Wavelets Beat Monkeys at Adversarial Robustness

Jingtong Su
Center for Data Science
New York University
js12196@nyu.edu

Julia Kempe
Center for Data Science and
Courant Institute of Mathematical Sciences
New York University
kempe@nyu.edu

Abstract

Research on improving the robustness of neural networks to adversarial noise - imperceptible malicious perturbations of the data - has received significant attention. Neural nets struggle to recognize corrupted images that are easily recognized by humans. The currently uncontested state-of-the-art defence to obtain robust deep neural networks is Adversarial Training (AT), but it consumes significantly more resources compared to standard training and trades off accuracy for robustness. An inspiring recent work [Dapello et al., 2020] aims to bring neurobiological tools to the question: How can we develop Neural Nets that robustly generalize like human vision? They design a network structure with a neural hidden first layer that mimics the primate primary visual cortex (V1), followed by a back-end structure adapted from current CNN vision models. This front-end layer, called VOneBlock, consists of a biologically inspired Gabor Filter Bank with fixed handcrafted "biologically constrained" weights, simple and complex cell non-linearities and a "V1 stochasticity generator" injecting randomness. It seems to achieve non-trivial adversarial robustness on standard vision benchmarks when tested on small perturbations.

Here we revisit this biologically inspired work, which heavily relies on handcrafted tuning of the parameters of the V1 unit based on neural responses derived from experimental records of macaque monkeys. We ask whether a principled parameter-free representation with inspiration from physics is able to achieve the same goal. We discover that the wavelet scattering transform can replace the complex V1-cortex and simple uniform Gaussian noise can take the role of neural stochasticity, to achieve adversarial robustness. In extensive experiments on the CIFAR-10 benchmark with adaptive adversarial attacks we show that: 1) Robustness of VOneBlock architectures is relatively weak (though non-zero) when the strength of the adversarial attack radius is set to commonly used benchmarks. 2) Replacing the front-end VOneBlock by an off-the-shelf parameter-free Scatternet followed by simple uniform Gaussian noise can achieve much more substantial adversarial robustness without adversarial training. Our work shows how physically inspired structures yield new insights into robustness that were previously only thought possible by meticulously mimicking the human cortex. Physics, rather than only neuroscience, can guide us towards more robust neural networks.

1 Introduction

Deep Neural Networks (DNNs) are shown to be extremely sensitive to test time perturbations [Szegedy et al., 2013, Goodfellow et al., 2014]. Take the object recognition task in the computer vision field as an example; an adversary can perturb the input image by only a few pixels values per pixel point to change the model prediction Su et al. [2019].

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

Formally, given a loss function $\mathcal{L}$, a parametric model $f(\cdot;\theta)$ with parameters θ, for an input sample x with label y an adversary will try to find the worst case nearby input point $x' = x + \delta$:

$$\delta = \arg\max_{||\delta|| \leq \epsilon} \mathcal{L}(f(x+\delta;\theta), y) \tag{1}$$

for some notion of ϵ-closeness, in computer vision usually given by either $||\cdot||_2$ or $||\cdot||_\infty$.

Defenses. To defend against such attacks, a rich body of research works has addressed the problem from several viewpoints (Papernot et al. [2016], Madry et al. [2017], Ilyas et al. [2019]). Among them, Adversarial Training (AT) (Madry et al. [2017], Goodfellow et al. [2014]), an iterative procedure that successively optimizes model parameters and computes worst-case augmentations for the training data, has become the gold standard baseline to achieve robustness. However, AT consumes significantly more resources to train, sacrifices test accuracy, and thus is limited in real-world applications (Shafahi et al. [2019], Wong et al. [2020], Tsipras et al. [2018]). Several works build upon AT to learn robust data embeddings ([Pang et al., 2020, Li et al., 2021]); however, there are limited works that aim to achieve adversarial robustness without adversarial training, by extracting robust representations from data directly. Yang et al. [2020] point out that several natural image datasets are separated, and thus imply the existence of robust and accurate classifiers with local Lipschitzness. Garg et al. [2018] propose a spectral-based function that is robust near the training set, and Awasthi et al. [2021] devise a robust PCA algorithm to project data in a principled way. These two methods operate on the training set, and no test time robustness was shown, compared to AT. For NLP tasks, [Jones et al., 2020] propose a robust encoding, by projecting words to a smaller and discrete space where similar inputs share exactly the same encoding.

Inspiration from Neuroscience. Arguably, current state-of-the art object recognition models based on convolutional neural nets (CNN) are inspired by the human visual system, and a series of research work has aimed to infuse computer vision with ideas from neuroscience, trying to more closely align neural networks and human vision (see e.g. [Kubilius et al., 2019, Lindsay and Miller, 2018, Geirhos et al., 2021]). [Dapello et al., 2020] are the first to address the robustness problem from this view in a principled structural way. Based on rich prior work on V1-modeling, they devise a systematic architecture, called the VOneBlock, as a preprocessing technique to the inputs of a back-end convolutional neural network. They claim models assisted by their block, collectively called "V1 models", can be more difficult to fool, and are even as robust as those trained with state-of-the-art Adversarial Training.

Our work. In this paper, we seek to investigate why and how V1 models function and explore whether insights from physics can help to simplify and extract the underlying principles. To this end, we first revisit the VOne block proposed in [Dapello et al., 2020] and test it on the CIFAR-10 dataset [Krizhevsky et al., 2009] with a relatively simple convolutional back-end, using both the gradient-based PGD-attack (Kurakin et al. [2017], Madry et al. [2018]) and adaptive attacks from the AutoAttack benchmark ([Croce and Hein, 2020]) with standard strength (ϵ=8/255 for ℓ_∞-attacks at which point most CNNs show 0% robust accuracy). Surprisingly, we reveal that V1 models are not as powerful as believed when tested with smaller attack radius as in Dapello et al. [2020]. We first show that under fair comparison, the V1 models are only slightly robust ($\sim$ 10% test robustness), and perform strictly worse than AT. We revisit the ablation study performed in [Dapello et al., 2020] in standard attack settings to show that when we remove any one of the components of the VOneBlock the model looses any robustness whatsoever, demonstrating that the delicate interplay of these various handcrafted features is necessary.

Enter wavelets. The scattering transform was originally introduced in the mathematics literature with follow-up works that appeared in the signal processing and computer science literature. Introduced by [Mallat, 2012], it combines wavelet multiscale decompositions with a deep convolutional architecture. More recently, it has been used in a number of scientific applications: intermittency in turbulence [Bruna et al., 2015], quantum chemistry and material science [Hirn et al., 2017, Eickenberg et al., 2018, Sinz et al., 2020], plasma physics [Glinsky et al., 2020], geography [Kavalerov et al., 2019], astrophysics [Allys et al., 2019, Saydjari et al., 2021, Regaldo-Saint Blancard et al., 2020], and cosmology [Cheng et al., 2020, Cheng and Ménard, 2021a] (see [Cheng and Ménard, 2021b]). The use of the scattering transform for data preprocessing to improve machine-learning tasks has recently emerged in the context of *differential privacy*. [Tramer and Boneh, 2021] show that ScatterNet used as a feature extractor improves upon the privacy-utility trade-off of deep learning.

We show that we can replace the biologically-inspired VOneBlock by the off-the-shelf ScatterNet Oyallon and Mallat [2015], Bruna and Mallat [2013], followed with injection of uniform Gaussian noise. The only one parameter we have tuned is the variance of the noise. We achieve surprising genuine robustness that far surpasses the VOneBlock in this regime. A cartoonish summary of our work is given in Figure 1.

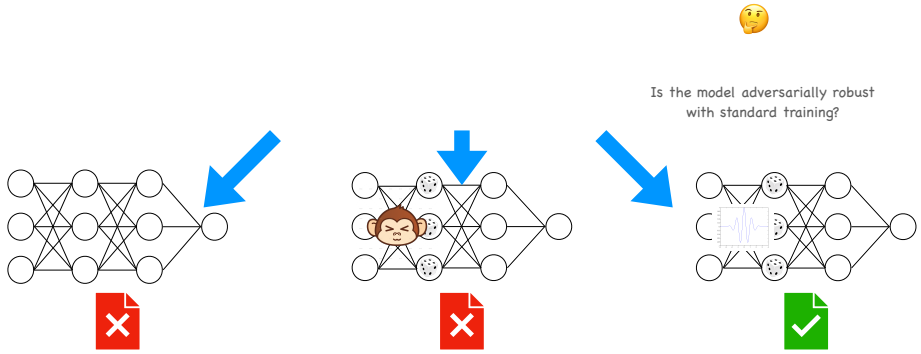

Figure 1: Illustration of model construction discussed in our paper. Only our Stochastic ScatterNet model retains genuine robustness. **Left:** Standard CNNs. **Middle:** the V1 model. The first layer is replaced with the Gabor Filter Bank (GFB) to approximate empirical primate V1 neural response, which requires massive experimental records to craft the weights, followed by injection of (neuronal) noise. **Right:** Our proposed Stochastic ScatterNet model. We use a mathematically- and physically-motivated wavelet transform in the first layer, followed by injection of uniform Gaussian noise.
We summarize our contributions as follows:

1. We replicate previous studies of the VOneBlock on CIFAR-10 for standard settings of attack strength and find that its robustness, while non-zero, falls below 10% (random guessing), and is considerably lower than for the adversarially-trained model (Table 1). Previous claims of AT-comparable PGD-robustness of the VOneBlock only hold when weakening the strength of the attacks.
2. We demonstrate that none of the components of VOne alone induce any robustness. We conclude the VOneBlock as a whole is the key towards slight robustness.
3. We replace the biologically-inspired VOne component with our Stochastic ScatterNet and show that it achieves significantly larger robustness (Table 2). This opens the route to systematic exploration of the wavelet scattering transform as a preprocessing tool to achieve human-like robust generalization.

2 Preliminaries

The V1 model. We adopt the biologically-inspired VOneBlock as proposed in [Dapello et al., 2020], which consists of a *fixed-weight* convolutional layer called the Gabor Filter Bank (GFB) with corresponding nonlinearty, and a stochastic layer called the V1 stochasticity: $\text{VOne}(x) = sto(\text{GFB}(x))$.

The mathematically parameterized GFB has its parameters tuned to approximate empirical primate V1 neural response data. It convolves the RGB input images with Gabor filters of multiple orientations, sizes, shapes, and spatial frequencies. To instantiate a VOneBlock, we use publicly available code that randomly samples (and then fixes) the values for the GFB parameters according to empirically observed distributions of preferred orientation, peak spatial frequency, and size/shape of receptive fields [De Valois et al., 1982a,b, Ringach, 2002].

A defining property of neuronal responses is their stochasticity. In awake monkeys, the spike train (corresponding to *activations*) for each trial is approximately Poisson: the spike count variance is equal to the mean. Thus, the VOne stochasticity - also referred to as *neuronal noise* - injects noise into the resulting activations z_i as

$$sto(z_i) = z_i + \frac{\mathcal{N}(0, |l(z_i)|)}{a}, \; l(z_i) = a \cdot z_i + b. \tag{2}$$

The affine transformation $l(z)$ serves to normalize mean activations to correspond to those of a population of primate V1 neurons measured in a 25ms time-window, and a and b are set accordingly (see App. A for more detail).

In our ablation studies for the VOne block we study several simplifications of the neuronal noise.

- V1-Magnitude Gaussian: We remove the affine transformation $l(z)$ (setting $a = 1$ and $b = 0$). This results in **magnitude-aware Gaussian noise** $sto_{MG}(z_i) = z_i + \mathcal{N}(0, |z_i|)$.
- V1-Gaussian: We use uniform **Gaussian noise** $Gau(z_i) = z_i + \mathcal{N}(0, \sigma^2)$ setting σ^2 to a constant.
- V1-None: We remove the injected noise altogether.

3 Experimental Results on CIFAR-10

Throughout our experiments, we use a simple three-layer CNN of width 64 with Max-Pooling as the back-end model. For the V1 model we replace its first layer with the VOne block. We use the ScatterNet implementation from *Kymatio* Andreux et al. [2020]. Full experimental details can be found in Appendix A. For adversarial robustness, we use FGSM Goodfellow et al. [2014] and the standard ℓ_∞ Projected Gradient Descent (PGD), with 20 steps, step size $\sigma = 2/255$ and budget $\epsilon = 8/255$. To check for genuine robustness, we apply the AutoAttack (AA) benchmark Croce and Hein [2020] with the same strength. Because we introduce stochasticity, we need to take extra care when attacking our models. See Appendix B.

Robustness of VOne: Table 1 shows test statistics of our V1 models, together with several variations of stochasticity, and the adversarially trained baseline. We find that robustness of V1 models is only slight, though genuine and non-zero, and much inferior to the AT model. This comparison was much more favorable in [Dapello et al., 2020], and is manifestly due to *weakening the strength of the attack* to $1/16$ of the standard magnitude. However, since the baseline CNN has *zero* robustness, we conclude that V1 models garner some robustness benefits. None of the VOne components separately give meaningful robustness, as discussed in Appendix C.

Table 1: V1 model v.s. AT baseline CIFAR-10 Test Performance (%). We boldface the **best** result.

			Robust	
Model	**Clean**	FGSM	PGD ℓ_∞ 20	AA
V1-Neuronal	58.66 ± 0.56	$\mathbf{18.07 \pm 1.54}$	$\mathbf{9.47 \pm 0.70}$	$\mathbf{27.57 \pm 0.80}$
V1-Magnitude Gaussian	63.44 ± 0.44	7.14 ± 0.58	0.71 ± 0.16	10.55 ± 0.87
V1-Std-0.35-Gaussian	63.01 ± 1.17	1.74 ± 0.32	0.01 ± 0.01	7.22 ± 0.60
V1-Std-0.20-Gaussian	63.75 ± 0.47	2.02 ± 0.20	0.00 ± 0.00	5.64 ± 0.41
V1-Std-0.15-Gaussian	64.07 ± 0.46	1.91 ± 0.14	0.03 ± 0.03	4.62 ± 0.13
V1-Std-0.10-Gaussian	63.86 ± 1.46	1.29 ± 0.14	0.05 ± 0.04	3.96 ± 0.28
V1-None	$\mathbf{64.42 \pm 1.13}$	1.51 ± 0.51	0.93 ± 0.60	0.00 ± 0.00
AT Baseline	58.07	33.94	31.49	26.18

Robustness of ScatterNet: Table 2 shows robustness of our Stochastic ScatterNet with uniform Gaussian noise of various magnitudes instead of the VOne module. Surprisingly, we observe that the wavelet scattering transform combined with simple uniform Gaussian noise can achieve much higher robustness than biologically-motivated V1-models.

Table 2: Stochastic ScatterNet Model CIFAR-10 Test Performance (%). We boldface the **best** result.

			Robust	
Stochasticity Type	**Clean**	FGSM	PGD ℓ_∞ 20	AA
Magnitude Gaussian	58.55 ± 0.21	16.98 ± 0.29	4.11 ± 0.14	22.90 ± 0.30
Std-0.35-Gaussian	36.02 ± 0.42	23.99 ± 0.64	22.98 ± 0.26	33.52 ± 0.38
Std-0.2-Gaussian	46.12 ± 0.36	26.16 ± 0.12	$\mathbf{24.84 \pm 0.57}$	39.88 ± 0.52
Std-0.15-Gaussian	50.83 ± 0.42	$\mathbf{26.38 \pm 0.44}$	24.03 ± 0.08	$\mathbf{41.43 \pm 0.40}$
Std-0.1-Gaussian	56.27 ± 0.59	24.45 ± 0.28	20.74 ± 0.24	39.75 ± 0.26
None	$\mathbf{76.44 \pm 0.30}$	6.03 ± 0.20	0.04 ± 0.04	0.01 ± 0.01

In particular, we observe that simple uniform Gaussian noise leads to higher robustness than neuronal noise of varying magnitude. Yet, similar to the VOne architecture, stochasticity seems indispensable for robustness.

4 Conclusion

In this paper, we show that the biologically-inspired VOneBlock is only slightly robust under standard attack settings, and none of its components can function by itself. Surprisingly, we observe that the wavelet scattering transform combined with simple uniform Gaussian noise can achieve much higher robustness than these well-motivated models which require mounts of experimental primate data to craft. In future work, we aim to explore how our stochastic ScatterNet fares for other dataset and back-end architectures and to gain further theoretical understanding of this surprising robustness.

We advocate to continue to study how bringing insights from physical models can assist vision tasks, in addition to biologically-inspired ones.

Broader Impact

The potential ethical aspects and future societal consequences of our work are aligned with those of computer vision and robustness. Our work empirically examines the effectiveness of robust networks inspired by biological and physical models and contributes to diminish potential susceptibility to malicious attacks. Moreover, we advocate a strategy that does not depend on experimental results from animals, thus avoiding potential ethical pitfalls.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. *Advances in Neural Information Processing Systems*, 33: 7779–7792, 2020.

Yao Li, Martin Renqiang Min, Thomas Lee, Wenchao Yu, Erik Kruus, Wei Wang, and Cho-Jui Hsieh. Towards robustness of deep neural networks via regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7496–7505, 2021.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Shivam Garg, Vatsal Sharan, Brian Hu Zhang, and Gregory Valiant. A spectral view of adversarially robust features. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10159–10169, 2018.

Pranjal Awasthi, Vaggos Chatziafratis, Xue Chen, and Aravindan Vijayaraghavan. Adversarially robust low dimensional representations. In *Conference on Learning Theory*, pages 237–325. PMLR, 2021.

Erik Jones, Robin Jia, Aditi Raghunathan, and Percy Liang. Robust encodings: A framework for combating adversarial typos. *arXiv preprint arXiv:2005.01229*, 2020.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

Grace W Lindsay and Kenneth D Miller. How biological attention mechanisms improve task performance in a large-scale visual system model. *ELife*, 7:e38105, 2018.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 2018.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65 (10):1331–1398, 2012.

Joan Bruna, Stéphane Mallat, Emmanuel Bacry, and Jean-François Muzy. Intermittent process analysis with scattering moments. *The Annals of Statistics*, 43(1):323–351, 2015.

Matthew Hirn, Stéphane Mallat, and Nicolas Poilvert. Wavelet scattering regression of quantum chemical energies. *Multiscale Modeling & Simulation*, 15(2):827–863, 2017.

Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, Stéphane Mallat, and Louis Thiry. Solid harmonic wavelet scattering for predictions of molecule properties. *The Journal of chemical physics*, 148(24):241732, 2018.

Paul Sinz, Michael W Swift, Xavier Brumwell, Jialin Liu, Kwang Jin Kim, Yue Qi, and Matthew Hirn. Wavelet scattering networks for atomistic systems with extrapolation of material properties. *The Journal of Chemical Physics*, 153(8):084109, 2020.

Michael E Glinsky, Thomas W Moore, William E Lewis, Matthew R Weis, Christopher A Jennings, David J Ampleford, Patrick F Knapp, Eric C Harding, Matthew R Gomez, and Adam J Harvey-Thompson. Quantification of maglif morphology using the mallat scattering transformation. *Physics of Plasmas*, 27(11):112703, 2020.

Ilya Kavalerov, Weilin Li, Wojciech Czaja, and Rama Chellappa. Three-dimensional fourier scattering transform and classification of hyperspectral images. *arXiv preprint arXiv:1906.06804*, 2019.

Erwan Allys, F Levrier, S Zhang, C Colling, B Regaldo-Saint Blancard, F Boulanger, P Hennebelle, and S Mallat. The rwst, a comprehensive statistical description of the non-gaussian structures in the ism. *Astronomy & Astrophysics*, 629:A115, 2019.

Andrew K Saydjari, Stephen KN Portillo, Zachary Slepian, Sule Kahraman, Blakesley Burkhart, and Douglas P Finkbeiner. Classification of magnetohydrodynamic simulations using wavelet scattering transforms. *The Astrophysical Journal*, 910(2):122, 2021.

Bruno Regaldo-Saint Blancard, François Levrier, Erwan Allys, Elena Bellomi, and François Boulanger. Statistical description of dust polarized emission from the diffuse interstellar medium-a rwst approach. *Astronomy & Astrophysics*, 642:A217, 2020.

Sihao Cheng, Yuan-Sen Ting, Brice Ménard, and Joan Bruna. A new approach to observational cosmology using the scattering transform. *Monthly Notices of the Royal Astronomical Society*, 499 (4):5902–5914, 2020.

Sihao Cheng and Brice Ménard. Weak lensing scattering transform: dark energy and neutrino mass sensitivity. *Monthly Notices of the Royal Astronomical Society*, 507(1):1012–1020, 2021a.

Sihao Cheng and Brice Ménard. How to quantify fields or textures? a guide to the scattering transform. *arXiv preprint arXiv:2112.01288*, 2021b.

Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021.

Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2865–2873, 2015.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Russell L De Valois, E William Yund, and Norva Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision research*, 22(5):531–544, 1982a.

Russell L De Valois, Duane G Albrecht, and Lisa G Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, 22(5):545–559, 1982b.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 2002.

Mathieu Andreux, Tomás Angles, Georgios Exarchakis, Roberto Leonarduzzi, Gaspar Rochette, Louis Thiry, John Zarka, Stéphane Mallat, Joakim Andén, Eugene Belilovsky, et al. Kymatio: Scattering transforms in python. *J. Mach. Learn. Res.*, 21(60):1–6, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

A Experimental Details

A.1 Training Strategy

Throughout our paper, we adopt the Adam optimizer Kingma and Ba [2014] to train the V1 models and the Stochastic ScatterNet models. The initial learning rate is fixed to 1e-3, and the number of epochs is fixed to 200. For AT baseline, we train the network with SGD, initial learning rate 1e-1, and decay the learning rate by 10 at the 100- and 150-th epoch. We run each experiment for 3 times, and report the average and standard deviation.

A.2 Model Structure

For the AT baseline, we adopt a simple 3-layer convolutional structure, with Max-Pooling and a linear layer in the end. All convolutional layers set the number of output channels to 64, and stride is set to 1. We replace the first layer with a V1 module, or a ScatterNet transform.

For the V1 block[1], we use its default setting, consisting of a 512x32x32 feature map. We set the number of input channels of the convolutional layer behind to 512 to fit this module. For the hyperparameters in Eq. (2), we take the default setting provided by the code of [Dapello et al., 2020] that imitates the mean stimulus response and spontaneous activity: $a = 0.35$ and $b = 0.07$.

For the ScatterNet, the output shape is 243x8x8 by default. We set the number of input channels of the convolutional layer behind to 243 to fit this module. We also cancel the Max-Pooling after the third layer, to keep the representation dimension at least at the standard, 64x4x4.

A detailed description is given in Table 3.

Table 3: Model illustration in our paper. The dimension of inputs/features and the layer structures are described top-down. We adopt the default setting of the V1 block and the ScatterNet module, leading to the number of output channels of 512 and 243. We adjust the 2nd and 3rd layer of the CNN to fit these feature shapes.

Simple CNN	**V1 Model**	**Stochastic ScatterNet**
3x32x32	3x32x32	3x32x32
Conv + MaxPool	V1 Module + Noise	ScatterNet + Noise
64x16x16	512x32x32	243x8x8
Conv + MaxPool	Conv + MaxPool	Conv + MaxPool
64x8x8	64x16x16	64x4x4
Conv + MaxPool	Conv + MaxPool	Conv
64x4x4	64x8x8	64x4x4
Linear	Linear	Linear
10	10	10

B Adversarial Attacks

Categorized by having access to the model parameters or not, the attack methods that try to solve Eq. 1 can be divided into white-box and black-box, respectively. The most popular attack baseline is a white-box attack, implemented by multi-step Projected Gradient Descent (PGD) [Madry et al., 2018, Kurakin et al., 2017]. Here the vector norm $||\cdot||$ is usually taken as $||\cdot||_\infty$ or $||\cdot||_2$, and we choose $||\cdot||_\infty$. We calculate the gradient iteratively

$$\nabla_x \mathcal{L}(f(x+\delta;\theta), y) \tag{3}$$

and update the adversarial example with the steepest direction according to the norm constraint:

$$x' \leftarrow \Pi_{\mathcal{B}}\{x' + \alpha \cdot \text{sign}[\nabla_x \mathcal{L}(f(x+\delta;\theta), y)]\}. \tag{4}$$

$\Pi_{\mathcal{B}}$ denotes the projection step, to ensure the generated adversarial example in each loop satisfies a certain norm-based constraint, *i.e.,* $\delta = ||x' - x|| \leq \epsilon$ for some ϵ. When the number of iterations is 1 this attack is called FGSM [Goodfellow et al., 2014].

To combat the effects of noise in the gradients, we adapted our attack such that at every PGD iteration, we take 10 gradient samples and move in the average direction to increase the loss, following Athalye et al. [2018].

We also evaluate robustness against the adaptive suite of the AutoAttack benchmark [Croce and Hein, 2020], which contains both black-box and white-box attacks and is considered a minimal set of attacks to establish genuine robustness. We use the AA option to include a gradient averaging mechanism, to produce reliable perturbations for models with stochasticity.

C None of the Components of VOne alone Induce Robustness

In this section, we perform ablation studies to understand the underlying behaviour of the V1 model, and in particular how each of its elements help the model gain robustness. We divide this examination

[1]We adopt the code from `https://github.com/dicarlolab/vonenet`

into three main parts. We check the utility of: 1) the Gabor Filter Bank, 2) the stochasticity, namely the neuronal/Gaussian noise, and 3) the affine transformation used with the neuronal noise.

Bio-Inspired Gabor Filter Bank Alone Does not Yield Robustness. First, we show the Gabor Filter Bank itself cannot extract features that are robust against adversarial perturbations. We simply turn off the stochasticity in the VOneBlock, and leave the GFB alone with the trainable back-end, training the model with *no noise*. The penultimate row of Table 1 shows the test-robustness: V1-None has neither PGD robustness nor AA robustness. The fact that stochasticity is indispensable matches the observations in [Dapello et al., 2020].

Stochasticity Alone is Insufficient. In order to evaluate the utility of stochasticity in a fair way, we remove the Gabor Filter Bank and instead add a *fixed-weight, randomly-initialized convolutional layer*. In other words, we replace the GFB whose weights are drawn from hand-crafted distributions and then fixed, with randomly drawn convolutional weights that are fixed, and retain the stochasticity alone. Table 4 shows the test results of this "random feature" stochastic model. We see that with stochasticity alone, the models have a little FGSM robustness, and slightly better AA robustness but failing PGD-robustness. Thus, we establish that stochasticity alone not sufficiently useful for robustness, at marginally at best.

Table 4: (Fixed-weights) Random Feature Stochastic Model CIFAR-10 Test Performance (%). We boldface the **best** result.

		Robust		
Stochasticity Type	**Clean**	FGSM	PGD ℓ_∞ 20	AA
Neuronal	68.26 ± 0.52	$\mathbf{5.81 \pm 0.31}$	0.07 ± 0.02	11.26 ± 0.44
Magnitude Gaussian	$\mathbf{69.67 \pm 0.96}$	4.74 ± 0.32	0.03 ± 0.03	$\mathbf{14.03 \pm 0.20}$
Std-0.35-Gaussian	64.53 ± 0.72	2.55 ± 0.05	0.04 ± 0.01	5.35 ± 0.08
None	68.66 ± 0.30	2.73 ± 0.50	$\mathbf{2.40 \pm 0.46}$	0.00 ± 0.00

Varying the stochasticity. To understand whether the biologically-inspired neuronal noise is indispensible for robustness of the VOne block, or whether other, more generic forms of stochasticity would suffice, we replace Neuronal noise with its handcrafted affine transformation by either Magnitude Gaussian (effectively removing the affine transformation) or standard Gaussians of varying magnitude. Results shown in Table 1 show that both the affine transformation and the magnitude-dependent variance of the noise seem indispensable. While removing the affine transformation retains at least a modicum of AA-robustness, it fails against PGD attacks. Making the Gaussians of uniform variance further degrades performance and leads to vanishing robustness.

Interestingly, for ScatterNets a similar variation of the stochastic noise, shown in Table 2, has a different effect. Here, uniform Gaussian noise yields much higher robustness than Magnitude Gaussian. We also see that stochasticity is important to achieve robustness using the wavelet scattering transform.