Supplementary Materials: Interpreting Adversarially Trained Convolutional Neural Networks

Tianyuan Zhang [1] **Zhanxing Zhu** [2 3 4]

1. Experiment Setup

1.1. Models

- **CIFAR-10**. We train a standard ResNet-18 (He et al., 2016) architecture, it has 4 groups of residual layers with filter sizes (64, 128, 256, 512) and 2 residual units.
- **Caltech-256 & Tiny ImageNet**. We use a ResNet-18 architecture using the code from pytorch(Paszke et al., 2017). Note that for models on Caltech-256 & Tiny ImageNet, we initialize them using ImageNet(Deng et al., 2009) pre-trained weighs provided by pytorch.

We evaluate the robustness of all our models using a l_∞ projected gradient descent adversary with $\epsilon = 8/255$, step size = 2 and number of iterations as 40.

1.2. Adversarial Training

We perform 9 types of adversarial training on each of the dataset. 7 of the 9 kinds of adversarial training are against a projected gradient descent (PGD) adversary(Madry et al., 2018), the other 2 are against FGSM adversary(Goodfellow et al., 2014).

1.2.1. Train against a projected gradient descent (PGD) adversary

We list value of ϵ for adversarial training of each dataset and l_p-norm. In all settings, PGD runs 20 iterations.

- **l_∞-norm bounded adversary**. For all of the three data set, pixel vaules range from 0 1, we train 4 adversarially trained CNNs with $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$, these four models are denoted as PGD-inf:1, 2, 4, 8 respectively, and steps size as 1/255, 1/255, 2/255, 4/255.
- **l_2-norm bounded adversary**. For Caltech-256 & Tiny ImageNet, the input size for our model is 224×224, we train three adversarially trained CNNs with $\epsilon \in \{4, 8, 12\}$, and these four models are denoted as PGD-l2: 4, 8, 12 respectively. Step sizes for these three models are 2/255, 4/255, 6/255. For CIFAR-10, where images are of size 32×32, the three adversarially trained CNNs have $\epsilon \in \{4/10, 8/10, 12/10\}$, but they are denoted in the same way and have the same step size as that in Caltech-256 & Tiny ImageNet.

1.2.2. Train against a FGSM adversary

ϵ for these two adversarially trained CNNs are $\epsilon \in \{4, 8\}$, and they are denoted as FGSM 4, 8 respectively.

2. Style-transferred test set

Following (Geirhos et al., 2019) we construct stylized test set for Caltech-256 and Tiny ImageNet by applying the AdaIn style transfer(Huang & Belongie, 2017) with a stylization coefficient of $\alpha = 1.0$ to every test image with the style of a randomly selected painting from [1]Kaggle's *Painter by numbers* dataset. we used source code provided by(Geirhos et al., 2019).

3. Experiments on Fourier-filtered datasets

(Jo & Bengio, 2017) showed deep neural networks tend to learn surface statistical regularities as opposed to high-level abstractions. Following them, we test the performance of different trained CNNs on the high-pass and low-pass filtered dataset to show their tendencies.

3.1. Fourier filtering setup

Following (Jo & Bengio, 2017) We construct three types of Fourier filtered version of test set.

- **The low frequency filtered version**. We use a radial mask in the Fourier domain to set higher frequency modes to zero.(low-pass filtering)

[1]School of EECS, Peking University, China [2]School of Mathematical Sciences, Peking University, China [3]Center for Data Science, Peking University [4]Beijing Institute of Big Data Research. Correspondence to: Zhanxing Zhu <zhanxing.zhu@pku.edu.cn>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]https://www.kaggle.com/c/painter-by-numbers/

- **The high frequency filtered version**. We use a radial mask in the Fourier domain to preserve only the higher frequency modes.(high-pass filtering)
- **The random filtered version**. We use a random mask in the Fourier domain to set each mode to 0 with probability p uniformly. The random mask is generated on the fly during the test.

3.2. Results

We measure generalization performance (accuracy on correctly classified images) of each model on these three filtered datasets from Caltech-256, results are listed in Table 1. AT-CNNs performs better on Low-pass filtered dataset and worse on High-pass filtered dataset. Results indicate that AT-CNNs make their predictions depend more on low-frequency information. This finding is consistent with our conclusions since local features such as textures are often considered as high-frequency information, and shapes and contours are more like low-frequency.

4. Detailed results

We the detailed results for our quantitative experiments here. Table 3, 2, 4 show the results of each models on test set with different saturation levels. Table 6, 5 list all the results of each models on test set after different path-shuffling operations.

5. Additional Figures

We show additional sensitive maps in Figure 1. We also compare the sensitive maps using **Grad** and **SmoothGrad** in Figure 2.

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Jo, J. and Bengio, Y. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Table 1. "Accuracy on correctly classified images" for different models on three Fourier-filtered Caltech-256 test sets.

DATA SET	THE LOW FREQUENCY FILTERED VERSION	THE HIGH FREQUENCY FILTERED VERSION	THE RANDOM FILTERED VERSION
STANDARD	15.8	16.5	73.5
UNDERFIT	14.5	17.6	62.2
PGD-l_∞:	71.1	3.6	73.4

Table 2. "Accuracy on correctly classified images" for different models on saturated Caltech-256 test set. It is easily observed AT-CNNs are much more robust to increasing saturation levels on Caltech-256.

SATURAION LEVEL	0.25	0.5	1	4	8	16	64	1024
STANDARD	28.62	57.45	85.20	90.13	65.37	42.37	23.45	20.03
UNDERFIT	31.84	**63.36**	**90.96**	84.51	57.51	38.58	26.00	23.08
PGD-l_∞: 8	32.84	53.47	82.72	86.45	**70.33**	**61.09**	**53.76**	**51.91**
PGD-l_∞: 4	31.99	57.74	85.18	87.95	**70.33**	58.38	48.16	45.45
PGD-l_∞: 2	**32.99**	60.75	87.75	89.35	68.78	51.99	40.69	37.83
PGD-l_∞: 1	32.67	61.85	89.36	**90.18**	69.07	50.05	37.98	34.80
PGD-l_2: 12	31.38	53.07	82.10	83.89	67.06	58.51	52.45	50.75
PGD-l_2: 8	32.82	56.65	85.01	86.09	68.90	58.75	51.59	49.30
PGD-l_2: 4	32.82	58.77	86.30	86.36	67.94	53.68	44.43	41.98
FGSM: 8	29.53	55.46	85.10	86.65	69.01	55.64	45.92	43.42
FGSM: 4	32.68	59.37	87.22	87.90	66.71	51.13	41.66	38.78

Table 3. "Accuracy on correctly classified images" for different models on saturated Tiny ImageNet test set. It is easily observed AT-CNNs are much more robust to increasing saturation levels on Tiny ImageNet.

SATURAION LEVEL	0.25	0.5	1	4	8	16	64	1024
STANDARD	7.24	25.88	72.52	72.73	25.38	8.24	2.62	1.93
UNDERFIT	7.34	25.44	69.80	60.67	18.01	6.72	3.16	2.65
PGD-l_∞: 8	11.07	29.08	67.11	74.53	49.8	40.16	35.44	33.96
PGD-l_∞: 4	12.44	33.53	72.94	75.75	46.38	32.12	24.92	22.65
PGD-l_∞: 2	12.09	34.85	75.77	76.15	41.35	25.20	16.93	14.52
PGD-l_∞: 1	11.30	**35.03**	76.85	78.63	40.48	21.37	12.70	10.81
PGD-l_2: 12	11.30	29.48	66.94	75.22	**52.26**	**42.11**	**37.20**	**35.85**
PGD-l_2: 8	12.42	32.78	71.94	75.15	47.92	35.66	29.55	27.90
PGD-l_2: 4	**12.63**	34.10	74.06	77.32	45.00	28.73	20.16	18.04
FGSM: 8	12.59	32.66	70.55	**81.53**	41.83	17.52	7.29	5.82
FGSM: 4	**12.63**	34.10	**74.06**	75.05	42.91	29.09	22.15	20.14

Table 4. "Accuracy on correctly classified images" for different models on saturated CIFAR-10 test set. It is easily observed AT-CNNs are much more robust to increasing saturation levels on CIFAR-10.

SATURAION LEVEL	0.25	0.5	1	4	8	16	64	1024
STANDARD	27.36	55.95	**91.03**	**93.12**	69.98	48.30	34.39	31.06
UNDERFIT	21.43	50.28	87.71	89.89	66.09	43.35	29.10	26.13
PGD-l_∞: 8	26.05	46.96	80.97	89.16	**75.46**	**69.08**	58.98	**64.64**
PGD-l_∞: 4	27.22	49.81	84.16	89.79	73.89	65.35	59.99	58.47
PGD-l_∞: 2	**28.32**	53.12	86.93	91.37	74.02	62.82	55.25	52.60
PGD-l_∞: 1	27.18	53.59	88.54	91.77	72.67	58.39	47.25	41.75
PGD-l_2: 12	25.99	46.92	81.72	88.44	73.92	66.03	60.98	59.41
PGD-l_2: 8	27.75	50.29	83.76	80.92	73.17	64.83	58.64	46.94
PGD-l_2: 4	27.26	51.17	85.78	90.08	73.12	61.50	52.04	48.79
FGSM: 8	25.50	46.11	81.72	87.67	74.22	67.12	**62.51**	61.32
FGSM: 4	26.39	**58.93**	84.30	89.02	73.47	64.43	58.80	56.82

Table 5. "Accuracy on correctly classified images" for different models on Patch-shuffled Caltech-256 test set. Results indicates that AT-CNNs are more sensitive to Patch-shuffle operations on Caltech-256.

DATA SET	2×2	4×4	8×8T
STANDARD	**84.76**	**51.50**	**10.84**
UNDERFIT	75.59	33.41	6.03
PGD-l_∞: 8	58.13	20.14	7.70
PGD-l_∞: 4	68.54	26.45	8.18
PGD-l_∞: 2	74.25	30.77	9.00
PGD-l_∞: 1	78.11	35.03	8.42
PGD-l_2: 12	58.25	21.03	7.85
PGD-l_2: 8	63.36	22.19	8.48
PGD-l_2: 4	69.65	28.21	7.72
FGSM: 8	64.48	22.94	8.07
FGSM: 4	70.50	28.41	6.03

Table 6. "Accuracy on correctly classified images" for different models on Patch-shuffled Tiny ImageNet test set. Results indicates that AT-CNNs are more sensitive to Patch-shuffle operations on Tiny ImageNet.

DATA SET	2×2	4×4	8×8T
STANDARD	**66.73**	**24.87**	4.48
UNDERFIT	59.22	23.62	4.38
PGD-l_∞: 8	41.08	16.05	6.83
PGD-l_∞: 4	49.54	18.23	6.30
PGD-l_∞: 2	55.96	19.95	5.61
PGD-l_∞: 1	60.19	23.24	6.08
PGD-l_2: 12	42.23	16.95	**7.66**
PGD-l_2: 8	47.67	16.28	6.50
PGD-l_2: 4	51.94	17.79	5.89
FGSM: 8	57.42	20.70	4.73
FGSM: 4	50.68	16.84	5.98

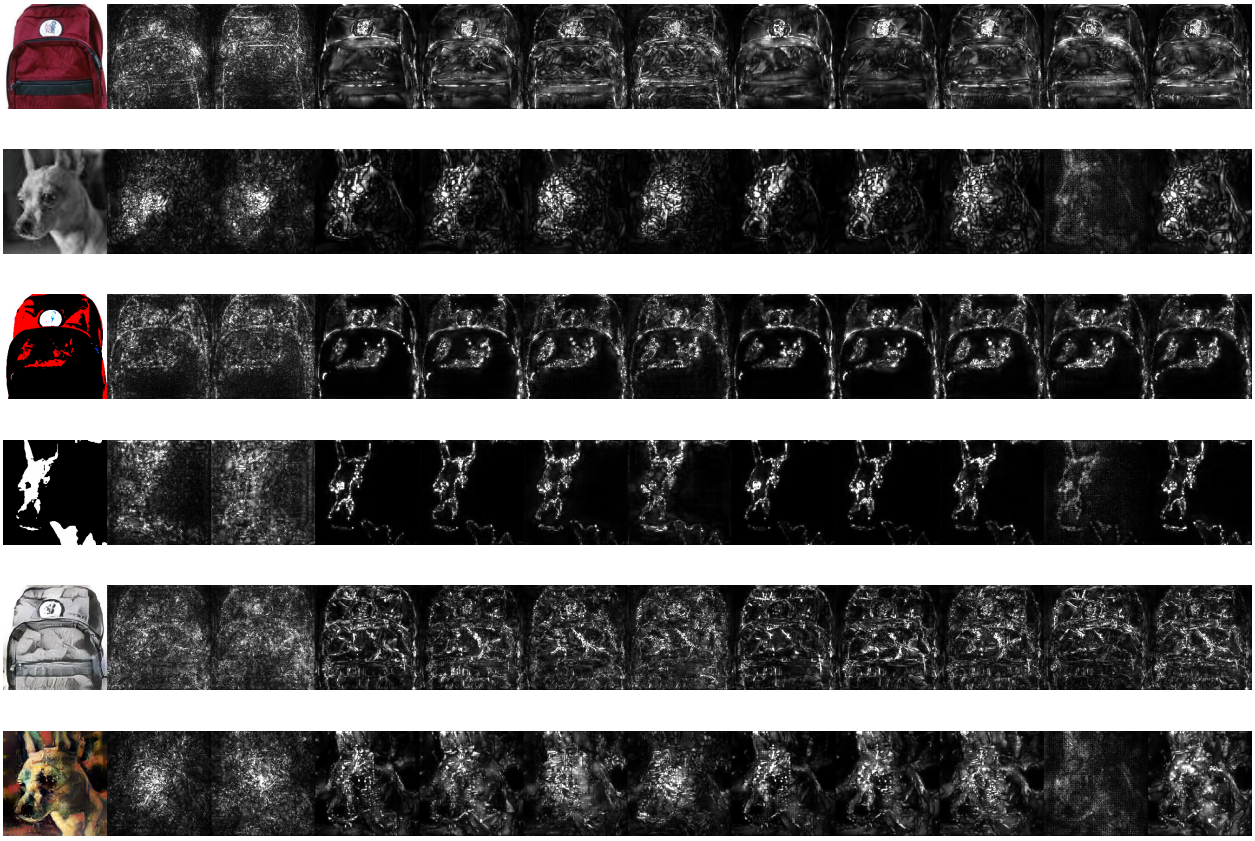

Figure 1. Visualization of Salience maps generated from **SmoothGrad** (Smilkov et al., 2017) for all 11 models. From left to right, Standard CNNs, underfitting CNNs, PGD-inf: 8, 4, 2, 1, PGD-L2: 12, 8, 4 and FGSM: 8, 4.

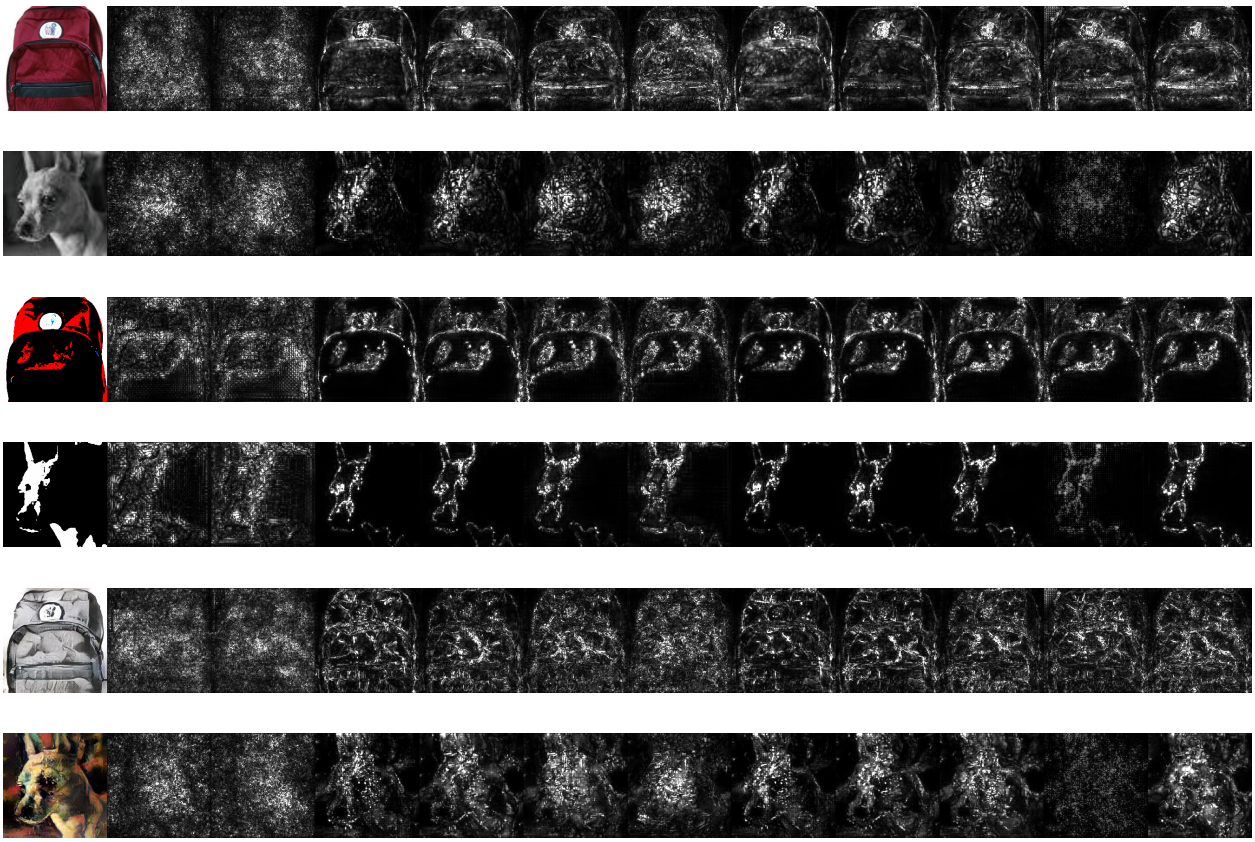

Figure 2. Visualization of Salience maps generated from **Grad** for all 11 models. From left to right, Standard CNNs, underfitting CNNs, PGD-inf: 8, 4, 2, 1, PGD-L2: 12, 8, 4 and FGSM: 8, 4. It's easily observed that sensitivity maps generated from **Grad** are more noisy compared with its smoothed variant **SmoothGrad**, especially for Standard CNNs and underfitting CNNs.