FIRST LINE OF DEFENSE: A ROBUST FIRST LAYER MITIGATES ADVERSARIAL ATTACKS

Janani Suresh*, Nancy Nayak+, Sheetal Kalyani*
*Indian Institute of Technology Madras,India +Imperial College,London
ee22s079@smail.iitm.ac.in, n.nayak@imperial.ac.uk, skalyani@ee.iitm.ac.in

Motivation

- Adversarial Training methods are computationally intensive
- Focus on architectural components - denoised smoothing, impact of topology, depth, and network-width
- Enhancing Native Robustness
  - Regularizing high-frequency filters
  - Adversarial Noise Filter (ANF) - the modified first layer inhibits the passage of adversarial noise

ANF

Increases the non-linearity in the architecture by combining the three operations:

- Larger kernels - smooth the features/noise
- More filters - better generalization
- Maxpool downsamples - and reduce the impact of adversarial noise

implicitly filters out the adversarial noise and reduces its propagation to other layers.

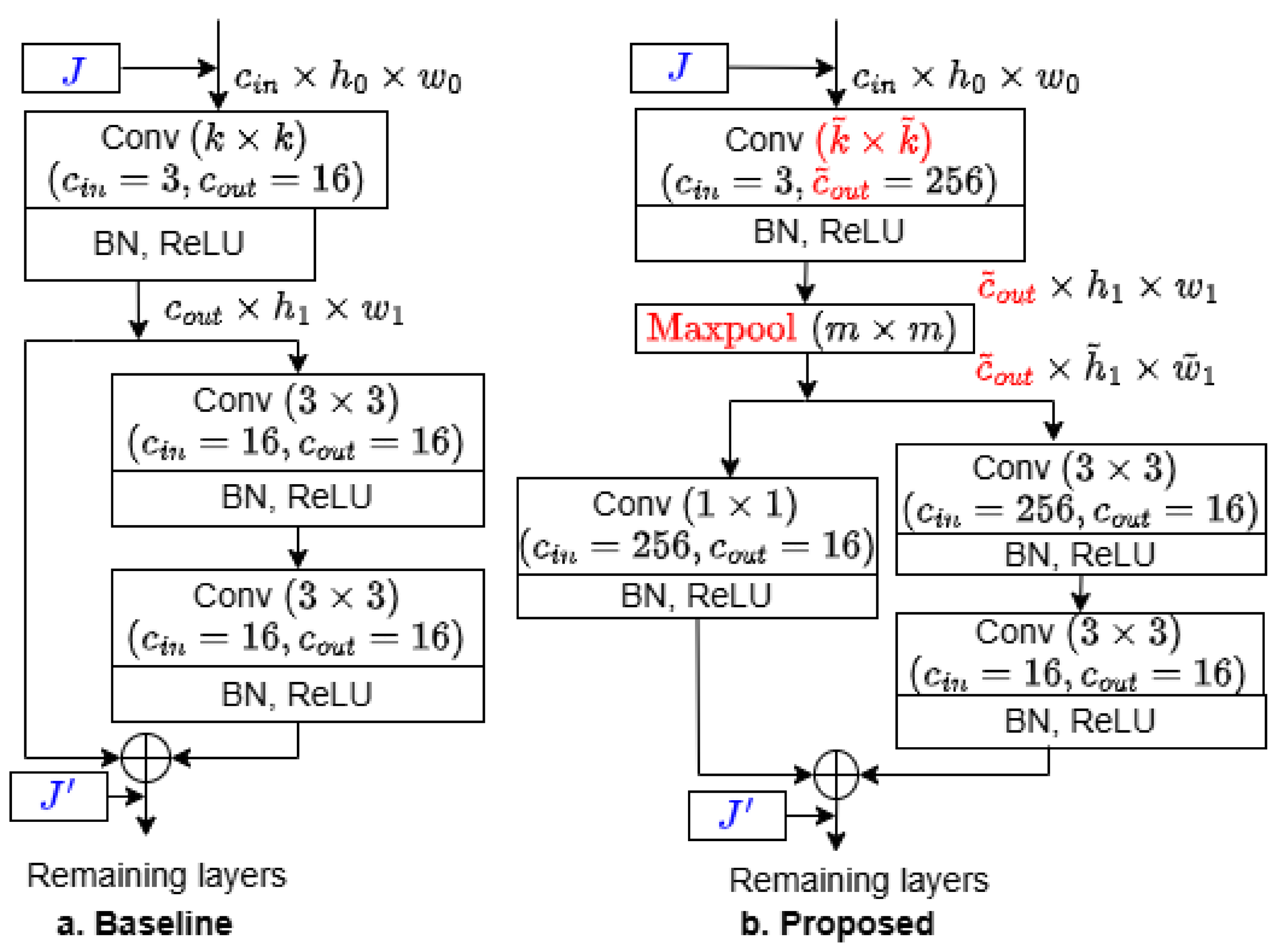


Figure 1: ANF as the first layer in ResNet20

Measure of denoising - mPSNR

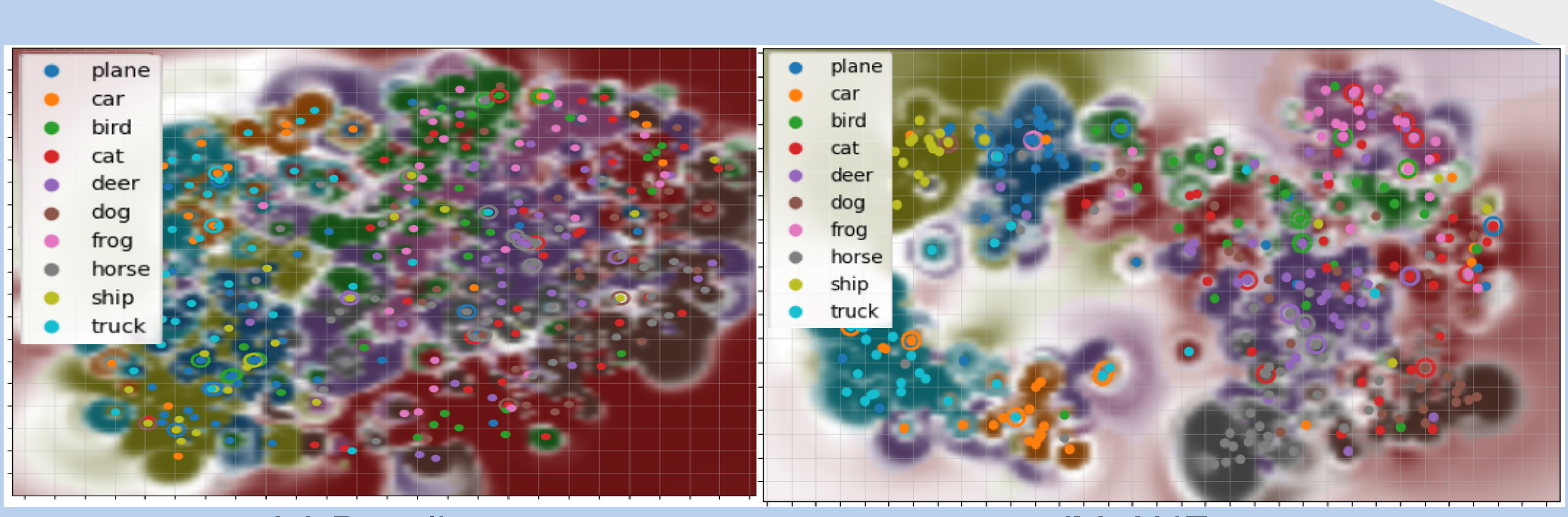
$$\text{invPSNR}_i = \frac{\sqrt{\frac{1}{DHW}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W}(h_{idhw} - \tilde{h}_{idhw})^2}}{\frac{1}{D}\sum_{d=1}^{D}\max_{1\le m\le H, 1\le n\le D}\, \mathbf{x}_{imn}^{d}}, \quad \text{mPSNR} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\text{invPSNR}_i}$$

Arch.	K	F	M	PGD	Clean acc	mPSNR at J	mPSNR at J'
Baseline	✗	✗	✗	27.22	91.26	160.42	22.66
Type 1	✗	✗	✓	**45.37**	88.35	158.18	**65.65**
Type 2	✗	✓	✗	29.91	91.16	160.23	24.23
Type 3	✗	✓	✓	**49.92**	89.72	156.63	**61.79**
Type 4	✓	✗	✗	45.54	85.68	156.08	59.08
Type 5	✓	✗	✓	**51.71**	80.99	153.06	**78.17**
Type 6	✓	✓	✗	40.12	86.64	157.47	29.80
Type 7	✓	✓	✓	**59.93**	83.09	151.62	**89.92**

Table 1: mPSNR in ResNet20 for CIFAR10. For column K, ✓increases the kernel size from 3×3 to 15×15; for column F, ✓increases filters from 16 to 256; for column M, ✓introduces a 5×5 maxpool operation.

Why does ANF work?

Visualization of the Decision Regions



(a) Baseline **(b)** ANF

Figure 2: Decision regions for ResNet20 with adversarial samples

- Most of the samples are misclassified and the decision regions are scattered with baseline, while ANF has sparse decision boundaries, making it more robust toward adversarial attacks.

Loss Surface Visualization with Adversarial Samples

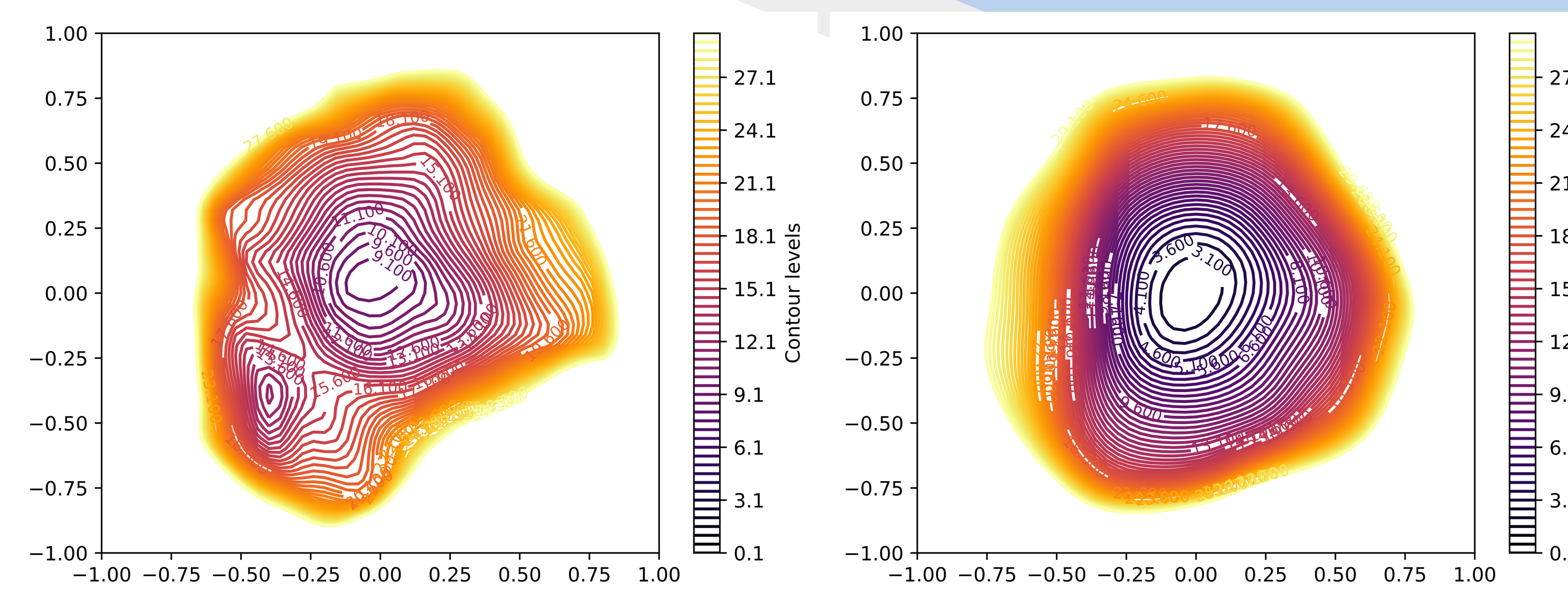


Figure 3: ResNet20 baseline (left) and ResNet20 with ANF (right)

- The loss surface looks smoother with ANF than baseline as the baseline has multiple minima compared to ANF having one distinct minima.

Frequency Spectrum of Unstructured Noise

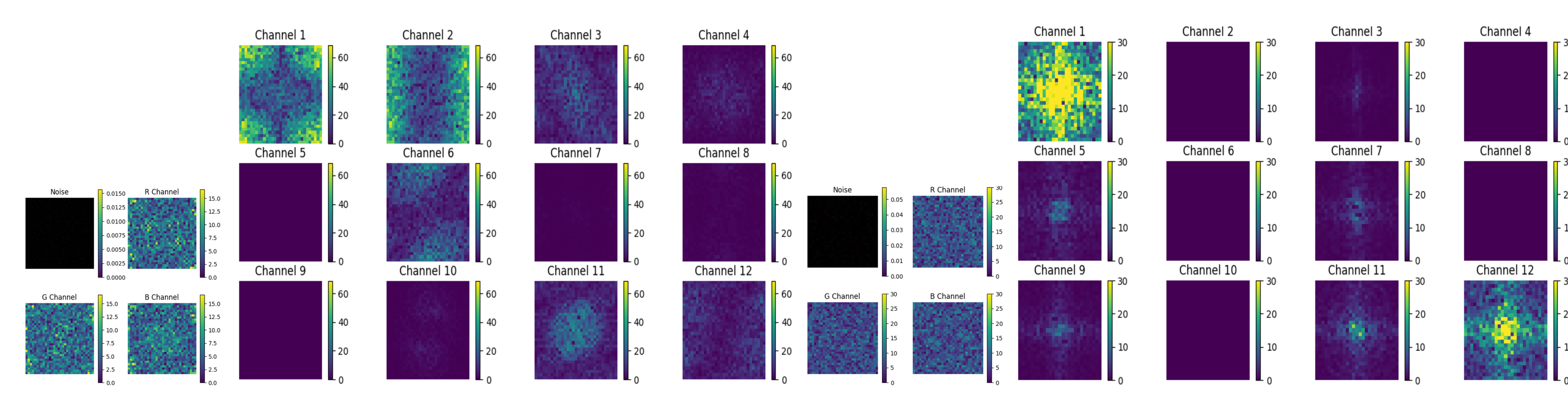


Figure 4: FFT of feature maps for unstructured noise at input and after first layer of ResNet20 - baseline (left) vs. ANF (right).

- ANF attenuates high-frequency components, lower intensity for high-frequency components

Feature Denoising with ANF

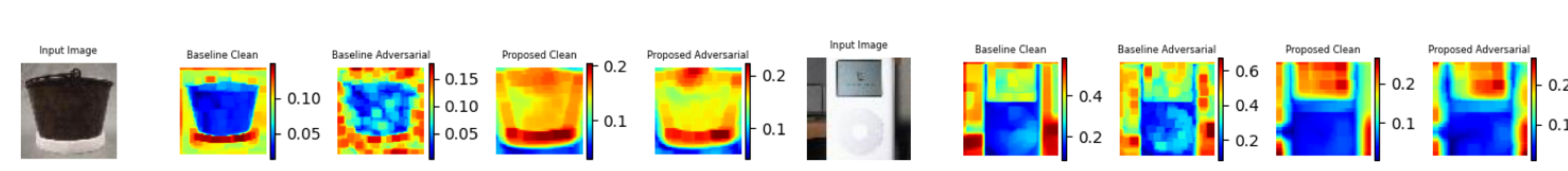


Figure 5: Feature maps of ResNet50 with TinyImagenet

- The ANF has smoothed out the feature maps compared to the baseline indicating that it can also mitigate adversarial noise within these maps

Results

Arch.	FGSM	PGD	AA	Corp-tnA	Clean acc
ResNet20 with CIFAR10					
Baseline	42.86	27.03	12.41	73.32	**91.26**
ANF	**59.56**	**59.98**	**55.14**	**78.43**	83.09
[1]	53.12	44.42	29.14	–	90.54
AT [1]	49.93	46.34	36.47	–	70.31
ResNet20 with CIFAR100					
Baseline	12.28	3.83	1.01	34.93	**65.34**
ANF	**26.8**	**26.43**	**21.58**	**48.13**	54.86
[1]	17.2	12.24	5.11	–	58.19
EfficientNet-B0 with CIFAR10					
Baseline	53.05	52.20	42.24	44.08	**92.29**
ANF	**64.95**	**66.23**	**62.27**	**80.18**	87.14
[1]	57.83	59.68	53.50	–	89.18
ResNet50 with ImageNet					
Baseline with AT	42.36	26.17	1.05	-	**64.37**
ANF with AT	**55.09**	**55.46**	**52.95**	-	61.67
AT [1]	36	37	24.32	-	58.09

Table 2: Comparison of ANF with baseline under adversarial attacks.

Key Findings

- The modified peak signal-to-noise ratio (mPSNR) values at the output of the ANF are higher
- The decision regions with ANF have better margins
- The visualized loss surfaces are smoother
- High-frequency components of noise are more attenuated
- Not only structured adversarial noise, architectures incorporating ANF exhibit better denoising in unstructured Gaussian noise compared to baseline architectures
- ANF smooths feature maps, suggesting its ability to mitigate adversarial noise

References

[1] J. Lukasik, P. Gavrikov, J. Keuper, and M. Keuper. Improving native CNN robustness with filter frequency regularization. *Transactions on Machine Learning Research*, 2023.