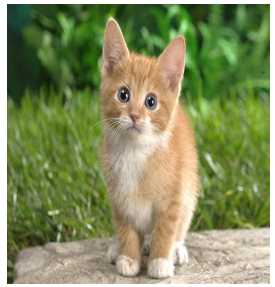
On Success and Simplicity: A Second Look at Transferable Targeted Attacks

Zhengyu Zhao, Zhuoran Liu, Martha Larson
Radboud University, Netherlands

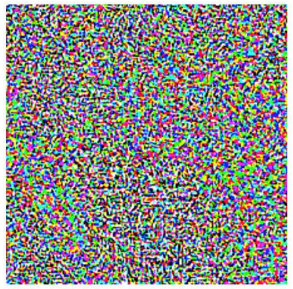
Non-targeted vs. targeted adversarial images

Radboud University

Tiger cat √

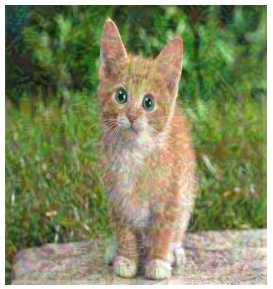


+



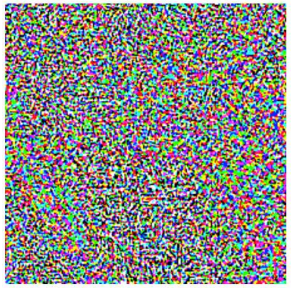
=

Persian cat ×



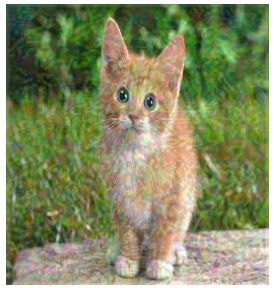
Non-targeted: any wrong class (relevant class is sufficient)

+



=

Airplane ×



Targeted: specific class (could be highly irrelevant)

Transferability of targeted adversarial images

Radboud University

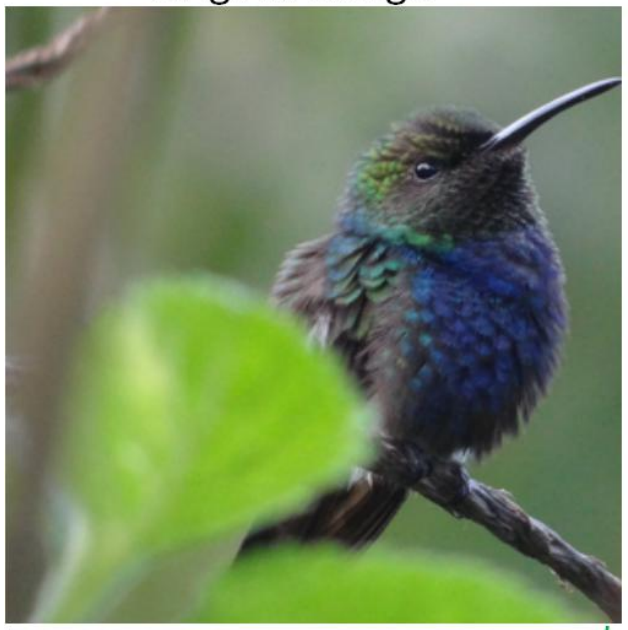
Source model (white box) : ResNet50

Target model (black box) : DenseNet121, VGG16, Inception-v3

Originl class: “hummingbird” Target class: “coffee mug”

X

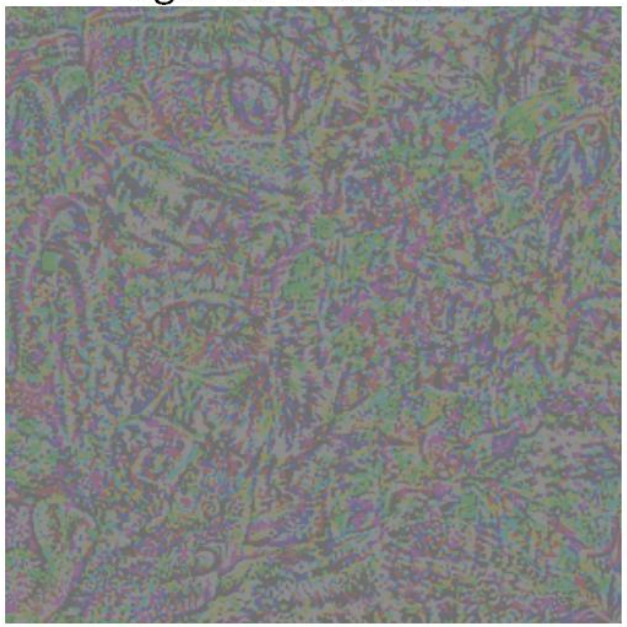
Original image



ResNet50: “hummingbird” √

+

Perturbation optimized against **ResNet50**



$L_\infty = 16/255$

=

$\dot{X}$

Transferable Adversarial image

ResNet50: “coffee mug” ×

DenseNet121: “coffee mug” ×

VGG16: “coffee mug” ×

Inception-v3: “coffee mug” ×

Existing targeted transfer methods

Radboud University

- Simple methods: (reputed to be) insufficient.
  - Gradient accumulation (MI[1], NI[2])
  - Data augmentation (TI[3], DI[4])

- Resource-intensive methods: SOTA.
  - Training *target-class-specific* classifiers (FDA[5,6])
  - Training *target-class-specific* generators (CDA[7], TTP[8])

1. Dong et al. *Boosting Adversarial Attacks with Momentum*. CVPR'18.
2. Lin et al. *Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks*. ICLR'20
3. Dong et al. *Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks*. CVPR'19
4. Xie et al. *Improving Transferability of Adversarial Examples with Input Diversity*. CVPR'19
5. Inkawhich et al. *Transferable Perturbations of Deep Feature Distributions*. ICLR'20
6. Inkawhich et al. *Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability*. NeurIPS'20
7. Naseer et al. *Cross-Domain Transferability of Adversarial Perturbations*. NeurIPS'19
8. Naseer et al. *On Generating Transferable Targeted Perturbation*. ICCV'21

Main message

Radboud University

Previous research: Simple methods << resource-intensive methods

Our investigation: Simple methods > resource-intensive methods

Transfer success rates (%)

Bound	Attack	D121	V16	D121-ens	V16-ens
$\epsilon = 16$	TTP [8]	**79.6**	**78.6**	92.9	89.6
	ours	75.9	72.5	**99.4**	**97.7**
$\epsilon = 8$	TTP [8]	37.5	46.7	63.2	66.2
	ours	**44.5**	**46.8**	**92.6**	**87.0**

8. Naseer et al. *On Generating Transferable Targeted Perturbation*. ICCV'21

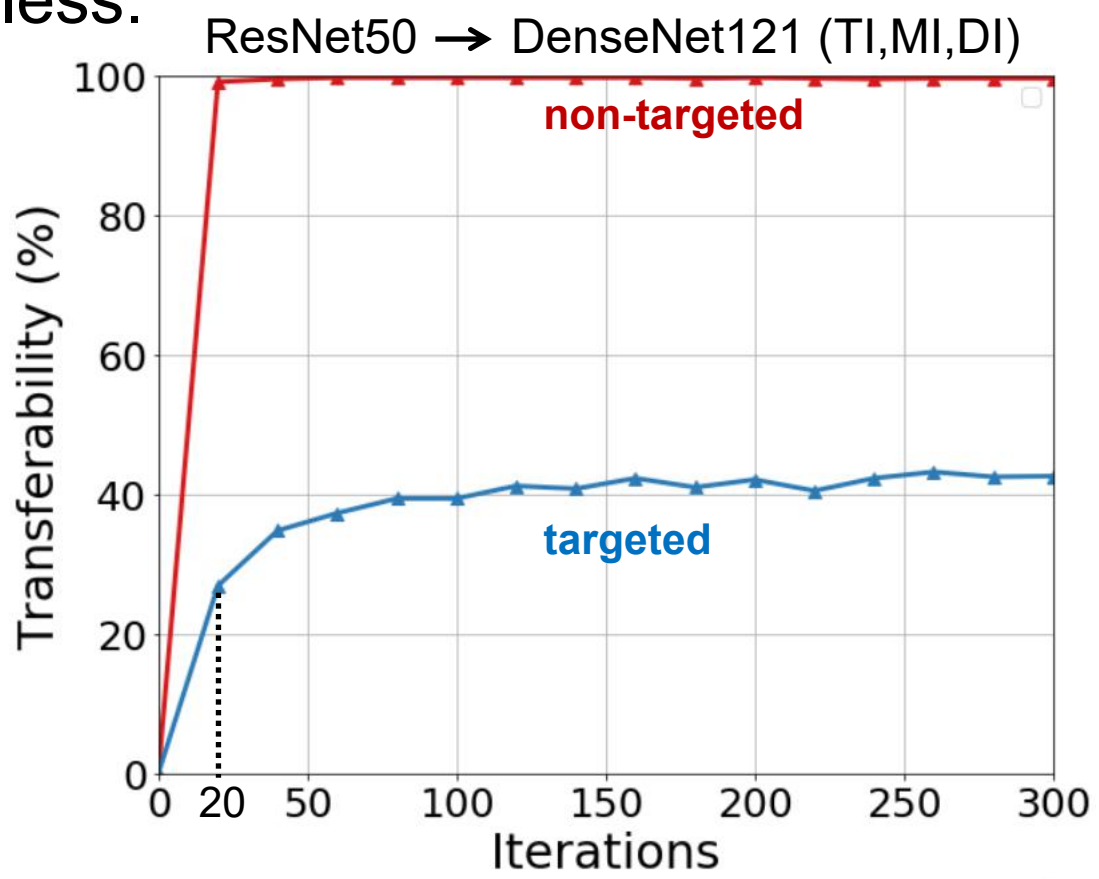
New insights into simple methods

Radboud University

1. Targeted transferability requires more iterations to converge.

→ Unreasonable evaluation (only <20 iterations).

- optimization perspective: meaningless.
- practical perspective: unrealistic.



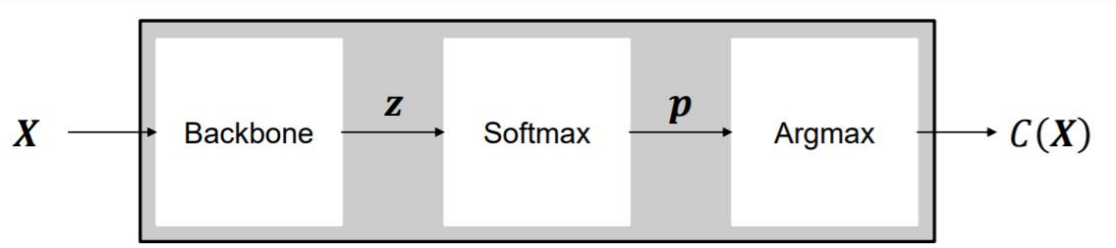
change perspective

New insights into simple methods

Radboud University

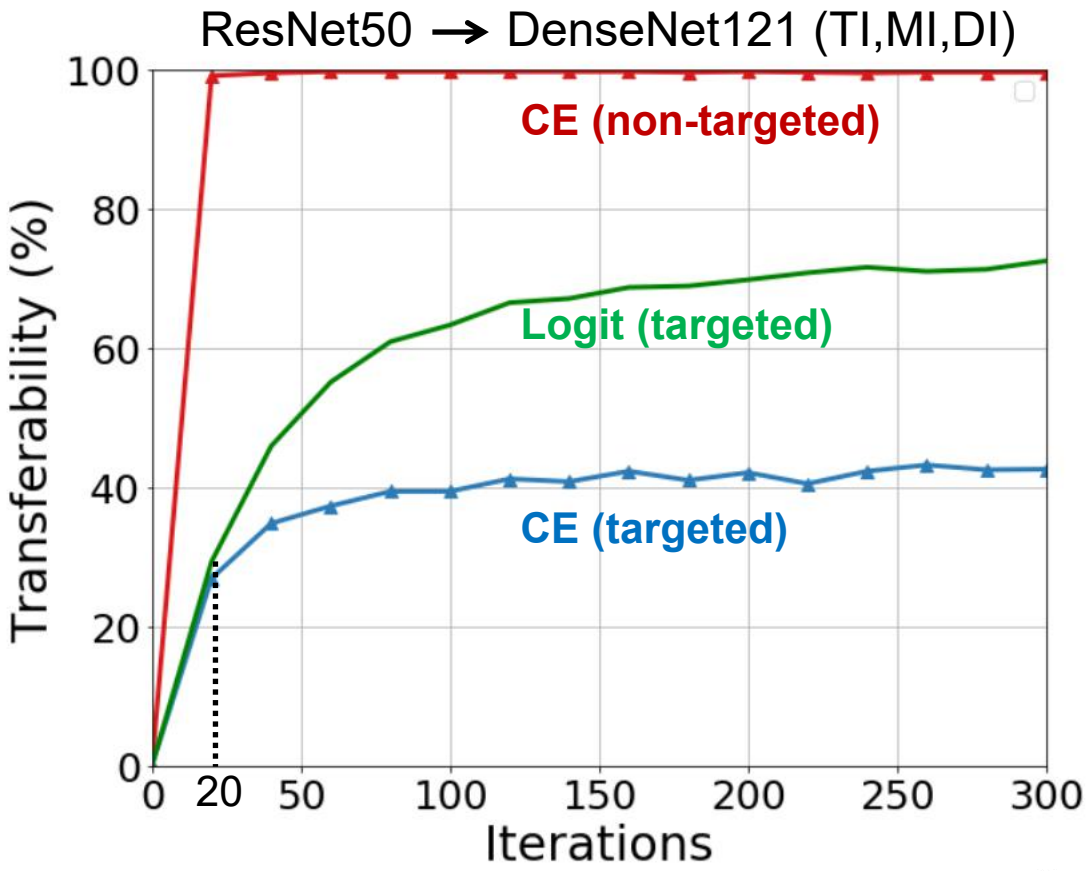
2. **Cross-Entropy (CE) loss** causes decreasing gradient problem.

→ We use a naive **Logit loss** (not novel but its advantage has not been recognized so far).



$$\textit{CE Loss} = -\log\left(\frac{e^{z_y}}{\sum_j e^{z_j}}\right) = -z_y + \log\sum_j e^{z_j}$$

$$\textit{Logit Loss} = -z_y$$



New realistic transfer scenarios

Radboud University

1. Ensemble transfer scenario with **low model similarity**.

2. Worse-case transfer scenario with **low-ranked targets**.

3. Transfer scenario on a **real-world system,** Google Cloud Vision API.

Scenario 1: ensemble transfer with low model similarity

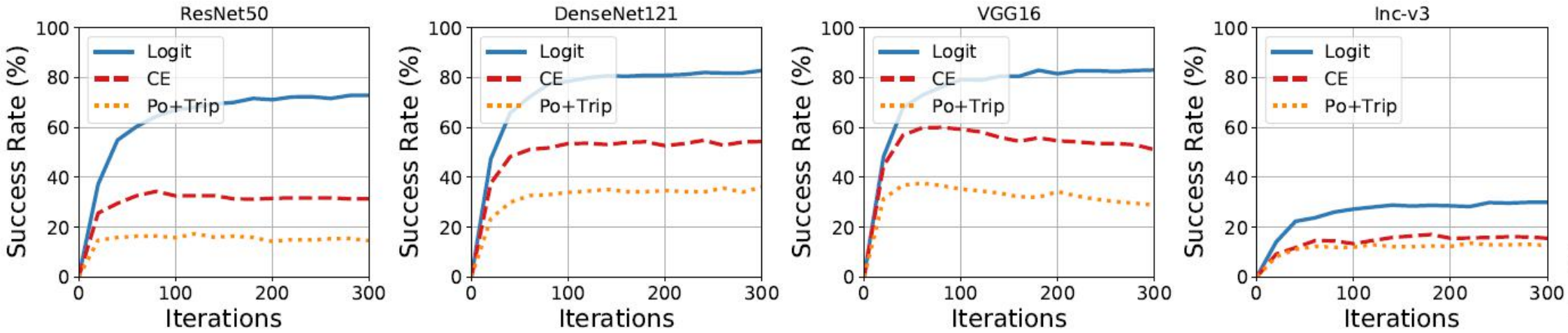
Radboud University

Equally high performance in ensemble transfer with high model similarity.

Attack	-Inc-v3	-Inc-v4	-IncRes-v2	-Res50	-Res101	-Res152	Average
CE	48.8/85.3	47.2/83.3	47.5/83.9	50.9/89.8	58.5/**93.2**	56.7/90.7	51.6/87.7
Po+Trip	**59.3**/84.4	**55.0**/82.4	51.4/80.8	56.9/85.0	60.5/87.9	57.6/85.7	56.8/84.4
Logit	56.4/**85.5**	52.9/**85.8**	**54.4/85.1**	**57.5/90.0**	**64.4**/91.4	**61.3/90.8**	**57.8/88.1**

↓

Logit loss largerly outperforms the others in ensemble transfer with low model similarity.



Scenario 2: worse case with low-ranked target classes

Radboud University

Targeted transfer is harder for lower-ranked target classes.

Attack	2nd	10th	200th	500th	800th	1000th
CE	**89.9**	76.7	49.7	43.1	37.0	25.1
Po+Trip	82.6	77.6	58.4	53.6	49.1	38.2
Logit	83.8	**81.3**	**75.0**	**71.0**	**65.1**	**52.8**

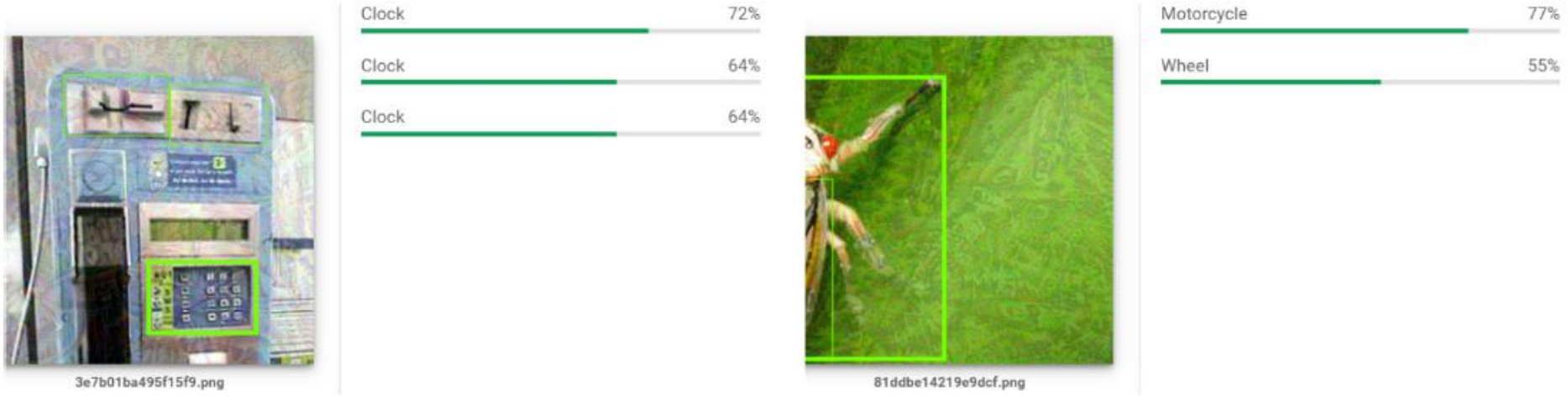
Scenario 3: real-world attack on Google Cloud Vision API

Radboud University

Logit achieves substantial success rates (%).

	CE	Po+Trip	Logit
Targeted	7	8	**18**
Non-targeted	**51**	44	**51**

Successful targeted adversarial images.



Target class: “analog clock”

Target class: “mountain bike”

Three future directions

Radboud University

Finding: Transferability on specific models (Inception) are very low.
→ 1. Understanding influence of model architectures on transferability.

Finding: Robust models may have different transfer properties.
→ 2. Exploring targeted transferability on robust models.

Finding: Simple and resource-intensive methods have different merits.
→ 3. Conducting a comprehensive comparison between these two types.

Radboud University

Thank you!