LCANets: Lateral Competition Improves Robustness Against Corruption and Attack

Michael A. Teti [1] Garrett T. Kenyon [1] Benjamin J. Migliori [1] Juston S. Moore [1]

Abstract

Although Convolutional Neural Networks (CNNs) achieve high accuracy on image recognition tasks, they lack robustness against realistic corruptions and fail catastrophically when deliberately attacked. Previous CNNs with representations similar to primary visual cortex (V1) were more robust to adversarial attacks on images than current adversarial defense techniques, but they required training on large-scale neural recordings or handcrafting neuroscientific models. Motivated by evidence that neural activity in V1 is sparse, we develop a class of hybrid CNNs, called LCANets, which feature a frontend that performs sparse coding via local lateral competition. We demonstrate that LCANets achieve competitive clean accuracy to standard CNNs on action and image recognition tasks and significantly greater accuracy under various image corruptions. We also perform the first adversarial attacks with full knowledge of a sparse coding CNN layer by attacking LCANets with white-box and black-box attacks, and we show that, contrary to previous hypotheses, sparse coding layers are not very robust to white-box attacks. Finally, we propose a way to use sparse coding layers as a plug-and-play robust frontend by showing that they significantly increase the robustness of adversarially-trained CNNs over corruptions and attacks.

1. Introduction

Convolutional Neural Networks (CNNs) are often considered a rough model of the ventral visual stream (Kubilius et al., 2019), where object recognition is thought to occur in primates. However, CNNs and biological visual systems behave very differently in practical performance. For instance, adversarial examples (Szegedy et al., 2014; Goodfellow et al., 2015) are very effective at causing a CNN to fail catastrophically while they remain indistinguishable from unperturbed images to humans (Elsayed et al., 2018).

Mounting research suggests that CNN image classifiers with representations similar to those in the primary visual cortex (V1) exhibit increased robustness to image corruptions and adversarial attacks (Li et al., 2019; Dapello et al., 2020; Safarani et al., 2021). These V1-like CNNs have been shown to be more robust than those trained with state-of-the-art adversarial defense techniques (Rusak et al., 2020). Current methods to develop V1-like CNNs involve replacing specific layers with hand-crafted neuroscientific models (Dapello et al., 2020) or jointly training with V1 responses to images (Li et al., 2019; Safarani et al., 2021). It is unclear how these techniques can be adapted to arbitrary datasets or used as a general purpose adversarial defense method, since neural datasets and knowledge of receptive field properties of sensory neurons to stimuli other than still images is limited.

In contrast, we develop a CNN frontend based on biologically plausible sparse coding models, such as the Locally Competitive Algorithm (LCA) (Rozell et al., 2008). Sparse coding models are a class of data-agnostic unsupervised models that have been shown to model neural responses in visual, auditory, and olfactory cortices (Rozell et al., 2008; Zylberberg et al., 2011; Terashima et al., 2013; Dodds & DeWeese, 2019; Jortner et al., 2007; Jürgensen et al., 2021). These models were originally developed based on longstanding neurophysiological evidence that the neural activity in V1 and other sensory areas is *sparse*, unlike the activations in a CNN (Barlow et al., 1961; Olshausen & Field, 1997; Vinje & Gallant, 2000; Foldiak, 2003; Poo & Isaacson, 2009; Hromádka et al., 2008; Yoshida & Ohki, 2020). LCA finds a faithful but sparse representation (i.e. few active neurons) of a given input by modeling the recurrent lateral competition observed in V1 (Blakemore et al., 1970), whereby neurons compete to represent an input by inhibiting neighboring neurons with similar receptive fields (Chettih & Harvey, 2019). LCA is unlike the classical model developed by Dapello et al. (Dapello et al., 2020), in which each neuron's response is entirely determined by the extent to which its preferred fea-

[1] Los Alamos National Laboratory, Los Alamos, NM, USA. Correspondence to: Michael Teti <mteti@lanl.gov>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

ture is present in the input (plus added stochasticity which is independent at each neuron). Instead, it is a non-classical model of V1, where a neuron's response is highly dependent on *dynamic lateral competition with surrounding neurons*.

Lateral inhibition/competition is thought to contribute to bottom-up attention, feature selectivity, contrast-invariant tuning, and noise filtering in V1 (Gajewska-Dendek et al., 2015; Crook et al., 1998; Sompolinsky & Shapley, 1997; Stemmler et al., 1995; Deneve et al., 1999; Mao & Massaquoi, 2007). Additionally, previous studies have reported that time-limited humans are fooled by adversarial examples, but humans with unlimited viewing time are not (Elsayed et al., 2018). Taken together, these findings have led to speculation that recurrent lateral competition plays an important role in the robustness of biological visual systems, as it takes time to unfold relative to feed-forward excitation. It is not presently clear what role recurrent lateral competition might play in a CNN or how it would impact clean or adversarial performance on standard classification tasks. Previous results indicate that LCA/sparse coding layers are able to filter out noise added to images (Springer et al., 2018; Kim et al., 2019; Ahmad & Scheinkman, 2019) and adversarial perturbations from attacks on standard CNNs (Springer et al., 2018; Sun et al., 2019; Nguyen et al., 2020; Kim et al., 2020) by encoding and then reconstructing the input image, but instances in which a sparse code was used as input to a neural network classifier and/or an attack was performed with full knowledge of the sparse coding layer are rare. In a recent work, Paiton et al. (Paiton et al., 2020) observed that shallow fully-connected classifiers which used LCA codes as input were more robust to white-box attacks than comparable networks, but they did not compare against any other defense methods. To our knowledge, this work introduces the first deep CNN classifiers with embedded recurrent, feature-specific lateral competition designed to implement convex sparse coding and an analysis of its robustness relative to standard robust CNNs.

Our Contributions. Here we develop LCANets, a class of hybrid CNNs which consist of a frontend with lateral competition implemented by the LCA sparse coding algorithm followed by a standard CNN architecture. First, we show that LCANets achieve competitive clean accuracy to current state-of-the-art defense methods on the UCF-101 and HMDB-51 action recognition datasets, as well as the CIFAR-10 image recognition dataset. We then show that LCANets are more robust to different image corruptions and a modern black-box attack with limited queries than state-of-the-art defense methods. Since the LCA layer is differentiable, we also attack LCANets in the first direct, white-box attack on sparse coding CNN layers, and we show that they are much less robust than previously thought. Finally, we show that lateral competition can be used to augment the robustness of adversarially-trained networks under both corruptions and adversarial attacks.

2. Background and Related Work

2.1. V1-Like CNNs

In a landmark study, Li et al. (Li et al., 2019) trained a V1-like ResNet-18 (He et al., 2016) to classify ImageNet by regularizing the network to be similar to experimentally measured neural activity in mouse V1. Their model was jointly trained to perform image classification and approximate the neural representational similarity between image pairs. This model was more robust to random noise and Projected Gradient Descent attacks (Madry et al., 2018b) than an undefended VGG-16 (Simonyan & Zisserman, 2015) on grayscale Cifar-10 images. Similarly, Safarani et al. (Safarani et al., 2021) trained a VGG-19 network to both predict responses of monkey V1 neurons to and classify Tiny ImageNet images, and they found this CNN was more robust to image corruptions than an undefended VGG-19 network.

Alternatively, Dapello et al. incorporated a Linear-Nonlinear-Poisson (LNP) model of primate V1 into the first layer of a ResNet-50 (He et al., 2016) by adding a biologically-constrained gabor filter bank, simple and complex cell nonlinearities, and V1 stochasticity (Dapello et al., 2020). Their VOneNet exhibited competitive performance on clean ImageNet examples and marginally better robustness than an adversarially trained ResNet-50 on average under image corruptions (Hendrycks & Dietterich, 2018) and a Projected Gradient Descent attack (Madry et al., 2018b). In a follow-up study, Dapello et al. illustrated how the V1 stochasticity component could be used to increase robustness in an audio classification network (Dapello et al., 2021), but it is unclear how the remaining VOneNet components can be applied to audio classification networks, for example.

2.2. Sparse Coding Defenses

LCA/sparse coding has previously been used to increase the performance of CNNs under corrupted or adversarial examples (Springer et al., 2018; Sun et al., 2019; Kim et al., 2019; Ahmad & Scheinkman, 2019; Nguyen et al., 2020; Kim et al., 2020). Most of these previous methods involved encoding and then reconstructing the input image prior to classification by the CNN, which denoises much of the perturbation in some cases. As a result, most previous methods studied the *transferability* of attacks to LCA/sparse coding from CNNs, rather than performing attacks against LCA/sparse coding directly. In the study that is closest to ours, (Paiton et al., 2020) observed that two and three-layer networks composed of fully-connected layers on top of an LCA code were more robust to Projected Gradient Descent (PGD) (Madry et al., 2018b) attacks than comparable undefended networks on MNIST and grayscale Cifar-10. To the

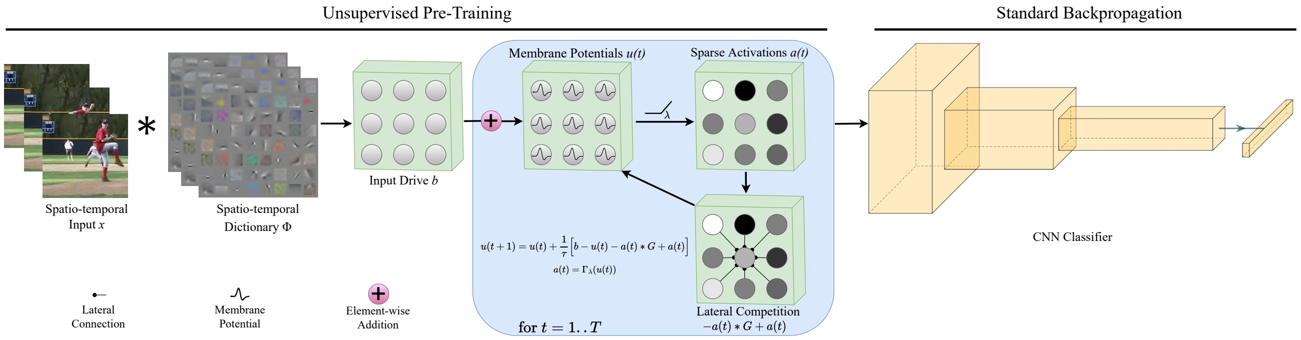

Unsupervised Pre-Training
Standard Backpropagation
Spatio-temporal Input x
Spatio-temporal Dictionary Φ
Input Drive b
Membrane Potentials $u(t)$
Sparse Activations $a(t)$
λ
$u(t+1) = u(t) + \frac{1}{\tau}\left[b - u(t) - a(t) * G + a(t)\right]$
$a(t) = \Gamma_\lambda(u(t))$
for $t = 1..T$
Lateral Competition
$-a(t) * G + a(t)$
CNN Classifier
Lateral Connection
Membrane Potential
Element-wise Addition

Figure 1. **LCANet architecture.** LCANets consist of an LCA block, which performs sparse coding via lateral competition, followed by a standard CNN. Since we first highlight the abilities of LCANets on action recognition, we depict an LCANet operating on spatiotemporal inputs with spatiotemporal features.

best of our knowledge, we are the first to perform a direct attack on an LCA/sparse coding-based network embedded in a standard deep CNN architecture, and we are also first to compare to other defense methods.

2.3. Divisive Normalization Networks

Divisive normalization (DN) is a somewhat similar mechanism to lateral competition in that a given neuron's output can be influenced by other neurons in the same layer (Carandini & Heeger, 2012). In contrast to the recurrent feature-specific competition in LCAs (and in V1 (Chettih & Harvey, 2019)), the weighting in DN is often learned and is not recurrent (Cornford et al., 2020; Burg et al., 2021). While DN has also been theorized as a model of V1 like LCA/sparse coding, there is not as much evidence to suggest that DN networks are robust in machine learning tasks as there is for LCA.

3. LCANet: A CNN with Recurrent Lateral Competition

Inspired by the sparse, robust representations present in the visual cortex, we developed the LCANet architecture (Figure 1). The key differences that distinguish LCANets from standard and previous V1-like robust CNNs are recurrent lateral competition and the ability to learn unsupervised features from data. LCANets consist of a locally-competitive algorithm (LCA) module at the input (i.e. a frontend) followed by standard CNN layers that generate a classification output. The LCA frontend is trained in an unsupervised manner to perform image reconstruction, and then frozen during the backpropagation training of the subsequent network layers. We chose LCA because it is a well-established implementation of Hopfield-style sparse coding, and because the thresholding mechanism provides flexible modeling choices that correspond to different L_P norms to induce sparsity.

Sparse coding models, in general, aim to find a faithful representation (code) of a given input using as few features (in the form of active neurons) as possible. This is a reconstruction minimization problem which can be defined as follows. We begin with an input $x \in \mathbb{R}^{C \times H \times W}$ and an overcomplete dictionary of convolutional features $\Phi \in \mathbb{R}^{M \times C \times k_H \times k_W}$, where C is the number of input channels, H is the image height, W is the image width, M is the number of convolutional features, and k_H, and k_W represent the spatial dimensions of each feature. We wish to obtain a representation $a \in \mathbb{R}^{M \times \left\lfloor \frac{H}{\text{stride}_H} \right\rfloor \times \left\lfloor \frac{W}{\text{stride}_W} \right\rfloor}$, where stride indicates the convolutional stride. a is a sparse code that represents the learned spatiotemporal reconstruction that is closest to the input x. The sparse coding problem (under the L_1 norm) involves solving the following penalized reconstruction problem:

$$\min_a \frac{1}{2} \|x - a \circledast \Phi\|_2^2 + \lambda \|a\|_1 \tag{1}$$

where $\circledast$ indicates the transpose convolution, Φ represents the previously learned dictionary elements, and λ determines the trade-off between reconstruction performance and the sparsity of the code.

To solve Equation 1, LCA implements leaky integrate-and-fire neurons with recurrent lateral competition (Rozell et al., 2008). Mathematically, the membrane potential dynamics of a neuron can be described by the following ordinary differential equation:

$$\dot{u}(t) = \frac{1}{\tau}\Big[b(t) - u(t) - a(t) * G + a(t)\Big] \tag{2}$$

where $u(t)$ is the neuron's membrane potential, τ is a time constant, $b(t) = x * \Phi$ is the neuron's input drive from the stimulus computed by taking the the convolution of the input with the dictionary, $G = \Phi * \Phi$ represents the pairwise feature similarity between each feature and every other feature, and $a(t) = \Gamma_\lambda(u(t))$ is the neuron's instantaneous firing rate computed by applying a soft threshold activation $T_\lambda(\cdot)$ with threshold λ to $u(t)$. Through this threshold, we also enforce that the firing rates $(a(t))$ are nonnegative, as in biological neurons. A desirable property of sparse coding is that Equation 1 is convex in a and in Φ individually (Garcia-Cardona & Wohlberg, 2018). We learn the dictionary Φ by coordinate ascent, solving for a given a batch of inputs using LCA and then updating Φ via SGD.

Rozell et al. (Rozell et al., 2008) showed that LCA systems satisfy the criterion for local asymptotic stability, i.e. that the system is inherently robust to perturbations up to some ϵ and will return to the equilibrium point in the limit as $t \rightarrow \infty$. Rozell further shows that as a consequence of this property, LCA systems will have unique equilibrium points, and an extremely high likelihood for each equilibrium point to be locally asymptotic in itself. This implies that for a given input, the system will trend to a distinct and stable solution. This is an extremely desirable property for a feature learning method, particularly one where the representation, or code, will be used to perform inference in a downstream task.

To complete the construction of the LCANet, the code a is then passed as the input to a standard CNN (see Section 4.1.1). Unlike previous approaches in which $\hat{x}$ was computed from a and then used as input to the CNN (Springer et al., 2018; Kim et al., 2019; Sun et al., 2019; Nguyen et al., 2020; Kim et al., 2020), *we never go back to the input space from the code*. In summary, the LCANet takes in a standard input and outputs a vector of class probabilities just like any standard CNN architecture. This construction gives the LCANet the benefit of robust description of inputs, as found in LCA applications, as well as the discriminative power of deeper CNNs.

4. Experiments and Results

We perform experiments on action and image recognition datasets using common image corruptions and adversarial attacks to test the robustness of LCANets. For baselines, we compare against standard ResNet models, adversarially-trained ResNet models, and VOneResNet models. In addition, we compare to a variant of LCANets (LCANet-F) which uses the convolutional features learned by LCANet without the lateral competition.

4.1. Action Recognition on UCF-101 and HMDB-51

4.1.1. LCANet Details

To construct an LCANet, we use the LCA-PyTorch package. We first perform unsupervised dictionary learning, as described in Section 3, for 10,000 steps to learn a dictionary of convolutional features on the Kinetics-700 dataset (Carreira et al., 2019). Since the data is spatio-temporal, we learn a spatio-temporal dictionary using 3D convolutions. The dictionary is initialized with random values from a Gaussian distribution. Table 1 shows the LCA hyperparameters used to learn the LCA dictionary. These hyperparameters are selected *before any models were trained* to match the known characteristics of V1 spatio-temporal receptive fields (STRFs) and increase computational efficiency while maintaining overcompleteness of the dictionary.

Next, the first layer of a 3D-ResNet50 (Kataoka et al., 2020) is replaced with the LCA layer to create the LCANet, and the LCA dictionary is fixed. The LCANet is then trained on the classification tasks like all other models as described in Section 4.1.3. We learn one dictionary with $\lambda = 0.5$, but we train three separate LCANets using this dictionary: LCANet0.1, LCANet0.5, and LCANet1.0 corresponding to λ values of 0.1 (97% sparse), 0.5 (99% sparse), and 1.0 (99.5% sparse), respectively. By increasing λ, we are decreasing the number of active neurons contributing to the representation of a given input, or equivalently, increasing the sparsity of the code inferred by LCA.

4.1.2. Baselines

LCANet-F) To quantify the contribution of lateral competition relative to the unsupervised LCA features in producing robust representations, we also train one final LCANet without lateral competition but with the LCA features (LCANet-F). In this model, we replace the first convolutional layer of 3D-ResNet50 with the LCA dictionary followed by a rectified linear activation, but no lateral competition is performed. This LCANet-F model is comparable to the VOneNet without complex cells or stochasticity, except the features are learned from the data.

3D-ResNet50) This model is a standard 3D-ResNet50, which we refer to as ResNet50. The 3D indicates that 3D convolutions are used in every convolutional layer as opposed to 2D convolutions. This model matches the previous state-of-the-art performance on the UCF-101 and HMDB-51 datasets (Kataoka et al., 2020).

3D-ResNet50-AT) To train the 3D-ResNet50-AT, which we refer to as ResNet50-AT, we perform standard adversarial training (Madry et al., 2018a) with an $||\delta||_\infty = \frac{13}{255} = 0.05$ constraint and 200 queries using the same attack used to test the models. All other training parameters were shared among all models and described in detail in Section 4.1.3. We observe that adversarial training was successful on this attack because the attack effectiveness dropped from about 75% to 5% over training.

VOneResNet50) Although the creation of arbitrary spatio-temporal Gabor filter banks has been demonstrated in the literature (Adelson & Bergen, 1985), knowledge of the parameters governing experimentally determined V1 STRFs is limited, as is the availability of V1 electrophysiological responses to spatio-temporal stimuli. As a result, the creation of biologically verifiable V1-like spatio-temporal convolution blocks for the purpose of increased robustness is very difficult. VOneNets (Dapello et al., 2020) represent the current successful implementation of this biological inspiration for spatial-only data, in which a VOneBlock frontend is programmed to match known biological receptive fields. While this method is not spatio-temporal, it provides an excellent point of comparison with state-of-the-art for our spatio-temporal LCANets. To isolate the contributions of the VOneBlock in our comparisons, we constructed a new VOneNet by replacing the first convolutional layer in a (2+1)D-ResNet50 (Tran et al., 2018) with the VOneBlock. We experimented with kernel sizes from 7×7 to 25×25, and we observed very small differences between them in terms of clean and adversarial accuracy. As a result, we report the best model with kernel size 11×11. We chose (2+1)D-ResNets because 3D convolutions are decomposed into a spatial convolution followed by a temporal convolution, thus allowing the spatial architecture of the VOneBlock to be used without conflating temporal effects. Configuring the VOneNet to handle spatiotemporal information in this way enables comparison with LCANets that natively exhibit V1-like representations learned from features of the data (including true spatiotemporal information).

4.1.3. Training and Testing Details

All models were implemented in PyTorch 1.10.1 on a high-performance computing node with eight NVIDIA GeForce RTX 2080 Ti GPUs, 80 CPU cores, and 754GB of memory. We slightly adapt the code developed by Hara et al. to train and test 3D-CNNs on large action recognition datasets (Hara et al., 2017; 2018a;b; Kataoka et al., 2020). The input to each model consists of 12 consecutive color video frames of spatial dimensions 64×64. Input video clips were augmented during training with random cropping and horizontal flipping. Classification models were first pre-trained on the Kinetics-700 dataset (Carreira et al., 2019). They were then fine-tuned and tested on the UCF-101 (Soomro et al., 2012) and HMDB-51 (Kuehne et al., 2011) datasets using the standard, train, validation, and test splits. All training hyperparameters, such as the batch size and learning rate were set to the values used in (Kataoka et al., 2020).

During testing, each 10 second video was cut into 12-frame input clips with a stride across time of 12, and center cropping was used. In preliminary experiments, we computed video-level predictions by averaging probabilities across all clips within a single video. We observed that video level accuracy for all models across clean and attacked datasets remains about 6% higher on UCF-101 and 9% higher on HMDB-51 than clip-level accuracy, which is consistent with the state-of-the-art (Kataoka et al., 2020). Since we also observe that video-level accuracy increased monotonically with clip-level accuracy for all models under all attacks, we report only the clip-level accuracy for simplicity. For both attacks and corruptions, we randomly down-select a test set of 1,500 video clips from each dataset to use as inputs to the models. We repeat this procedure three times with three random seeds (the same random seed was shared across models).

4.1.4. LCANets Are More Robust to Image Corruptions

No image augmentation was performed during training. We evaluate the robustness of LCANets against three common image corruptions:[1] additive Gaussian noise, Gaussian blurring, and random erasure. These specific corruptions were chosen because they are categorically different from one another, and each can be caused by some underlying physical process or sensor noise at some point during acquisition. They were also chosen before any experiments were performed or any models were trained. All corruptions were performed before image normalization.

Additive Gaussian Noise) The most striking observation in the presence of additive Gaussian noise is the approximately 20% gain in performance of LCANet1.0 relative to all other models for both UCF-101 and HMDB-51 (Figure 2). Across both datasets, LCANet0.5 also displays significant robustness after LCANet1.0, followed by ResNet50-AT. The LCANet-F model was also relatively robust to the three levels of noise. On the other hand, VOneNet's performance decreased sharply with increasing noise levels, only surpassing the undefended ResNet50. The relative success of LCANet-F over VOneNet highlights the robustness of V1-like features learned through sparse dictionary learning on data.

Gaussian Blur) It has been shown that CNNs commonly exploit the high frequency information in images, which limits their robustness (Wang et al., 2020). Blurring the image removes much of this high frequency information. As as result, we would expect the performance of non-robust CNNs to decrease more rapidly than robust CNNs as the blurring becomes more severe. Under this corruption, the performance of all models decreases relatively uniformly (Figure 3). At more severe levels of blurring, however, LCANets still maintain a small advantage over the robust models and a significant advantange over the base 3D-ResNet50, followed by VOneNet. This indicates that LCANets and VOneNets are potentially relying less on the non-robust high frequency information than standard CNNs. On the other hand, the relatively poor performance of the ResNet50-AT suggests that adversarially trained CNNs may still exploit high-frequency information that is removed by blurring. Similar to the

[1] See supplementary material for examples of corrupted video frames.

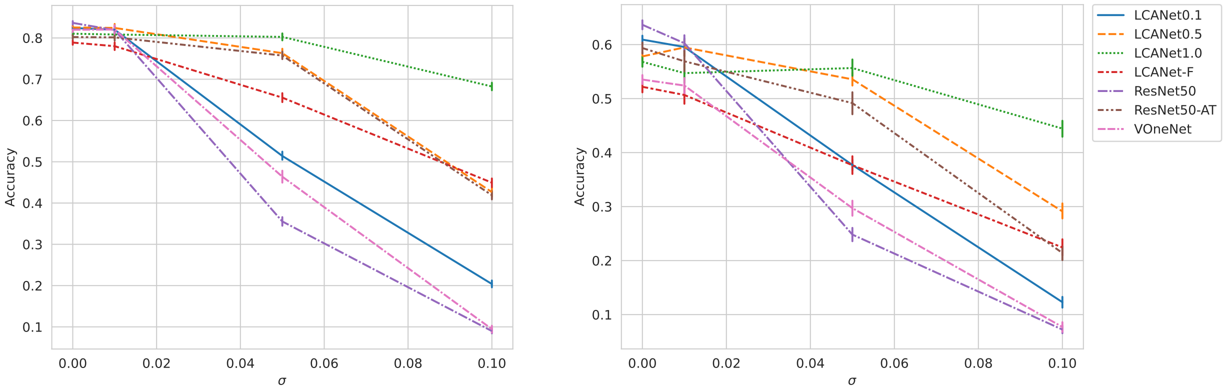


Figure 2. **LCANets are significantly more robust to Gaussian noise than state-of-the-art robust models.** LCANets significantly outperform all other models under additive Gaussian noise on both the UCF-101 (left) and HMDB-51 (right) datasets. The unsupervised LCA features appear to contribute significantly to this robustness, as indicated by the moderate robustness exhibited by LCANet-F.

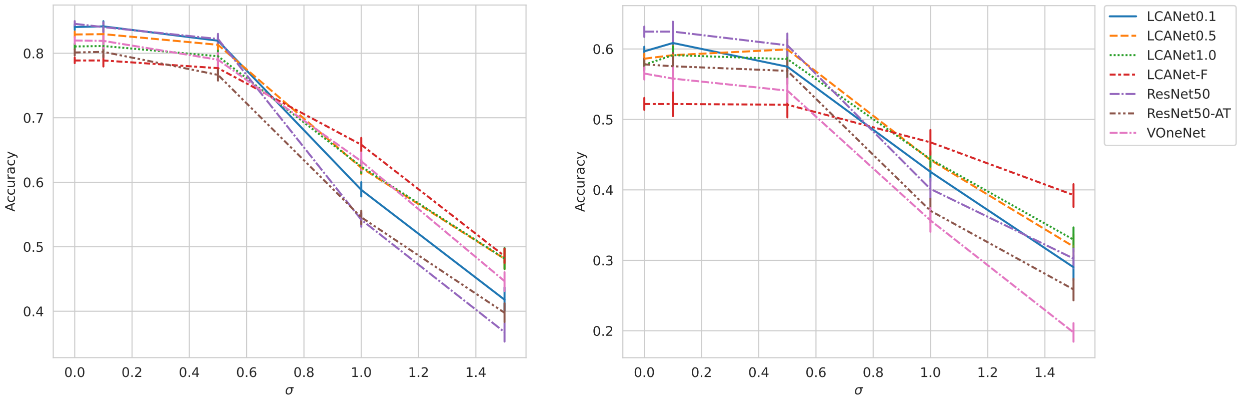


Figure 3. **LCANets are more robust to Gaussian blurring than state-of-the-art robust models.** Although accuracy decreases uniformly over all models on the UCF-101 (left) and HMDB-51 (right) datasets, LCANets maintain a small gain in accuracy over robust models under increased blurring. Strikingly, the unsupervised LCA features exhibit the most robustness of all models tested, as the LCANet-F model outperforms or matches the second most robust model on this corruption.

Gaussian noise corruption, the LCANet-F model outperforms the VOneNet (and all other models on HMDB-51) again suggesting that unsupervised V1-like features learned on data via sparse dictionary learning are more robust than input-layer features learned by standard CNN training.

Random Erasure) The final corruption we consider is random erasure which was first formulated as a data augmentation technique (Zhong et al., 2017). Erasure can be introduced through occlusion or sensor error. At each level of corruption, there does not appear to be any striking differences in model performance. LCANets achieve moderate robustness, lagging behind the top performing models by only a few percentage points. Their decreased performance likely results from the inability of the LCA neurons to charge up enough to get over threshold under lateral competition since they are receiving zero excitation from large portions of the input. Future work should address this corruption by modeling it as a missing data problem.

4.1.5. LCANets are Robust to High-Strength Black-Box Attacks

To evaluate the robustness of LCANets against adversarial examples, we employ a query-efficient black-box attack (Ilyas et al., 2018). We chose to evaluate against a untargeted black-box attack because it has previously been shown that the majority of state-of-the-art CNN defense methods developed on white-box attacks give a false sense of security through gradient obfuscation (Athalye et al., 2018), and they may be susceptible to certain black-box attacks (Mahmood et al., 2021). Black-box attacks also represent a more realistic threat, as it is unlikely that an adversary would have full access to a deployed model's parameters and structure.

As reported in (Dapello et al., 2020) for white-box attacks, VOneNet exhibits high robustness to this black-box attack (Figure 4), maintaining the highest accuracy on average across all values of ϵ. Although not the most robust across all attack strengths, LCANets perform well against this

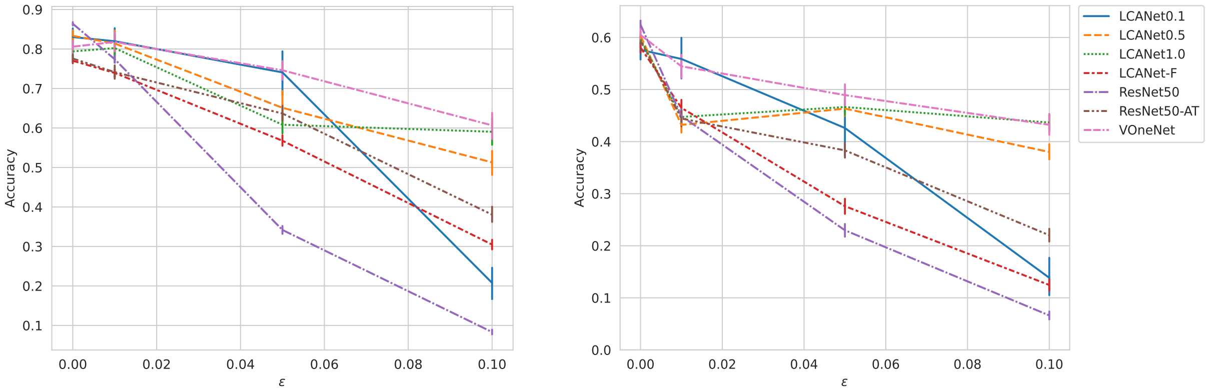


Figure 4. **LCANets are as robust to high strength black-box attack as state-of-the-art defense methods.** Although VOneNets are the most robust on average over all three attack strengths of this L_∞ constrained black-box attack (Ilyas et al., 2018), LCANets perform about the same or only a few percentage points worse while also outperforming adversarially trained ResNets on UCF-101 (left) and HMDB-51 (right).

attack, lagging behind VOneNet at lower ϵ by only a few percentage points and matching VOneNet accuracy under high ϵ. LCANet0.1 is competitive until ϵ reaches the value of λ it uses, at which point the accuracy plummets. This is expected, and likely results from the attack being able to drive many weakly active or inactive neurons over threshold. Not surprisingly, the ResNet50-AT is relatively competitive for values of ϵ at or below the value used during adversarial training. Finally, we can see that for adversarial robustness, lateral competition plays a larger role than robust features, indicated by the poor performance of LCANet-F relative to LCANet1.0 and LCANet0.5.

4.1.6. LCANets Are Susceptible to White-Box Adversarial Attacks

Since the LCA layer is differentiable, we are able to perform white-box attacks directly through the entire LCANet. Specifically, we use the projected gradient descent (PGD) attack (Madry et al., 2018a), which is a standard white-box attack used to perform adversarial training and test a network's robustness. We use the Adversarial Robustness Toolbox to implement the attack with an L_∞ constraint on the UCF-101 dataset. Following (Dapello et al., 2020), we set the number of attack iterations to 64 and the step size to $\epsilon/32$. Since the VOneNet has a stochastic layer, we follow (Athalye et al., 2018) and take the average gradient over 10 passes through the model to compute the update at each attack iteration when attacking this model.

We find that the adversarially-trained CNN performs much better than all other CNNs tested (Figure 5). The VOneNet exhibited the next best performance, followed by the LCANets in order from highest to lowest λ. Although the LCANets are not nearly as robust as the adversarially-trained ResNet, they are still more robust than the standard ResNet50. We observe a similar trend on the CIFAR-10 dataset (Figure 13).

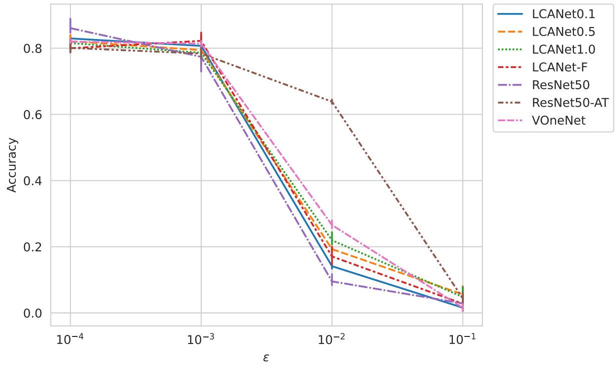


Figure 5. **LCANets are susceptible to white-box attacks.** Here, we see that the adversarially-trained CNN (ResNet50-AT) performs significantly better under a PGD attack than all other models on the UCF-101 dataset, which is to be expected. Although the LCANets are not nearly as robust as the adversarially-trained CNN, they are still significantly more robust than a standard CNN.

4.2. Image Recognition on CIFAR-10

4.2.1. Training and Testing Details

Much of the training details and model parameters are the same as in Section 4.1. In this image recognition task we use ResNet18 as the CNN backbone and train each model for 60 epochs on the standard training set with a batch size of 128, the one cycle learning rate scheduler with max learning rate of 0.12 (Dong et al., 2015), and horizontal flipping and random cropping augmentation. The LCA hyperparameters used for this task are in Table 2. We also compare against three adversarially-trained CNNs on this task by performing adversarial training with $\epsilon = 0.25$, $\epsilon = 0.5$, and $\epsilon = 1.0$ under an L_2 constraint by following the procedure outlined here. In most of the following experiments, we will only display results from the ResNet18 adversarially-trained with $\epsilon = 0.5$ (which we term ResNet18-AT0.5) since that was the best-performing of the three based on the average accuracy over the clean and adversarial test sets. The performance of all three adversarially-trained models under L_2 and L_∞ PGD attacks is displayed in Figure 11. Overall, we find that these results confirm the finding observed in the action recognition experiments.

4.2.2. LCANets Are Very Robust to Image Corruptions

Here, we use the CIFAR-10-C dataset to test the robustness of LCANets against different types of noise and corruptions. The CIFAR-10-C dataset originally has 20 different corruptions, each with 5 different severity levels. We observe that a good portion of the corruptions do not reduce the accuracy of any of our models significantly. As a result, we test on only those corruptions that caused the accuracy of the standard ResNet18 model to drop 10% or more from clean images to the highest severity level,[2] which leaves 11 corruptions: contrast, defocus blur, fog, gaussian blur, gaussian noise, impulse noise, motion blur, saturate, shot noise, speckle noise, and zoom blur.

Under most corruptions, LCANets perform at least as well as all other models tested here, with the most significant performance increases at the highest severity levels (See Figure 12 in the Appendix). The most striking difference between LCANets and all other models appears in the contrast corruption, under which LCANet performance remains unaffected while all other models suffer dramatic degradation in accuracy. LCANets also remain the top-performing models under all of the blurring corruptions and the fog corruption, which is in contrast to the ResNet18-AT0.5 in particular. The VOneNet's performance was relatively poor under most corruptions we tested, which confirms our action recognition experiments, and is similar to the results reported in (Dapello et al., 2020). Overall, we observe that the most robust LCANet's performance is more than 6% higher than all other models over the corruptions we considered here.

4.2.3. Augmenting Adversarial Training With LCA Frontends

So far, we have shown that LCANets are very robust to corruptions and a black-box attack, but they are less robust to a white-box attack than adversarially-trained CNNs. On the other hand, adversarially-trained CNNs were less robust to image corruptions than LCANets. Here, we highlight the versatility of the robust LCA frontend by combining it with the robust backbone of adversarially-trained CNNs, and we show that these adversarially-trained LCANets are extremely robust to both white-box adversarial attacks and image corruptions. To train this joint model, we initialize the CNN layers with the weights from an adversarially-trained CNN, in this case ResNet18-AT0.5. The first convolution layer is then removed and replaced with a new one which will take in the LCA feature maps and produce a representation that is useful to the adversarially-trained backbone. All layers in this hybrid model except for the LCA layer are then finetuned with adversarial training for 20 epochs to obtain competitive accuracy. Since the LCA layer is not updated, the network is still afforded robustness on corruptions, but it is also robust to white-box adversarial attacks (Figure 6).

5. Discussion

In this work, we take inspiration from the primary visual cortex and develop hybrid CNNs called LCANets, which leverage a frontend with sparsity and lateral competition to produce robust V1-like representations for downstream classification by a CNN. Through our experiments, we demonstrate competitive clean accuracy and state-of-the-art robustness to image corruptions. By performing the first direct adversarial attacks on a sparse coding CNN layer, we observe that sparse CNN layers are not as robust to adversarial attacks as previously thought. By combining the LCA frontend with adversarial training, we are able to produce CNNs that are very robust to both corruptions and standard adversarial attacks.

One interesting result is the large discrepancy between LCANet (and VOneNet) performance on the white-box and black-box attacks. One possible explanation for this difference regarding LCANet is that the exploration used in this black-box attack is small enough that it may be difficult for the attack to break LCA out of an equilibrium point, especially with limited queries. It is possible that 1,000 or 10,000 queries, as (Ilyas et al., 2018) suggest, would be more effective, although we do see that the attack is reasonably effective even with the number of queries we used. Future work can investigate these hypotheses more closely and compare other black-box attacks against LCANets to see if these results are specific to this attack or they apply more generally to an entire class of black-box attacks.

Biological visual systems operate seamlessly under a wide array of environments and conditions, but the same cannot be said for current CNNs. Here, we show that by incorporating just one computational element present in biological visual systems into a single CNN layer, we can greatly increase the robustness to different image corruptions and noise. In particular, we saw that LCANets are not vulnerable at all to changing contrasts, whereas all other models performed very poorly as more extreme manipulations in contrast were introduced. This is a well-known characteristic of V1 simple cells, which LCA has previously been

[2] This was determined before any other models were tested.

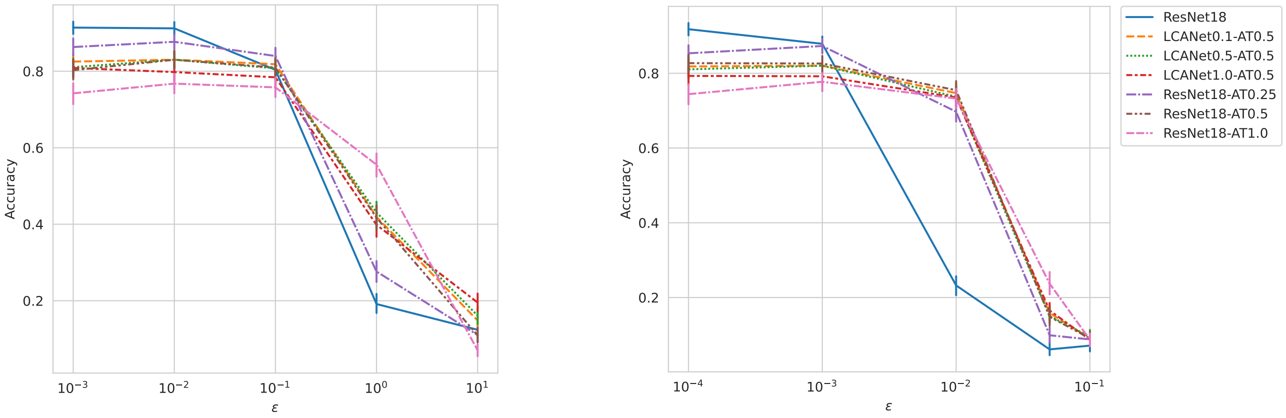


Figure 6. **Adversarially-trained LCANets achieve the same robustness as adversarially-trained CNNs.** Outfitting an adversarially-trained CNN with an LCA frontend does not reduce adversarial robustness under an L_2 (left) or an L_∞ (right) constraint.

shown to exhibit as well (Zhu & Rozell, 2013). Future work can focus on adding computational mechanisms which model other response characteristics of V1 neurons, which may also lead to increases in robustness.

By isolating the LCA features from the lateral competition, we saw that the features learned by LCA are relatively robust to many of the corruptions used, while the lateral competition affords the LCANet additional robustness under certain corruptions an adversarial attacks. Under corruptions, the LCA features were more robust than the current state-of-the-art V1-like CNN (Dapello et al., 2020), but contain less than half the complexity. In addition, this isolation may provide relevant hypotheses for computational neuroscience as well. For example, it is possible to use neuro-active chemicals that can impact lateral connectivity within specific regions to study the interplay between V1 receptive fields and lateral competition when biological visual systems are subjected to corrupted or adversarial stimuli. This could potentially further our understanding of visual processing in V1, which could then be used to produce more robust CNNs.

One limitation of our method is the large amount of time and computation required to perform hundreds of LCA iterations per forward pass of the LCANet. This adds significant time during training and testing compared to standard CNNs or VOneNets. It also potentially limits LCANets to datasets where the input dimensionality is either relatively small or can be reduced to a manageable size, or applications where inference speed is not critical. Future work can focus on developing more efficient biologically plausible sparse coding models or adapting LCANets to run on accelerated hardware, for example neuromorphic chips on which LCAs have previously been implemented.

6. Conclusion

Previous work has demonstrated the desirability of V1-like properties in CNNs. However, these techniques require the collection of large-scale neural recordings to stimuli that is similar to that in the desired task, or specialized neuroscientific models based on decades of experimental findings. We develop hybrid CNNs called LCANets with a biologically plausible frontend which performs sparse coding via the LCA algorithm and learns robust V1-like features via unsupervised dictionary learning. Using these LCANets, we test current hypotheses about the role of sparse coding in robustness against corruptions and adversarial attacks, and we show how our LCA frontend can easily be incorporated into other robust CNNs for further gains in robustness. Our results present a way of reducing the need for experimentally-measured data or handcrafted neuroscientific models while maintaining a V1-like representation with state-of-the-art robustness to common corruptions and adversarial attacks. A consequence of this is the potential to apply our techniques to non-traditional data modalities beyond natural biological sensing such as vision, for example autonomous driving and automated medical diagnostics where robustness is critical.

Acknowledgements

MT would like to thank the Center for Nonlinear Studies at Los Alamos National Laboratory (LANL) for supporting this work via a graduate fellowship. This work was also supported by the Laboratory Directed Research and Development program of LANL under project number 20210043DR, by NSF CRCNS, and by DOE ASCR. LANL is operated by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy (Contract No. 89233218CNA000001).

References

Adelson, E. H. and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

Ahmad, S. and Scheinkman, L. How can we be so dense? the benefits of using highly sparse representations. *arXiv preprint arXiv:1903.11257*, 2019.

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

Barlow, H. B. et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01), 1961.

Blakemore, C., Carpenter, R. H., and Georgeson, M. A. Lateral inhibition between orientation detectors in the human visual system. *Nature*, 228(5266):37–39, 1970.

Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., and Ecker, A. S. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028, 2021.

Carandini, M. and Heeger, D. J. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.

Carreira, J., Noland, E., Hillier, C., and Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

Chettih, S. N. and Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in v1. *Nature*, 567 (7748):334–340, 2019.

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., and Richards, B. A. Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. In *International Conference on Learning Representations*, 2020.

Crook, J. M., Kisvárday, Z. F., and Eysel, U. T. Evidence for a contribution of lateral inhibition to orientation tuning and direction selectivity in cat visual cortex: reversible inactivation of functionally characterized sites combined with neuroanatomical tracing techniques. *European Journal of Neuroscience*, 10(6):2056–2075, 1998.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., and DiCarlo, J. J. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13073–13087. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/98b17f068d5d9b7668e19fb8ae470841-Paper.pdf.

Dapello, J., Feather, J., Le, H., Marques, T., Cox, D., McDermott, J., DiCarlo, J. J., and Chung, S. Neural population geometry reveals the role of stochasticity in robust perception. *Advances in Neural Information Processing Systems*, 34, 2021.

Deneve, S., Pouget, A., and Latham, P. E. Divisive normalization, line attractor networks and ideal observers. In *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, pp. 104–110, Cambridge, MA, USA, 1999. MIT Press. ISBN 0262112450.

Dodds, E. M. and DeWeese, M. R. On the sparse structure of natural sounds and natural images: similarities, differences, and implications for neural coding. *Frontiers in computational neuroscience*, 13:39, 2019.

Dong, C., Loy, C. C., He, K., and Tang, X. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. Adversarial examples that fool both computer vision and time-limited humans. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/8562ae5e286544710b2e7ebe9858833b-Paper.pdf.

Foldiak, P. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

Gajewska-Dendek, E., Wróbel, A., and Suffczynski, P. Lateral inhibition as the organizer of the bottom-up attentional modulation in the primary visual cortex. *BMC Neuroscience*, 16(1):1–2, 2015.

Garcia-Cardona, C. and Wohlberg, B. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Transactions on Computational Imaging*, 4(3):366–381, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *stat*, 1050:20, 2015.

Hara, K., Kataoka, H., and Satoh, Y. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3154–3160, 2017.

Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018a.

Hara, K., Kataoka, H., and Satoh, Y. Towards good practice for action recognition with spatiotemporal 3d convolutions. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2516–2521. IEEE, 2018b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Hromádka, T., DeWeese, M. R., and Zador, A. M. Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS biology*, 6(1):e16, 2008.

Ilyas, A., Engstrom, L., and Madry, A. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2018.

Jortner, R. A., Farivar, S. S., and Laurent, G. A simple connectivity scheme for sparse coding in an olfactory system. *Journal of Neuroscience*, 27(7):1659–1669, 2007.

Jürgensen, A.-M., Khalili, A., Chicca, E., Indiveri, G., and Nawrot, M. P. A neuromorphic model of olfactory processing and sparse coding in the drosophila larva brain. *Neuromorphic Computing and Engineering*, 1(2):024008, 2021.

Kataoka, H., Wakamiya, T., Hara, K., and Satoh, Y. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020.

Kim, E., Yarnall, J., Shah, P., and Kenyon, G. T. A neuromorphic sparse coding defense to adversarial images. In *Proceedings of the International Conference on Neuromorphic Systems*, pp. 1–8, 2019.

Kim, E., Rego, J., Watkins, Y., and Kenyon, G. T. Modeling biological immunity to adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Krotov, D. and Hopfield, J. J. Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, 116(16):7723–7731, 2019.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32:12805–12816, 2019.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pp. 2556–2563. IEEE, 2011.

Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F., Pitkow, Z., and Tolias, A. Learning from brains how to regularize machines. *Advances in Neural Information Processing Systems*, 32: 9529–9539, 2019.

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018b. URL https://openreview.net/forum?id=rJzIBfZAb.

Mahmood, K., Gurevin, D., van Dijk, M., and Nguyen, P. H. Beware the black-box: on the robustness of recent defenses to adversarial examples. *Entropy*, 23(10):1359, 2021.

Mao, Z.-H. and Massaquoi, S. G. Dynamics of winner-take-all competition in recurrent neural networks with lateral inhibition. *IEEE transactions on neural networks*, 18(1): 55–69, 2007.

Nguyen, N. T. T., Moore, J. S., and Kenyon, G. T. Using models of cortical development based on sparse coding to discriminate between real and synthetically-generated faces. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pp. 1–7, 2020. doi: 10.1109/AIPR50011.2020.9425143.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Paiton, D. M., Frye, C. G., Lundquist, S. Y., Bowen, J. D., Zarcone, R., and Olshausen, B. A. Selectivity and robustness of sparse coding networks. *Journal of Vision*, 20(12):10–10, 11 2020. ISSN 1534-7362. doi: 10.1167/jov.20.12.10. URL https://doi.org/10.1167/jov.20.12.10.

Poo, C. and Isaacson, J. S. Odor representations in olfactory cortex:"sparse" coding, global inhibition, and oscillations. *Neuron*, 62(6):850–861, 2009.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20 (10):2526–2563, 2008.

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., and Brendel, W. A simple way to make neural networks robust against diverse image corruptions, 2020.

Safarani, S., Nix, A., Willeke, K. F., Cadena, S. A., Restivo, K., Denfield, G., Tolias, A. S., and Sinz, F. H. Towards robust vision by multi-task learning on monkey visual cortex. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Sompolinsky, H. and Shapley, R. New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology*, 7(4):514–522, 1997. ISSN 0959-4388. doi: https://doi.org/10.1016/S0959-4388(97)80031-1. URL https://www.sciencedirect.com/science/article/pii/S0959438897800311.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Springer, J. M., Strauss, C. S., Thresher, A. M., Kim, E., and Kenyon, G. T. Classifiers based on deep sparse coding architectures are robust to deep learning transferable examples. *arXiv preprint arXiv:1811.07211*, 2018.

Stemmler, M., Usher, M., and Niebur, E. Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics. *Science*, 269(5232):1877–1880, 1995.

Sun, B., Tsai, N.-h., Liu, F., Yu, R., and Su, H. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11447–11456, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Terashima, H., Hosoya, H., Tani, T., Ichinohe, N., and Okada, M. Sparse coding of harmonic vocalization in monkey auditory cortex. *Neurocomputing*, 103:14–21, 2013.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Vinje, W. E. and Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

Wang, H., Wu, X., Huang, Z., and Xing, E. P. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020.

Yoshida, T. and Ohki, K. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature communications*, 11(1):1–19, 2020.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. arxiv. *arXiv preprint arXiv:1708.04896*, 2017.

Zhu, M. and Rozell, C. J. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLOS Computational Biology*, 9(8):1–15, 08 2013. doi: 10.1371/journal.pcbi.1003191. URL https://doi.org/10.1371/journal.pcbi.1003191.

Zylberberg, J., Murphy, J. T., and DeWeese, M. R. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS computational biology*, 7(10):e1002250, 2011.

A. Supplementary Information

A.1. LCA Hyperparameters

Table 1. **LCA Hyperparameters on UCF-101 and HMDB-51.**

Hyperparameter	Value
k_H	9
k_W	9
k_T	5
M	256
λ	0.5
τ	250
stride_H	2
stride_W	2
stride_T	1
LCA iterations	600

Table 2. **LCA hyperparameters on CIFAR-10.**

Hyperparameter	Value
k_H	7
k_W	7
M	128
λ	0.5
τ	100
stride_H	2
stride_W	2
LCA iterations	600

A.2. Corruptions on Action Recognition

A.2.1. ADDITIVE GAUSSIAN NOISE

Additive Gaussian noise is typically used to model thermal noise (AKA Johnson-Nyquist noise), which is present in all electrical circuits. Although current image denoising methods are relatively good at removing this noise, they are not perfect. As a result, robust models should be able to deal with at least a small amount of Gaussian noise, especially for critical applications. To test each model's robustness to additive Gaussian noise, we add random values from a Gaussian distribution to each input video clip while varying the standard deviation (Figure 7).

A.2.2. GAUSSIAN BLUR

To perform this blurring, we use the GaussianBlur function available in torchvision with a kernel size of 5 and varying values of sigma (Figure 8).

A.2.3. RANDOMERASURE

Random erasure is applied by randomly selecting a rectangle within a frame and changing all pixel values in the rectangle uniformly such that they have a value of zero after normalization (Figure 9). This was performed on each frame individually. To perform this corruption, we use the RandomErasing function in torchvision.

A.3. Black-Box Attack Details

For the black-box attack, we use an L_∞ constraint with the same hyperparameters used in (Ilyas et al., 2018). In our preliminary experiments using this attack, we initially set the number of queries to 1,500 but observed that the attack

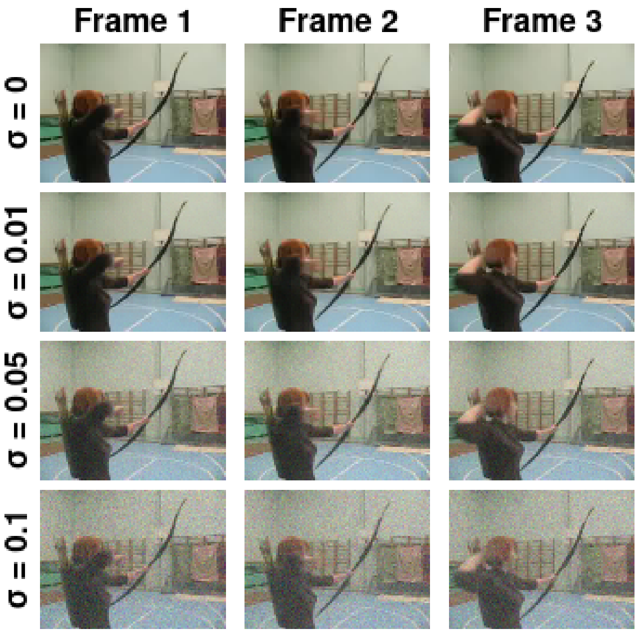


Figure 7. **Additive Gaussian Noise.** Random values taken from a Gaussian distribution are added to each frame. Although we use clips of 12 consecutive frames to train and test each model, three frames are shown here for illustration.

typically plateaus or reaches 100% effectiveness before 500 model queries for all models and ϵ values. As a result, we set the maximum number of queries to 400 for all models.

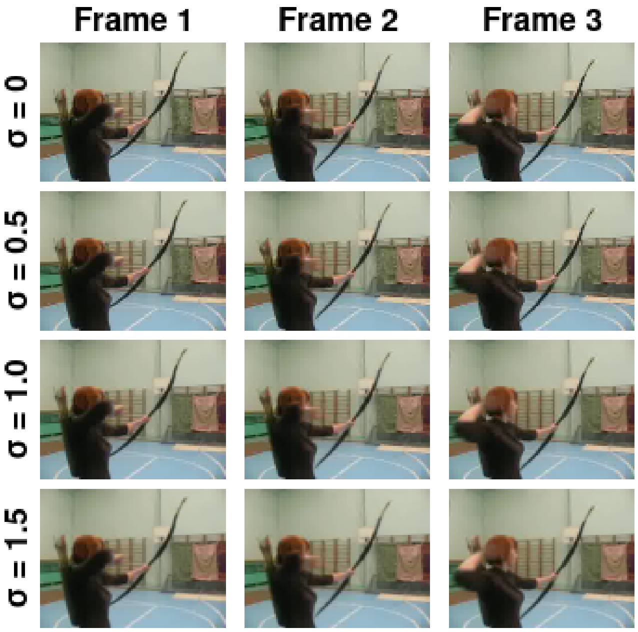


Figure 8. **Gaussian Blur.** A Gaussian blur was applied to each frame with kernel size 5. The σ parameter determines the contribution of neighboring pixels as a function of distance from the center pixel. Although we use clips of 12 consecutive frames to train and test each model, three frames are shown here for illustration.

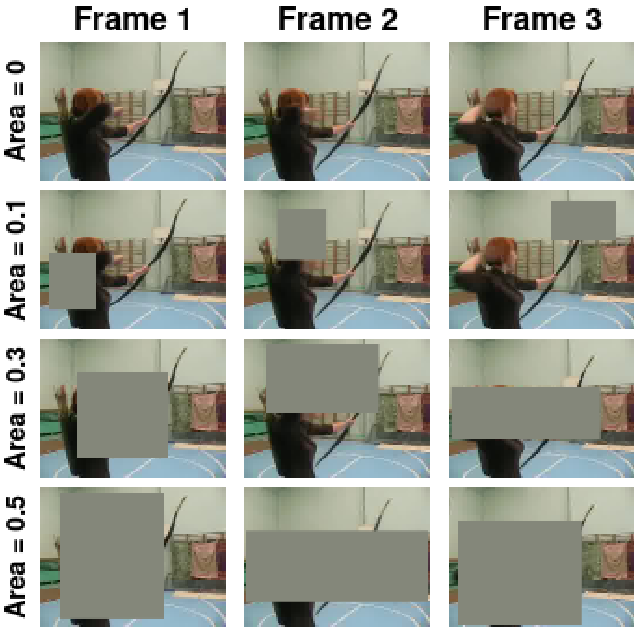


Figure 9. **Random Erasure.** The random erasure corruption randomly chooses a rectangle from each frame and replaces the pixel values within that rectangle such that it is zero after normalization. Area refers to the proportion of the image that is erased. Although we use clips of 12 consecutive frames to train and test each model, three frames are shown here for illustration.

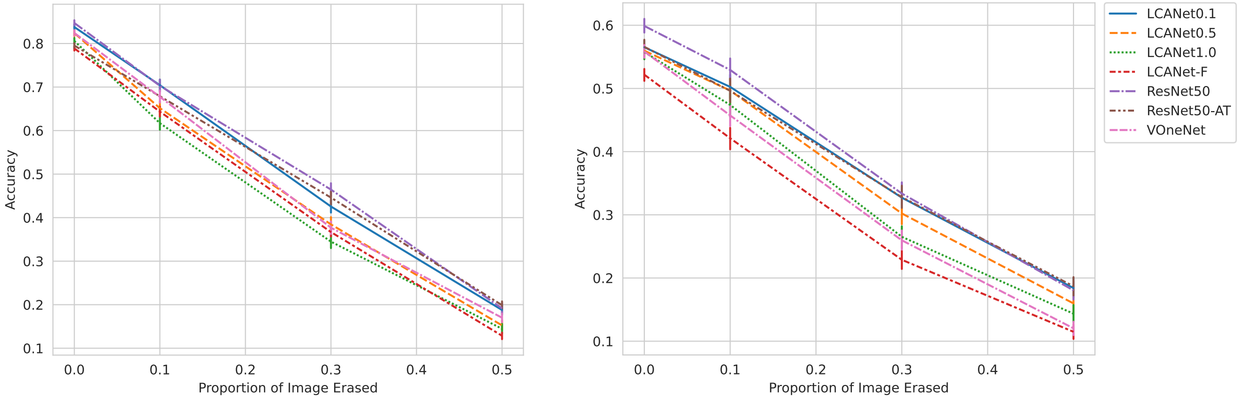
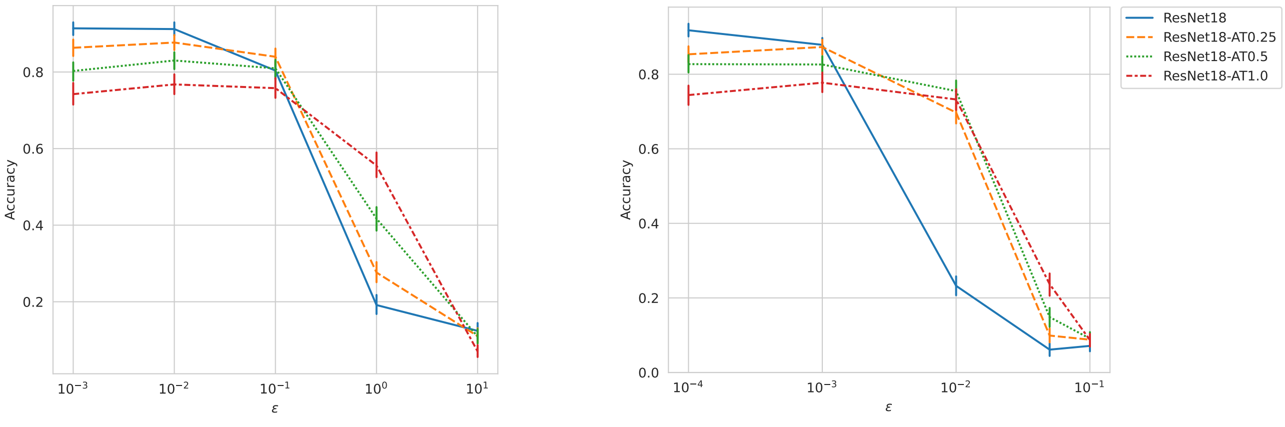


Figure 10. **LCANets maintain competitive accuracy under random erasure corruption.** Across models, performance decreases uniformly as a higher proportion of the image is erased on UCF-101 (left) and HMDB-51 (right). Although ResNet50 and ResNet50-AT exhibit the highest accuracy under varying levels of erasure, LCANet accuracy is only a few percentage points below the top performing models.

A.4. Adversarially-Trained CNN Performance

Figure 11 illustrates the performance of the three adversarially-trained models under a PGD attack. We follow (Dapello et al., 2020) and use 20 PGD iterations during adversarial training and 64 PGD iterations during the attacks.

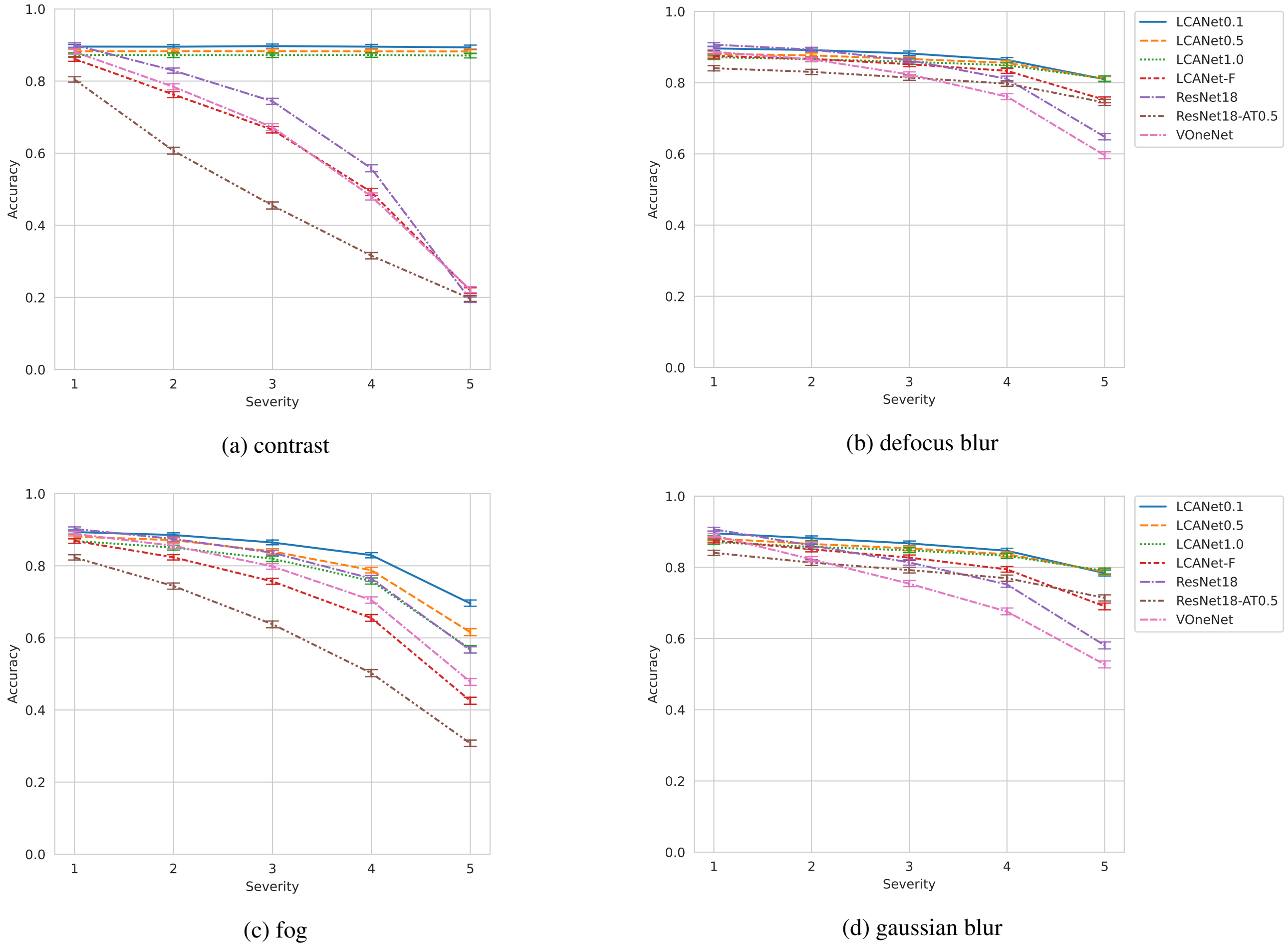


(a) L_2

(b) L_∞

Figure 11. **The performance of the three adversarially-trained ResNets used in the white-box attacks in Section 4.2 against a PGD with an L_2 and L_∞ constraint.**

A.5. CIFAR-10-C Results



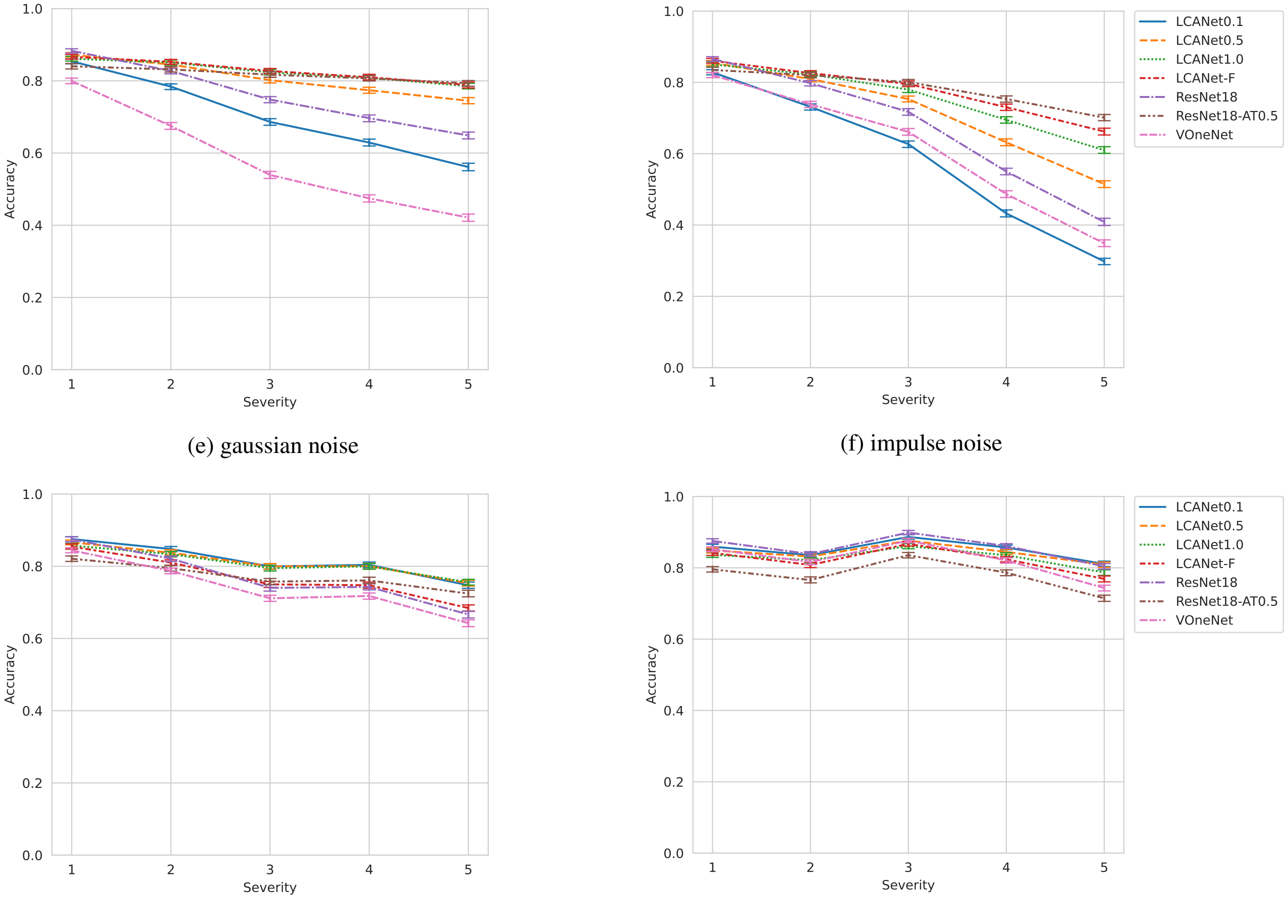
(a) contrast

(b) defocus blur

(c) fog

(d) gaussian blur

Figure 12. **LCANets are significantly more robust to different image corruptions.**



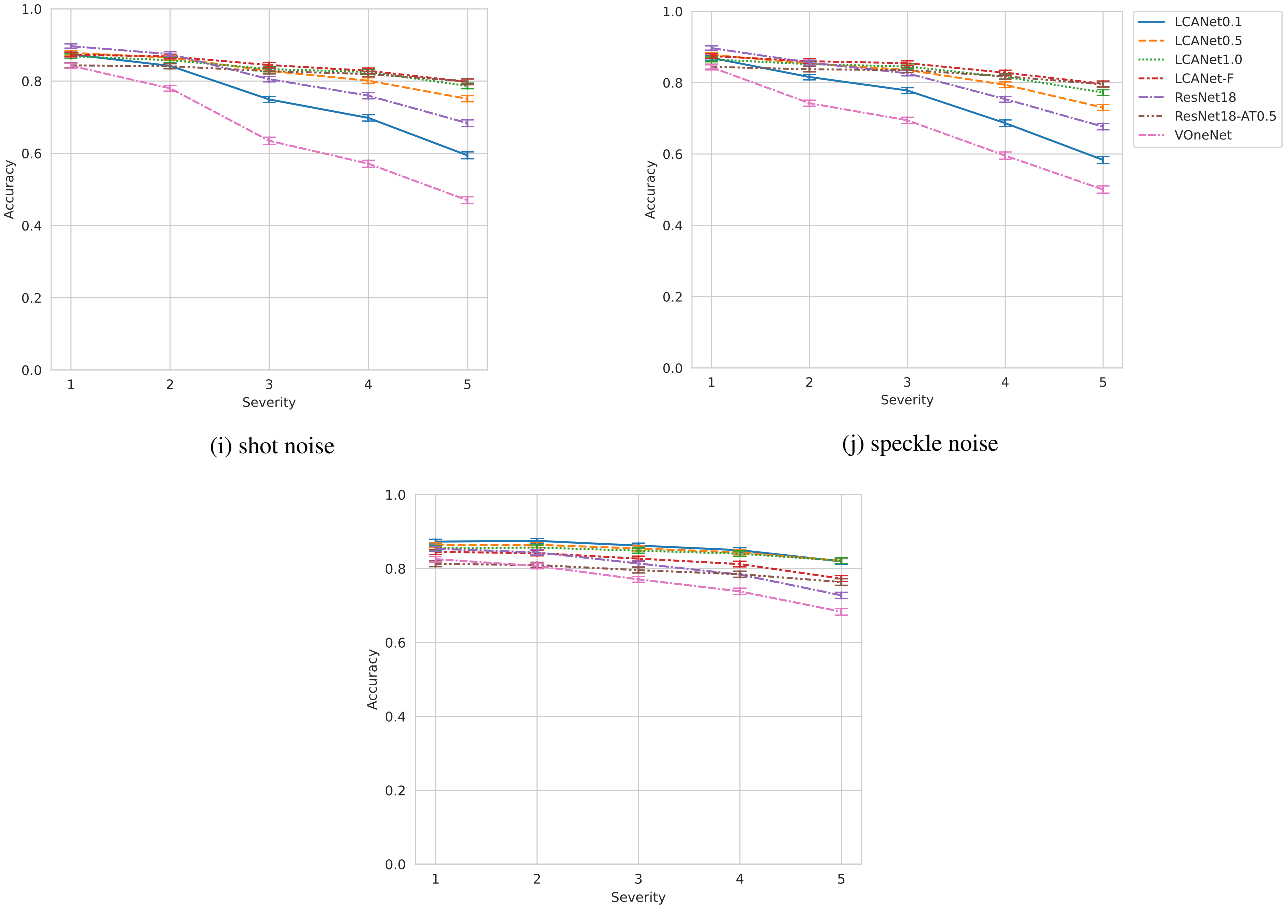
(e) gaussian noise

(f) impulse noise

(g) motion blur

(h) saturate

Figure 12. **LCANets are significantly more robust to different image corruptions (cont.).**



(i) shot noise

(j) speckle noise

(k) zoom blur

Figure 12. **LCANets are significantly more robust to image corruptions (cont.).**

A.6. White-Box Attack on CIFAR-10

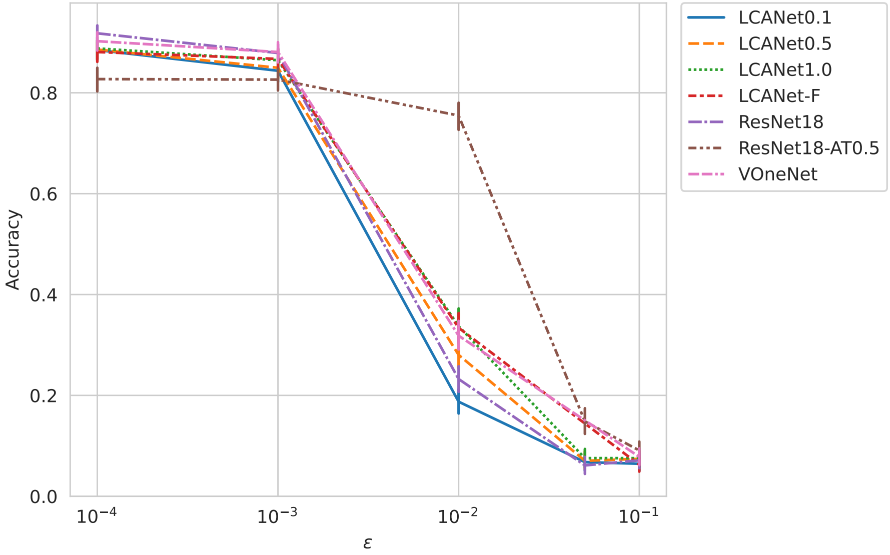
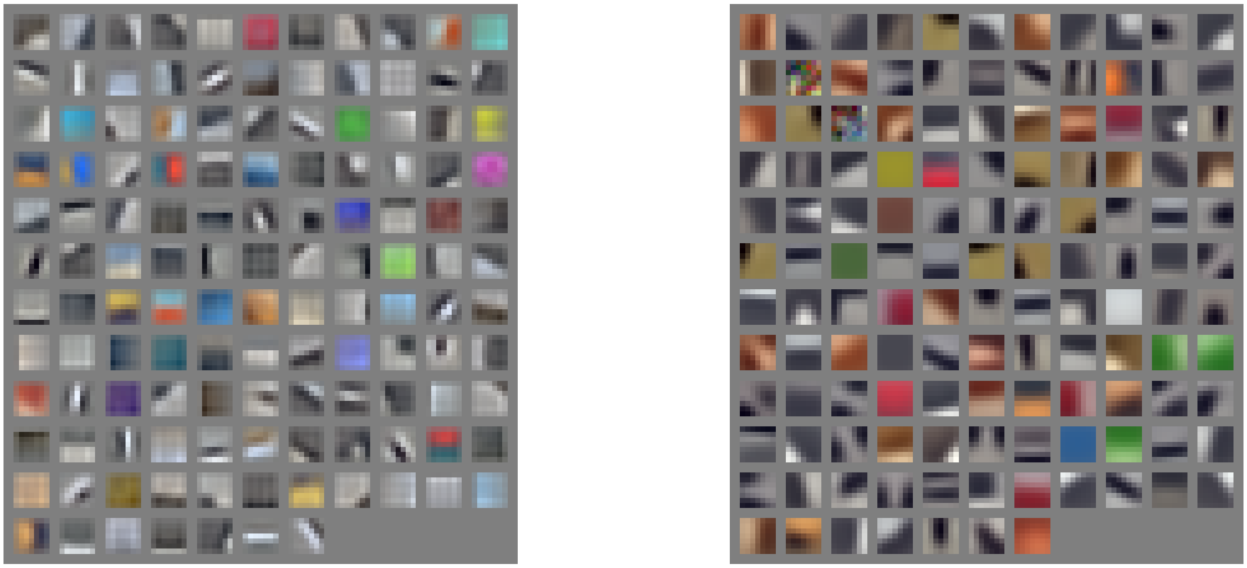


Figure 13. **LCANets are susceptible to white-box attacks on CIFAR-10.**

A.7. Unsupervised LCA Features

Various algorithms have been proposed to learn features in an unsupervised fashion from data, many of which are even biologically inspired. One such algorithm, referred to as the Krotov-Hebbian learning rule (Krotov & Hopfield, 2019), performs local Hebbian learning with the inclusion of global inhibition, which is in contrast to the local inhibition present in LCA. The features learned by the Krotov-Hebbian rule have recently been shown to produce good representations for downstream image classifiers as well (Krotov & Hopfield, 2019). Therefore, we use this learning rule to learn unsupervised features with the intent to determine how robust our features are relative to a comparable biologically-plausible unsupervised feature learning algorithm. When we put both sets of features (Figure 14) at the frontend of a ResNet18 and train it on the CIFAR-10 dataset, we observe that they do not perform significantly different, even over corruptions and adversarial attacks.



(a) LCA Features

(b) Krotov-Hebbian Features

Figure 14. **Comparison of the LCA and Krotov-Hebbian unsupervised features.**