm 0.7}$	$\mathbf{70.2}_{\pm 0.7}$	$\mathbf{79.4}_{\pm 0.6}$	$\mathbf{91.3}_{\pm 0.1}$	$\mathbf{98.7}_{\pm 0.3}$	$\mathbf{96.6}_{\pm 0.1}$	$\mathbf{98.3}_{\pm 0.0}$	$\mathbf{99.8}_{\pm 0.0}$	$\mathbf{43.6}_{\pm 2.1}$	$\mathbf{32.9}_{\pm 1.2}$	$\mathbf{45.3}_{\pm 2.1}$	$\mathbf{64.5}_{\pm 0.9}$	$\mathbf{85.4}_{\pm 1.4}$	$\mathbf{77.2}_{\pm 9.2}$	$\mathbf{86.8}_{\pm 1.9}$	$\mathbf{97.2}_{\pm 0.3}$
ST+MI	$58.4_{\pm 2.6}$	$45.0_{\pm 2.9}$	$58.7_{\pm 4.0}$	$76.7_{\pm 2.3}$	$82.9_{\pm 3.1}$	$75.6_{\pm 4.1}$	$83.5_{\pm 3.6}$	$94.6_{\pm 1.2}$	$22.4_{\pm 1.8}$	$15.4_{\pm 1.9}$	$25.6_{\pm 2.8}$	$46.0_{\pm 3.9}$	$44.8_{\pm 4.1}$	$40.2_{\pm 5.9}$	$50.4_{\pm 6.3}$	$76.1_{\pm 4.5}$
SAM+MI	$82.3_{\pm 3.9}$	$69.6_{\pm 5.3}$	$80.5_{\pm 0.1}$	$91.7_{\pm 1.5}$	$98.1_{\pm 0.9}$	$95.5_{\pm 1.9}$	$97.8_{\pm 0.1}$	$99.7_{\pm 0.1}$	$46.7_{\pm 7.0}$	$33.2_{\pm 5.7}$	$49.0_{\pm 1.1}$	$69.6_{\pm 3.8}$	$84.0_{\pm 6.9}$	$76.9_{\pm 5.8}$	$84.8_{\pm 1.2}$	$96.4_{\pm 0.9}$
SAM&JR+MI	$\mathbf{85.0}_{\pm 0.4}$	$\mathbf{73.8}_{\pm 0.5}$	$\mathbf{83.3}_{\pm 0.5}$	$\mathbf{93.6}_{\pm 0.1}$	$\mathbf{98.9}_{\pm 0.2}$	$\mathbf{97.2}_{\pm 0.2}$	$\mathbf{98.7}_{\pm 0.1}$	$\mathbf{99.8}_{\pm 0.1}$	$\mathbf{52.4}_{\pm 2.0}$	$\mathbf{38.4}_{\pm 1.5}$	$\mathbf{54.0}_{\pm 1.7}$	$\mathbf{74.5}_{\pm 0.4}$	$\mathbf{89.6}_{\pm 1.6}$	$\mathbf{82.9}_{\pm 1.5}$	$\mathbf{90.6}_{\pm 2.1}$	$\mathbf{98.3}_{\pm 0.3}$
ST+DIM	$70.2_{\pm 3.8}$	$59.2_{\pm 2.7}$	$69.2_{\pm 3.5}$	$83.6_{\pm 3.1}$	$91.9_{\pm 2.2}$	$88.3_{\pm 2.8}$	$91.7_{\pm 2.2}$	$97.2_{\pm 1.3}$	$33.6_{\pm 4.8}$	$25.7_{\pm 3.2}$	$36.6_{\pm 4.2}$	$54.9_{\pm 6.3}$	$65.4_{\pm 6.3}$	$62.0_{\pm 6.2}$	$69.9_{\pm 5.3}$	$87.5_{\pm 5.3}$
SAM+DIM	$\mathbf{84.4}_{\pm 2.6}$	$74.7_{\pm 2.7}$	$\mathbf{82.2}_{\pm 1.8}$	$91.8_{\pm 2.2}$	$98.9_{\pm 0.4}$	$97.7_{\pm 0.7}$	$\mathbf{99.1}_{\pm 0.8}$	$99.7_{\pm 0.2}$	$50.4_{\pm 4.6}$	$38.9_{\pm 3.2}$	$51.3_{\pm 4.2}$	$67.8_{\pm 5.8}$	$89.3_{\pm 3.5}$	$84.2_{\pm 2.2}$	$89.6_{\pm 3.0}$	$96.8_{\pm 1.9}$
SAM&JR+DIM	$83.6_{\pm 1.6}$	$\mathbf{74.8}_{\pm 1.0}$	$81.8_{\pm 1.4}$	$\mathbf{92.1}_{\pm 0.6}$	$\mathbf{99.2}_{\pm 0.2}$	$\mathbf{98.2}_{\pm 0.3}$	$99.0_{\pm 0.1}$	$\mathbf{99.8}_{\pm 0.1}$	$\mathbf{51.6}_{\pm 3.9}$	$\mathbf{40.7}_{\pm 2.4}$	$\mathbf{52.7}_{\pm 2.9}$	$\mathbf{70.2}_{\pm 2.0}$	$\mathbf{91.5}_{\pm 2.9}$	$\mathbf{86.9}_{\pm 2.8}$	$\mathbf{91.8}_{\pm 2.6}$	$\mathbf{98.1}_{\pm 0.8}$

A good surrogate is even better than an ensemble of diverse surrogates



LGV (Large Geometric Vicinity): obtaining a sufficient amount of neighbors with standard SGD, then iteratively attacking them.

Our experimental results suggest that:

- More neighbors may ***harm*** the transferability.
- More **"good"** neighbors are preferable.

Transfer ASRs under different surrogates w/wo LGV and superior fine-tuning mechanisms.

	Untargeted								**Targeted**							
	4/255				8/255				4/255				8/255			
	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121	Res-50	VGG16	Inc-V3	Dense-121
ST	55.0$_{\pm 2.4}$	43.3$_{\pm 2.1}$	54.9$_{\pm 2.5}$	71.7$_{\pm 1.2}$	79.6$_{\pm 2.6}$	71.3$_{\pm 3.2}$	80.2$_{\pm 3.8}$	93.1$_{\pm 1.3}$	19.0$_{\pm 1.2}$	13.2$_{\pm 1.0}$	21.1$_{\pm 1.5}$	36.4$_{\pm 1.8}$	41.1$_{\pm 4.7}$	34.3$_{\pm 3.7}$	45.8$_{\pm 4.2}$	70.8$_{\pm 4.6}$
ST+LGV$_{\text{SGD}}$	70.7$_{\pm 2.2}$	57.1$_{\pm 1.1}$	70.1$_{\pm 0.9}$	87.1$_{\pm 0.6}$	94.0$_{\pm 2.0}$	89.1$_{\pm 2.5}$	94.3$_{\pm 1.8}$	99.3$_{\pm 0.2}$	29.3$_{\pm 2.0}$	20.4$_{\pm 1.3}$	32.8$_{\pm 2.1}$	53.4$_{\pm 2.1}$	65.6$_{\pm 4.2}$	56.8$_{\pm 3.2}$	71.3$_{\pm 2.5}$	92.0$_{\pm 1.2}$
ST+LGV$_{\textit{SAM}}$	79.5$_{\pm 0.8}$	66.8$_{\pm 1.1}$	79.8$_{\pm 1.0}$	92.3$_{\pm 0.2}$	98.6$_{\pm 0.2}$	96.0$_{\pm 0.5}$	98.7$_{\pm 0.2}$	**99.9**$_{\pm 0.0}$	39.4$_{\pm 1.0}$	27.7$_{\pm 0.5}$	44.4$_{\pm 1.7}$	65.7$_{\pm 0.9}$	82.5$_{\pm 1.7}$	74.2$_{\pm 1.0}$	87.4$_{\pm 0.8}$	97.8$_{\pm 0.2}$
ST+LGV$_{\textit{SAM\&JR}}$	**82.0**$_{\pm 1.5}$	**69.5**$_{\pm 1.7}$	**81.3**$_{\pm 1.3}$	**93.4**$_{\pm 0.3}$	**99.1**$_{\pm 0.3}$	**97.1**$_{\pm 0.7}$	**99.0**$_{\pm 0.2}$	**99.9**$_{\pm 0.0}$	**42.9**$_{\pm 2.6}$	**30.3**$_{\pm 1.5}$	**46.7**$_{\pm 1.4}$	**68.1**$_{\pm 1.0}$	**87.1**$_{\pm 3.5}$	**79.1**$_{\pm 3.4}$	**90.2**$_{\pm 1.7}$	**98.5**$_{\pm 0.3}$
SAM	76.8$_{\pm 4.7}$	65.0$_{\pm 5.3}$	74.9$_{\pm 0.6}$	87.4$_{\pm 1.2}$	97.5$_{\pm 1.2}$	94.3$_{\pm 2.2}$	97.0$_{\pm 0.2}$	99.5$_{\pm 0.1}$	37.1$_{\pm 6.4}$	27.2$_{\pm 5.0}$	38.5$_{\pm 2.0}$	55.8$_{\pm 3.7}$	77.2$_{\pm 9.2}$	68.3$_{\pm 8.4}$	78.4$_{\pm 2.7}$	93.4$_{\pm 0.9}$
SAM+LGV$_{\text{SGD}}$	70.8$_{\pm 6.1}$	57.2$_{\pm 6.5}$	69.2$_{\pm 0.7}$	84.2$_{\pm 0.7}$	93.1$_{\pm 3.9}$	87.6$_{\pm 5.3}$	93.2$_{\pm 1.4}$	98.6$_{\pm 0.3}$	26.7$_{\pm 1.9}$	18.7$_{\pm 0.1}$	31.5$_{\pm 2.2}$	51.3$_{\pm 2.6}$	63.6$_{\pm 14.1}$	53.6$_{\pm 11.8}$	66.4$_{\pm 5.4}$	87.6$_{\pm 2.2}$
SAM+LGV$_{\textit{SAM}}$	81.5$_{\pm 4.0}$	69.5$_{\pm 4.8}$	79.8$_{\pm 1.3}$	91.6$_{\pm 1.3}$	98.9$_{\pm 0.7}$	96.9$_{\pm 1.3}$	98.7$_{\pm 0.1}$	**99.9**$_{\pm 0.0}$	42.1$_{\pm 7.3}$	30.5$_{\pm 5.5}$	44.3$_{\pm 1.6}$	63.4$_{\pm 1.9}$	86.2$_{\pm 6.7}$	78.4$_{\pm 6.6}$	87.9$_{\pm 1.1}$	97.5$_{\pm 0.4}$
SAM+LGV$_{\textit{SAM\&JR}}$	**82.7**$_{\pm 4.2}$	**71.2**$_{\pm 5.1}$	**80.9**$_{\pm 0.6}$	**92.1**$_{\pm 1.2}$	**99.1**$_{\pm 0.6}$	**97.4**$_{\pm 1.3}$	**99.0**$_{\pm 0.3}$	**99.9**$_{\pm 0.0}$	**44.7**$_{\pm 7.7}$	**32.6**$_{\pm 5.9}$	**46.6**$_{\pm 0.6}$	**65.6**$_{\pm 1.7}$	**87.6**$_{\pm 6.9}$	**79.8**$_{\pm 7.1}$	**89.1**$_{\pm 2.0}$	**97.9**$_{\pm 0.5}$
SAM&JR	81.2$_{\pm 0.7}$	70.2$_{\pm 0.7}$	79.4$_{\pm 0.6}$	91.3$_{\pm 0.1}$	98.7$_{\pm 0.3}$	96.6$_{\pm 0.1}$	98.3$_{\pm 0.0}$	99.8$_{\pm 0.0}$	43.6$_{\pm 2.1}$	32.9$_{\pm 1.2}$	45.3$_{\pm 2.1}$	64.5$_{\pm 0.9}$	85.4$_{\pm 1.4}$	77.2$_{\pm 2.1}$	86.8$_{\pm 1.9}$	97.2$_{\pm 0.3}$
SAM&JR+LGV$_{\text{SGD}}$	73.8$_{\pm 1.8}$	61.1$_{\pm 1.6}$	73.1$_{\pm 1.8}$	88.3$_{\pm 0.5}$	95.8$_{\pm 1.1}$	90.9$_{\pm 1.9}$	95.6$_{\pm 1.1}$	99.4$_{\pm 0.1}$	35.6$_{\pm 2.9}$	25.1$_{\pm 2.4}$	39.0$_{\pm 2.7}$	60.2$_{\pm 1.7}$	70.5$_{\pm 5.0}$	61.1$_{\pm 5.9}$	74.5$_{\pm 5.7}$	93.6$_{\pm 1.5}$
SAM&JR+LGV$_{\textit{SAM}}$	82.5$_{\pm 0.7}$	71.2$_{\pm 1.0}$	81.0$_{\pm 0.4}$	93.0$_{\pm 0.2}$	99.1$_{\pm 0.1}$	97.4$_{\pm 0.2}$	98.9$_{\pm 0.1}$	**99.9**$_{\pm 0.1}$	44.0$_{\pm 1.5}$	32.2$_{\pm 1.1}$	46.8$_{\pm 1.4}$	67.6$_{\pm 0.8}$	87.6$_{\pm 1.2}$	80.2$_{\pm 1.7}$	89.7$_{\pm 1.2}$	98.4$_{\pm 0.1}$
SAM&JR+LGV$_{\textit{SAM\&JR}}$	**83.8**$_{\pm 0.8}$	**72.9**$_{\pm 1.2}$	**82.5**$_{\pm 0.9}$	**93.7**$_{\pm 0.1}$	**99.3**$_{\pm 0.1}$	**97.8**$_{\pm 0.3}$	**99.3**$_{\pm 0.1}$	**99.9**$_{\pm 0.0}$	**46.9**$_{\pm 1.4}$	**34.8**$_{\pm 1.2}$	**49.5**$_{\pm 1.0}$	**70.0**$_{\pm 0.5}$	**90.0**$_{\pm 0.9}$	**83.2**$_{\pm 0.7}$	**91.8**$_{\pm 0.7}$	**98.8**$_{\pm 0.1}$

An extensive examination of conclusions in the literature

TABLE 1: The overview of our interactions with literature in the field.

Existing conclusions and viewpoints	**Our observations and inferences**	**Relation**
Stronger regularized (smoother) models provide better surrogates on average [9].	(1) *AT* with large budget yields smoother models that degrade transferability. In Sec. 2.1. (2) Stronger regularizations cannot always outperform less smooth solutions like *SAM*. In Sec. 4.3.	Partly conflicting
AT and data augmentation do not show strong correlations to transfer attacks in the "real-world " environment [40].	(1) *AT* with small budget benefits transfer attack while large budget hinders it. (2) Data augmentation generally impairs transfer attacks, especially for stronger augmentations. In Sec. 6, **Q6**.	Conflicting
Surrogate models with better generalization behavior could result in more transferable AEs [60].	Data augmentations that yield surrogates with the best generalization perform the worst in transfer attacks. In Sec. 6, **Q4**.	Conflicting
Attacking multiple surrogates from a sufficiently large geometry vicinity (LGV) benefits transferability [22].	Attacking multiple surrogates from arbitrary LGV of a single superior surrogate may degrade transferability. In Sec. 5.2.	Partly conflicting
Regularizing pressure transfers from the weight space to the input space. [11].	This transfer effect exists, yet is marginal and unstable. In Sec. 4.2.	Partly conflicting
The poor transferability of ViT is because existing attacks are not strong enough to fully exploit its potential [44].	The transferability of ViT may have been restrained by its default training paradigm. In Sec. 6, **Q5**	Parallel
Model complexity (the number of local optima in loss surface) correlates with transferability [9].	A smoother model is expected to have less and wider local optima in a finite space. In Sec. 6, **Q3**	Causal
AEs lie off the underlying manifold of clean data [20].	Adversarial training causes data distribution shift induced by off-manifold AEs, thus impairing gradient similarity.	Dependent
(1) Attacking an ensemble of surrogates in the distribution found by *Bayes* learning improves the transferability [32]. (2) *SAM* can be seen as a relaxation of Bayes [42].	*SAM* yields general input gradient alignment towards every training solution. Attacking *SAM* solution significantly improves transferability. In Sec. 6, **Q7**.	Matching

Summary

- We investigate the complex trade-off between model smoothness and gradient similarity under various training mechanisms.
- We propose a general method for boosting adversarial transferability via training superior surrogates.
- We present a series of conclusions regarding adversarial transferability.