Adversarially Robust Models may not Transfer Better: Sufficient Conditions for Domain Transferability from the View of Regularization

Xiaojun Xu*, Jacky Yibo Zhang*, Evelyn Ma, Danny Son, Oluwasanmi Koyejo, Bo Li

University of Illinois at Urbana-Champaign; * Equal contribution

ILLINOIS

Motivation

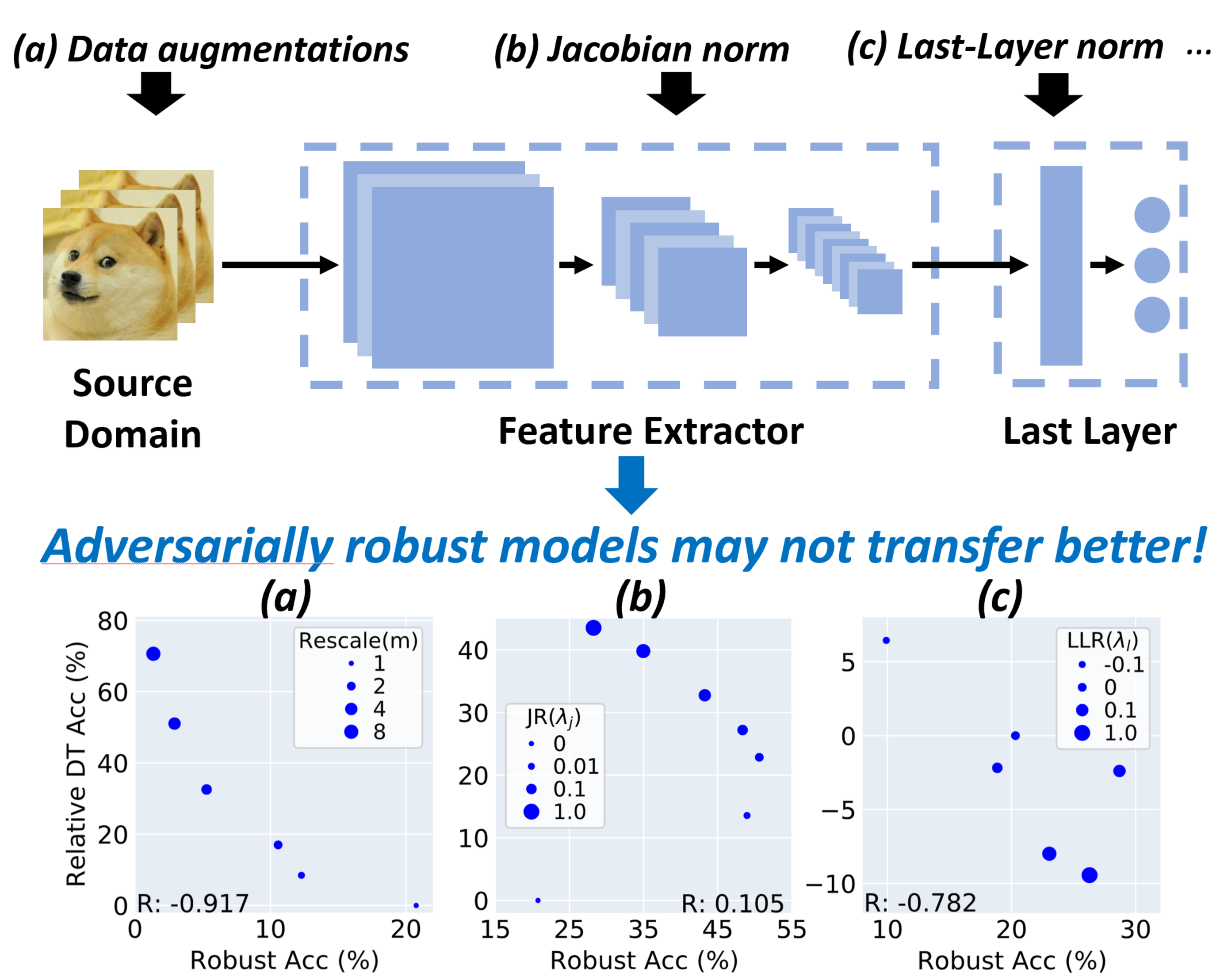
It is observed that adversarially robust models transfer better [1, 2].

Questions

- **(Q1)** Is it really that adversarially robust models can transfer better?
- **(Q2)** If not, what properties affect domain transferability more than robustness?
- **(Q3)** How to explain their empirical findings?

[1] Salman et al. "Do adversarially robust imagenet models transfer better?." NeurIPS 2020.
[2] Utrera et al. "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification." ICLR 2021.

Adversarially Robust Model may not Transfer Better (Answer to Q1)



Theoretical Result

Improving adversarial robustness is neither necessary nor sufficient for improving domain transferability!

Empirical Result

More robust models may even transfer worse!

Regularization Affects Domain Transferability (Answer to Q2&3)

High-level Idea of Theoretical Analysis

- We define a novel *pseudometric* to characterize the distance between two distributions.
- We formally define the *relative domain transfer loss*. The *smaller* the loss, the *better* the relative domain transferability.
- With the two key definitions, we prove:

Sketch of Theorems.

Shrinking the function class of the source model will decrease a tight upper bound on the relative domain transferability loss.

Q2 Answer: It is expected that stronger regularization during source model training leads to better relative domain transferability (target domain performance relative to source domain performance).

Q3 Answer: adversarial training => training with regularization => better transferability.

Data Augmentations (DAs) as Regularization

Data augmentations can be viewed as regularizations, and thus improving domain transferability.

- **Can be viewed as Regularization**: Adversarial training, Gaussian blur, rescale, etc.
- **Cannot be viewed as Regularization**: Rotation, Translation, etc.
- More analysis in our paper!

Empirical Evaluation Settings

Pipeline

Step 1: Train $g_s \circ f$ on the source domain.
Step 2: Fix f and finetune $g_t \circ f$ on the target domain.
Domain pairs: (CIFAR-10 -> SVHN) and (ImageNet -> CIFAR-10).

Metrics

Relative domain transfer accuracy:

Substract the value on vanilla model (constant) so that the comparison can be shown.

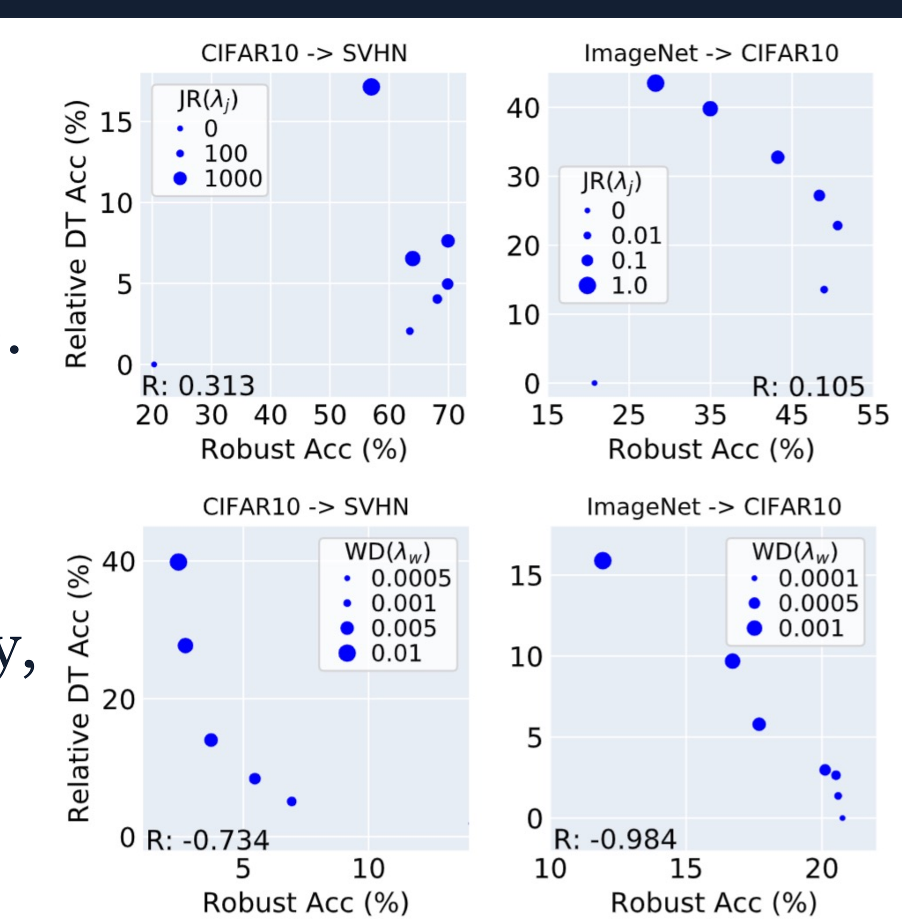
$$\text{DT Acc} = (acc_{tgt} - acc_{src}) - \boxed{(acc_{tgt}^v - acc_{src}^v)}$$

Robust Accuracy: accuracy under PGD attack ($\ell_2, \epsilon = 0.25$, 20 steps) on source domain.

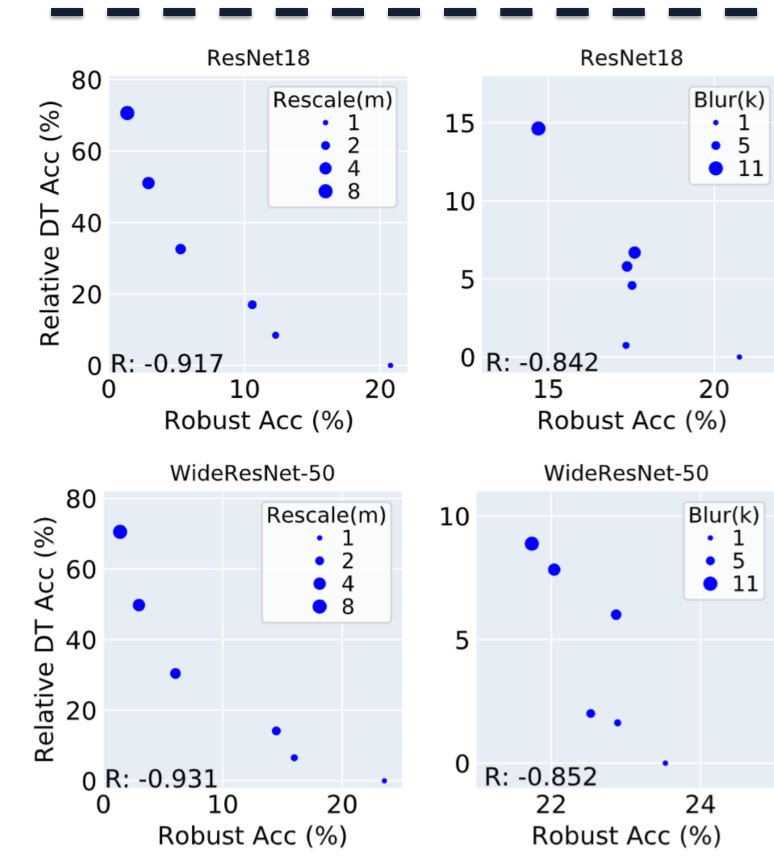
Empirical Evaluation

Impact of Regularizations

Jacobian Regularization (JR) with λ_j.
Weight Decay (WD) with λ_w.



Conclusion: stronger regularizer leads to better domain transferability, while robustness does not improve.



Impact of Data Augmentations

Rescaling: rescale to m times smaller.
Blurring: Gaussian blur with kernel size k.

Conclusion: stronger augmentation leads to better domain transferability, while robustness drops.

More Results (see our paper):

- Other regularizations (orthogonal training, last-layer regularizing).
- Other augmentations (Gaussian blurring, posterizing).
- DAs that cannot be viewed as regularization (rotation, translation)
- Results of absolute DT accuracy and other model architectures.