# Adversarially Robust Models may not Transfer Better: Domain Transferability from the View of Regularization

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

**University of Illinois Urbana-Champaign**

*Equal Contribution

# An Interesting Observation…

- Salman et al. "Do adversarially robust imagenet models transfer better?." *NeurIPS* 2020.

**adversarially robust (trained) models transfer better.**

- Utrera et al. "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification." *ICLR* 2021.

**adversarially robust (trained) models transfer better.**

# Still Many Questions…

## Motivation of Our Work

- Is it really that adversarially more robust models transfer better?
- If not, what properties affect domain transferability better?
- How to explain their empirical findings?

Our work aims to address these questions

# Theoretical Analysis

# Is it really that adversarially more robust models transfer better?

- In this work, we show that

Improving adversarial robustness is neither necessary nor sufficient for improving domain transferability without any additional assumptions!

Also observed in experiments with real data!

# What Properties Affect Domain Transferability Better than Robustness?

- Our Theorems show that
  - ➢ Shrinking the function class of the source model monotonically decrease a tight upper bound on the *relative domain transferability loss* (target domain loss value minus source domain loss value).
  - ➢ It is reasonable to expect that stronger regularization during source model training leads to better *relative* domain transferability (target domain performance relative to source domain performance).

# What Data Augmentations Can be Viewed as Regularization?

- Can be viewed as regularization
  - ➢ Adversarial training
  - ➢ Gaussian blur, rescale, etc.
- Can **not** be viewed as regularization
  - ➢ Rotation
  - ➢ Translation

Analysis about a wide range of data augmentations in paper

# How to explain their empirical findings?

- Our results suggest a more nuanced explanation of the “adversarially trained (robust) models transfer better:”

  adversarial training => training with regularization => better transferability.

# Experiment Results

**How does different regularization/augmentation affect domain transferability and robustness empirically?**

# Experiment Setting

## Pipeline and Metrics

- Pipeline: 1) Train $g_s \circ f$ on the source domain. 2) Fix $f$ and finetune $g_t \circ f$ on the target domain.
- Used (source, target) domain pair: (CIFAR-10, SVHN) and (ImageNet, CIFAR-10).

# Experiment Setting

## Pipeline and Metrics

- Pipeline: 1) Train $g_s \circ f$ on the source domain. 2) Fix $f$ and finetune $g_t \circ f$ on the target domain.
- Used (source, target) domain pair: (CIFAR-10, SVHN) and (ImageNet, CIFAR-10).
- Metrics:
  - Relative Domain Transfer Accuracy: $\text{DT Acc} = (acc_{tgt} - acc_{src}) - (acc^v_{tgt} - acc^v_{src})$

    Substract the value on vanilla model (constant) so that the comparison can be shown.
  - Robust Accuracy: accuracy under PGD attack ($\ell_2, \epsilon = 0.25$, 20 steps) on source domain.

# Experiment Results

## Jacobian Regularization and Weight Decay

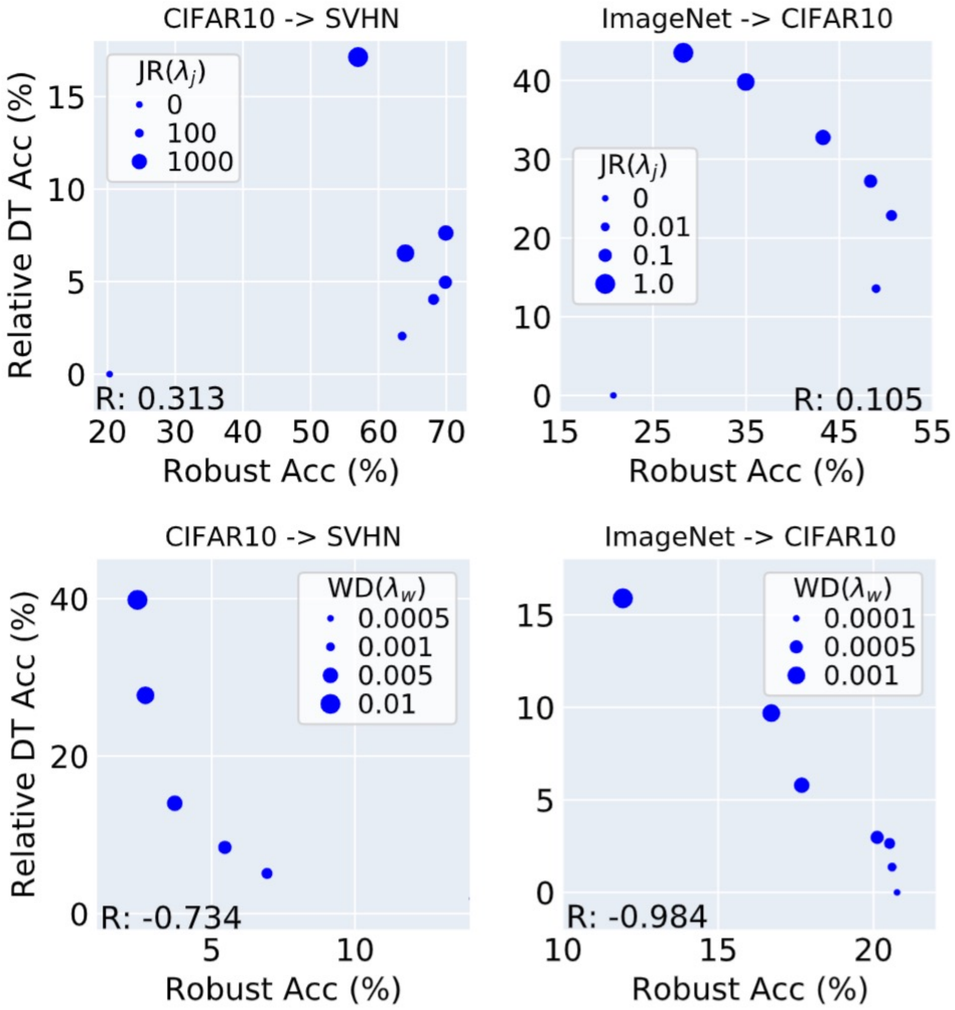
- Jacobian Regularization (JR) with $\lambda_j$:

$$L_{JR}(g_s \circ f, x, y) = L_{CE}(g_s \circ f, x, y) + \lambda_j \cdot ||J(f, x)||_F^2$$

- Weight Decay (WD) with $\lambda_w$.

- Stronger regularizer (larger dots) leads to better domain transferability!

- Robustness does not improve with the better transferability.

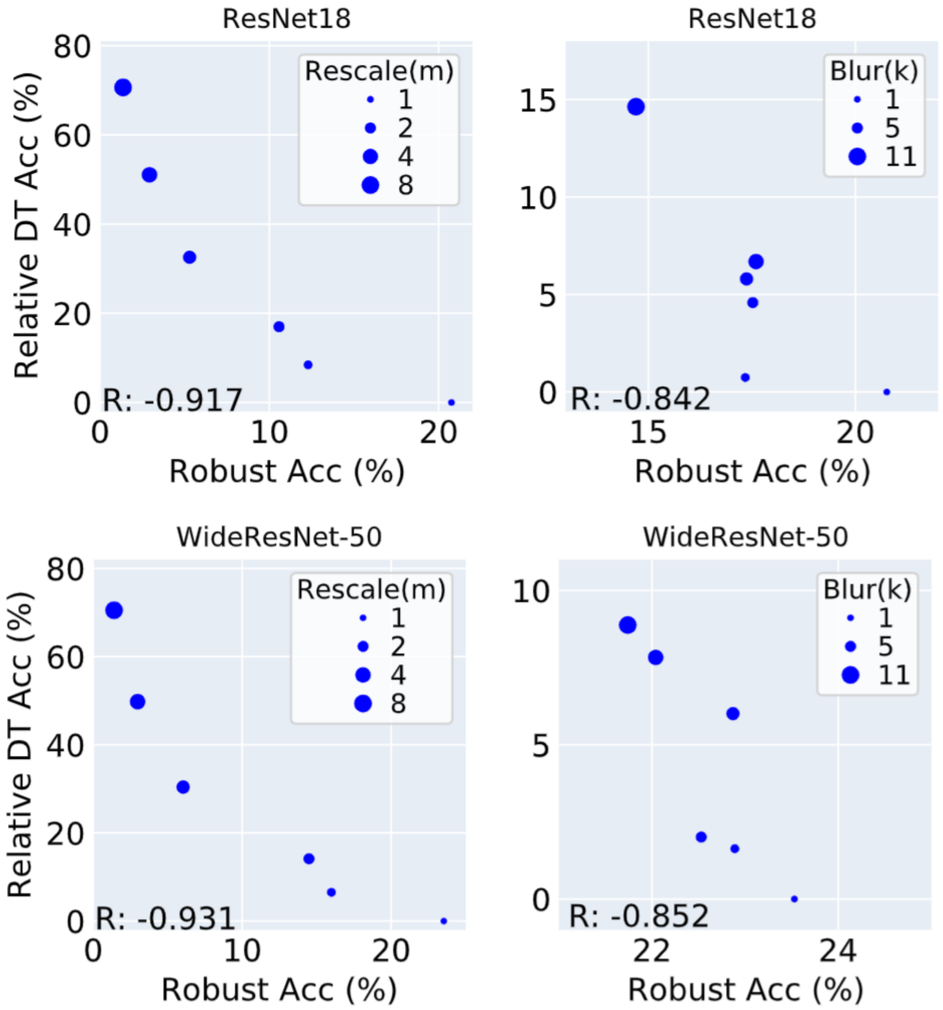


# Experiment Results

## Rescaling and Blurring

- Rescale: rescale to $m$ times smaller.
- Blur: Gaussian blur with kernel size $k$.

- Stronger augmentation leads to better domain transferability.
- Robustness on the source domain drops with strong augmentation.

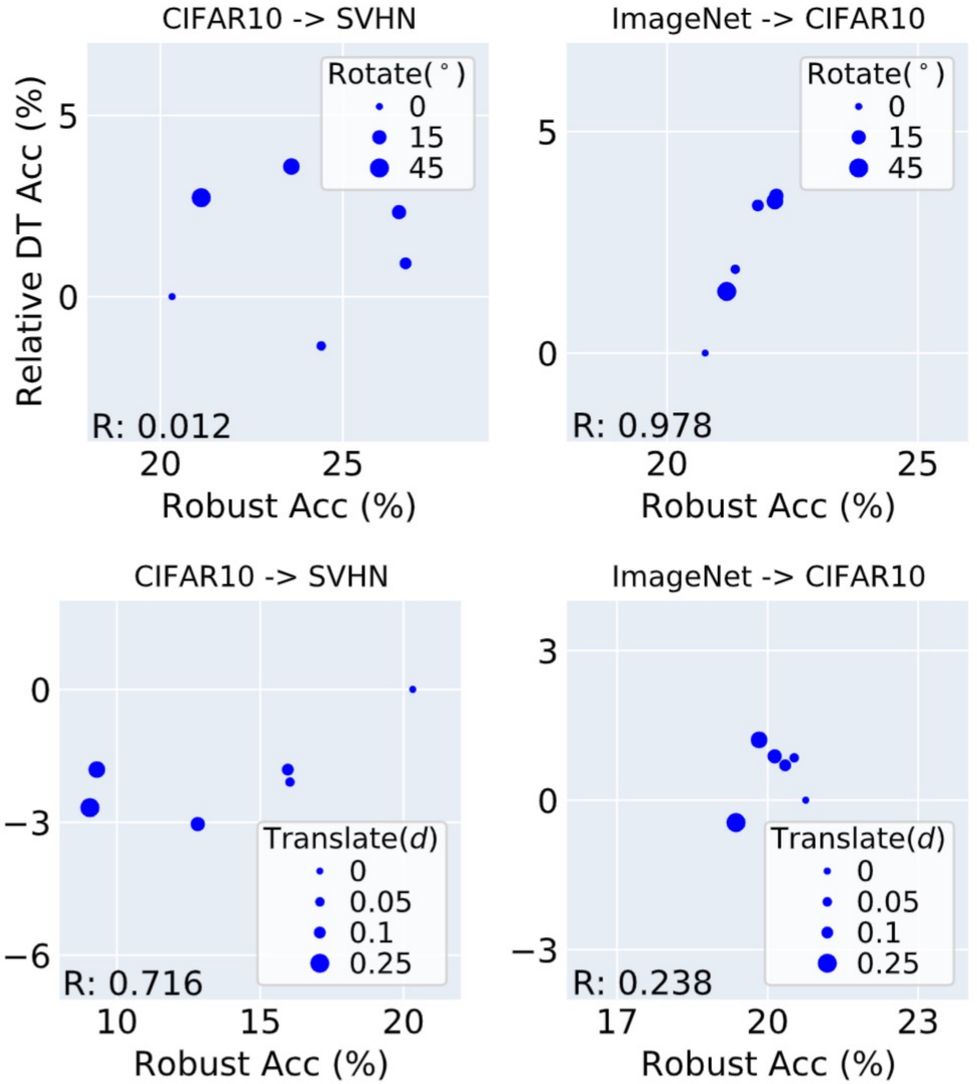


# Experiment Results

## Augmentations that Cannot be Viewed as Regularizations

- Rotation: rotate by certain degrees.
- Translation: translate with proportion $d$ horizontally and vertically.

- We can see that these augmentations do not have a significant impact on the domain transferability!



# Conclusion

# Conclusion

- Takeaways:
  1. Robustness is neither necessary nor sufficient for domain transferability.
  2. Stronger regularization leads to better relative domain transferability.

- See our paper for more detail: https://arxiv.org/pdf/2202.01832.pdf

- Thanks for listening!