LINKING AVERAGE- AND WORST-CASE PERTURBATION ROBUSTNESS VIA CLASS SELECTIVITY AND DIMENSIONALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Representational sparsity is known to affect robustness to input perturbations in deep neural networks (DNNs), but less is known about how the semantic content of representations affects robustness. Class selectivity—the variability of a unit's responses across data classes or dimensions—is one way of quantifying the sparsity of semantic representations. Given recent evidence that class selectivity may not be necessary for, and in some cases can impair generalization, we sought to investigate whether it also confers robustness (or vulnerability) to perturbations of input data. We found that class selectivity leads to increased vulnerability to average-case (naturalistic) perturbations in ResNet18 and ResNet20, as measured using Tiny ImageNetC and CIFAR10C, respectively. Networks regularized to have lower levels of class selectivity are more robust to average-case perturbations, while networks with higher class selectivity are more vulnerable. In contrast, we found that class selectivity increases robustness to multiple types of worst-case (i.e. white box adversarial) perturbations, suggesting that while decreasing class selectivity is helpful for average-case perturbations, it is harmful for worst-case perturbations. To explain this difference, we studied the dimensionality of the networks' representations: we found that the dimensionality of early-layer representations is inversely proportional to a network's class selectivity, and that adversarial samples cause a larger increase in early-layer dimensionality than corrupted samples. We also found that the input-unit gradient was more variable across samples and units in high-selectivity networks compared to low-selectivity networks. These results lead to the conclusion that units participate more consistently in low-selectivity regimes compared to high-selectivity regimes, effectively creating a larger attack surface and hence vulnerability to worst-case perturbations.

1 INTRODUCTION

Methods for understanding deep neural networks (DNNs) often attempt to find individual neurons or small sets of neurons that are representative of a network's decision (Erhan et al., 2009; Zeiler and Fergus, 2014; Karpathy et al., 2016; Amjad et al., 2018; Lillian et al., 2018; Dhamdhere et al., 2019; Olah et al., 2020). Selectivity in individual units (i.e. variability in a neuron's activations across semantically-relevant data features) has been of particular interest to researchers trying to better understand deep neural networks (DNNs) (Zhou et al., 2015; Olah et al., 2017; Morcos et al., 2018; Zhou et al., 2018; Meyes et al., 2019; Na et al., 2019; Zhou et al., 2019; Rafegas et al., 2019; Bau et al., 2020; Leavitt and Morcos, 2020). However, recent work has shown that selective neurons can be irrelevant, or even detrimental to network performance, emphasizing the importance of examining distributed representations for understanding DNNs (Morcos et al., 2018; Donnelly and Roegiest, 2019; Dalvi et al., 2019b; Leavitt and Morcos, 2020).

In parallel, work on robustness seeks to build models that are robust to perturbed inputs (Szegedy et al., 2013; Carlini and Wagner, 2017a;b; Vasiljevic et al., 2016; Kurakin et al., 2017; Gilmer et al., 2018; Zheng et al., 2016). Hendrycks and Dietterich (2019) distinguish between two types of robustness: corruption robustness, which measures a classifier's performance on low-quality or naturalistically-perturbed inputs—and thus is an "average-case" measure—and adversarial robustness,

which measures a classifier's performance on small, additive perturbations that are tailored to the classifier—and thus is a "worst-case" measure.[1]

Research on robustness has been predominantly focused on worst-case perturbations, which is affected by weight and activation sparsity (Madry et al., 2018; Balda et al., 2020; Ye et al., 2018; Guo et al., 2018; Dhillon et al., 2018) and representational dimensionality (Langeberg et al., 2019; Sanyal et al., 2020; Nayebi and Ganguli, 2017). But less is known about the mechanisms underlying average-case perturbation robustness and its common factors with worst-case robustness. Some techniques for improving worst-case robustness also improve average-case robustness (Hendrycks and Dietterich, 2019; Ford et al., 2019), thus it is possible that sparsity and representational dimensionality also contribute to average-case robustness. Because class selectivity regularization provides a method for controlling the sparsity of semantic representations and can also improve test accuracy on unperturbed data (Leavitt and Morcos, 2020), we sought to investigate whether it could be utilized to improve perturbation robustness and elucidate the factors responsible for it.

In this work we pursue a series of experiments investigating the causal role of selectivity in robustness to worst-case and average-case perturbations in DNNs. To do so, we used a recently-developed class selectivity regularizer (Leavitt and Morcos, 2020) to directly modify the amount of class selectivity learned by DNNs, and examined how this affected the DNNs' robustness to worst-case and average-case perturbations. Our findings are as follows:

- Networks regularized to have lower levels of class selectivity are more robust to average-case perturbations, while networks with higher class selectivity are generally less robust to average-case perturbations, as measured in ResNets using the Tiny ImageNetC and CIFAR10C datasets. The corruption robustness imparted by regularizing against class selectivity was consistent across nearly all tested corruptions.
- In contrast to its impact on average-case perturbations, decreasing class selectivity *reduces* robustness to worst-case perturbations in both tested models, as assessed using gradient-based white-box attacks.
- The variability of the input-unit gradient across samples and units is proportional to a network's overall class selectivity, indicating that high variability in perturbability within and across units may facilitate worst-case perturbation robustness.
- The dimensionality of activation changes caused by corruption markedly increases in early layers for both perturbation types, but is larger for worst-case perturbations and low-selectivity networks. This implies that representational dimensionality may present a trade-off between worst-case and average-case perturbation robustness.

Our results demonstrate that changing class selectivity, and hence the sparsity of semantic representations, can confer robustness to average-case or worst-case perturbations, but not both simultaneously. They also highlight the roles of input-unit gradient variability and representational dimensionality in mediating this trade-off.

2 Related Work

2.1 Perturbation Robustness

The most commonly studied form of robustness in DNNs is robustness to adversarial attacks, in which an input is perturbed in a manner that maximizes the change in the network's output while attempting to minimize or maintain below some threshold the magnitude of the change to the input (Serban et al., 2019; Warde-Farley and Goodfellow, 2017) . Because white-box adversarial attacks are optimized to best confuse a given network, robustness to adversarial attacks are a "worst-case" measure of robustness. Two factors that have been proposed to account for DNN robustness to worst-case perturbations are particularly relevant to the present study: sparsity and dimensionality.

[1] We use the terms "worst-case perturbation" and "average-case perturbation" instead of "adversarial attack" and "corruption", respectively, because this usage is more general and dispenses with the implied categorical distinction of using seemingly-unrelated terms. Also note that while Hendrycks and Dietterich (2019) assign specific and distinct meanings to "perturbation" and "corruption", we use the term "perturbation" more generally to refer to any change to an input.

Multiple studies have linked activation and weight sparsity with robustness to worst-case perturbations. Adversarial training improves worst-case robustness Goodfellow et al. (2015); Huang et al. (2016) and results in sparser weight matrices (Madry et al., 2018; Balda et al., 2020). Methods for increasing the sparsity of weight matrices (Ye et al., 2018; Guo et al., 2018) and activations (Dhillon et al., 2018) likewise improve worst-case robustness, indicating that the weight sparsity caused by worst-case perturbation training is not simply a side-effect.

Researchers have also attempted to understand the nature of worst-case robustness from a perspective complementary to that of sparsity: dimensionality. Like sparsity, worst-case perturbation training reduces the rank of weight matrices and representations, and regularizing weight matrices and representations to be low-rank can improve worst-case perturbation robustness (Langeberg et al., 2019; Sanyal et al., 2020; Nayebi and Ganguli, 2017). Taken together, these studies support the notion that networks with low-dimensional representations are more robust to worst-case perturbations.

Comparatively less research has been conducted to understand the factors underlying average-case robustness. Certain techniques for improving worst-case perturbation robustness also help against average-case perturbations (Hendrycks and Dietterich, 2019; Geirhos et al., 2018; Ford et al., 2019). Examining the frequency domain has elucidated one mechanism: worst-case perturbations for "baseline" models tend to be in the high frequency domain, and improvements in average-case robustness resulting from worst-case robustness training are at least partially ascribable to models becoming less reliant on high-frequency information (Yin et al., 2019; Tsuzuku and Sato, 2019; Geirhos et al., 2018). But it remains unknown whether other factors such as sparsity and dimensionality link these two forms of robustness.

2.2 Class selectivity

One technique that has been of particular interest to researchers trying to better understand deep (and biological) neural networks is examining the selectivity of individual units (Zhou et al., 2015; Olah et al., 2017; Morcos et al., 2018; Zhou et al., 2018; Meyes et al., 2019; Na et al., 2019; Zhou et al., 2019; Rafegas et al., 2019; Bau et al., 2020; Leavitt and Morcos, 2020; Sherrington, 1906; Kandel et al., 2000). Evidence regarding the importance of selectivity has mostly relied on single unit ablation, and has been equivocal (Radford et al., 2017; Morcos et al., 2018; Amjad et al., 2018; Zhou et al., 2018; Donnelly and Roegiest, 2019; Dalvi et al., 2019a). However Leavitt and Morcos (2020) examined the causal role of single unit selectivity in network performance using an approach that sidesteps the limitations of single unit ablation. By regularizing for or against class selectivity in the loss function, they found that reducing class selectivity has little negative impact on—and can even improve—test accuracy in CNNs trained on image recognition tasks, but that increasing class selectivity has significant negative effects on test accuracy. However, their study focused on examining the effects of class selectivity on test accuracy in unperturbed (clean) inputs. Thus it remains unknown how class selectivity affects robustness to perturbed inputs, and whether class selectivity can serve as or elucidate a link between worst-case and average-case robustness.

3 Approach

A detailed description of our approach is provided in Appendix A.1.

Models and training protocols Our experiments were performed on ResNet18 (He et al., 2016) trained on Tiny ImageNet (Fei-Fei et al., 2015), and ResNet20 (He et al., 2016) trained on CIFAR10 (Krizhevsky, 2009). We focus primarily on the results for ResNet18 trained on Tiny ImageNet in the main text for space, though results were qualitatively similar for ResNet20 trained on CIFAR10. Experimental results were obtained with model parameters from the epoch that achieved the highest validation set accuracy over the training epochs, and 20 replicate models with different random seeds were run for each hyperparameter set.

Class selectivity index Following (Leavitt and Morcos, 2020). A unit's class selectivity index is calculated as follows: At every ReLU, the activation in response to a single sample was averaged across all elements of the filter map (which we refer to as a "unit"). The class-conditional mean activation was then calculated across all samples in the clean test set, and the class selectivity index (SI) was calculated as follows:

$$SI = \frac{\mu_{max} - \mu_{-max}}{\mu_{max} + \mu_{-max}} \tag{1}$$

where μ_{max} is the largest class-conditional mean activation and μ_{-max} is the mean response to the remaining (i.e. non-μ_{max}) classes. The selectivity index ranges from 0 to 1. A unit with identical average activity for all classes would have a selectivity of 0, and a unit that only responds to a single class would have a selectivity of 1.

As Morcos et al. (2018) note, the selectivity index is not a perfect measure of information content in single units. For example, a unit with a litte bit of information about many classes would have a low selectivity index. However, it identifies units that are class-selective similarly to prior studies (Zhou et al., 2018). Most importantly, it is differentiable with respect to the model parameters.

Class selectivity regularization We used (Leavitt and Morcos, 2020)'s class selectivity regularizer to control the levels of class selectivity learned by units in a network during training. Class selectivity regularization is achieved by minimizing the following loss function during training:

$$loss = -\sum_{c}^{C} y_c \cdot \log(\hat{y_c}) - \alpha \mu_{SI} \tag{2}$$

The left-hand term in the loss function is the standard classification cross-entropy, where c is the class index, C is the number of classes, y_c is the true class label, and $\hat{y_c}$ is the predicted class probability. The right-hand component of the loss function, $-\alpha\mu_{SI}$, is the class selectivity regularizer. The regularizer consists of two terms: the selectivity term,

$$\mu_{SI} = \frac{1}{L}\sum_{l}^{L}\frac{1}{U}\sum_{u}^{U} SI_u \tag{3}$$

where l is a convolutional layer, L is number of layers, u is a unit, U is the number of units in a given layer, and SI_u is the class selectivity index of unit u. The selectivity term of the regularizer is obtained by computing the selectivity index for each unit in a layer, then computing the mean selectivity index across units within each layer, then computing the mean selectivity index across layers. Computing the mean within layers before computing the mean across layers (as compared to computing the mean across all units in the network) mitigates the biases induced by the larger numbers of units in deeper layers. The other term in the regularizer is α, the regularization scale, which determines whether class selectivity is promoted or discouraged. Negative values of α discourage class selectivity in individual units and positive values encourage it. The magnitude of α controls the contribution of the selectivity term to the overall loss. During training, the class selectivity index was computed for each minibatch. The final (output) layer was not subject to selectivity regularization or included in our analyses because by definition, the output layer must be class selective in a classification task.

Measuring average-case robustness To evaluate robustness to average-case perturbations, we tested our networks on CIFAR10C and Tiny ImageNetC, two benchmark datasets consisting of the CIFAR10 or Tiny ImageNet data, respectively, to which a set of naturalistic corruptions have been applied (Hendrycks and Dietterich, 2019, Figure A1). . We average across all corruption types and severities (see Appendix A.1.2 for details) when reporting corrupted test accuracy.

Measuring worst-case robustness We tested our models' worst-case (i.e. adversarial) robustness using two methods. The fast gradient sign method (FGSM) (Goodfellow et al., 2015) is a simple attack that computes the gradient of the loss with respect to the input image, then scales the image's pixels (within some bound) in the direction that increases the loss. The second method, projected gradient descent (PGD) (Kurakin et al., 2016; Madry et al., 2018), is an iterated version of FGSM. We used a step size of 0.0001 and an l_∞ norm perturbation budget (ϵ) of 16/255.

Computing the stability of units and layers To quantify variation in networks' perturbability, we first computed the l_2 norm of the input-unit gradient for each unit u in a network. We then computed the mean (μ_u) and standard deviation (σ_u) of the norm across samples for each unit. σ_u/μ_u yields the coefficient of variation (Everitt, 2002) for a unit (CV_u), a measure of variation in perturbability for individual units. We also quantified the variation *across* units in a layer by computing the standard deviation of μ_u across units in a layer l, $\sigma(\mu_u) = \sigma_l$, and dividing this by the corresponding mean across units $\mu(\mu_u) = \mu_l$, to yield the CV across units $\sigma_l/\mu_l = CV_l$.

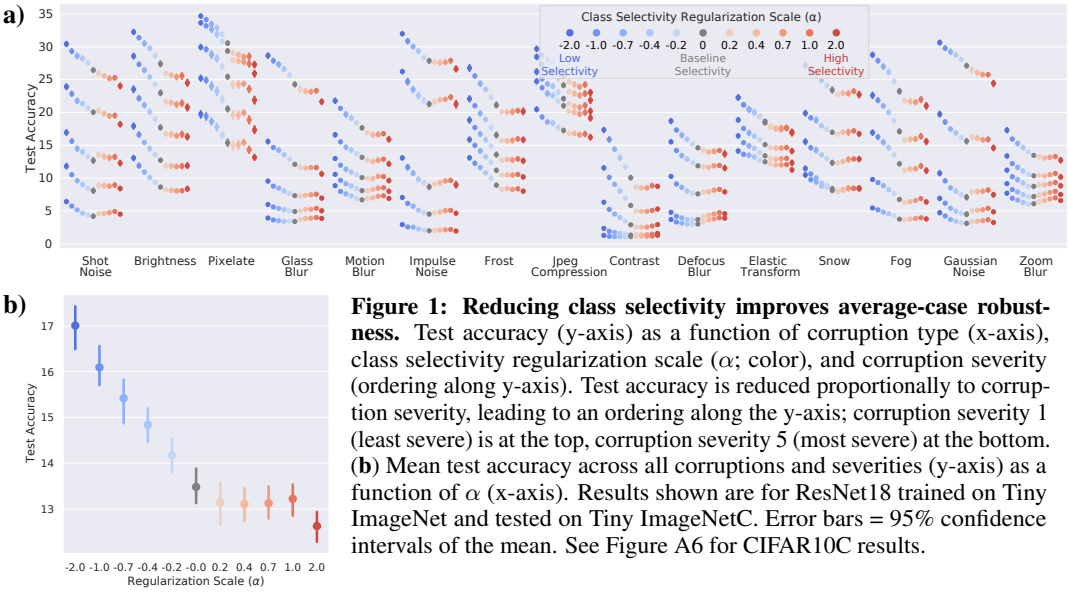

Figure 1: Reducing class selectivity improves average-case robustness. Test accuracy (y-axis) as a function of corruption type (x-axis), class selectivity regularization scale (α; color), and corruption severity (ordering along y-axis). Test accuracy is reduced proportionally to corruption severity, leading to an ordering along the y-axis; corruption severity 1 (least severe) is at the top, corruption severity 5 (most severe) at the bottom. (**b**) Mean test accuracy across all corruptions and severities (y-axis) as a function of α (x-axis). Results shown are for ResNet18 trained on Tiny ImageNet and tested on Tiny ImageNetC. Error bars = 95% confidence intervals of the mean. See Figure A6 for CIFAR10C results.

4 Results

4.1 Average-case robustness is inversely proportional to class selectivity

Certain kinds of sparsity—including reliance on single directions (Morcos et al., 2018), and the semantic sparsity measured by class selectivity (Leavitt and Morcos, 2020)—have been shown to impair network performance. We sought to extend this question to robustness: how does the sparsity of semantic representations affect robustness to average-case perturbations of the input data? We used a recently-introduced method (Leavitt and Morcos (2020); Approach 3) to modulate the amount of class selectivity learned by DNNs (Figure A2 demonstrates effects of selectivity regularization). We then examined how this affected performance on Tiny ImageNetC and CIFAR10C, two benchmark datasets for average-case corruptions (Approach 3).

Changing the level of class selectivity across neurons in a network could have one of three possible effects on corruption robustness: If concentrating semantic representations into fewer neurons (i.e. promoting semantic sparsity) provides fewer potent dimensions on which perturbed inputs can act, then increasing class selectivity should confer networks with robustness to average-case perturbations, while reducing class selectivity should render networks more vulnerable. Alternatively, if distributing semantic representations across more units (i.e. reducing sparsity) dilutes the changes induced by perturbed inputs, then reducing class selectivity should increase a network's robustness to average-case perturbations, while increasing class selectivity should reduce robustness.

We found that decreasing class selectivity leads to increased robustness to average-case perturbations for both ResNet18 tested on Tiny ImageNetC (Figure 1) and ResNet20 tested on CIFAR10C (Figure A6). In ResNet18, we found that mean test accuracy on corrupted inputs increases as class selectivity decreases (Figure 1), with test accuracy reaching a maximum at regularization scale $\alpha = -2.0$ (mean test accuracy across corruptions and severities at $\alpha_{-2.0}$ =17), representing a 3.5 percentage point (pp) increase relative to no selectivity regularization (i.e. α_0; test accuracy at $\alpha_0 = 13.5$). In contrast, regularizing to increase class selectivity has either no effect or a negative impact on corruption robustness. Corrupted test accuracy remains relatively stable until $\alpha = 1.0$, after which point it declines. The results are qualitatively similar for ResNet20 tested on CIFAR10C (Figure A6), except the vulnerability to corruption caused by increasing selectivity is even more dramatic. We also controlled for the effect of baseline clean accuracy across models with different class selectivity regularization scales and found similar results (Appendix A.3).

We also found that the effect of class selectivity on perturbed robustness is consistent across corruption types. In ResNet18, regularizing against selectivity improves perturbation robustness for 12 out of 15 corruptions types in Tiny ImageNetC (Figure 1a; Figure A4), and for 14 out of 19 corruption types in CIFAR10C for ResNet20 (Figure A6a, Figure A7). Together these results demonstrate that reduced class selectivity confers robustness to average-case perturbations, implying that distributing semantic representations across neurons—i.e. low sparsity—dilutes the changes induced by average-case perturbations.

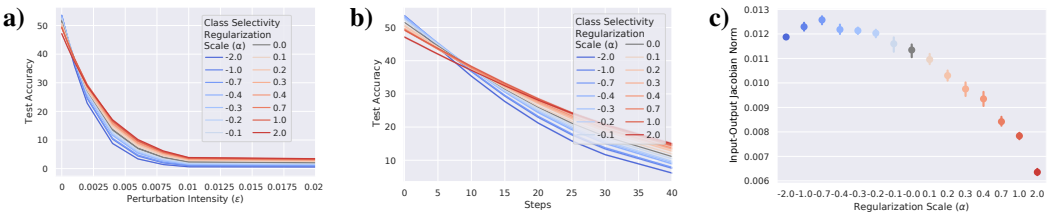

Figure 2: Class selectivity imparts worst-case perturbation robustness. (**a**) Test accuracy (y-axis) as a function of perturbation intensity (ϵ; x-axis) and class selectivity regularization scale (α; color) for the FGSM attack. (**b**) Test accuracy (y-axis) as a function of attack optimization steps (x-axis) and α for the PGD attack. (**c**) Network stability, as measured with l_2 norm of the input-output Jacobian (y-axis), as a function of α (x-axis). All results are for ResNet18 trained on Tiny ImageNet. Shaded region and bars = 95% confidence interval of the mean. See Figure A9 for ResNet20 results.

4.2 Class selectivity imparts worst-case perturbation robustness

We showed that the sparsity of a network's semantic representations, as measured with class selectivity, is causally related to a network's robustness to average-case perturbations. But how does the sparsity of semantic representations affect *worst-case* robustness? We addressed this question by testing our class selectivity-regularized networks on inputs that had been perturbed using using one of two gradient-based methods (see Approach 3).

If distributing semantic representations across units provides more dimensions upon which a worst-case perturbation is potent, then worst-case perturbation robustness should be proportional to class selectivity. However, if increasing the sparsity of semantic representations creates more responsive individual, then worst-case robustness should be inversely proportional to class selectivity.

Unlike average-case perturbations, decreasing class selectivity *decreases* robustness to worst-case perturbations for both ResNet18 trained on Tiny ImageNet and ResNet20 trained on CIFAR10 (Figures 2 and A9). For small perturbations, the effects of class selectivity regularization on test accuracy appear to overwhelm the effects of perturbations. But as the magnitude of perturbation increases, a stark ordering emerges: test accuracy monotonically decreases as a function of class selectivity in ResNet18 for both FGSM (Figure 2a) and PGD attacks (Figure 2b). The ordering is also present for ResNet20, though less consistent for the two networks with the highest class selectivity ($\alpha = 0.7$ and $\alpha = 1.0$). However, increasing class selectivity is much more damaging to test accuracy in ResNet20 trained on CIFAR10 compared to ResNet18 trained on Tiny ImageNet (Leavitt and Morcos, 2020, Figure A2), so the the substantial performance deficits of extreme selectivity in ResNet20 likely mask the perturbation-robustness. This result demonstrates that networks with sparse semantic representations are less vulnerable to worst-case perturbation than networks with distributed semantic representations. We also verified that the worst-case robustness of high-selectivity networks is not fully explained by gradient-masking (Athalye et al., 2018, Appendix A.4).

Networks whose outputs are more stable to small input perturbations are known to have improved generalization performance and worst-case perturbation robustness (Drucker and Le Cun, 1992; Novak et al., 2018; Sokolic et al., 2017; Rifai et al., 2011; Hoffman et al., 2019). To examine whether increasing class selectivity improves worst-case perturbation robustness by increasing network stability, we analyzed each network's input-output Jacobian, which is proportional to its stability—a large-magnitude Jacobian means that a small change to the network's input will cause a large change to its output. If class selectivity induces worst-case robustness by increasing network stability, then networks with higher class selectivity should have smaller Jacobians. But if increased class selectivity induces adversarial robustness through alternative mechanisms, then class selectivity should have no effect on the Jacobian. We found that the l_2 norm of the input-output Jacobian is inversely proportional to class selectivity for both ResNet18 (Figure 2c) and ResNet20 (Figure A9c), indicating that distributed semantic representations are more vulnerable to worst-case perturbation because they are less stable than sparse semantic representations.

4.3 Variability of the input-unit gradient across samples and units

We observed that the input-output Jacobian is proportional to worst-case vulnerability and inversely proportional to class selectivity, but focusing on input-output stability potentially overlooks phenomena present in hidden layers and units. If class selectivity imparts worst-case robustness by making individual units less reliably perturbable—because each unit is highly tuned to a particular subset of images—then we should expect to see more variation across input-unit gradients for units in high-selectivity networks compared to units in low-selectivity networks. Alternatively, worst-case ro-

bustness in high-selectivity networks could be achieved by reducing both the magnitude and variation of units' perturbability, in which case we would expect to observe lower variation across input-unit gradients for units in high-selectivity networks compared to low-selectivity networks.

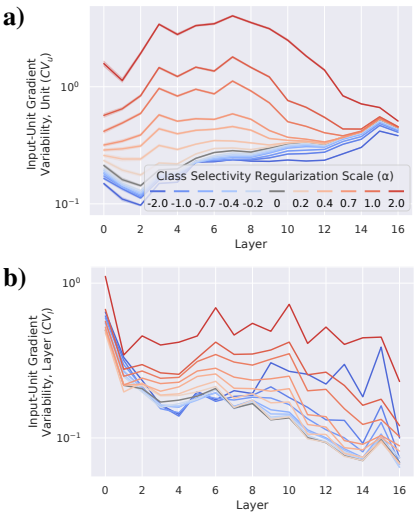

Figure 3: Class selectivity causes higher variation in perturbability within and across units (**a**) Coefficient of variation of input-unit gradient for each unit (CV_u; see Approach 3; y-axis) as a function of layer (x-axis). (**b**) CV of input-unit gradient across units in a layer (CV_l; y-axis) as a function of layer (x-axis). Results shown are for ResNet18 trained on Tiny ImageNet. Shaded regions = 95% confidence intervals of the mean. See Appendix A.5 for results for ResNet20 trained on CIFAR10.

We quantified variation in unit perturbability using the coefficient of variation of the input-unit gradient across samples for each unit (CV_u; Approach 3). The CV is a measure of variability that normalizes the standard deviation of a quantity by the mean. A large CV indicates high variability, a small CV indicates low variability. To quantify variation in perturbability *across* units, we computed the CV across units in each layer, (CV_l; Approach 3).

We found that units in high-selectivity networks exhibited greater variation in their perturbability than units in low-selectivity networks, both within individual units and across units in each layer. This effect was present in both ResNet18 trained on Tiny ImageNet (Figure 3) and ResNet20 trained on CIFAR10 (Figure A12), although the effect was less consistent for across-unit variability in later layers in ResNet18 (Figure 3b). Interestingly, class selectivity affects both the numerator (σ) and denominator (μ) of the CV calculation for both the CV across samples and CV across units (Appendix A.5). These results indicate that that high class selectivity imparts worst-case robustness by increasing the variation in perturbability within and across units, while the worst-case vulnerability associated with low class selectivity results from more consistently perturbable units. It is worth noting that the inverse can be stated with regards to average-case robustness: low variation in perturbability both within and across units in low-selectivity networks is associated with robustness to average-case perturbations, despite the these units (and networks) being more perturbable on average.

4.4 Dimensionality in early layers predicts perturbation vulnerability

Prior research has elucidated the mechanisms of worst-case perturbations using the framework of dimensionality (Langeberg et al., 2019; Nayebi and Ganguli, 2017; Sanyal et al., 2020), hence investigating the dimensionality of the changes induced by average-case and worst-case perturbations could reveal a common factor linking them to each other and to class selectivity.

One possible explanation for the discrepancy between class selectivity's impact on worst-case and average-case corruptions is that different corruption types impact representations with varying dimensionalities. For example, if only a few neurons are needed to change a network's decision, the dimensionality of the change in the representation due to perturbation might be very low, as only a few units need to be modified. We thus measured dimensionality using a straightforward, linear method: we applied Principal Component Analysis (PCA) to the activation matrices of each layer in our networks and computed the number of components necessary to explain 95% of the variance (and replicated our results with different variance thresholds; Appendix A.1.4; Appendix A.6). We first examined the dimensionality of the representations of the clean test data. If the sparsity of semantic representations is reflected in dimensionality, then networks with more class selectivity should have lower-dimensional representations than networks with less class selectivity. Alternatively, if high-selectivity representations are of similar dimensionality to low-selectivity representations—though perhaps occupying different sub-spaces—then dimensionality would be unaffected by class selectivity.

We found that the sparsity of a DNN's semantic representations corresponds directly to the dimensionality of those representations. Dimensionality is inversely proportional to class selectivity in early ResNet18 layers ($\leq$layer 9; Figure 4a), and across all of ResNet20 (Figure A15d). Networks with

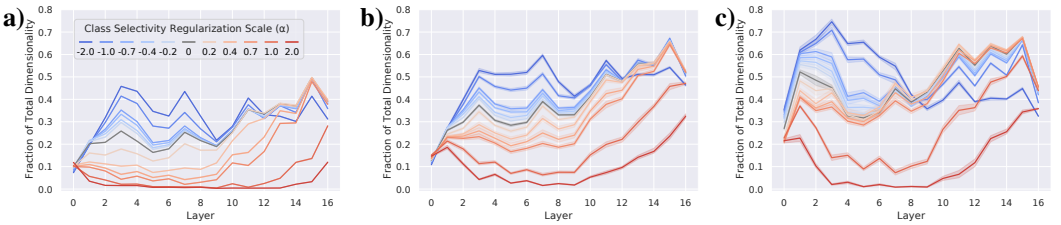

Figure 4: Dimensionality in early layers predicts worst-case vulnerability. (**a**) Fraction of dimensionality (y-axis; see Appendix A.1.4) as a function of layer (x-axis). (**b**) Dimensionality of difference between clean and average-case perturbation activations (y-axis) as a function of layer (x-axis). (**c**) Dimensionality of difference between clean and worst-case perturbation activations (y-axis) as a function of layer (x-axis). Results shown are for ResNet18 trained on Tiny ImageNet. See Appendix A15 for ResNet20.

higher class selectivity tend to have lower dimensionality, and networks with lower class selectivity tend to have higher dimensionality. These results show that the sparsity of a network's semantic representations is indeed reflected in those representations' dimensionality.

We next examined the dimensionality of perturbation-induced changes in representations by subtracting the perturbed activation matrix from the clean activation matrix and computing the dimensionality of this "difference matrix" (see Appendix A.1.4). Intuitively, this metric quantifies the dimensionality of the change in the representation caused by perturbing the input. If it is small, the perturbation impacts fewer units, while if it is large, more units are impacted. Interestingly, we found that the dimensionality of the changes in activations induced by both average-case (Figure 4b) and worst-case perturbations (Figure 4c) was notably higher for networks with reduced class-selectivity, suggesting that decreasing class selectivity causes changes in input to become more distributed.

We found that the activation changes caused by average-case perturbations are higher-dimensional than the representations of the clean data in both ResNet18 (compare Figures 4b and 4a) and ResNet20 (Figures A15e and A15d), and that this effect is inversely proportional to class selectivity (Figures 4b and A15e); the increase in dimensionality from average-case perturbations was more pronounced in low-selectivity networks than in high-selectivity networks. These results indicate that class selectivity not only predicts the dimensionality of a representation, but also the change in dimensionality induced by an average-case perturbation.

Notably, however, the increase in early-layer dimensionality was much larger for worst-case perturbations than average-case perturbations (Figure 4c; Figure A15f) . These results indicate that, while the changes in dimensionality induced by both naturalistic and adversarial perturbations are proportional to the dimensionality of the network's representations, these changes do not consistently project onto coding-relevant dimensions of the representations. Indeed, the larger change in early-layer dimensionality caused by worst-case perturbations likely reflects targeted projection onto coding-relevant dimensions and provides intuition as to why low-selectivity networks are more susceptible to worst-case perturbations.

5 Discussion

Our results demonstrate that changes in the sparsity of semantic representations, as measured with class selectivity, induce a trade-off between robustness to average-case vs. worst-case perturbations: highly-distributed semantic representations confer robustness to average-case perturbations, but their increased dimensionality and consistent perturbability result in vulnerability to worst-case perturbations. In contrast, sparse semantic representations yield low-dimensional representations and inconsistently-perturbable units, imparting worst-case robustness. Furthermore, the dimensionality of the difference in early-layer activations between clean and perturbed samples is larger for worst-case perturbations than for average-case perturbations. More generally, our results link average-case and worst-case perturbation robustness through class selectivity and representational dimensionality.

We hesitate to generalize too broadly about our findings, as they are limited to CNNs trained on image classification tasks. It is possible that the results we report here are specific to our models and/or datasets, and also may not extend to other tasks. And while class selectivity regularization appears effective at imparting average-case perturbation robustness, we hope future work will more comprehensively assess the viability of its adoption into the deep learning toolkit for this purpose.

REFERENCES

Rana Ali Amjad, Kairen Liu, and Bernhard C. Geiger. Understanding Individual Neuron Importance Using Information Theory. April 2018. URL https://arxiv.org/abs/1804.06679v3.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In International Conference on Machine Learning, pages 274–283. PMLR, July 2018. URL http://proceedings.mlr.press/v80/athalye18a.html. ISSN: 2640-3498.

Emilio Balda, Niklas Koep, Arash Behboodi, and Rudolf Mathar. Adversarial Risk Bounds through Sparsity based Compression. In International Conference on Artificial Intelligence and Statistics, pages 3816–3825. PMLR, June 2020. URL http://proceedings.mlr.press/v108/balda20a.html. ISSN: 2640-3498.

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences, September 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1907375117. URL https://www.pnas.org/content/early/2020/08/31/1907375117. Publisher: National Academy of Sciences Section: Physical Sciences.

Nicholas Carlini and David Wagner. Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec '17, pages 3–14, Dallas, Texas, USA, November 2017a. Association for Computing Machinery. ISBN 978-1-4503-5202-4. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57, May 2017b. doi: 10.1109/SP.2017.49. ISSN: 2375-1207.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What Is One Grain of Sand in the Desert? Analyzing Individual Neurons in Deep NLP Models. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):6309–6317, July 2019a. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33016309. URL https://aaai.org/ojs/index.php/AAAI/article/view/4592.

Fahim Dalvi, Avery Nortonsmith, Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, and James Glass. NeuroX: A Toolkit for Analyzing Individual Neurons in Neural Networks. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):9851–9852, July 2019b. ISSN 2374-3468. doi: 10.1609/aaai.v33i01.33019851. URL https://www.aaai.org/ojs/index.php/AAAI/article/view/5063.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. How Important is a Neuron. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=SylKoo0cKm.

Guneet S. Dhillon, Kamyar Azizzadenesheli, Zachary C. Lipton, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, and Animashree Anandkumar. Stochastic Activation Pruning for Robust Adversarial Defense. February 2018. URL https://openreview.net/forum?id=H1uR4GZRZ.

Jonathan Donnelly and Adam Roegiest. On Interpretability and Feature Representations: An Analysis of the Sentiment Neuron. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, Advances in Information Retrieval, Lecture Notes in Computer Science, pages 795–802, Cham, 2019. Springer International Publishing. ISBN 978-3-030-15712-8. doi: 10.1007/978-3-030-15712-8_55.

Harris Drucker and Yann Le Cun. Improving generalization performance using double backpropagation. IEEE Transactions on Neural Networks, 3(6):991–997, 1992.

Dumitru Erhan, Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Visualizing Higher-Layer Features of a Deep Network. 2009.

Brian Everitt. The Cambridge Dictionary of Statistics, volume 106. Cambridge University Press, Cambridge, 2002.

Li Fei-Fei, Andrej Karpathy, and Justin Johnson. Tiny imagenet visual recognition challenge, 2015. URL https://tiny-imagenet.herokuapp.com/.

Nicolas Ford, Justin Gilmer, Nicholas Carlini, and Ekin Cubuk. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In International Conference on Machine Learning, pages 2280–2289. PMLR, May 2019. URL http://proceedings.mlr.press/v97/gilmer19a.html. ISSN: 2640-3498.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. September 2018. URL https://openreview.net/forum?id=Bygh9j09KX.

Justin Gilmer, Ryan P. Adams, Ian Goodfellow, David Andersen, and George E. Dahl. Motivating the Rules of the Game for Adversarial Example Research. July 2018. URL https://arxiv.org/abs/1807.06732v2.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [cs, stat], March 2015. URL http://arxiv.org/abs/1412.6572. arXiv: 1412.6572.

Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse DNNs with Improved Adversarial Robustness. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 242–251. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7308-sparse-dnns-with-improved-adversarial-robustness.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations. arXiv:1807.01697 [cs, stat], April 2019. URL http://arxiv.org/abs/1807.01697. arXiv: 1807.01697.

Judy Hoffman, Daniel A. Roberts, and Sho Yaida. Robust Learning with Jacobian Regularization. arXiv:1908.02729 [cs, stat], August 2019. URL http://arxiv.org/abs/1908.02729. arXiv: 1908.02729.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvari. Learning with a Strong Adversary. arXiv:1511.03034 [cs], January 2016. URL http://arxiv.org/abs/1511.03034. arXiv: 1511.03034.

Yerlan Idelbayev. akamaster/pytorch_resnet_cifar10, January 2020. URL https://github.com/akamaster/pytorch_resnet_cifar10. original-date: 2018-01-15T09:50:56Z.

E R Kandel, J H Schwartz, and Jessica Chao. Principles of neural science. McGraw-Hill, New York, 2000.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and Understanding Recurrent Networks. In International Conference on Learning Representations, page 11, 2016.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. November 2016. URL https://openreview.net/forum?id=S1OufnIlx.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL https://openreview.net/forum?id=BJm4T4Kgx.

Peter Langeberg, Emilio Rafael Balda, Arash Behboodi, and Rudolf Mathar. On the Effect of Low-Rank Weights on Adversarial Robustness of Neural Networks. arXiv:1901.10371 [cs, stat], January 2019. URL http://arxiv.org/abs/1901.10371. arXiv: 1901.10371.

Matthew L. Leavitt and Ari Morcos. Selectivity considered harmful: evaluating the causal impact of class selectivity in DNNs. arXiv:2003.01262 [cs, q-bio, stat], March 2020. URL http://arxiv.org/abs/2003.01262. arXiv: 2003.01262.

Peter E. Lillian, Richard Meyes, and Tobias Meisen. Ablation of a Robot's Brain: Neural Networks Under a Knife. December 2018. URL https://arxiv.org/abs/1812.05687v2.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. February 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation Studies in Artificial Neural Networks. arXiv:1901.08644 [cs, q-bio], February 2019. URL http://arxiv.org/abs/1901.08644. arXiv: 1901.08644.

Ari S. Morcos, David G. T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=r1iuQjxCZ.

Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. Discovery of Natural Language Concepts in Individual Units of CNNs. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=S1EERs09YQ.

Aran Nayebi and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv:1703.09202 [cs, q-bio, stat], March 2017. URL http://arxiv.org/abs/1703.09202. arXiv: 1703.09202.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and Generalization in Neural Networks: an Empirical Study. arXiv:1802.08760 [cs, stat], June 2018. URL http://arxiv.org/abs/1802.08760. arXiv: 1802.08760.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. Distill, 2(11):e7, November 2017. ISSN 2476-0757. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. Distill, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to Generate Reviews and Discovering Sentiment. arXiv:1704.01444 [cs], April 2017. URL http://arxiv.org/abs/1704.01444. arXiv: 1704.01444.

Ivet Rafegas, Maria Vanrell, Luis A. Alexandre, and Guillem Arias. Understanding trained CNNs by indexing neuron selectivity. Pattern Recognition Letters, page S0167865519302909, October 2019. ISSN 01678655. doi: 10.1016/j.patrec.2019.10.013. URL http://arxiv.org/abs/1702.00382. arXiv: 1702.00382.

Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: explicit invariance during feature extraction. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML'11, pages 833–840, Bellevue, Washington, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5.

Amartya Sanyal, Varun Kanade, Philip H. S. Torr, and Puneet K. Dokania. Robustness via Deep Low-Rank Representations. arXiv:1804.07090 [cs, stat], February 2020. URL http://arxiv.org/abs/1804.07090. arXiv: 1804.07090.

Alexandru Constantin Serban, Erik Poll, and Joost Visser. Adversarial Examples - A Complete Characterisation of the Phenomenon. arXiv:1810.01185 [cs], February 2019. URL http://arxiv.org/abs/1810.01185. arXiv: 1810.01185.

Charles S. Sherrington. The integrative action of the nervous system. The integrative action of the nervous system. Yale University Press, New Haven, CT, US, 1906. doi: 10.1037/13798-000.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust Large Margin Deep Neural Networks. IEEE Transactions on Signal Processing, 65(16):4265–4280, August 2017. ISSN 1053-587X, 1941-0476. doi: 10.1109/TSP.2017.2708039. URL http://arxiv.org/abs/1605.08254. arXiv: 1605.08254.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. December 2013. URL https://arxiv.org/abs/1312.6199v4.

Yusuke Tsuzuku and Issei Sato. On the Structural Sensitivity of Deep Convolutional Networks to the Directions of Fourier Basis Functions. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 51–60, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00014. URL https://ieeexplore.ieee.org/document/8954086/.

Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. Examining the Impact of Blur on Recognition by Convolutional Networks. November 2016. URL https://arxiv.org/abs/1611.05760v2.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1 0 Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. arXiv:1907.10121 [physics], July 2019. URL http://arxiv.org/abs/1907.10121. arXiv: 1907.10121.

David Warde-Farley and Ian Goodfellow. Adversarial Perturbations of Deep Neural Networks. In Perturbations, Optimization, and Statistics, pages 311–342. MITP, 2017. ISBN 978-0-262-33793-9. URL https://ieeexplore.ieee.org/document/8093865. Conference Name: Perturbations, Optimization, and Statistics.

Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C. Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (September 2017), September 2017. URL https://doi.org/10.5281/zenodo.883859.

Shaokai Ye, Siyue Wang, Xiao Wang, Bo Yuan, Wujie Wen, and Xue Lin. Defending DNN Adversarial Attacks with Pruning and Logits Augmentation. February 2018. URL https://openreview.net/forum?id=S1qI2FJDM.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pages 13276–13286. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9483-a-fourier-perspective-on-model-robustness-in-computer-vision.pdf.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the Robustness of Deep Neural Networks via Stability Training. April 2016. URL https://arxiv.org/abs/1604.04326v1.

B. Zhou, D. Bau, A. Oliva, and A. Torralba. Interpreting Deep Visual Representations via Network Dissection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9):2131–2145, September 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2858759.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object Detectors Emerge in Deep Scene CNNs. In International Conference on Learning Representations, April 2015. URL http://arxiv.org/abs/1412.6856. arXiv: 1412.6856.

Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Revisiting the Importance of Individual Units in CNNs via Ablation. arXiv:1806.02891 [cs], June 2018. URL http://arxiv.org/abs/1806.02891. arXiv: 1806.02891.

A APPENDIX

A.1 DETAILED APPROACH

Unless otherwise noted: all experimental results were derived from the corrupted or adversarial test set with the parameters from the epoch that achieved the highest clean validation set accuracy over the training epochs; 20 replicates with different random seeds were run for each hyperparameter set; error bars and shaded regions denote bootstrapped 95% confidence intervals; selectivity regularization was not applied to the final (output) layer, nor was the final layer included in any of our analyses.

A.1.1 MODELS

All models were trained using stochastic gradient descent (SGD) with momentum = 0.9 and weight decay = 0.0001. The maxpool layer after the first batchnorm layer in ResNet18 (see He et al. (2016)) was removed because of the smaller size of Tiny Imagenet images compared to standard ImageNet images (64x64 vs. 256x256, respectively). ResNet18 were trained for 90 epochs with a minibatch size of 4096 samples with a learning rate of 0.1, multiplied (annealed) by 0.1 at epochs 35, 50, 65, and 80. .

ResNet20 (code modified from Idelbayev (2020)) were trained for 200 epochs using a minibatch size of 256 samples and a learning rate of 0.1, annealed by 0.1 at epochs 100 and 150.

A.1.2 DATASETS

Tiny Imagenet (Fei-Fei et al., 2015) consists of 500 training images and 50 images for each of its 200 classes. We used the validation set for testing and created a new validation set by taking 50 images per class from the training set, selected randomly for each seed. We split the 50k CIFAR10 training samples into a 45k sample training set and a 5k validation set, similar to our approach with Tiny Imagenet.

All experimental results were derived from the test set with the parameters from the epoch that achieved the highest validation set accuracy over the training epochs. 20 replicates with different random seeds were run for each hyperparameter set. Selectivity regularization was not applied to the final (output) layer, nor was the final layer included any of our analyses.

CIFAR10C consists of a dataset in which 19 different naturalistic corruptions have been applied to the CIFAR10 test set at 5 different levels of severity. Tiny ImageNetC also has 5 levels of corruption severity, but consists of 15 corruptions.

We would like to note that Tiny ImageNetC does not use the Tiny ImageNet test data. While the two datasets were created using the same data generation procedure—cropping and scaling images from the same 200 ImageNet classes—they differ in the specific ImageNet images they use. It is possible that the images used to create Tiny ImageNetC are out-of-distribution with regards to the Tiny ImageNet training data, in which case our results from testing on Tiny ImageNetC actually underestimate the corruption robustness of our networks. The creators of Tiny ImageNetC kindly provided the clean (uncorrupted) Tiny ImageNetC data necessary for the dimensionality analysis, which relies on matches corrupted and clean data samples.

A.1.3 SOFTWARE

Experiments were conducted using PyTorch (Paszke et al., 2019), analyzed using the SciPy ecosystem (Virtanen et al., 2019), and visualized using Seaborn (Waskom et al., 2017).

A.1.4 QUANTIFYING DIMENSIONALITY

We quantified the dimensionality of a layer's representations by applying PCA to the layer's activation matrix for the clean test data and counting the number of dimensions necessary to explain 95% of the variance, then dividing by the total number of dimensions (i.e. the fraction of total dimensionality; we also replicated our results using the fraction of total dimensionality necessary to explain 90% and 99% of the variance). The same procedure was applied to compute the dimensionality of perturbation-induced changes in representations, except the activations for a perturbed data set were subtracted

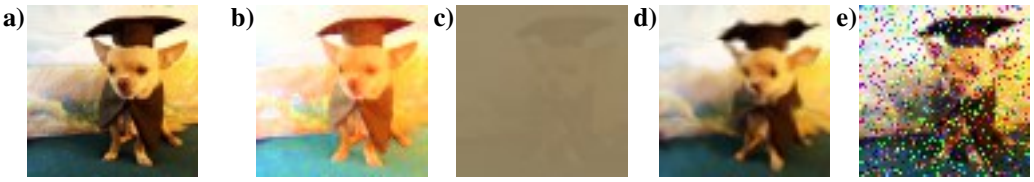

Figure A1: Example naturalistic corruptions from the Tiny ImageNetC dataset. (**a**) Clean (no corruption). (**b**) Brightness. (**c**) Contrast. (**d**) Elastic transform. (**e**) Shot noise. All corruptions are shown at severity level 5/5.

from the corresponding clean activations prior to applying PCA. For average-case perturbations, we performed this analysis for every corruption type and severity, and for the worst-case perturbations we used PGD with 40 steps.

A.2 Effects of class selectivity regularization on test accuracy

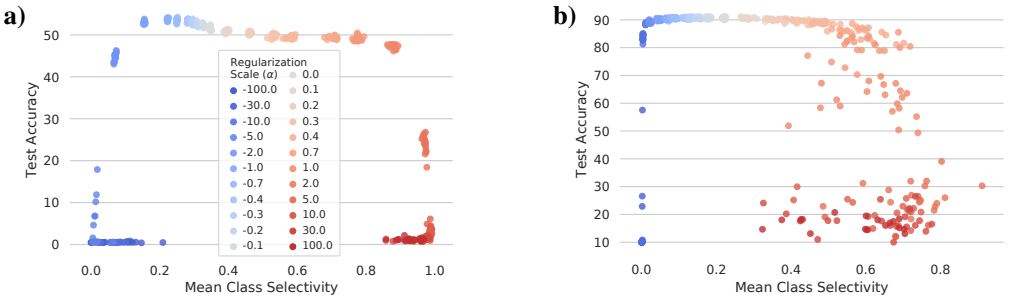

Figure A2: Effects of class selectivity regularization on test accuracy. Replicated as in Leavitt and Morcos (2020). (**a**) Test accuracy (y-axis) as a function of mean class selectivity (x-axis) for ResNet18 trained on Tiny ImageNet. α denotes the sign and intensity of class selectivity regularization. Negative α lowers selectivity, positive α increases selectivity, and the magnitude of α changes the strength of the effect. Each data point represents the mean class selectivity across all units in a single trained model. (**b**) Same as (**a**), but for ResNet20 trained on CIFAR10.

A.3 Additional results for average-case perturbation robustness

Because modifying class selectivity can affect performance on clean (unperturbed) inputs (Leavitt and Morcos (2020); Figure A2), it is possible that the effects we observe of class selectivity on perturbed test accuracy are not caused by changes in perturbation robustness per se, but simply by changes in baseline model accuracy. We controlled for this by normalizing each model's perturbed test accuracy by its clean (unperturbed) test accuracy. The results are generally consistent even after controlling for clean test accuracy, although increasing class selectivity does not cause the same deficits in as measured using non-normalized perturbed test accuracy in ResNet18 trained on Tiny ImageNet (Figure A3a. Interestingly, in ResNet20 trained on CIFAR10, normalizing perturbed test accuracy reveals a more dramatic improvement in perturbation robustness caused by reducing class selectivity (Figure A6c).

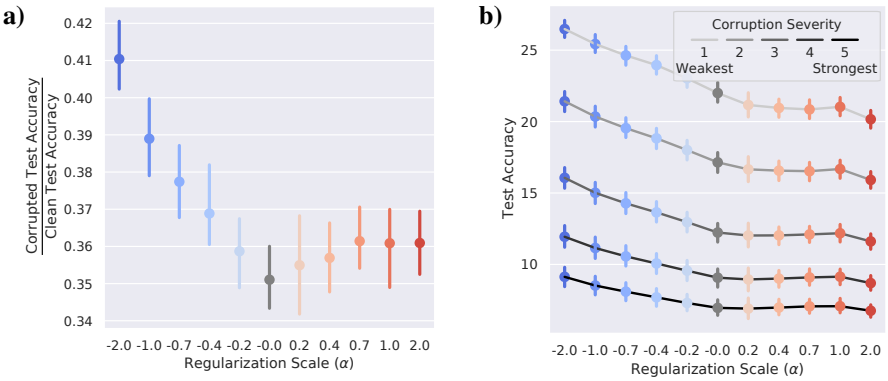

Figure A3: Controlling for clean test accuracy, and effect of corruption severity across corruptions. (**a**) Corrupted test accuracy normalized by clean test accuracy (y-axis) as a function of class selectivity regularization scale (α; x-axis). Negative α lowers selectivity, positive α increases selectivity, and the magnitude of α changes the strength of the effect. Normalized perturbed test accuracy appears higher in networks with high class selectivity (large α), but this is likely due to a floor effect: clean test accuracy is already much closer to the lower bound—chance—in networks with very high class selectivity, which may reflect a different performance regime, making direct comparison difficult. (**b**) Mean test accuracy across all corruptions (y-axis) as a function of α (x-axis) for different corruption severities (ordering along y-axis; shade of connecting line). Error bars indicate 95% confidence intervals of the mean.

Figure A4: Mean test accuracy across corruption intensities for each corruption type for ResNet18 tested on Tiny ImageNetC. Test accuracy (y-axis) as a function of corruption type (x-axis) and class selectivity regularization scale (α, color). Reducing class selectivity improves robustness against all 15/15 corruption types.

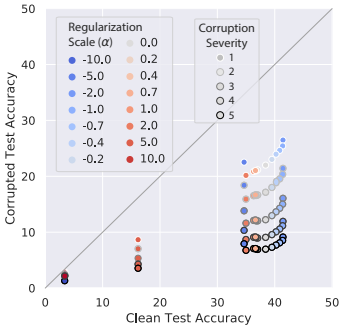

Figure A5: Trade-off between clean and perturbed test accuracy in ResNet18 tested on Tiny ImageNetC. Clean test accuracy (x-axis) vs. perturbed test accuracy (y-axis) for different corruption severities (border color) and regularization scales (α, fill color). Mean is computed across all corruption types. Error bars = 95% confidence intervals of the mean.

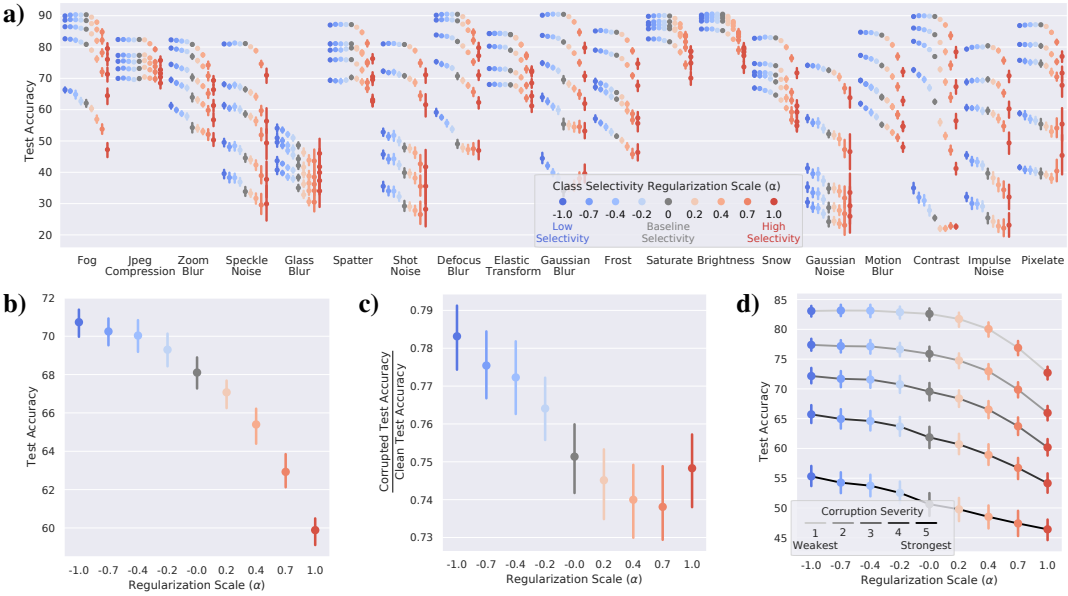

Figure A6: Reducing class selectivity confers robustness to average-case perturbations in ResNet20 trained on CIFAR10. (**a**) Test accuracy (y-axis) as a function of corruption type (x-axis), class selectivity regularization scale (α; color), and corruption severity (ordering along y-axis). Test accuracy is reduced proportionally to corruption severity, leading to an ordering along the y-axis, with corruption severity 1 (least severe) at the top and corruption severity 5 (most severe) at the bottom. Negative α lowers selectivity, positive α increases selectivity, and the magnitude of α changes the strength of the effect (see Figure **??** and Approach 3). (**b**) Mean test accuracy across all corruptions and severities (y-axis) as a function of α (x-axis). (**c**) Corrupted test accuracy normalized by clean test accuracy (y-axis) as a function of class selectivity regularization scale (α; x-axis). Negative α lowers selectivity, positive α increases selectivity, and the magnitude of α changes the strength of the effect. Normalized perturbed test accuracy appears higher in networks with higher class selectivity (larger α), but this is likely due to a floor effect: clean test accuracy is already much closer to the lower bound—chance—in networks with very high class selectivity, which may reflect a different performance regime, making direct comparison difficult. (**d**) Mean test accuracy across all corruptions (y-axis) as a function of α (x-axis) for different corruption severities (ordering along y-axis; shade of connecting line). Error bars indicate 95% confidence intervals of the mean.

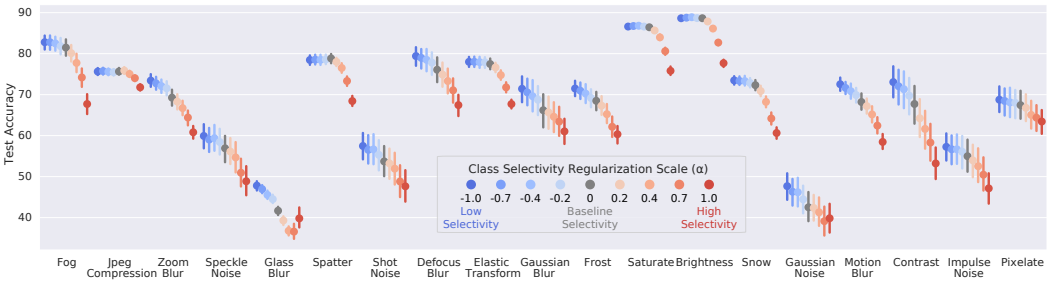

Figure A7: Mean test accuracy across corruption intensities for each corruption type for ResNet20 tested on CIFAR10C. Test accuracy (y-axis) as a function of corruption type (x-axis) and class selectivity regularization scale (α, color). Reducing class selectivity improves robustness against 14/19 corruption types. Error bars = 95% confidence intervals of the mean.

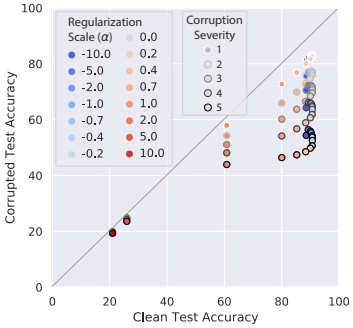

Figure A8: Trade-off between clean and corrupted test accuracy in ResNet20 tested on CIFAR10C. Clean test accuracy (x-axis) vs. corrupted test accuracy (y-axis) for different corruption severities (border color) and regularization scales (α, fill color). Mean is computed across all corruption types.

A.4 WORST-CASE PERTURBATION ROBUSTNESS

We also confirmed that the worst-case robustness of high-selectivity networks was not simply due to gradient-masking (Athalye et al., 2018) by generating worst-case perturbations using each of the replicate models trained with no selectivity regularization ($\alpha = 0$), then testing selectivity-regularized models on these samples. We found that high-selectivity models were less vulnerable to the $\alpha = 0$ samples than low-selectivity models for high-intensity perturbations (Appendix A10, indicating that gradient-masking does not fully account for the worst-case robustness of high-selectivity models.

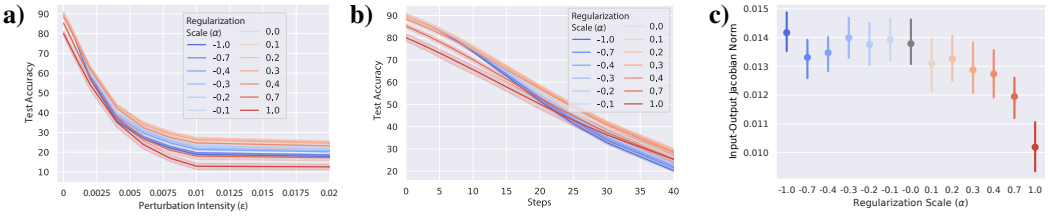

Figure A9: Reducing class selectivity increases worst-case perturbation vulnerability in ResNet20 trained on CIFAR10. (**a**) Test accuracy (y-axis) as a function of perturbation intensity (ϵ; x-axis) and class selectivity regularization scale (α; color) for the FGSM attack. (**b**) Test accuracy (y-axis) as a function of adversarial optimization steps (x-axis) and α (color) for the PGD attack. (**c**) Network stability, as measured with norm of the input-output Jacobian (y-axis) as a function of α (x-axis).

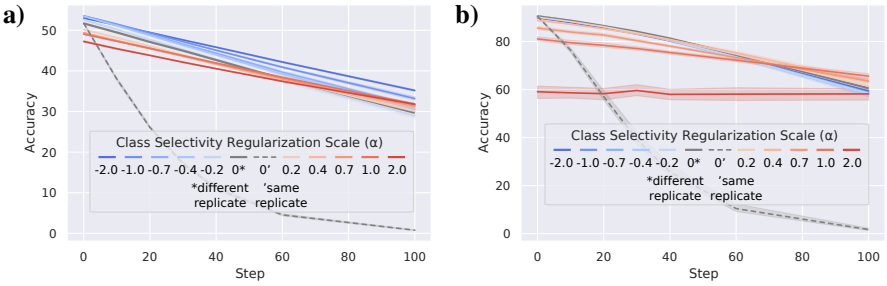

Figure A10: Gradient masking does not fully account for worst-case perturbation robustness conferred by increased class selectivity. (**a**) Test Accuracy (y-axis) as a function of adversarial optimization steps (x-axis) and class selectivity regularization scale (α; color) when tested on adversarial examples generated using $\alpha = 0$. Because 20 replicate networks were trained for each value of α (see Approach 3), models trained with $alpha = 0$ could be tested on adversarial examples generated from a different replicate $\alpha = 0$ network ("different replicate"; solid line) or adversarial samples generated from their own parameters ("same replicate"; dashed line). Data shown are for ResNet18 trained on Tiny ImageNet (**b**). Same as (**a**), but for ResNet20 trained on CIFAR10.

A.5 STABILITY TO INPUT PERTURBATIONS IN UNITS AND LAYERS

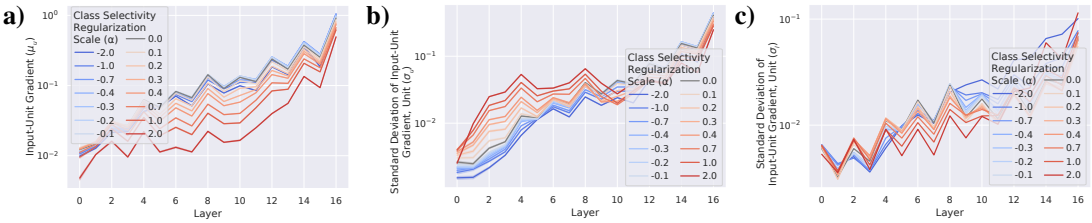

Figure A11: Mean and standard deviation of input-unit gradient in ResNet18 trained on Tiny ImageNet. (**a**) Input-unit gradient (y-axis) as a function of layer (x-axis). (**b**) Standard deviation of input-unit gradient for each unit (σ_u; y-axis) as a function of layer (x-axis). (**c**) Standard deviation of input-unit gradient across units in a layer (σ_l; y-axis) as a function of layer (x-axis). Shaded region = 95% confidence interval of the mean.

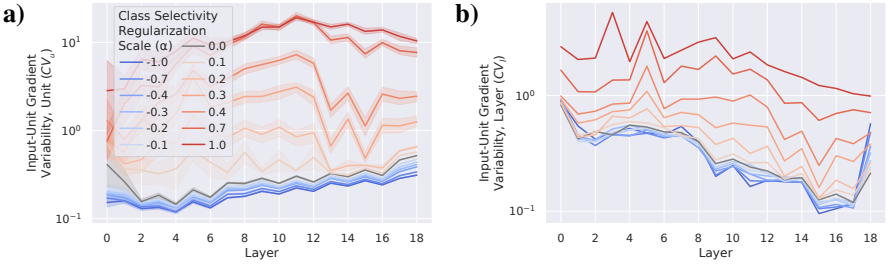

Figure A12: Class selectivity causes higher variation in perturbability within and across units in ResNet20 trained on CIFAR10. (**a**) Coefficient of variation of input-unit gradient for each unit (CV_u; see Approach 3; y-axis) as a function of layer (x-axis). (**b**) CV of input-unit gradient across units in a layer (CV_l; y-axis) as a function of layer (x-axis). Shaded regions = 95% confidence interval of the mean.

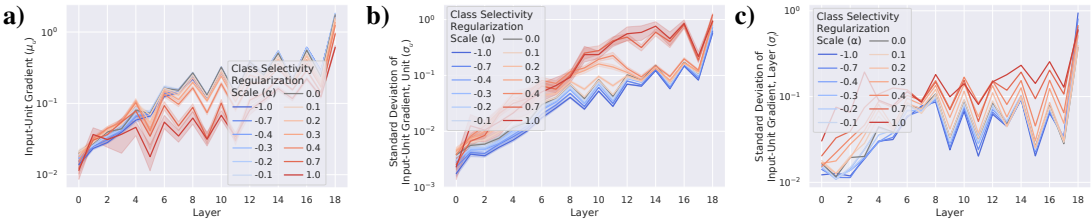

Figure A13: Mean and standard deviation of input-unit gradient in ResNet20 trained on CIFAR10. (**a**) Input-unit gradient (y-axis) as a function of layer (x-axis). (**b**) Standard deviation of input-unit gradient for each unit (σ_u; y-axis) as a function of layer (x-axis). (**c**) Standard deviation of input-unit gradient across units in a layer (σ_l; y-axis) as a function of layer (x-axis). Shaded region = 95% confidence interval of the mean.

A.6 REPRESENTATIONAL DIMENSIONALITY

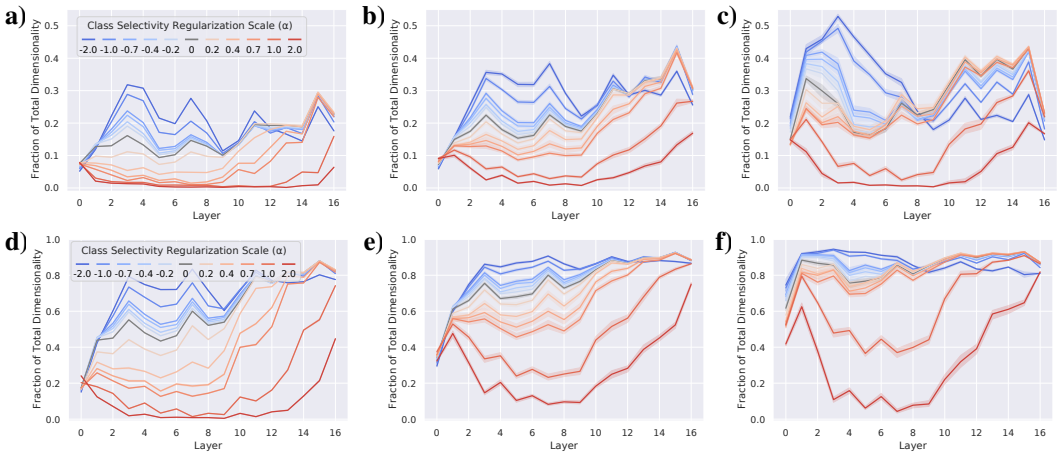

Figure A14: Dimensionality in early layers predicts worst-case vulnerability in ResNet18 trained on Tiny ImageNet. Identical to Figure 4, but dimensionality is computed as the number of principal components needed to explain 90% of variance in **(a)** - **(c)**, and 99% of variance in **(d)** - **(f)**. (**a**) Fraction of dimensionality (y-axis; see Appendix A.1.4) as a function of layer (x-axis). (**b**) Dimensionality of difference between clean and average-case perturbation activations (y-axis) as a function of layer (x-axis). (**c**) Dimensionality of difference between clean and worst-case perturbation activations (y-axis) as a function of layer (x-axis). **(d)** - **(f)**, identical to **(a)** - **(c)**, but for 99% explained variance threshold.

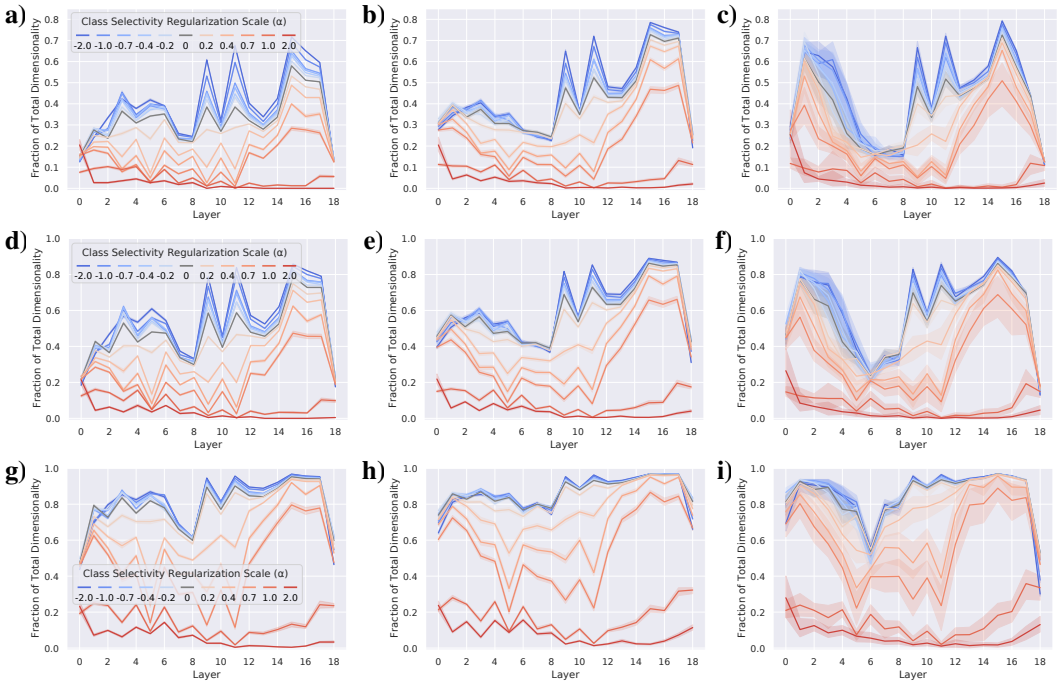

Figure A15: Dimensionality in early layers predicts worst-case vulnerability in ResNet20 trained on CIFAR10. Identical to Figure 4, but for ResNet20 trained on CIFAR10. (**a**) Fraction of dimensionality needed to explain 90% of variance (y-axis; see Appendix A.1.4) as a function of layer (x-axis). (**b**) Dimensionality of difference between clean and average-case perturbation activations (y-axis) as a function of layer (x-axis). (**c**) Dimensionality of difference between clean and worst-case perturbation activations (y-axis) as a function of layer (x-axis). **(d)** - **(f)**, identical to **(a)** - **(c)**, but for 95% explained variance threshold. **(g)** - **(i)**, identical to **(a)** - **(c)**, but for 99% explained variance threshold.