Adversarial Defense by Suppressing High Frequencies

5th place of Alibaba Adversarial AI Challenge

Zhendong Zhang and Xiaolong Liang

August 7, 2019

Introduction

Definition: Adversarial perturbations

- small perturbations applied to clean data
- misleading the predictions of deep neural networks (DNNs)

Definition: Adversarial defense

Learning DNNs which are robust to adversarial perturbations (our goal)

Alibaba Adersarial AI Challenge (AAAC)

- Dataset
  - 110 thousands of images with 110 categories for electric business
- Evaluation
  - denote δ as perturbations by black box attackers, P_y as the predicted labels

$$score = \begin{cases} 0, & P_y \neq y \\ mean(\|\delta\|_2), & P_y = y \end{cases}$$

  - average *score* over all images and attackers

Motivations

- DNNs trained on ImageNet bias towards textures or high frequencies [1].
- DNNs are mainly fooled by textures. That is, adversarial perturbations are high frequencies.
- Most information on clean images converge on low frequencies.

$\rightarrow$ Suppressing high frequencies reduces the effects of adversarial perturbations while keeps most information on clean images. It also reduces the bias towards textures.

DNNs bias towards textures or high frequencies

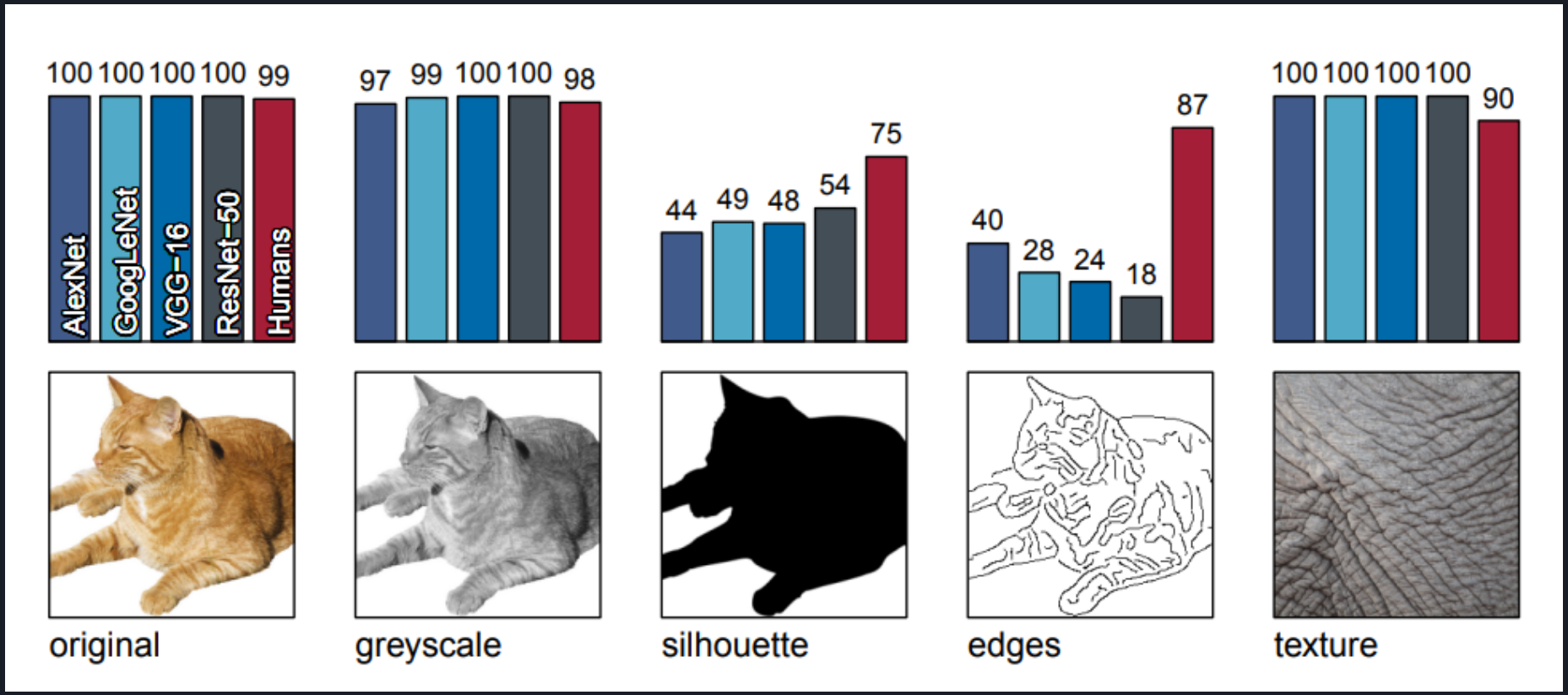


Figure: *Accuracy for cat. Copied from [1].*

Statistics in frequency domain

Discrete Fourier transform (DFT) for an image $\mathbf{x} \in \mathcal{R}^{M \times N}$:

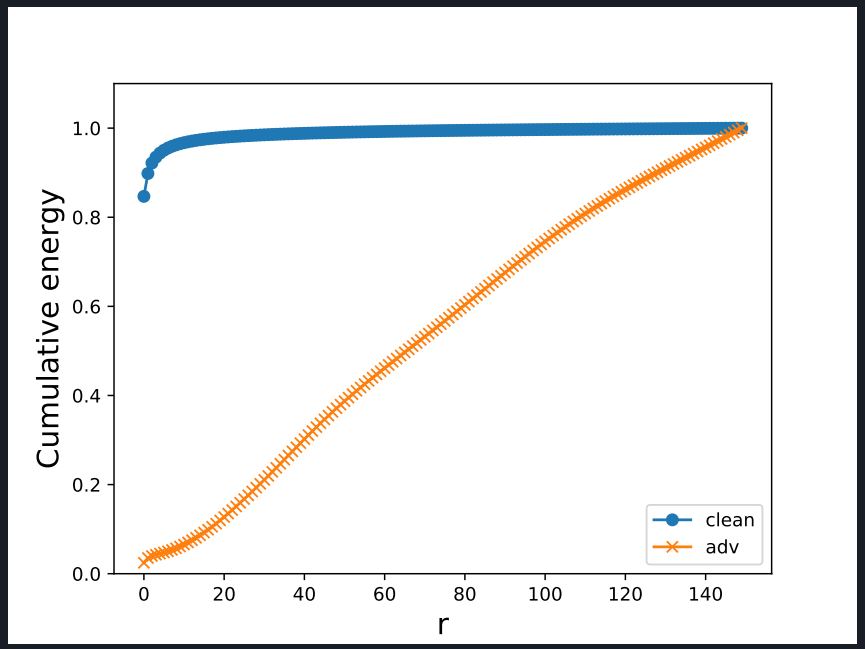
$$\hat{\mathbf{x}}_{u,v} = \sum_{a=0}^{M-1} \sum_{b=0}^{N-1} \mathbf{x}_{a,b} e^{-j2\pi\left(\frac{u}{M}a + \frac{v}{N}b\right)}$$

Definition: Cumulative spectral energy

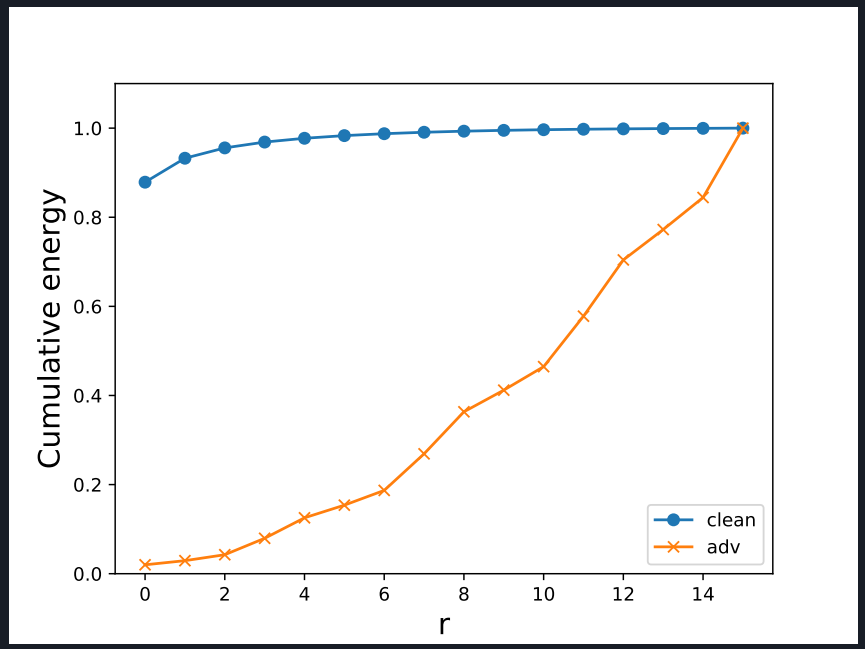
$$CSE(r) = \sum_{u=-r}^{r} \sum_{v=-r}^{r} \hat{\mathbf{x}}^*_{u,v} \hat{\mathbf{x}}_{u,v}$$

i.e. energy within radius r in frequency domain.

Clean images converge on low frequencies while adversarial perturbations are nearly uniform.



(a) AAAC



(b) CIFAR-10

Figure: *Normalized $CSE(r)$ for clean images (blue) and adversarial perturbations (orange).*

Solutions

- High frequency suppressing module
- Adversarial training
- Model ensembles

High frequency suppressing module

Given an image $\mathbf{x}$, transform it into frequency domain, remove high frequencies, then transform back.

$$\mathbf{x}_o \leftarrow \mathcal{F}^{-1}\left(\mathcal{M} \odot \mathcal{F}(\mathbf{x})\right)$$

where $\mathcal{F}$ means DFT, $\odot$ is element-wise multiplication and $\mathcal{M}$ is a box window with radius r.

$$\mathcal{M}_{u,v} = \begin{cases} 1, & 0 <= |u|, |v| <= r \\ 0, & else \end{cases}$$

Then $\mathbf{x}_o$ is processed by a standard DNN.

High frequency suppressing module

- **efficient**: the computational costs are dominated by DFT. $\mathcal{O}(n^2 \log n)$ for an $n \times n$ image.
- **differentiable**: since DFT is differentiable, it's possible to optimize with adversarial training jointly.
- **controllable**: based on Parseval theory

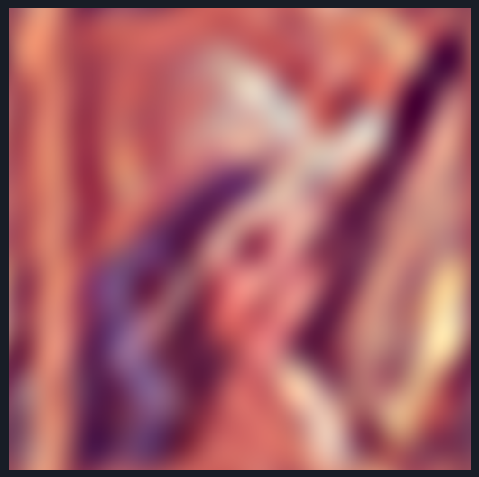
$$\|\mathbf{x} - \mathbf{x}_o\|_2^2 = \|\hat{\mathbf{x}} - \mathcal{M} \odot \hat{\mathbf{x}}\|_2^2$$

distance between $\mathbf{x}$ and $\mathbf{x}_o$ are easily controlled by varying r of the box window.

High frequency suppressing module

(a) original

(b) $r = 4$



(c) $r = 8$

(d) $r = 16$

(e) $r = 32$

(f) $r = 64$

Adversarial training

optimizing DNNs w.r.t both clean samples and adversarial samples [2]. $\mathcal{L}$ is the cross-entropy loss.

$$\mathcal{L}(f(\mathbf{x}), y) + \beta \max_{\|\delta\|<\epsilon} \mathcal{L}(f(\mathbf{x}+\delta), y)$$

TRADES [3] is a novel adversarial training method which encourages the output to be smooth:

$$\mathcal{L}(f(\mathbf{x}), y) + \beta \max_{\|\delta\|<\epsilon} \mathcal{L}(f(\mathbf{x}), f(\mathbf{x}+\delta))$$

We use TRADES because it has a better tradeoff between robustness and accuracy.

Model ensembles

The smaller the radius r, the more robust w.r.t adversarial samples while the less accurate w.r.t clean samples.
It is reasonable to integrate models with different r.
Our final solution integrates models with

$$r \in \{8, 12, 16, 24\}$$

Each model contains a high frequency suppressing module and a standard ResNet-18.

Results

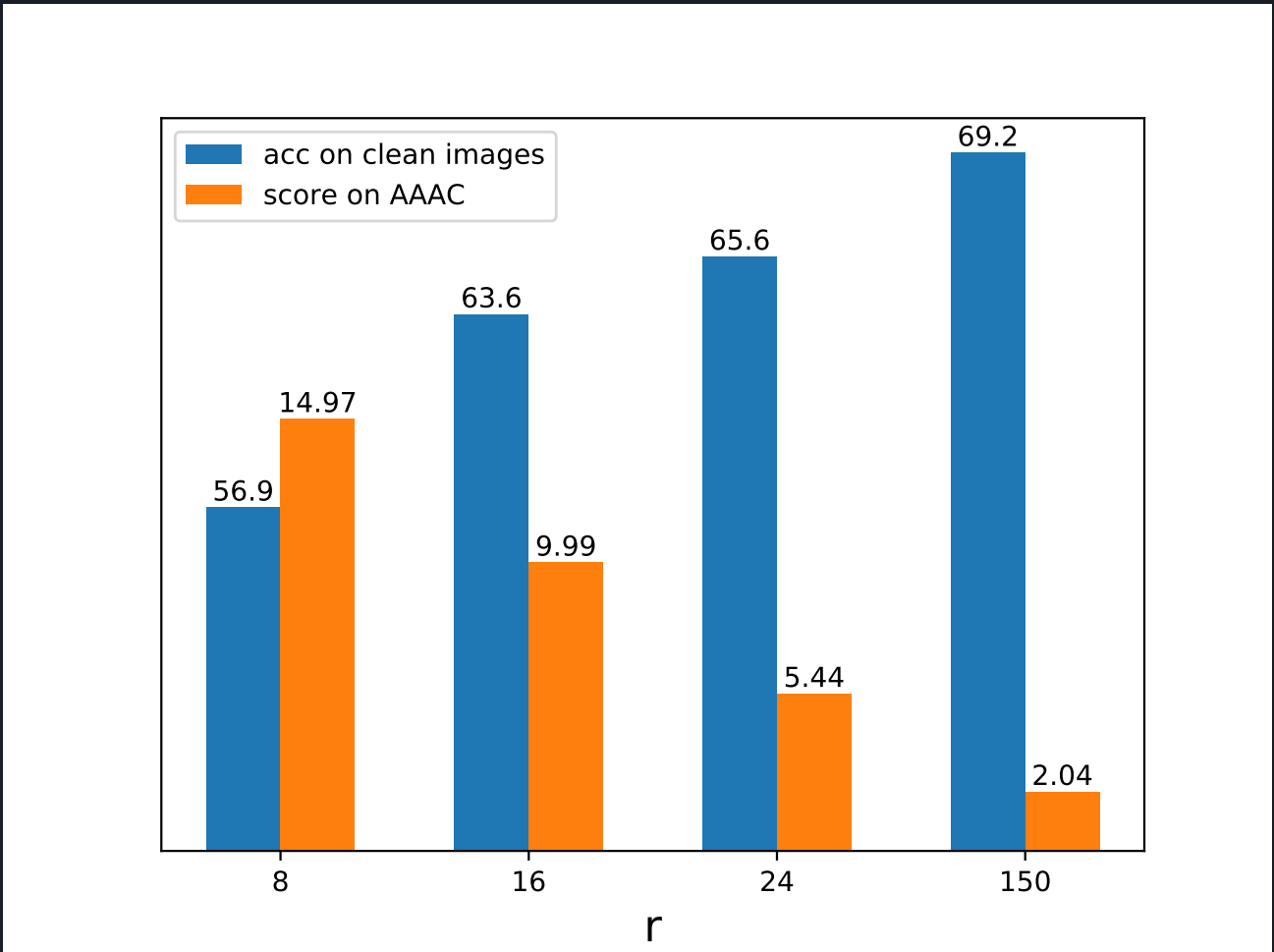


Figure: *Trade-off between robustness and accuracy w.r.t radius.*

Results

1st $\rightarrow$ 20.13

VS

ours $\rightarrow$ 19.75

Suppression module	Adversarial training	Model ensemble	Score
$\times$	$\times$	$\times$	2.04
$\times$	$\surd$	$\times$	9.99
$\surd$	$\times$	$\times$	14.97
$\surd$	$\surd$	$\times$	19.05
$\surd$	$\surd$	$\surd$	19.75

Table: *Ablation study for three strategies and their combinations.*

Results

- Our high frequency suppressing module is faster than a convolutional layer. Thus the overall inference time (without ensembles) is nearly the same as the standard ResNet-18's.
- The size of the checkpoint file is only 44 megabit.
- There are no specific assumptions or adjustments for AAAC dataset. Thus our method is easy to transfer.

Key references

1. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. **ICLR2019**.
2. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. **ICLR2018**.
3. Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. arXiv.