香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

A Singular Value Perspective on Model Robustness[1]

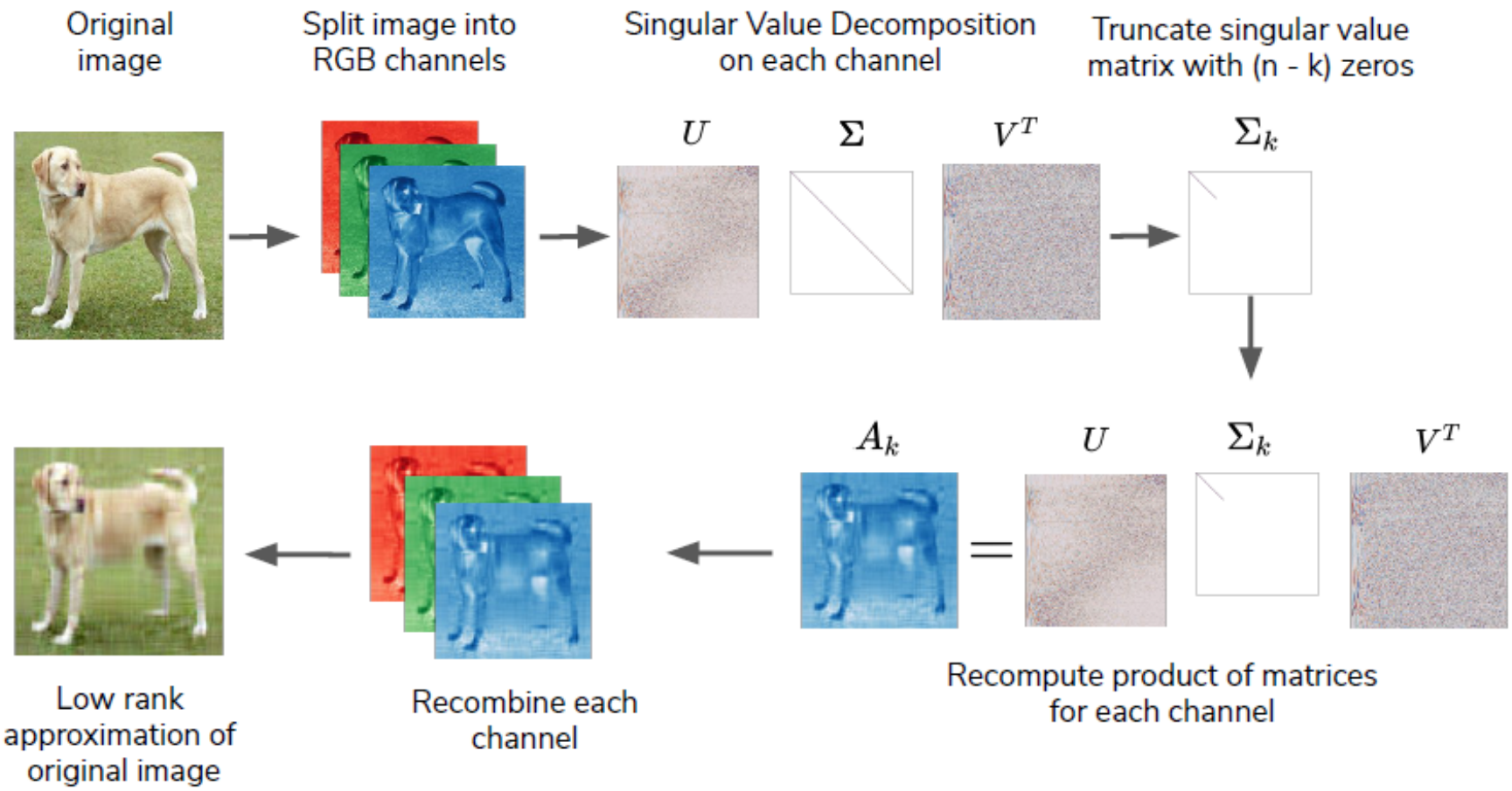
Rongjun Tang

2022.01.26

[1] Jere, Malhar, Maghav Kumar, and Farinaz Koushanfar. "A Singular Value Perspective on Model Robustness." arXiv preprint arXiv:2012.03516 (2020).

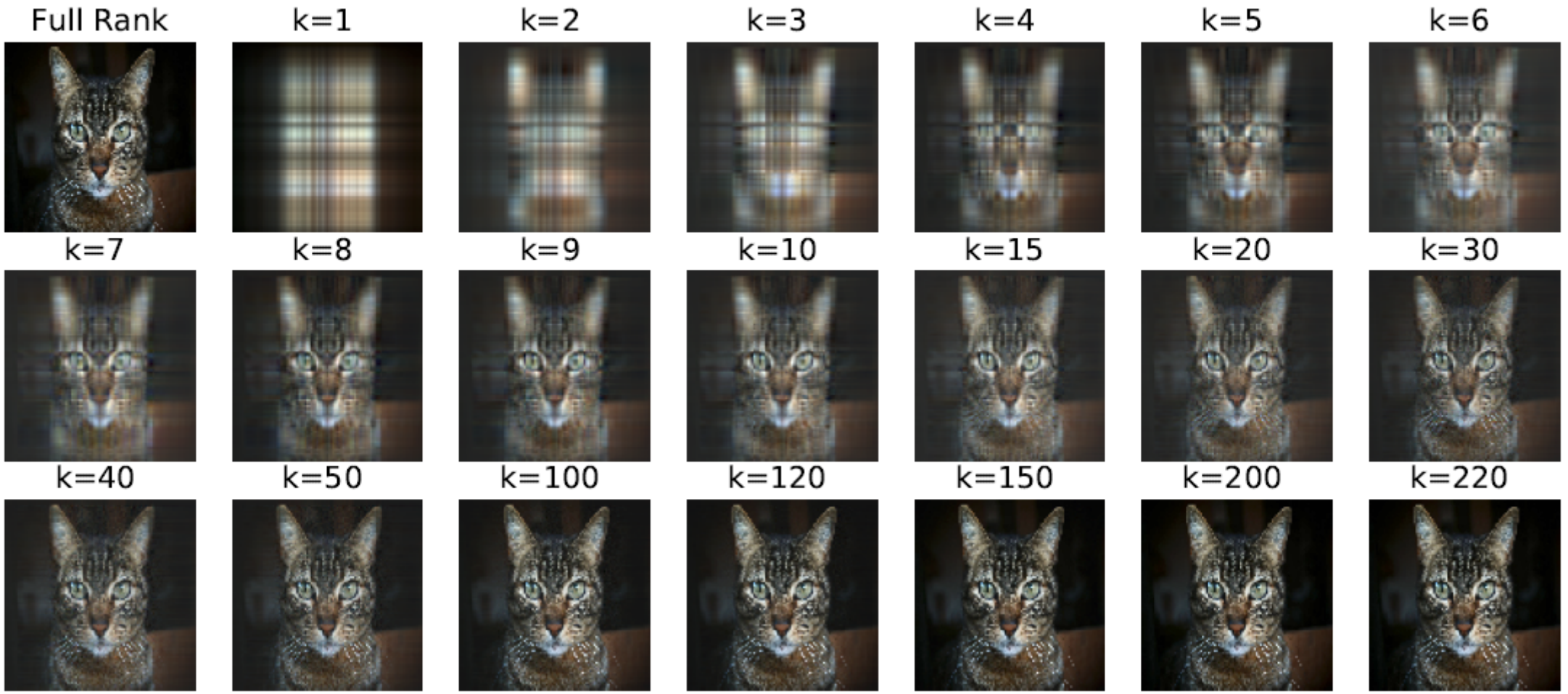
Intro: SVD of Images

Generating a rank-k image via truncated SVD.



Intro: Human Perception on Re-composed Image

Higher rank means more details on image.



Purpose of the Paper

1. Explore the learning preference of a normal and an adversarial training neural network in a singular value perspective.

2. Take the rank information into consideration when calculating gradients.

Note: adversarial training defends against adversarial samples generated by L_∞ PGD attack.

PGD is an iterative adversarial attack method that seeks to generate a targeted adversarial sample x' from an original image x with maximum perturbation limit ϵ.

Rank Dependence of CNNs

Algorithm 1: Finding the accuracies of a model $f(\cdot)$ for a batch of images $x_{1:N}$, where $x_i \in [0,1]^{w \times h \times c}$ as a function of the input rank.

Result: Accuracies for each rank $k = 0 : w$

$full\_rank\_preds \leftarrow f(x_{1:N})$

$rank\_k\_acc \leftarrow zeros(w+1)$

for $k = 0 : w$ **do**

  $rank\_k\_x = zeros\_like(x_{1:N})$

  for $i = 1 : N$ **do**

    for $channel = 1 : c$ **do**

      $u, \sigma, v = SVD(x[i][channel])$

      $\sigma[k : w] = 0$

      $rank\_k\_x[i][channel] = u\ diag(\sigma)\ v$

    end

  end

  $rank\_k\_acc[k] = (f(rank\_k\_x) == full\_rank\_preds)$

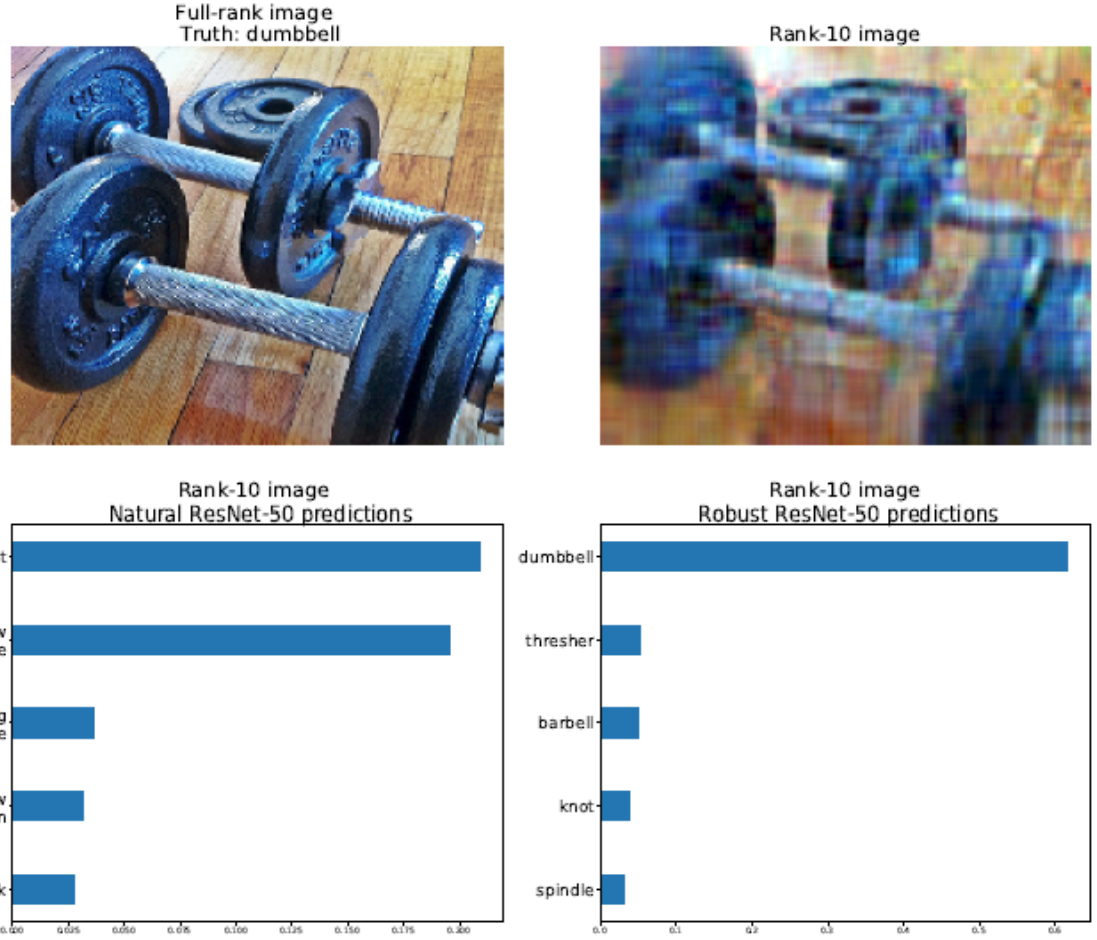
end

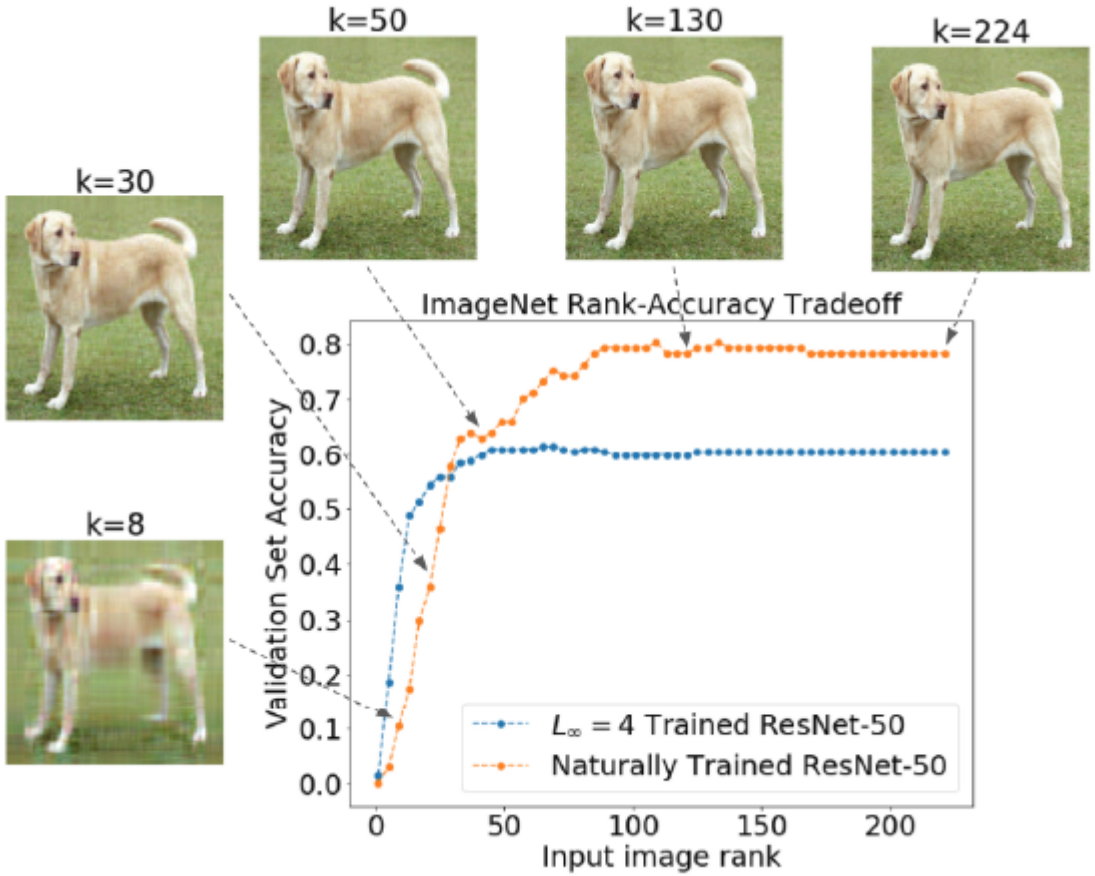
return $rank\_k\_acc$

Explore the dependence of image rank and classifier accuracy for ResNet-50.

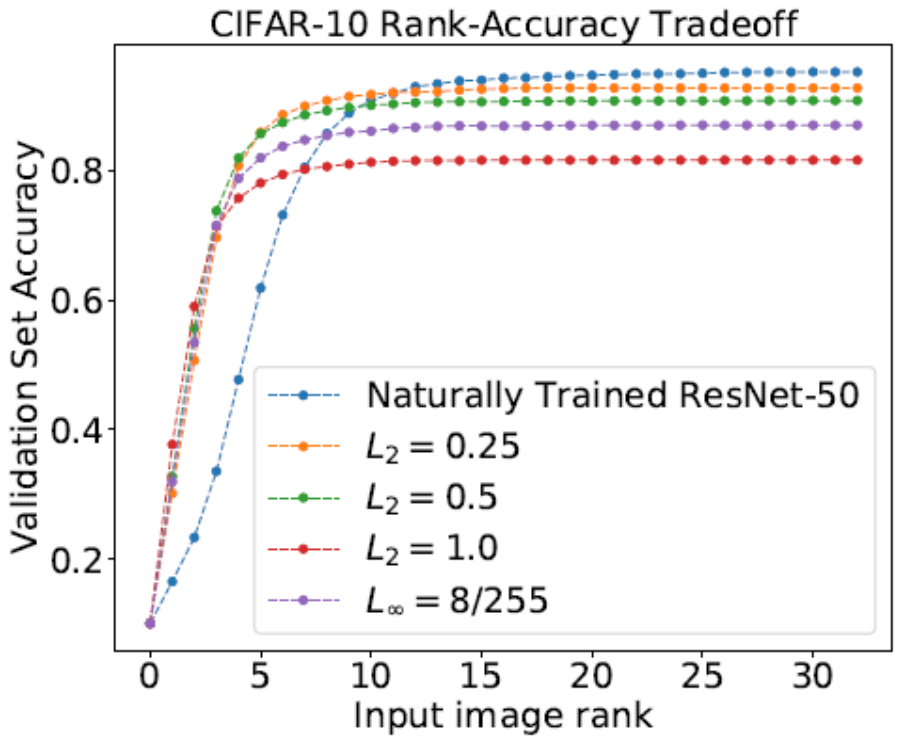
ImageNet and CIFAR10 datasets; 1000 images

Rank Dependence of CNNs

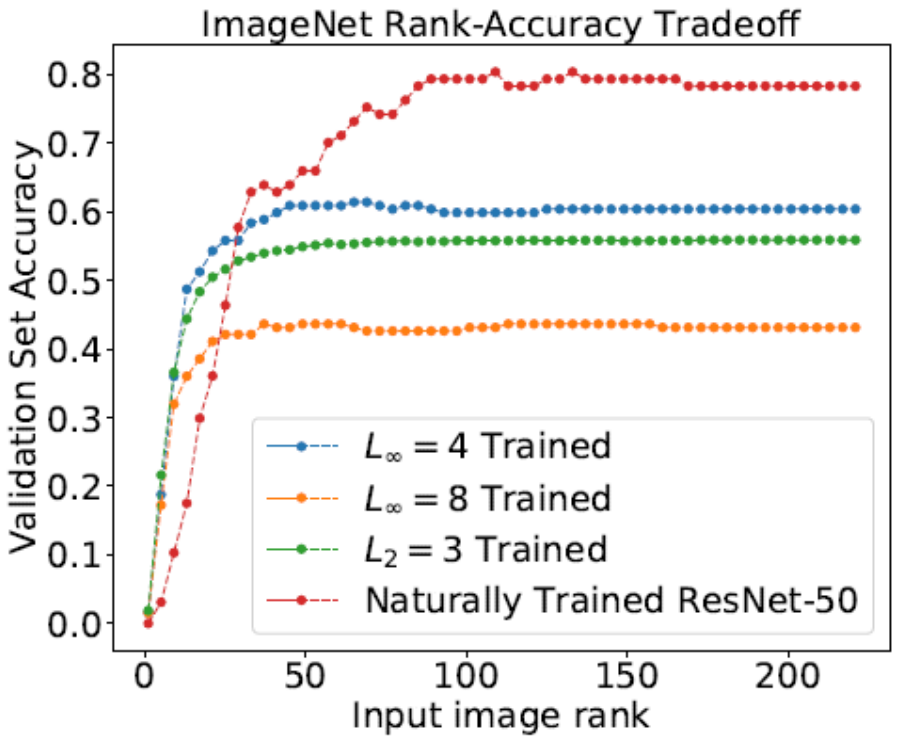




Rank Dependence of CNNs



(a)



(b)

Rank Integrated Gradients (RIG)

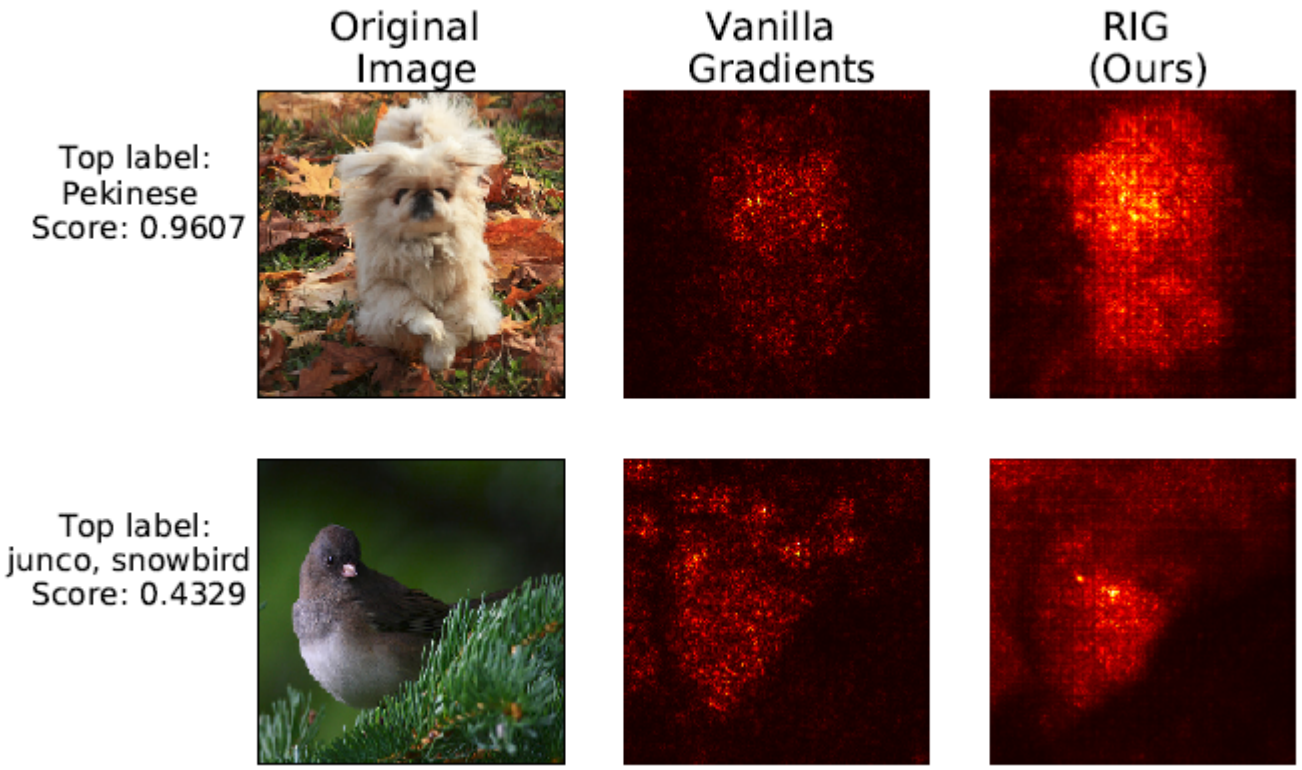
Visualize the rank-dependency of CNNs.

The proposed technique weighs low-rank representations with their contributions to the gradients of an input for the top class predicted on the full-rank input.

$$RIG(x, f, i) = \sum_{k=1}^{w} \frac{w-k}{w} \times \left(\frac{\partial f(x_k)}{\partial x_k}\right)_i$$

Rank Integrated Gradients (RIG)

RIG requires no modification to the model and is extremely easy to implement, requiring less than 10 lines of PyTorch code and using a few calls to the gradient operation, thereby allowing even novice practitioners to easily apply the technique.



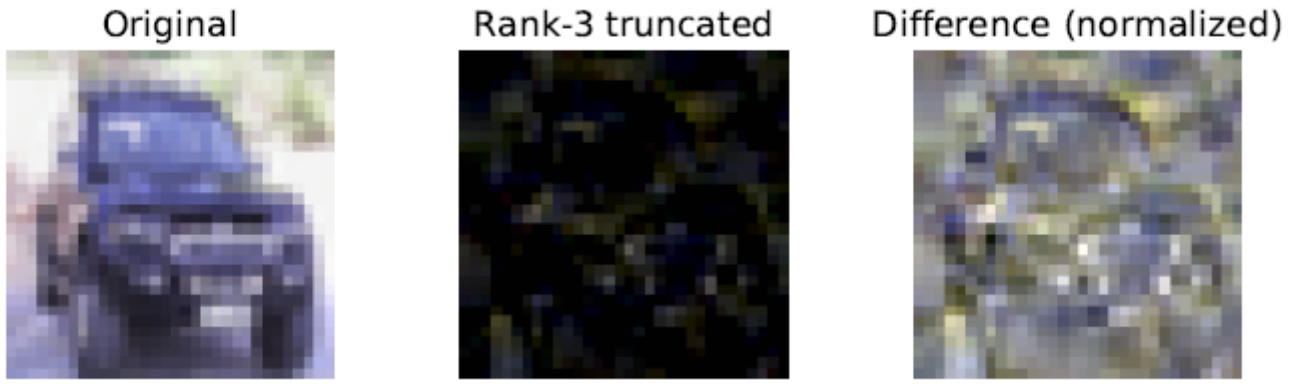
More Results

Robustness of low-rank CIFAR-10 and ImageNet-trained ResNet-50 models.

	CIFAR-10			ImageNet			
Trained on	**Attack success rate**	**Recovery rate**	**Top-1 accuracy**	**Trained on**	**Attack success rate**	**Recovery rate**	**Top-1 accuracy**
Full rank	99.93%	0.03%	95.21%	**Full rank**	95.87%	0.01%	78.35%
20	99.70%	0.15%	95.41%	**100**	81.81%	7.07%	73.99%
10	99.43%	0.19%	94.90%	**50**	73.99%	8.57%	70.16%
5	99.27%	0.34%	91.54%	**30**	73.19%	5.15%	69.07%

For CIFAR10, low-rank representations are sufficient to achieve test accuracy of more than 90% for full-rank CIFAR-10 test sets.
For ImageNet, its data consists of a large number of imperceptible, high-rank features that do not contain semantically meaningful content but contribute to test accuracy. Also, meaningful for adversarial attack!

More Results



Reverse rank truncated CIFAR-10 image. Middle column corresponds to images where the 3 largest singular values were set to 0, with the difference between the original image and truncated image on the right column.
After training with the rank-3 truncated data, we get a full-rank test accuracy of 28.02%.

Take Home Messages

Naturally trained CNNs place a large importance on human-imperceptible higher-rank components, and that adversarial retraining increases reliance on human-aligned lower-ranked components.

Moreover, higher rank information also generalizes to test data.

Neural networks trained on low-rank images are more adversarially robust than their naturally trained counterparts for ImageNet, but not hold on CIFAR10 (image size?).

What about backdoored data? The trigger information to be higher rank? Singular value information maybe used in attack and defense?