Neural Population Geometry Reveals the Role of Stochasticity in Robust Perception

Anonymous Author(s)
Affiliation
Address
email

Abstract

Adversarial examples are often cited by neuroscientists and machine learning researchers as an example of how computational models diverge from biological sensory systems. Recent work has proposed adding biologically-inspired components to visual neural networks as a way to improve their adversarial robustness. One surprisingly effective component for reducing adversarial vulnerability is response stochasticity, like that exhibited in biological neural recordings. Here, using recently developed geometrical techniques from computational neuroscience, we investigate how adversarial perturbations influence the internal representations of standard, adversarially-trained, and biologically-inspired stochastic networks. We find distinct geometric signatures for each type of network, revealing different mechanisms for achieving robust representations. Next, we generalize these results to the auditory domain, showing that neural stochasticity also makes auditory models more robust to adversarial perturbations. Geometric analysis of the stochastic networks reveal overlap between representations of clean and adversarially perturbed stimuli, and quantitatively demonstrate that competing geometric effects of stochasticity mediate a tradeoff between adversarial and clean performance. Our results shed light on the mechanisms of robust perception and the potential benefits of stochastic responses, paving the way for increasingly robust models and a better understanding of the role stochasticity plays in biological computation.

1 Introduction and related work

In recent years, artificial neural networks (ANNs) have come to dominate both visual object recognition and auditory recognition tasks [1, 2, 3], establishing them as leading candidate models of human perception [4, 5, 6]. These models have outperformed previous models at predicting neural activity and behavior [7, 8]. However, they still exhibit many non-human-like traits [9, 10]. One such failure is in the existence of adversarial perturbations – small changes to stimuli explicitly crafted to fool a model while remaining imperceptible to humans [11, 12, 13] – which demonstrate the fragility of some ANNs as models of biological perception. Many attempts have been made to address this issue and create ANNs that are more robust to adversarial attacks [14, 15, 16]. Recently, Dapello et. al. discovered that one such method, known as adversarial training, not only reduces the network's adversarial vulnerability but also yields network representations that are more similar to those in the primate primary visual cortex [17]. Following this thread, the authors developed VOneNets, a class of networks that simulates the primate primary visual cortex at the front of a convolutional neural network, and shows improved robustness to adversarial attacks without adversarial training.

A key component of VOneNet was the inclusion of stochastic representations during both training and inference. This stochasticity is inspired by biological sensory neurons, which exhibit trial-to-trial variability across presentations of the same stimulus [18]. The basis of this stochasticity and its

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

implications for information processing are open questions in neuroscience [19]. Previous work in machine learning has also reported that adding noise can improve the adversarial robustness of a model: Liu et al. [14] used random noise applied to pixels and intermediate layers to improve adversarial robustness, and Cohen et al. [16] demonstrated a method to transform a model with Gaussian noise in the pixel space to one with certified robustness to attacks of a given strength. Unlike VOneNets, both [14] and [16] rely on using multiple inference passes with different noise samples to improve robustness.

Here, using recently developed manifold analysis techniques from computational neuroscience [20] we investigate the neural population geometry [21] of standard, adversarially-trained, and biologically-inspired stochastic networks in response to clean and adversarially-perturbed examples. We investigate both the visual and auditory domains. We present several key findings:

- Using manifold analysis, we demonstrate that adversarially-trained networks and stochastic VOneNets have distinct geometric signatures in their internal representations. Adversarially trained networks shrink the size of the manifold formed by perturbations around a clean image. By contrast, network stochasticity increases the size of such manifolds.
- We demonstrate the generality of our findings by translating the results to a novel biologically-inspired auditory ANN, StochCochResNet50, that includes stochastic responses. Stochastic representations make auditory networks more robust to adversarial perturbations, and the underlying neural population geometry is largely consistent with observations in vision networks.
- Manifold analyses of the stochastic networks reveal overlap between the representations of adversarial examples and clean stimuli. We quantitatively survey the stochasticity conditions leading to the overlap, and map a competing geometric effect mediating a trade-off between clean and adversarial performance.

2 Methods and experimental setup

2.1 Replica-based manifold analysis

In this paper, we use replica mean-field theoretic manifold analysis (MFTMA) [20, 22], which formally connects the linear decodability of object manifolds, defined as a set of stimulus-evoked representations in neural state space [23], to their geometrical properties. Here we provide a brief description of the key quantities (See SM 2 and [20] for a complete treatment of MFTMA). Using this framework, we analyze the size, shape, and distribution of object manifolds as they are transformed throughout the network, gaining insight into the neural population geometry underlying performance of the classification task. Specifically, given P object manifolds (measured as feature point clouds) in N dimensions, the MFTMA analysis returns capacity (a measure of linear separability) and the manifold dimensions, radii, and center correlations associated with capacity estimation.

Manifold capacity ($\alpha = P/N$) refers to the maximum number of object manifolds (P_{max}) that can be linearly separated given N features, and characterizes the linearly-decodable object information per feature dimension. MFTMA estimates α through measures of the manifold dimension (D_M), manifold radius (R_M), and manifold center correlation, which refer to dimensionality, size, and distribution of object manifolds relevant for the linear classification. In our analysis, we combine the manifold dimension and radii together into a single **manifold width** measure, defined as $R_M \cdot \sqrt{D_M}$, which captures the general width of the convex hull of a manifold [1]. Manifold width formally links the linear separability of object manifolds with their underlying geometric structure. Specifically, small values of manifold width yield more linearly separable manifold geometry. While manifold center correlation plays a role in determining capacity, we did not find it insightful for discriminating model defenses; see SM for measured values.

In this work, we characterize two types of object manifolds:

1. **Class manifolds**: each manifold is defined by measuring activations evoked by multiple examples drawn from a specific class of visual or auditory stimuli (i.e., object or word identity) (Figure 1A). The variability within the manifold can come from different exemplars

[1] Full values of D_M, R_M are provided in SM.

within the class, but may also be influenced by adversarial perturbations, and/or neural noise in stochastic networks.

2. **Exemplar manifolds**: each manifold is defined by measuring activations evoked by multiple instances of a single exemplar (i.e., image or utterance) with the manifold variability reflecting either norm-bounded adversarial perturbations[2] (Figure 2A) and/or the influence of stochastic representations.

2.2 Quantifying object manifold overlap

The MFTMA framework provides a geometric description of exemplar manifolds evoked by adversarial and clean stimuli in the case of many object manifolds. However, we also aim to directly characterize the amount of overlap between the adversarial and clean exemplar manifolds generated from the same stimuli. To do so, we use a notion of object manifold overlap defined as the generalization error of a linear Support Vector Machine (SVM) fit to separate two sets of representations, and report the chance-normalized error rate on the heldout data. A normalized error rate of 1 means that two representations are completely overlapped, and 0 means that they are completely separable and thus not overlapped. We use the SVM implementation from `scikit-learn` with a train/test split of 0.8/0.2. As manifold inseparability as measured by SVM is only one possible description of manifold overlap, we further detail another analysis used pairwise distance distribution in Section SM 3.

2.3 Adversarial attacks

For performing adversarial attacks, we use either single-step fast gradient sign method (FGSM) with a random starting location or multi-step projected gradient descent (PGD). Unless otherwise specified, we use L_∞ norm constrained attacks. All attacks are untargeted and performed on a model-by-model basis. When the goal is to evaluate adversarial accuracy in newly created networks with stochastic internal representations, we use ensemble-PGD, where each step is in the mean direction of k samples of the noisy gradient to ensure a useful signal [24].

3 Manifold analysis of robustness in ImageNet-trained networks

We first investigate the geometric properties of previous networks that perform the ImageNet classification task and have enhanced adversarial robustness. We aim to compare the geometric signatures of robust networks trained with adversarial training to those trained with stochastic activations.

3.1 Models and Dataset

We use data sampled from the ImageNet [25] test set. For class manifold analysis, the clean stimulus set consists of 50 classes, with each class containing 50 unique exemplar images for a total of 2500 unique images. For exemplar manifold analysis, 100 unique images are sampled from the ImageNet test set and each is perturbed with FGSM 50 times for 5000 unique images. We analyzed three publicly available ImageNet-trained models, including ResNet50 with standard ImageNet training, ResNet50 adversarially-trained on ImageNet with an $L_\infty = 4/255$ penalty (ATResNet50), and VOneResNet50, a ResNet50-based model with the first conv-relu-maxpool layers replaced by a linear-nonlinear-Poisson model front-end (called the VOneBlock) with Gabor filters and noise fitted to primate neuronal data. In addition, we analyze a novel variant of VOneResNet50 (GVOneResNet50) with additive Gaussian noise scaled to match the mean over all units in the output of the VOneBlock in response to a set of reference stimuli. GVOneResNet50 performs similarly to the original VOneResNet50 in clean and adversarial conditions (see SM 1.1.1 for more details). While the first two models are deterministic, VOneResNet50 and GVOneResNet50 have stochastic internal representations at the output of the VOneBlock. Additional model details are in section SM 1.1.

3.2 MFTMA reveals unique robustness strategies for VOneNet and adversarial training

[2] We replicate many experiments for non-adversarial but random perturbations within the ϵ-sized ball to directly compare model geometry measured on the same stimuli. Results are similar and provided in SM.

Our analysis begins by observing the effects of norm-bounded, gradient-based adversarial attacks on the class manifold geometry of representations in ResNet50, ATResNet50, VOneResNet50, and GVOneResNet50. For all experiments, Images perturbed with PGD L_∞ constraints of $\epsilon \in [0, 1/1020, 1/255]$ are shown to a model and intermediate representations are extracted from the network, randomly projected down to 5000 features, and analyzed with MFTMA (Figure 1A). Much like has been previously observed using MFTMA analysis of deep neural networks performing object recognition [22], models develop separable representations in later layers, with capacity peaking in the final layer before classification.

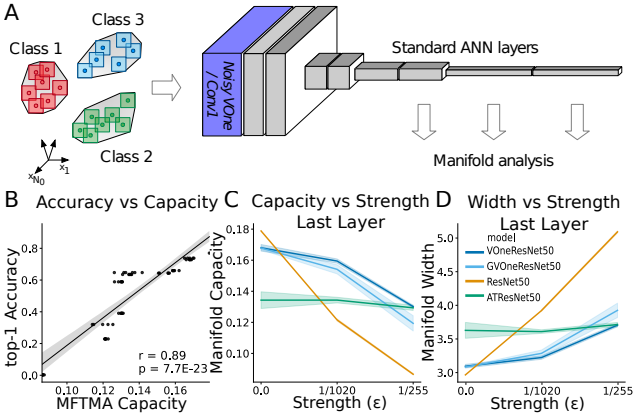

Figure 1: **Geometry and capacity of clean and adversarially-perturbed class manifolds in ImageNet neural networks** (**A**) Clean or adversarially-perturbed stimuli grouped into class manifolds are provided as input to a model, internal representations are extracted, and MFTMA analysis is applied. (**B**) MFTMA capacity from the final model layer reflects accuracy from clean and adversarial images. (**C**) Manifold capacity and (**D**) manifold width for the last model layer is plotted against attack strength, revealing enhanced representational invariance in ATResNet50 (green) and to a lesser degree VOneResNet50 (dark blue) and GVOneResNet50 (light blue), while undefended ResNet50 (orange) is least invariant. Error bars are standard deviation (STD) across 5 seeds in all plots.

We first focus on this penultimate layer, where class manifold representations have become linearly separable and capacity peaks. For the full layer-wise trajectories, see SM 4.3. We empirically confirm that capacity of the penultimate layer is predictive of the top-1 accuracy across models and attack strengths (Figure 1B), indicating that geometry of class manifold representations are sensitive to the attacks we are investigating. MFTMA exposes how capacity (Figure 1C) and class manifold width (Figure 1D) vary in the four models as a function of perturbation strength. Adversarial attacks cause the width of class manifolds to grow most rapidly in the undefended ResNet50, to a lesser degree in both VOneNets, and least of all in ATResNet50.

Next, to investigate how the models represent adversarial perturbations around individual images, we introduce another approach to use MFTMA to investigate the mechanisms of adversarial robustness. Instead of class manifolds, we consider exemplar manifolds consisting of points sampled using FGSM with L_∞ constraints of $\epsilon \in [0, 1/1020, 1/255, 4/255, 8/255]$ from a random starting in the epsilon-sized ball around 100 exemplar images, with the goal of tracing how these ϵ-sized ball exemplar manifolds develop as they travel through subsequent layers of our networks of interest (Figure 2A).

Restricting our analysis to the layer-wise trajectory of $\epsilon = 8/255$ sized ball exemplar manifolds (Figure 2B), the most salient feature is how distinct the trajectories are for the ATResNet50 and VOneNets. Our results indicate the defense mechanism induced through adversarial training generally shrinks the epsilon-sized ball around an exemplar as it propagates through the network, effectively mapping small perturbations around the clean image to a single point in later layers of the network. By contrast, the ϵ-sized ball is not shrunk in stochastically operating VOneNets or the standard ResNet50. In fact, at the noise layer in VOneNets, the ϵ-sized ball exemplar manifolds become highly entangled, with capacity dropping dramatically before recovering in downstream layers.

Figure 2C demonstrates that when we normalize the capacity of the adversarially-induced exemplar manifolds by the capacity for representations of the same unperturbed images (clean exemplar manifolds), we observe that VOneNets are far more stable than the standard ResNet50 network. At lower attack strengths where VOneNets' accuracies are minimally degraded, their normalized capacity remains stable, indicating that they represent the perturbed images with approximately the same capacity as clean images. In all experiments, we find similar geometric signatures for the Gaussian and Poisson VOneNets. For simplicity in the rest of the paper we focus on stochastic representations with Gaussian noise, but future work could extend the results to the Poisson noise case.

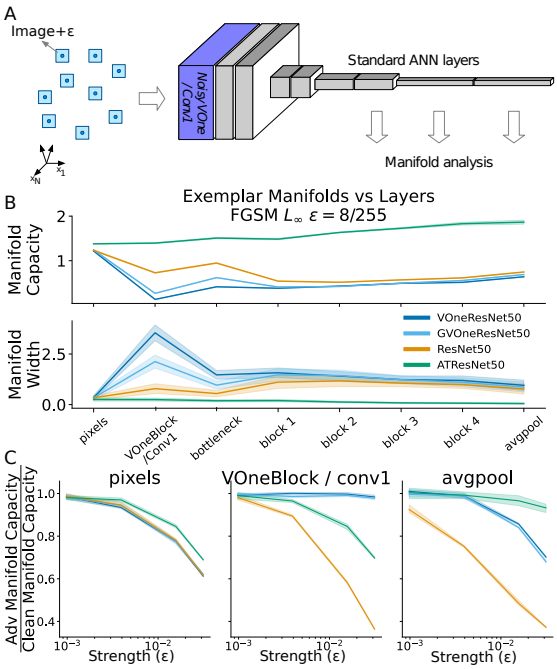

Figure 2: **Geometry and capacity of adversarially-perturbed exemplar manifolds in neural networks** (**A**) Each manifold consists of a set of adversarially-perturbed examples within an ϵ-sized ball around a single exemplar image. (**B**) Mean capacity and manifold width as a function of layer depth; ATResNet50 shrinks the absolute size of ϵ-sized ball manifolds, while undefended ResNet50, VOneResNet50, and GVOneResNet50 do not. (**C**) Adversarial representation capacity normalized by clean representation capacity is shown for (left to right) pixels, first convolutional layer (the VOneBlock for VOneNets), and the last layer before the softmax output layer (performing average pooling). VOneResNet50, GVOneResNet50, and ATResNet50 show less of a capacity decrease for low-ϵ perturbed image manifolds compared to the standard ResNet50. Error bars represent STD across 5 seeds in all plots; unnormalized capacity detailed in SM 4.2.

4 Stochastic representations improve robustness in auditory networks

The results from the previous section highlight distinct geometric profiles for adversarially trained models and for models with stochastic responses. To demonstrate that the observed neural population geometry generalizes across modalities, we compare biologically-inspired and adversarially-trained auditory models trained to perform a speech recognition task. Our focus here is not on generating an auditory model that is fully defended against adversarial attacks, but rather to test whether the influence of stochastic representations and geometric trends observed in the previous section generalizes across domains. We thus focus on L_∞ attacks acknowledging that L_p audio attacks that successfully change a network's prediction are often audible to human listeners [26, 27].

4.1 Models and dataset

Auditory models are trained to perform the word recognition task in the Word-Speaker-Noise dataset introduced in [10]. Networks learn to distinguish the word present in the middle of a two-second speech clip from 793 word classes. Word class manifolds are constructed from 50 unique words with 50 unique speakers saying the word, drawn from the Wall Street Journal Corpus [28] (2500 unique speech clips). Exemplar manifolds are measured from a random selection of 100 example clips from the class manifold dataset, with 50 samples measured for each clip.

Auditory models contain a biologically-inspired 'cochleagram' representation [29, 30], followed by a ResNet50 architecture. The cochleagram consists of differentiable operations, allowing generation of adversarial examples in the waveform. Unlike VOneNets, the architecture maintains the conv-relu-maxpoool before the first residual block, on the grounds that the cochleagram models the ear rather than primary auditory cortex. For stochastic models, Gaussian noise with standard deviation σ is added after the cochleagram representation (Figure 3A), and we refer to this model as StochCochResNet50. A comparison network without noise ($\sigma = 0$) is similarly trained and evaluated (CochResNet50). A robust model is achieved with adversarial training using L_∞ perturbations with $\epsilon = 10^{-3}$ and maximum of 5 steps for the attack (ATCochResNet50). Training details are presented in SM 1.2.

4.2 Adversarial robustness in auditory models with stochastic cochleagrams

When setting the level of noise in the VOneBlock, neural responses from Macaque V1 to a specific image set were used for tuning the relative amplitude of the stochastic representations. However, this type of neural data is not readily available for all sensory areas. We empirically found an optimal level of noise for our auditory models by varying the level of additive Gaussian noise in

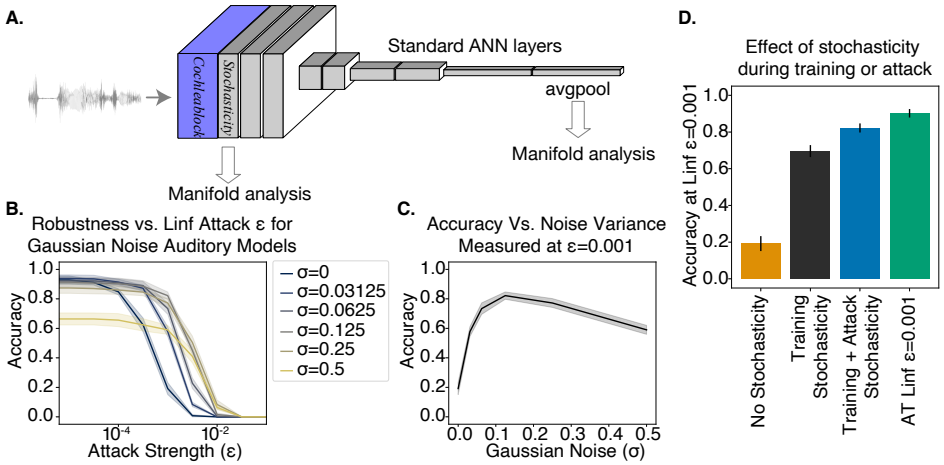

Figure 3: **Robustness of auditory networks trained with stochastic cochlear representations** (**A**) Depiction of auditory network with stochastic cochleagram. (**B**) Performance of auditory networks trained with varying stochasticity levels evaluated on L_∞ adversarial attacks, averaged over 100 randomly chosen test examples of clean speech. Error bars are STD across 5 sets of test stimuli. (**C**) An intermediate level of Gaussian noise yields the best adversarial robustness. (**D**) Stochastic activations during training but not during testing yields improvements in adversarial robustness, and maintaining stochasticity during adversarial generation and testing further increases robustness.

the stochastic cochleagram. The robustness of these models to L_∞ perturbations was evaluated, varying the allowable size of the perturbation. To ensure a reliable signal for adversarial attack generation, model gradients are sampled eight times for each PGD iteration. The resulting accuracy for each model is shown in Figure 3B. As noise is increased the model's robustness to adversarial perturbations increases, but only up to a point, yielding a peak in the robustness curve (Figure 3C). The accuracy for StochCochResNet50 with noise during the training and attack was only slightly worse than ATCochResNet50 (Figure 3D). Much like what was observed in VOneResNet50 [17], including noise only during training but not during adversarial evaluation significantly increases adversarial robustness. This suggests that the adversarial performance cannot only be explained by masking adversarial perturbations during the attack and instead reflects downstream representational changes from training with stochasticity. The additional performance boost when stochasticity is included during inference suggests a secondary defense driven by the stochastic activations during evaluation. These results show that by including stochasticity in a biologically-inspired peripheral model we can improve a model's performance on adversarial examples and by extension learn a more human-like representation.

4.3 Neural population geometry of auditory networks

We investigated whether the neural population geometry used by StochCochResNet50 is similar to that observed in the visual domain. As with the ImageNet experiments, clean or perturbed audio is presented to the model and intermediate representations are extracted, randomly projected down to 5000 features if the layer has greater than 5000 of activations, and analyzed with MFTMA. To examine the key geometric factors associated with an increase in capacity, we look at class manifold width verses center correlations, which together determine the capacity estimate. The two geometric properties show a similar dependence for the auditory (Figure 4A) and visual (Figure 4B) networks when varying the ϵ-sized perturbations used to construct the class manifolds. This points to the same underlying mechanism for the classification degradation in the presence of adversarial vulnerability – specifically that the class manifolds become larger and more correlated when adversarial perturbations are present.

An analysis of the ϵ-sized ball exemplar manifolds reveals the geometry of StochCochResNet50 with noise during training but not during evaluation is similar to that of ATCochResNet50, while the StochCochResNet50 with noise during the attack has a much lower capacity for the adversarial exemplar manifolds at late stages of the network, much like VOneResNet50 (Figure 4C). Like seen in VOneResNet50, when the ϵ-sized ball exemplar manifold capacity is normalized by the clean

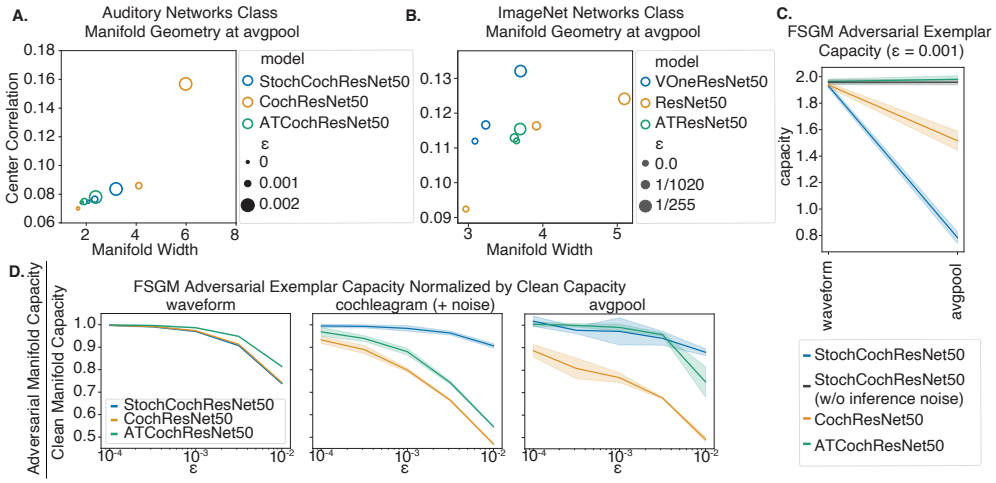

Figure 4: **Geometry of auditory models** Geometry of class manifolds for Auditory (**A**) and ImageNet (**B**) networks when computed from clean or adversarial stimuli. (**C**) Capacity of adversarial exemplar manifolds for auditory models. Evaluating the capacity for the adversarial exemplar manifolds with stochasticity present during inference yields significantly lower capacity at the avgpool layer of the network, while the model trained, but not tested, with stochasticity looks similar to the model achieved through adversarial training. Error bars are STD across 4 seeds. (**D**) Adversarial exemplar capacity normalized by the clean manifold capacity as a function of adversarial perturbation size. Unnormalized capacity detailed in SM 7.1. Error bars are STD across 4 random seeds.

manifold capacity, the normalized capacity for StochCochResNet50 is stable with increasing epsilon by the stochastic cochleagram layer (Figure 4D). This suggests that StochCochResNet50 represents the adversarially-perturbed sounds with the same manifold capacity and width as the clean sounds. By comparison, ATCochResNet50 does not show invariance to the adversarially perturbed stimuli until the avgpool layer.

5 Manifold overlap and the opposing effects of noise on robust classification

Results from the previous experiments indicate that stochastic representations provide a defense against adversarial perturbations. We hypothesize that for small perturbations, activation elicited from adversarial examples significantly overlap with the exemplar manifold formed by the stochastic representation of a clean image, which is present during both training and inference, effectively eliminating sensitivity to small adversarial perturbations. Here, we directly investigate whether small adversarial perturbations indeed overlap with multiple stochastic representations evoked from the same stimulus, and investigate how this overlap trades off with task performance. To do so, we test whether an SVM can separate the adversarial exemplar manifolds from the clean exemplar manifolds generated for the same stimulus. Finally, we analyze the opposing effects of noise on clean and adversarial performance by investigating a CIFAR model trained with different noise levels.

5.1 CIFAR model and datasets

We include investigation how the level of noise changes the manifold geometry in a smaller model trained on the CIFAR-10 dataset [31] with an architecture similar to ResNet18 [2], with the first conv-relu-maxpool layers replaced by fixed-weight Gabor filters and biological-inspired activation functions adapted from the Gaussian VOneNet[3]. In SM 5, we repeat the experiments with Poisson-like stochastic activations as well, where the trends are similar.

5.2 Low perturbation strength adversarial examples overlap with stochastic representations

We measure the overlap between ϵ-sized ball adversarial exemplar manifolds and clean exemplar manifolds. For all models, the adversarial exemplar manifold is generated by running FGSM L_∞,

[3]More details about the model architecture and training can be found in SM 1.3

varying the strength of the adversarial attack. The clean exemplar manifold is generated by measuring multiple stochastic representations from the same natural stimulus. We train an SVM to separate the adversarial exemplar manifold from the clean exemplar manifold (Figure 5). For CIFAR SVM experiments, the stimulus set includes 20 unique image exemplars, each with 1000 samples. For VOneResNet50 and StochCochResnet50 the stimulus set includes 10 unique stimulus exemplars, each with 5000 samples, and the features are downsampled to 5000 dimensions. As hypothesized, the manifolds become more linearly separable and less overlapped as the adversarial attack strength increases.

5.3 Adversarial robustness requires balancing adversarial and class information

Although increasing the variance of stochastic representations may hide larger adversarial perturbations, high noise levels may also reduce capacity for class manifolds and decrease model accuracy. We hypothesize that the optimal noise level for model accuracy and robustness requires a balance between these two factors (Figure 6A). We empirically test this hypothesis, first by showing that the adverarsarial and clean exemplar manifolds become less linearly separable with increasing noise variance (Figure 6B). We also evaluate the performance of the CIFAR model with varying noise levels across different adversarial attack strengths (Figure 6C). As the the variance of the stochastic responses increase the clean performance decreases, while performance under adversarial attack depends non-linearly on the noise variance. Using MFTMA we examine the effect of varying levels of stochasticity on class manifold capacity and related geometric properties. We generated clean exemplar and class manifolds from a stimulus set of 100 unique images, each with 50 noise samples. Figure 6D shows the dependence of manifold capacity on neural noise level. As the noise level increases, both class and exemplar manifolds become more entangled and capacity drops. In addition, Figure 6E shows that as the noise level increases, the manifold width increases. Thus, both the capacity, which characterizes manifold linear separability, and the geometry demonstrate that increasing noise levels makes manifolds larger, more entangled and less linearly separable. This type of optima is also observed when choosing the level of stochasticity for the auditory networks in Figure 3C, and additional auditory analysis of the noise level is found in section SM 8.

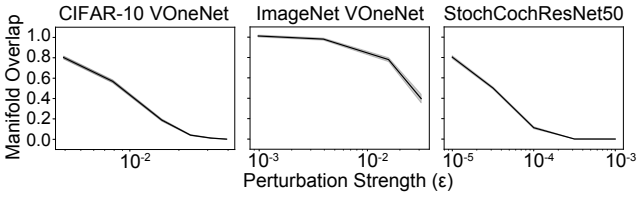

Figure 5: **Adversarial exemplar manifolds overlap with clean exemplar manifolds at smaller attack strengths.** A binary SVM trained to separate adversarial and clean exemplar manifolds extracted from the stochastic representation show that clean and adversarial exemplar manifolds measured from the same stimulus become less linearly separable as the attack strength decreases. Experiments are presented for CIFAR VOneNet(Error bar is SEM across 20 images and 6 random seeds), ImageNet VOneResNet50 (Error bar is SEM across 10 images and 5 random seeds) and StochCochResNet50 (Error bar is SEM across 10 image and 5 seeds).

6 Discussion

Using recently developed techniques to analyze the neural population geometry of a variety of networks on clean and adversarially perturbed stimuli, our work provides new insights into mechanisms underlying adversarial robustness. We show key geometrical differences between networks trained with adversarial training and VOneNets operating with biologically-inspired stochastic neural representations. Further, our novel auditory models provide a non-trivial demonstration of the usefulness of stochastic representations for defending against adversarial attacks. Auditory models typically contain analogous biologically-inspired fixed front-end representations (specifically, modeling the human peripheral auditory system), but these operations are quite distinct from those in the visual system. Further, in contrast to visual neural networks, auditory networks are typically trained with additive noise on the input [8, 10] as additive noise is ubiquitous in real-world auditory signals, with concurrent sounds summing together. Focusing on stochasticity, we show that there exists an overlap between the stochastic representations of clean and adversarially perturbed stimuli, helping to explain why the stochastic networks generalize better to perturbations below a certain threshold. Finally, we

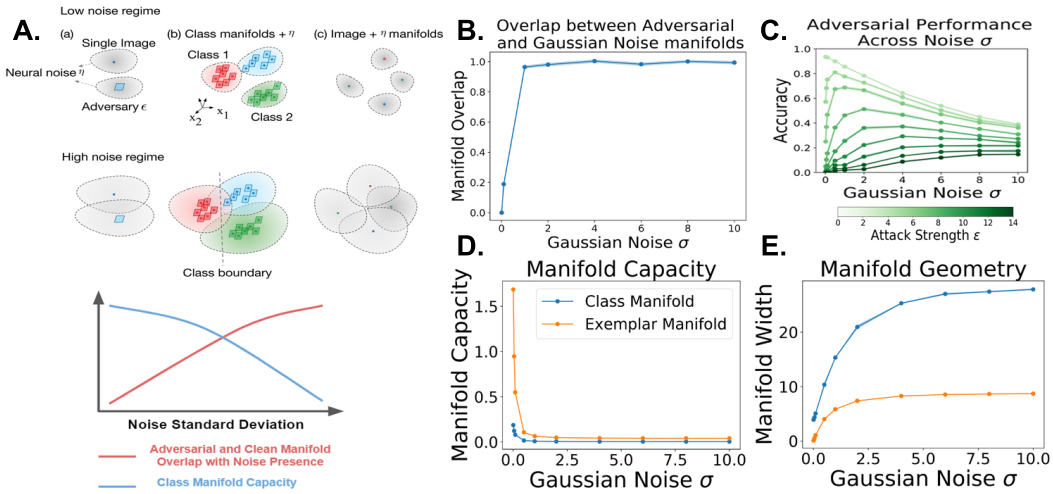

Figure 6: **Stochastic representations induce opposing geometrical properties that determine a model's adversarial performance** (**A**) Illustrations of the dependence of clean and adversarial exemplar manifold geometry on the level of stochasticity in the representation. In the low-noise regime, clean and adversarial exemplar manifolds do not overlap significantly, but class manifolds are well separated. In the high-noise regime, clean and adversarial exemplar manifolds are more overlapped, but the class manifolds also becomes more entangled (lower capacity). (**B**) Binary SVMs are trained to separate adversarial and clean exemplar manifolds extracted from same image in CIFAR models, and show that clean and adversarial exemplar manifolds become less linearly separable as the noise level increases. Error bar is SEM across 20 images and 6 random seeds. (**C**) Dependence between model performance and noise level at multiple attack strengths. Error bar is SEM calculated across 6 random seeds.(**D**) Manifold capacity decreases as the neural noise level increases. MFTMA is performed for both class manifold and image manifold. Error bar is SEM from 6 random seeds. (**E**) Manifold width increases as the noise level increases. Error bar is SEM from 6 random seeds.

isolate and geometrically quantify two competing effects of stocastic representations on network performance: while stochasticity increases the overlap between the distribution of activations around a clean exemplar and the distribution of activations from adversaries, an increase in stochasticity also creates more overlap between class manifolds, making them less separable, and ultimately reducing peak clean network performance.

Here, we focused on linear methods to probe the mechanisms of adversarial robustness. Future work could benefit from more sophisticated methods to measure manifold overlap, and also extend analysis to more naturally occurring corruptions [32]. Further, while we focus on the case of noise injected after fixed biologically inspired filters, an interesting future direction is to explore what types of representations benefit from the addition of noise; for instance [17] found there was not as strong of a protective effect with neural noise added to a standard convolutional filter. We leave the exact role of the neurally plausible filters, as used in both VOneResNet50 and StochCochRenet50, as a promising direction for future work. The presence of noise at all stages of the brain, also absent from our current work, suggests that if we can resolve when and where stochasticity is useful, this defense may be extensible to additional network layers for greater gains in adversarial robustness.

In theoretical neuroscience, much of the discussion on the role of stochasticity in neural coding has centered around the efficient representation of uncertainties associated with task-relevant variables [33, 34]. Our work adds to this line of research, by using geometry to demonstrate how it improves the neural population's robustness to adversarial perturbations unseen during training in deep networks. We hope that our work will motivate mechanistic explanations of structure underlying biologically plausible robust computation in ANNs through the lens of geometry, and further identification of biological constraints that might inspire favorable changes in task-efficient neural representation.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.

[3] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[4] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[5] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[6] Adam H Marblestone, Greg Wayne, and Konrad P Kording. Toward an integration of deep learning and neuroscience. *Frontiers in computational neuroscience*, 10:94, 2016.

[7] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[8] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[9] 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. *arXiv preprint arXiv:1811.12231*, 2018.

[10] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, pages 10078–10089, 2019.

[11] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. December 2013.

[12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. August 2016.

[13] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

[14] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. December 2017.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. June 2017.

[16] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. February 2019.

[17] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In *NeurIPS*, 2020.

[18] William R Softky and Christof Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *Journal of neuroscience*, 13(1):334–350, 1993.

[19] A Aldo Faisal, Luc PJ Selen, and Daniel M Wolpert. Noise in the nervous system. *Nature reviews neuroscience*, 9(4):292–303, 2008.

[20] SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Classification and geometry of general perceptual manifolds. *Physical Review X*, 8(3):031003, 2018.

[21] SueYeon Chung and LF Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *arXiv preprint arXiv:2104.07059*, 2021.

[22] Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):1–13, 2020.

[23] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[24] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. February 2018.

[25] J Deng, W Dong, R Socher, L Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[26] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *arXiv preprint arXiv:1808.05665*, 2018.

[27] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*, pages 5231–5240. PMLR, 2019.

[28] Douglas B Paul and Janet M Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, pages 357–362. Association for Computational Linguistics, 1992.

[29] Brian R Glasberg and Brian CJ Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[30] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

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

[32] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. March 2019.

[33] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[34] Wei Ji Ma and Mehrdad Jazayeri. Neural coding of uncertainty and probability. *Annual review of neuroscience*, 37:205–220, 2014.

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] The claims in the abstract and intro are directly supported by the sections 3, 4, 5.
   (b) Did you describe the limitations of your work? [Yes] In the discussion section 6, we describe the main limitations of our approach.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not believe that there will be any negative societal impacts to a better understanding of the mechanisms of adversarial robustness.

(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] While we draw heavily on existing theoretical results, we do not present any novel proofs or theorems.
   (b) Did you include complete proofs of all theoretical results? [N/A] While we draw heavily on existing theoretical results, we do not present any novel proofs or theorems.
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] We include sufficient instructions to reproduce all of our work, and links to relevant libraries where appropriate. If our paper is accepted, we will make a public repository with code to replicate our results.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Each experimental section details the models and datasets used, and the supplemental section further discusses hyperparameters / etc
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All plots include error bars over multiple random seeds.
   (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)? [Yes] The supplemental section details what compute was necessary for each model and dataset. We will release the name of the computing cluster if our paper is accepted.
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] We cite the creators of ImageNet, CIFAR, Word-Speaker-Noise, and all public models.
   (b) Did you mention the license of the assets? [Yes] All models we create will be released as open source with the MIT License, present in the public repository upon acceptance.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We have not currently included any new assets but we will include new models as public resources if our paper is accepted.
   (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? [No] The Spoken Wikipedia Corpus (used as one part of the Word-Speaker-Noise dataset) contains some offensive articles, however the authors of [10] report removing the articles before assembling the Word-Speaker-Noise dataset, and none of the words used for training have offensive content.
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]