# FIXED INTER-NEURON COVARIABILITY INDUCES ADVERSARIAL ROBUSTNESS

Muhammad A. Shah, Bhiksha Raj

Language Technologies Institute Carnegie Mellon University Pittsburgh, PA

#### ABSTRACT

The vulnerability to adversarial perturbations is a major flaw of Deep Neural Networks (DNNs) that raises question about their reliability when in real-world scenarios. On the other hand, human perception, which DNNs are supposed to emulate, is highly robust to such perturbations, indicating that there may be certain features of the human perception that make it robust but are not represented in the current class of DNNs. One such feature is that the activity of biological neurons is correlated and the structure of this correlation tends to be rather rigid over long spans of times, even if it hampers performance and learning. We hypothesize that integrating such constraints on the activations of a DNN would improve its adversarial robustness, and, to test this hypothesis, we have developed the Self-Consistent Activation (SCA) layer, which comprises of neurons whose activations are consistent with each other, as they conform to a fixed, but learned, covariability pattern. When evaluated on image and sound recognition tasks, the models with a SCA layer achieved high accuracy, and exhibited significantly greater robustness than multi-layer perceptron models to state-of-the-art Auto-PGD adversarial attacks without being trained on adversarially perturbed data.

Index Terms— Adversarial Robustness, Deep Learning, Biologically inspired

## 1. INTRODUCTION

While Deep Neural Networks (DNNs) have advanced the state of the art in many fields of Artificial Intelligence (AI), their vulnerability to adversarial perturbation – subtle distortions that are semantically irrelevant to humans but can change the response of DNNs when added to their inputs [1, 2] – raises concerns about their reliability when deployed in high-stakes real world applications such as self-driving cars, and biometric authentication systems. These concerns will need to be resolved for the wide-spread integration of AI into human society to be possible.

Over the years, the development of methods to make DNNs more robust to adversarial perturbations has become an active area of research. A large fraction of the proposed methods rely heavily on machine learning and statistical analysis techniques [3, 4, 5, 6, 7, 8], with the most popular and successful of them being adversarial training in which involves training on adversarially perturbed data [3]. Despite their effectiveness, these methods have two major shortcommings: they trade-off accuracy on natural data for adversarial robustness, and their effectiveness deteriorates if presented with perturbations other than the precise type of adversarial perturbation that they are designed to defend against [9, 10, 11], for instance a technique designed to defend against low-magnitude perturbations applied to the whole input may not be effective against high-magnitude perturbations applied to a very localized region of the input [9].

In contrast, human perception is highly accurate and naturally robust to a wide variety of such perturbations without ever being exposed to them, and therefore it may be beneficial, as has been the case throughout the history of neural networks, to seek inspiration from biological perceptual systems. Following this reasoning, several biologically-inspired methods have been proposed for defending DNNs against adversarial perturbations. These methods generally computationalize, and integrate into DNNs, some mechanism of biological perception that is as yet unrepresented in modern DNNs, but is hypothesized to contribute to its robustness [12, 13, 14, 15]. The improvement in robustness that these methods yield is relatively modest, compared to the non-biologically inspired methods mentioned earlier, however, it usually does not come at the cost of accuracy and tends to generalize better to other types of perturbations.

Building on this body of work, in this paper we investigate the role of inflexible inter-neuron covariability structures in making perception more robust. It has been observed that in the brain the spiking activity of individual neurons tends to be correlated [16, 17] and, the structure of this correlation tends to be inflexible over long periods of time even if it limits performance and learning [18]. We hypothesize that integrating such a mechanism into DNNs may improve their robustness because it will restrict the space of adversarial perturbations to only those that give rise to activations that respect a fixed covariability structure. In other words, forcing the intermediate activations to conform to a fixed covariability structure prevents the adversarial perturbation from causing arbitrary changes to the intermediate activations and thus misclassifications.

Integrating this behaviour into DNNs is not straightforward, because, unlike biological neurons, neurons in a DNN output a deterministic real number, which raises the question of how can we simulate correlated spiking in a system that is neither stochastic nor produces spiking activity? A solution presents itself if we consider the outputs of the artificial neuron as the frequency of an underlying spike train. If the spiking activity of a group of neurons is correlated, then the frequency of their spikes may also be correlated. Therefore, we use spiking frequency as a proxy for the spikes trains and, since the spiking frequency is represented in a DNN by the outputs of the neurons, we impose a fixed covariability structure on the latter.

To simulate a fixed inter-neuron covariability pattern, we have developed the Self-Consistent Activation (SCA) layer, which comprises of neurons whose activations are consistent with each other as they conform to a fixed covariability pattern. The SCA layer first computes the feed forward activations for the neurons based on the input, and then iteratively optimizes these activations to make them conform to a fixed, but learned, covariability pattern.

We evaluated the effectiveness of the SCA layers on image and sound classification tasks. For image classification we used MNIST [19] and Fashion-MNIST (FMNIST) [20], and for sound classification we used SpeechCommands [21]. Compared to a Multi-Layer

Perceptron (MLP), we find that the inter-neuron correlations in models with a SCA layer are more invariant to adversarial peturbations, thus indicating that the SCA layer does indeed make the inter-neuron covariance structure more inflexible as we intended. Furthermore, models with the SCA layer achieved similar if not better accuracy on the unperturbed data, and significantly higher accuracy on adversarially perturbed data generated by Auto-PGD [22] - a state-of-the-art white box adversarial attack. We evaluated the accuracy of the models under adversarial perturbations of several sizes and found that on average the model with the SCA layer achieved an absolute improvement of 4%, 5% and 6%, and a relative improvement of 117%, 155% and 45%, compared to the MLP model on FMNIST, Speech-Commands and MNIST, respectively. Similar trends were observed when the models were trained on adversarially perturbed data, with the SCA model achieving significantly higher accuracy on the clean and adversarially perturbed data on all datasets except MNIST.

#### 2. RELATED WORK

Non-Biologically Inspired Methods for Robustness: Most of the proposed methods for making DNNs robust to adversarial perturbations are not inspired from biology but rather rely on machine learning and statistical analysis techniques. Perhaps the most successful of these methods is adversarial training [3, 23], which adversarially perturbs each training minibatch by using Projected Gradient Descent to modify the data in the minibatch such that the loss is maximized. Over the years several improvements to the basic adversarial training algorithm have been proposed, with each modifying different parts of the model training pipeline, such as data augmentation [24], and regularization [25]. Another popular group of methods concentrate on creating models that are provably robust against adversarial perturbations and are accompanied by formal guarantees of the form: with probability at least  $1 - \delta$ , where  $\delta$  is small, the model's output will not change if a perturbation having norm at most  $\epsilon$  is added to a given input [4, 26, 27, 28]. These methods have two major shortcommings: they trade-off accuracy on natural data for adversarial robustness, and their effectiveness deteriorates if presented with perturbations other than the precise type of adversarial perturbation that they are designed to defend against [9, 10, 11].

Biologically Inspired Methods for Robustness: These methods generally involve developing computational analogues of biological process that are absent from common DNNs. Examples of such processes are predictive coding [12, 23], associative memory [29] biologically constrained visual filters, nonlinearities and stochasticity [13], foveation [14, 30, 31, 6], and non-uniform retinal sampling and cortical fixations [15]. Perhaps most closely related to our work are the methods based on predictive coding theory, which posits that the brain is a hierarchical Bayesian network in which the output of each layer is the maximum aposteriori estimate of the activations given the output of the previous layer and the next layer [32]. Integrating predictive coding layers within a model has been shown to improve its adversarial robustness [12, 33]. This approach is similar to ours in so far as it maximizes some notion of consistency between activations. Under predictive coding the activations become consistent when the input can be reconstructed perfectly from them, while in our work they become consistent when their covariability structure matches the one learned during training.

## 3. THE COVARIABILITY OF DNN ACTIVATIONS

We hypothesize that the inflexible covariability structure of neuronal activations that is observed in the animal brain contributes to the

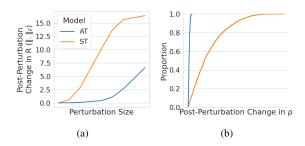


Fig. 1: (a) The Frobenius norm of the change in the correlation matrix of the activations of neurons in the penultimate layer of a MLP trained on clean (ST) and adversarially perturbed (AT) FMNIST images when adversarial perturbations of different sizes are added to the input; (b) CDF of the change in correlation between neuron pairs when adversarial perturbation of  $\ell_{\infty}$  norm 0.1 is added.

robustness of biological vision. As a preliminary step in our investigation of this hypothesis we determine if there is a relationship between the inflexibility of the correlation matrix of a DNN's activations, which we consider a proxy of its covariability structure, and its robustness to adversarial perturbations. To this end, we analyse the correlation structure between the neural activations of a DNN in response to data perturbed with perturbations of different sizes. To this end, we train two 5-layer MLP models on FMNIST, one on clean data and the other via adversarial training, and compute the correlation between the activations of the penultimate layer in response to clean and adversarially perturbed images. We use  $\mathbf{R}_{\epsilon}$  to refer to the correlation matrix produced by data perturbed by perturbations of  $\ell_{\infty}$  norm  $\epsilon$ . We quantify the overall change in the correlation structure as  $\|\mathbf{R}_0 - \mathbf{R}_{\epsilon}\|_F$ , where  $\|\cdot\|_F$  is the Frobenius norm and plot this quantity for several values of  $\epsilon$  in Figure 1a. We can see that the correlation structure of the adversarially trained MLP is much more invariant to adversarial perturbations compared to the correlation structure of the MLP trained on clean data. It is only after the size of the perturbation become very large does the correlation structure of the adversarially trained model begin to change significantly. To verify that the change in the norm is not caused due to a small number of neurons, we compute the absolute change in the correlation of each neuron pair due to the addition of adversarial perturbations of size 0.1, and plot the cumulative frequency curve shown in Figure 1b. We see that the curve for the adversarially trained model is significantly shifted to the left of the curve for the model trained on clean data indicating that the correlation between most, if not all, the pairs of neurons has not changed significantly.

From these observations we can infer that the invariance of the inter-neuron covariability structure across different perturbations of the input is related to the adversarial robustness of the model. If this relationship is causal then constraining the inter-neuron covariability structure should induce adversarial robustness. To determine if this is indeed the case we design a neural network layer that explicitly optimizes its activations to make them conform to a fixed covariability structure. We then include this layer in a DNN model and evaluate its robustness against state-of-the-art adversarial attacks.

## 4. SELF-CONSISTENT ACTIVATION LAYER

We have developed the Self-Consistent (SCA) Activation Layer to simulate an inflexible inter-neuron covariability structure. At a high-level, the SCA layer computes its output as  $SCA(\mathbf{x}) = g_C(\mathbf{a_x})$  where  $\mathbf{x} \in \mathbb{R}^{d_\mathbf{x}}$  is the input,  $\mathbf{a_x} = f(\mathbf{x}) \in \mathbb{R}^{d_\mathbf{a}}$  is the feed-forward

activation vector, and  $g_{\mathcal{C}}(\mathbf{a}_{\mathbf{x}})$  is the projection of  $\mathbf{a}_{\mathbf{x}}$  onto  $\mathcal{C}$ , the subspace comprised of the vectors that respect the learned covariance structure. If we consider covariability to be a linear relationship, like covariance, then  $g_{\mathcal{C}}$  would simply be a linear projection. However, to allow for more complex inter-neuron interaction, in this paper we have decided to adopt the following non-linear form for  $g_{\mathcal{C}}$ :

$$\arg\min_{\mathbf{a_x}} \|\mathbf{a_x} - \phi(\mathbf{W}_g \mathbf{a_x} + \mathbf{b}_g)\|_2^2 + \lambda \|\mathbf{x} - \mathbf{W}_h \mathbf{a_x} - \mathbf{b}_h\|_2^2, (1)$$

where  $\phi = ReLU$ ,  $\mathbf{W}_g \in \mathbb{R}^{d_\mathbf{a} \times d_\mathbf{a}}$ ,  $\mathbf{W}_h \in \mathbb{R}^{d_\mathbf{x} \times d_\mathbf{a}}$ ,  $\mathbf{b}_g \in \mathbb{R}^{d_\mathbf{a}}$  and  $\mathbf{b}_h \in \mathbb{R}^{d_\mathbf{x}}$ . The first term represents the distance between the activation and its projection onto  $\mathcal{C}$ , while the second term represents the information about  $\mathbf{x}$  that is not carried by  $\mathbf{a}_\mathbf{x}$ . The latter is added as a regularizer to prevent degenerate solutions, like  $\mathbf{a}_\mathbf{x} = 0$ , in which  $\mathbf{a}_\mathbf{x}$  carries no information about  $\mathbf{x}$ , and  $\lambda$  is a scalar that controls the strength of the regularization. We set the diagonal of  $\mathbf{W}_g$  to zero to prevent it from becoming the identity matrix and we perform the minimization using batch gradient descent. The exact sequence of operations performed by the SCA layer is shown in Algorithm 1.

## Algorithm 1 SCA Layer

```
1: \mathbf{u} \leftarrow f(\mathbf{x})

2: \mathbf{for}\ t : 1 \rightarrow T\ \mathbf{do}

3: \mathbf{a_x} \leftarrow \phi(\mathbf{u})

4: J \leftarrow \|\mathbf{a_x} - \phi(\mathbf{W}_g \mathbf{a_x} + \mathbf{b}_g)\|_2^2 + \lambda \|\mathbf{x} - \mathbf{W}_h \mathbf{a_x} - \mathbf{b}_h\|_2^2

5: \mathbf{u} \leftarrow \mathbf{a_x} - \eta \nabla_{\mathbf{a_x}} J

6: \mathbf{end}\ \mathbf{for}

7: \mathbf{a_x} \leftarrow \phi(\mathbf{u})
```

## 5. EVALUATION

## 5.1. Experimental Setup

# 5.1.1. Datasets

We evaluate the performance of SCA layers on image and audio classification tasks. For image classification we use the MNIST [19] and Fashion MNIST (FMNIST) [20] datasets, which contain 60,000  $28 \times 28$  black-and-white images of handwritten digits and 10 types of clothes, respectively. From both MNIST and FMNIST, we used 45,000 images for training, 5000 for evaluation and 10,000 for testing. For the audio classification task we use a subset of SpeechCommands dataset [21], which contains about 40,000 1 second, 16KHz recordings of humans vocalizing digits 0 to 9. We use 31,000 recordings for training, 4,000 recordings for validation and 4,000 recordings for testing.

#### 5.1.2. Data Preprocessing

For the image datasets, we flatten the image into a vector, which is then normalized by subtracting 0.5 and then dividing by 0.5. The audio data is pre-processed by first downsampling to 8KHz. Then 128 Mel-Frequency Cepstral Coefficients (MFCCs) are computed from a log mel spectrogram having 512 FFT points computed over a 64 ms sliding window with a stride of 32ms. By retaining only the first 16 MFCCs we obtain a  $16\times251$  matrix for each 1s audio recording. The matrix is then flattened, and normalized by subtracting -0.96 and then dividing by 9.2 (the mean and standard deviation computed over the validation set).

Fig. 2: Schematics of the MLP and SCA models. L s are affine projections with the superscripts representing the output dimension. The output dimension of  $L_2$  is set to 384 in MLPs trained on MNIST and FMNIST, and to 2048 in MLPs trained on SpeechCommands.  $\phi = ReLU$ ,  $\sigma = \text{dropout} \circ \phi$ ,  $\mathcal{J}$  is the loss from eq. (1).

It is important to note here that the adversarial perturbations are computed on and applied to the original input, prior to pre-processing. All the above mentioned pre-processing steps are implemented using PyTorch and torchaudio, and are fully differentiable, therefore the gradients can be propagated back to the original input and used to compute adversarial perturbations.

#### 5.1.3. Models

We compare the performance of models containing SCA layers (SCA model) to MLP models having comparable architectures and number of parameters. The schematics of these models are shown in Figure 2. The SCA model performs T=16 optimization steps. The probability of dropout is set to 0.2 for all models.

The models are optimized using the Adam optimizer using a learning rate of 0.001 and a batch size of 256 for up to 100 epochs. The learning rate is halved if the loss on the validation set does not decrease for 5 epochs, and if it does not decrease for 20 epochs the training is stopped early. All the results presented below are averaged over 5 trials with different random seeds.

## 5.2. Results

#### 5.2.1. Analysis of Activation Covariability Structure

To verify that SCA layers increase the invariance of the inter-neuron correlation structure we analyse the correlation between the activations of the penultimate layer using the method introduced in 3. Specifically, we compute the correlation matrix  $\mathbf{R}_{\epsilon}$  from the activations of the penultimate layer of the SCA model and MLP in response to 1000 data samples, perturbed by adversarial perturbations of  $\ell_{\infty}$  norm  $\epsilon$ . We compute  $\mathbf{R}_{\epsilon}$  for each dataset using several values of  $\epsilon$ . For each dataset and  $\epsilon$  we then compute  $\|\mathbf{R}_0 - \mathbf{R}_{\epsilon}\|_F$  to represent the overall change in the correlation structure due to the addition of adversarial perturbation of size  $\epsilon$ . Figure 3 shows this quantity for the SCA model and MLP on each dataset. We see that in every case the correlation structure of the SCA model changes more slowly than the MLP, and thus is more invariant to adversarial perturbation. This result shows that the SCA layer indeed produces the intended effect of constraining the covariability structure of neural activations.

## 5.2.2. Robustness of Models Trained on Clean Data

We evaluate the adversarial robustness of the SCA models by training them on *clean* data and computing their classification accuracy on adversarially perturbed data. The adversarial perturbations are computed by running AutoAttack [22] with different bounds on the maximum  $\ell_{\infty}$  norm of the perturbation.

The accuracy achieved by the SCA models and the baseline MLPs under adversarial attack is shown in Table 1. We see that

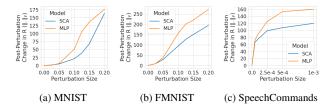


Fig. 3: The Frobenius norm of the change in the correlation matrix of the penultimate layer activations from the SCA model and MLP due to the addition of adversarial perturbations of different  $\ell_{\infty}$  norms.

Model	Perturbation Sizes $(\ell_{\infty})$							
	0.0	0.05	0.1	0.125	0.15	0.2		
	MNIST							
SCA	97.9	88.1	54.8	32.3	15.7	1.7		
MLP	98.4	90.0	52.6	28.2	12.8	1.4		
	FMNIST							
SCA	89.4*	$46.9^{*}$	$12.7^*$	6.1*	2.8*	0.1		
MLP	88.7	39.2	7.7	3.1	1.0	0.0		
	Perturbation Sizes $(\ell_{\infty})$							
	0.0	5e-5	1e-4	2.5e-4	5e-4	1e-3		
	SpeechCommands							
SCA	90.1*	45.1*	32.2*	10.3*	2.4*	$0.5^{*}$		
MLP	88.2	38.4	20.7	3.7	0.8	0.1		
						*p < 0.05		

Table 1: The accuracy (%) achieved by MLP and the SCA models

on all the three datasets the SCA model is *significantly* more robust than the MLP to adversarial perturbations, achieving higher accuracy on perturbations of all sizes. Averaged across all perturbation sizes, the SCA model achieves an absolute improvement in accuracy of 4%, 5% and 6%, and a relative improvement of 117%, 155% and 45%, compared to the MLP model on FMNIST, SpeechCommands and MNIST, respectively. The above results clearly show that SCA layers are much more robust to adversarial attacks than layers of perceptrons, and that their robustness is not limited to a particular type of data but generalizes across data complexity and modality.

#### 5.2.3. Robustness of Adversarially Trained Models

Given the SCA model is naturally more robust to adversarial perturbations, we hypothesize that an adversarially trained SCA model will also be more robust than an adversarially trained MLP. To verify this hypothesis, we adversarially train MLP and SCA models using the method from [3]: for each mini-batch we computer adversarial pertuabtions having  $\ell_{\infty}$ -norm at most  $\epsilon$  by performing 7 iterations of PGD with the step size  $\epsilon/4$ , where  $\epsilon$  is set to 0.3 for MNIST and FMNIST, and 2e-4 for SpeechCommands.

Table 2 shows the accuracy of the adversarially trained models under Auto-PGD attack [22]. The SCA model achieves greater accuracy than the MLP model for all perturbation sizes on Speech-Commands, and for perturbation size greater than 0.3 on FMNIST. Averaged across all perturbation sizes, the SCA model achieves an absolute improvement in accuracy of 1.7% and 2.4%, and a relative improvement of 49% and 12.6%, compared to the MLP model on FMNIST and SpeechCommands, respectively. On the other hand, we find that the adversarially trained MLP performs better than the SCA model under large perturbations on MNIST. The fact that the SCA model is significantly more robust than the MLP on more complex datasets, particularly SpeechCommands, but not on the simpler

Model	Perturbation Sizes $(\ell_{\infty})$							
	0.0	0.2	0.3	0.325	0.35	0.375		
	MNIST							
SCA	96.9*	88.0*	70.4	56.3	29.9	7.9		
MLP	94.7	85.8	<b>74.9</b> *	<b>64.6</b> *	34.4	$11.0^{*}$		
	FMNIST							
SCA	70.4*	51.5	35.1	22.4*	$7.0^{*}$	2.3*		
MLP	67.8	<b>53.5</b> *	42.4*	10.6	3.1	1.3		
	Perturbation Sizes $(\ell_{\infty})$							
	0.0	1e-4	2.5e-4	5e-4	1e-3	2e-3		
	SpeechCommands							
SCA	67.3	61.2*	56.4*	46.9*	30.9*	14.2*		
MLP	66.3	59.6	54.6	43.6	26.8	11.4		
						p < 0.05		

**Table 2**: The accuracy achieved by adversarially trained MLP and SCA models under adversarial perturbations of various  $\ell_{\infty}$  norms. The norm of the perturbations used during training is 0.3 for MNIST and FMNIST models, and 2.4e-4 for the SpeechCommands models.

-		N						
		Number of Steps						
		1	2	4	8	16		
	clean	89.3	89.5	89.8	89.1	89.5		
	perturbed	35.8	35.8	35.7	40.3	49.3		

**Table 3:** The accuracy of SCA models with different number of self-consistency optimizing steps (T) on clean and adversarially perturbed data. The size of adversarial perturbations is set to 0.05.

MNIST, may indicate that, under adversarial training, the utility of SCA layers is realized only in more complex tasks.

## 5.2.4. Impact of Number of Steps

To determine the impact of changing the number of steps, T, in Algorithm 1, we train models with different values of T using only clean data from FMNIST. The accuracy of these models on clean and perturbed data is shown in Table 3. We observe that increasing T does not impact clean accuracy but robustness significantly improves as T is increased beyond 8. This indicates that the more self-consistent the activations become the more robust the model gets.

## 6. CONCLUSION

In this paper we investigate the impact of the biological phenomenon of inflexible inter-neuron covariability on robustness to adversarial perturbation. To this end we develop the SCA layer, which as neural network layer that comprises of neurons whose activations conform to a fixed, but learned, covariability pattern. We demonstrate that DNNs with the SCA layer tend to be significantly more robust to adversarial perturbations than conventional MLPs without ever being trained on adversarially perturbed data, achieving an average improvement of 102% in accuracy relative to the MLP on image and audio data perturbed by state-of-the-art adversarial attack methods. These results indicate that constraining the inter-neuron covariability structure does indeed make DNNs more robust. More generally, these results lend credence to the approach of seeking inspiration from biology for developing better and more reliable AI systems.

#### 7. REFERENCES

- [1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.
- [2] 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*, vol. 32, 2019.
- [3] 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.
- [4] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter, "Certified adversarial robustness via randomized smoothing," in *Interna*tional Conference on Machine Learning. PMLR, 2019.
- [5] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J Zico Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," *NeurIPS*, vol. 33, pp. 21945–21957, 2020
- [6] Muhammad A Shah and Bhiksha Raj, "Less is more: Training on low-fidelity images improves robustness to adversarial attacks," in *Submitted to ICLR*., 2023, under review.
- [7] Muhammad A Shah, Raphael Olivier, and Bhiksha Raj, "Towards adversarial robustness via compact feature representations," in *ICASSP*. IEEE, 2021.
- [8] Raphael Olivier, Bhiksha Raj, and Muhammad Shah, "High-frequency adversarial defense for speech and audio," in *ICASSP*. IEEE, 2021, pp. 2995–2999.
- [9] Sander Joos, Tim Van hamme, Davy Preuveneers, and Wouter Joosen, "Adversarial robustness is not enough: Practical limitations for securing facial authentication," in *Proceedings of* the 2022 ACM on International Workshop on Security and Privacy Analytics, 2022, pp. 2–12.
- [10] Yash Sharma and Pin-Yu Chen, "Attacking the madry defense model with l\_1-based adversarial examples," arXiv preprint arXiv:1710.10733, 2017.
- [11] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel, "Towards the first adversarially robust neural network model on mnist," arXiv preprint arXiv:1805.09190, 2018.
- [12] Dylan M Paiton, Charles G Frye, Sheng Y Lundquist, Joel D Bowen, Ryan Zarcone, and Bruno A Olshausen, "Selectivity and robustness of sparse coding networks," *Journal of vision*, 2020.
- [13] 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," *NeurIPS*, vol. 33, pp. 13073–13087, 2020.
- [14] Aditya Jonnalagadda, William Yang Wang, B.S. Manjunath, and Miguel Eckstein, "Foveater: Foveated transformer for image classification," 2022.
- [15] Manish Reddy Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio, "Biologically inspired mechanisms for adversarial robustness," *NeurIPS*, vol. 33, pp. 2135–2146, 2020.
- [16] Jay A Hennig, Emily R Oby, Darby M Losey, Aaron P Batista, M Yu Byron, and Steven M Chase, "How learning unfolds in the brain: toward an optimization view," *Neuron*, 2021.

- [17] Patrick T Sadtler, Kristin M Quick, Matthew D Golub, Steven M Chase, Stephen I Ryu, Elizabeth C Tyler-Kabara, Byron M Yu, and Aaron P Batista, "Neural constraints on learning," *Nature*, 2014.
- [18] Matthew D Golub, Patrick T Sadtler, Emily R Oby, Kristin M Quick, Stephen I Ryu, Elizabeth C Tyler-Kabara, Aaron P Batista, Steven M Chase, and Byron M Yu, "Learning by neural reassociation." Nature neuroscience, 2018.
- [19] Yann LeCun, Corinna Cortes, and CJ Burges, "Mnist handwritten digit database," ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2010.
- [20] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.
- [21] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," arXiv preprint arXiv:1804.03209, 2018.
- [22] Francesco Croce and Matthias Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameterfree attacks," in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.
- [23] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang, "Recent advances in adversarial training for adversarial robustness," arXiv preprint arXiv:2102.01356, 2021.
- [24] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann, "Data augmentation can improve robustness," *NeurIPS*, 2021.
- [25] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*. PMLR, 2019, pp. 7472–7482.
- [26] Marc Fischer, Maximilian Baader, and Martin Vechev, "Certified defense to image transformations via randomized smoothing," *NeurIPS*, 2020.
- [27] Aounon Kumar and Tom Goldstein, "Center smoothing: Certified robustness for networks with structured outputs," *NeurIPS*, 2021.
- [28] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin, "Certified adversarial robustness with additive noise," *Advances in neural information processing systems*, vol. 32, 2019.
- [29] Dmitry Krotov and John Hopfield, "Dense associative memory is robust to adversarial inputs," *Neural computation*, 2018.
- [30] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao, "Foveation-based mechanisms alleviate adversarial examples," arXiv preprint arXiv:1511.06292, 2015.
- [31] Jonathan M Gant, Andrzej Banburski, and Arturo Deza, "Evaluating the adversarial robustness of a foveated texture transform module in a cnn," in SVRHM Workshop@ NeurIPS, 2021.
- [32] Rajesh PN Rao and Dana H Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extraclassical receptive-field effects," *Nature neuroscience*, 1999.
- [33] Bhavin Choksi, Milad Mozafari, Callum Biggs O'May, Benjamin Ador, Andrea Alamia, and Rufin VanRullen, "Predify: Augmenting deep neural networks with brain-inspired predictive coding dynamics," *NeurIPS*, 2021.