CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection

Hanshu Yan [1] **Jingfeng Zhang** [2] **Gang Niu** [2] **Jiashi Feng** [1] **Vincent Y. F. Tan** [1 3] **Masashi Sugiyama** [2 4]

Abstract

We investigate the adversarial robustness of CNNs from the perspective of channel-wise activations. By comparing normally trained and adversarially trained models, we observe that adversarial training (AT) robustifies CNNs by aligning the channel-wise activations of adversarial data with those of their natural counterparts. However, the channels that are *negatively-relevant* (NR) to predictions are still over-activated when processing adversarial data. Besides, we also observe that AT does not result in similar robustness for all classes. For the robust classes, channels with larger activation magnitudes are usually more *positively-relevant* (PR) to predictions, but this alignment does not hold for the non-robust classes. Given these observations, we hypothesize that suppressing NR channels and aligning PR ones with their relevances further enhances the robustness of CNNs under AT. To examine this hypothesis, we introduce a novel mechanism, *i.e.*, Channel-wise Importance-based Feature Selection (CIFS). The CIFS manipulates channels' activations of certain layers by generating non-negative multipliers to these channels based on their relevances to predictions. Extensive experiments on benchmark datasets including CIFAR10 and SVHN clearly verify the hypothesis and CIFS's effectiveness of robustifying CNNs.

1. Introduction

Convolutional neural networks (CNNs) have achieved tremendous successes in real-world applications, such as autonomous vehicles (Grigorescu et al., 2020; Hu et al., 2020) and computer-aided medical diagnoses (Trebeschi et al., 2017; Shu et al., 2020). However, CNNs have be shown vulnerable to well-crafted (and even minute) adversarial perturbations to inputs (Szegedy et al., 2014; Goodfellow et al., 2015; Ilyas et al., 2019). This has become hazardous in high-stakes applications such as medical diagnoses and autonomous vehicles.

Recently, many empirical defense methods have been proposed to secure CNNs against these adversarial perturbations, such as adversarial training (AT) (Madry et al., 2018), input/feature denoising (Xie et al., 2019; Du et al., 2020) and defensive distillation (Papernot et al., 2016). AT (Madry et al., 2018), which generates adversarial data on the fly for training CNNs, has emerged as one of the most successful methods. AT effectively robustifies CNNs but leads to a clear drop in the accuracies for natural data (Tsipras et al., 2019) and suffers from the problem of overfitting to adversarial data used for training (Rice et al., 2020; Zhang et al., 2021; Chen et al., 2021). To ameliorate these problems, researchers have proposed variants of AT, including TRADES (Zhang et al., 2019) and Friendly-Adversarial-Training (FAT) (Zhang et al., 2020). To further robustify CNNs under AT, many works attempt to propose novel defense mechanism to mitigate the effects of adversarial data on features (Xie et al., 2019; Du et al., 2020; Xu et al., 2019). For example, Xie et al. (2019) found that adversarial data result in abnormal activations in the feature maps and performed feature denoising to remove the adversarial effects. Most of these works improved robustness by identifying and suppressing abnormalities *at certain positions* across channels (commonly referred to as feature maps in CNNs), whereas the other direction, namely, the connection between robustness and irregular activations of certain *entire channels*, has received scant attention.

Since channels of CNNs' deeper layers are capable of extracting semantic characteristic features (Zeiler & Fergus, 2014), the process of making predictions usually relies heavily on aggregating information from various channels (Bach et al., 2015). As such, anomalous activations of certain channels may result in incorrect predictions. Thus, it is imperative to explore which channels are *entirely* irregularly

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore; [2]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan; [3]Department of Mathematics, National University of Singapore, Singapore; [4]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. Correspondence to: Jingfeng Zhang <jingfeng.Zhang@riken.jp>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

activated by adversarial data and which channels' activations benefit or degrade robustness. By utilizing this connection, we will be able to further enhance the robustness of CNNs via suppressing or promoting certain vulnerable or reliable channels respectively.

In this work, we attempt to build such a connection by comparing the channel-wise activations of non-robust (normally trained) and robustified (adversarially trained) CNNs. The *channel-wise activations* are defined as the average activation magnitudes of all features within channels (Bai et al., 2021). To identify what types of channels appear to be abnormal under attacks, we regard *channels' relevances* to prediction results (formally defined in Equation (1) as g^l) as the gradients of the corresponding logits *w.r.t* channel-wise activations. The channels, whose relevances to prediction results are positive or negative ($g^l_{[i]} > 0$ or $g^l_{[i]} \leq 0$), are called *positively-relevant* (PR) or *negatively-relevant* (NR) channels.

On the one hand, we observe that, AT robustifies CNNs by aligning adversarial data's channel-wise activations with those of natural data. However, we find that the NR channels of adversarially trained CNNs are still over-activated by adversarial data (see Figure 1c). Thus, we wonder: *If we suppress NR channels during AT to facilitate the alignment of channel's activations, will it benefit CNNs' robustness?* On the other hand, we find that adversarially trained classification models do not enjoy similar robustness across all the classes (see Figure 1c and 1d). For classes with relatively good robustness, channels' activations usually align well with their relevances, i.e., channels with larger activations are more PR to labels. Given this phenomenon, a natural question arises: *If we align channels' activations with their relevances during AT, will it improve the robustness of CNNs?* Regarding these two questions, we propose a unified hypothesis on robustness enhancement, denoted as $\mathcal{H}$: Suppressing NR channels and aligning channels' activations with their relevances to prediction results benefit the robustness of CNNs.

To examine this hypothesis, we propose a novel mechanism, called Channel-wise Importance-based Feature Selection (CIFS), which adjusts channels' activations with an *importance mask* generated from channels' relevances. For a certain layer, CIFS first takes as input the representation of a data point at this layer and makes a *raw prediction* for the data point by a *probe network*. The probe serves as the surrogate for the subsequent classifier (the composition of subsequent layers) in the backbone and is jointly trained with the backbone under supervision of true labels. Then, CIFS computes the gradients of the sum of the top-k logits *w.r.t.* the channels' activations. We can obtain the *relevance* of each channel to the top-k prediction results by accumulating the gradients within the channel. Finally, CIFS generates a mask of importance scores for each channel by mapping channels' relevances monotonically to non-negative values. Through extensive experiments, we answer the two questions in the affirmative and confirm hypothesis $\mathcal{H}$. Indeed, our results show that CIFS clearly enhances the adversarial robustness of CNNs.

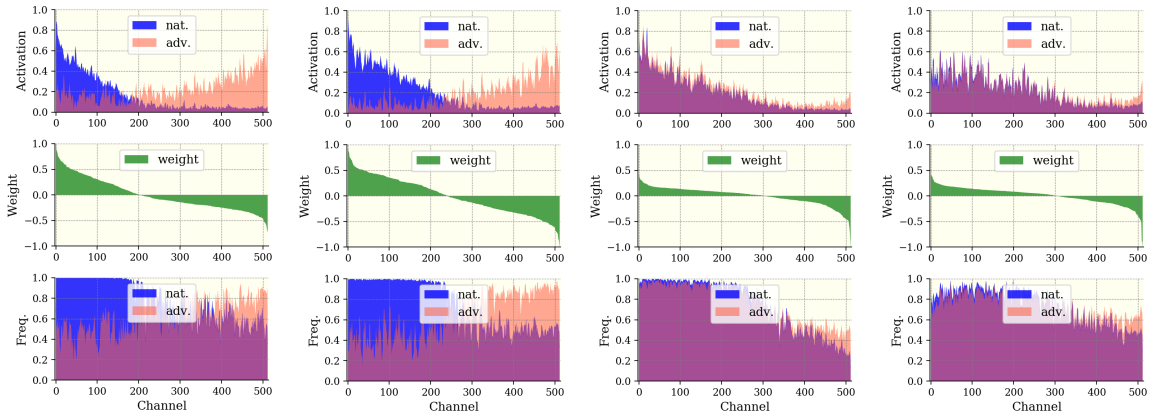
We comprehensively evaluate the robustness of CIFS-modified CNNs on benchmark datasets against various attacks. On the CIFAR10 dataset, CIFS improves the robustness of the ResNet-18 by 4 percentage points against the PGD-100 attack. We also observe that CIFS ameliorates the overfitting during AT. In particular, the robustness at the last epoch is close to that at the best epoch. Finally, we conduct an ablation study to further understand how various elements of CIFS affect the robustness enhancement, such as the top-k feedback and architectures of the probe network.

2. Related Works

This section briefly reviews relevant adversarial defense methods from two perspectives: adversarial training (AT)-based defense and robust network architecture design.

AT-based Defense Adversarial training (AT) defends against adversarial attacks by utilizing adversarially generated data in model training (Goodfellow et al., 2015), formulated as a minimax optimization problem. Recently variants of AT (Cai et al., 2018; Wang et al., 2019; 2020; Wu et al., 2020; Zhang et al., 2021) have been proposed. For example, the Misclassification-Aware-AdveRsarial-Training (Wang et al., 2020) modifies the process of generating adversarial data by simultaneously applying the misclassified natural data, together with the adversarial data for model training. Recent works have shown AT robustifies CNNs but degrades the natural accuracy (Tsipras et al., 2019; Zhang et al., 2019; Lamb et al., 2019). To achieve a better trade-off, Zhang et al. (2019) decomposed the adversarial prediction error into the natural error and boundary error and proposed TRADES to control both terms at the same time. Besides, inspired by curriculum learning (Cai et al., 2018; Bengio et al., 2009), Zhang et al. (2020) proposed FAT to train models with increasingly adversarial data, which enhances generalization without sacrificing robustness.

In addition, some works introduced various types of regularization for training models, such as layer-wise feature matching (Sankaranarayanan et al., 2018; Liao et al., 2018; Kannan et al., 2018), low-rank representations (Sanyal et al., 2020; Mustafa et al., 2019), attention map alignment (Xu et al., 2019), and Lipschitz regularity(Virmaux & Scaman, 2018; Cissé et al., 2017). These types of regularization can work in conjunction with AT and benefit the models' robustness.



(a) Normal / "automobile" (b) Normal / "cat" (c) Adv. / "automobile" (d) Adv. / "cat"

Figure 1. The magnitudes of channel-wise activations (top) at the penultimate layer, their activated frequencies (bottom), and the weights of the last linear layer (middle) *vs.* channel indices. #/$*$ means the CNN is trained in the # way and we plot the average activations of data from the class $*$. *The weights corresponding to class $*$ are sorted in the descending order and can indicate channels' relevances to class $*$*. The activation magnitudes and the activated frequencies of natural and adversarial (PGD-20) data are plotted according to the indices of the sorted weights. The robust accuracies on the whole CIFAR10 are 0% for the normally trained ResNet-18 model and 46.6% for the adversarially trained one (69.0% for the "automobile" class and 16.7% for the "cat" class).

Robust Network Design Other than robust training strategies, some works explored robust network architectures (Yan et al., 2020; Hsieh et al., 2019). For instance, the work by Yan et al. (2020) showed neural ODE-based models are inherently more robust than conventional CNN models; Guo et al. (2018) demonstrated that appropriately designed higher model sparsity implies better robustness of nonlinear networks. Another line of works defended against adversarial attacks via gradient obfuscation, such as random or non-differentiable image/feature transformations (Xie et al., 2018; Du et al., 2020; Dhillon et al., 2018; Xiao et al., 2020). However, they have been shown to be insecure to adaptive attacks (Athalye et al., 2018; Tramer et al., 2020). Recently, many researchers have attempted to develop novel mechanisms for robustness enhancement. Xie et al. (2019) performed feature denoising to remove the adversarial effects on feature maps. Zoran et al. (2020) utilized the spatial attention mechanism to identify highlight important regions of feature maps. Most of these works manipulated CNNs' intermediate representations in the *spatial domain*, whereas our work studies the adversarial robustness from the channel-wise activation perspective.

Channel-wise Activation Suppressing (CAS) (Bai et al., 2021), the most relevant work to ours, also studied the channel-wise activations of adversarial data. It showed channels are activated more uniformly by adversarial data compared to the natural ones, and AT improves the robustness by attempting to align the distributions of channels' activations of natural and adversarial data. However, there are still some channels that are over-activated by adversarial data. To suppress these channels, the authors proposed CAS to adjust channels' activations based on their importance. Although CAS empirically suppresses certain channels, the authors did not show that the suppressed channels correspond to the target ones; this means the primary objective of CAS may not have been met. Thus, there is no guarantee CAS can enhance the robustness of CNNs (see Section 4.1 for further evidence on this). *In contrast*, our work first builds a connection between robustness and channels' activations via their relevances to predictions. Then, the proposed CIFS can explicitly control channels' activations based on their relevances. Finally, experiments demonstrate the effectiveness of CIFS on robustness enhancement.

3. Channel-wise Importance-based Feature Selection

In this section, we first study the adversarial robustness by comparing channels' activations of non-robust (normally trained) and robustified (adversarially trained) CNNs. Based on our observations of AT's effects, we propose a hypothesis on robustness enhancement via the adjustment of channels' activations (Section 3.1). To examine this hypothesis, we then develop a novel mechanism, CIFS (Section 3.2), to manipulate channels' activation levels according to their relevances to predictions . Finally, we verify the proposed hypothesis through extensive experiments (Section 3.3).

3.1. Non-robust CNNs vs. Robustified CNNs: a Channel-wise Activation Perspective

We compare a non-robust ResNet-18 (He et al., 2016) model with an AT-robustified one on the CIFAR10 dataset (Krizhevsky, 2009). In ResNet-18, the representations of penultimate layer are spatially averaged for each channel, then fed into the last linear layer for making predictions.

Thus, the weights of the last linear layer indicate channels' relevances to predictions (according to the definition of channels' relevances in Introduction). We visualize the channel-wise activation magnitudes, the activated frequencies (counted via a threshold of 1% of the largest magnitude among all channels) in the penultimate layer for both natural and adversarial data, as well as the weights of the last linear layer in Figure 1. The details of implementation are provided in Appendix A.1.

From Figures 1a and 1b, we observe, for a non-robust ResNet-18, the activation distribution of the adversarial data is obviously mismatched with that of the natural data: natural data activate channels that are PR to predictions with high values and high frequency, while adversarial data tend to amplify the NR ones. From Figures 1c and 1d, we observe that AT robustifies the model by aligning the activation distribution of adversarial data with that of natural data. Specifically, when dealing with adversarial data, AT boosts the activation magnitudes of PR channels while suppressing the activations of NR ones. However, we observe that, for many NR channels (e.g., around 150 channels from 350^{th} to 512^{th}), the activations of adversarial data are much higher than those of natural data. These over-activations decrease the prediction scores corresponding to their true categories. Given this observation, we wonder (Q1): *if we suppress these NR channels to regularize the freedom of adversarial perturbations, will it further improve the model's robustness upon AT*?

Besides, an adversarially trained model does not enjoy similar robustness for all classes, i.e., the robust accuracy of a certain class may be much higher than another (e.g., "automobile" with 69.0% vs. "cat" with 16.7% against PGD-20). Comparing the activations of these two classes (Figures 1c and 1d), we observe that, for the class with strong robustness (e.g., "automobile"), channels' activations align better with their relevances to labels, i.e., the channel with a greater extent of activation usually corresponds to a larger weight in the linear layer. In contrast, this alignment does not hold for the class with relatively poor robustness (e.g., for class "cat", the most activated channels, lying between the 26^{th} and 125^{th}, are sub-PR to predictions). Given this phenomenon, we may ask another question (Q2): *If we scale channels' activations based on their relevances to predictions, will it improve the model's robustness?*

Considering the two questions above, we propose the unified hypothesis $\mathcal{H}$, as stated in the Introduction.

3.2. Importance-based Channel Adjustment

To examine the hypothesis $\mathcal{H}$, one needs a systematic approach to manipulate the channels, viz. selecting channels via suppressing NR ones but promoting PR ones. To this end, we introduce a mechanism, dubbed as Channel-wise

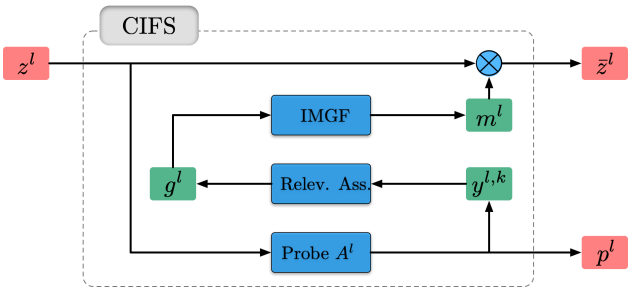


Figure 2. CIFS: *1)* Probe Network A^l first makes a raw prediction p^l for z^l. *2)* Channels' relevances g^l are assessed (Relev. Ass.) based on the gradients of the top-k prediction results $y^{l,k}$. *3)* The IMGF generates an importance mask m^l from g^l for channel adjustment.

Importance-based Feature Selection (CIFS). CIFS modifies layers of CNNs by adjusting channels' activations with importance scores that are generated from the channels' relevances to predictions.

For clearly state CIFS, we first introduce some notations: For a K-category classification problem, let $(X, Y) \sim P_{XY}$ denote the pair of the random input and its label, where $X \in \mathcal{X} \subset \mathbb{R}^{n_X}$ and $Y \in \mathcal{Y} = \{0, 1, \ldots, K-1\}$. We design an L-layer CNN-based classification model to make accurate predictions for data sampled from P_{XY}. The l^{th} layer is denoted by $f^l(\cdot)$ and parametrized by $\theta^l \in \Theta^l$; the mapping from the input to the l^{th} layer's output is denoted by $f^{[l]} = f^l \circ f^{l-1} \circ \cdots \circ f^1$ and the combination of all the first l layers' parameters is denoted by $\theta^{[l]}$, i.e., $\theta^{[l]} = (\theta^1, \ldots, \theta^l)$. Let us examine the l^{th} layer where an input x is transformed into a high-dimensional representation $z^l = f^{[l]}(x) \in \mathbb{R}^{n_{\text{C}}^l \times n_{\text{F}}^l}$; z^l has n_{C}^l channels and each channel is a feature vector of length n_{F}^l. With these notations, we elaborate the details of CIFS in ***three*** steps (as shown in Figure 2).

1) Surrogate Raw Prediction: To assess channels' relevances to predictions, a naive strategy is to compute the gradients of the final prediction with respect to z^l, i.e., $\nabla_{z^l} f^{[l+1:L]}(z^l)$, where $f^{[l+1:L]} = f^L \circ \cdots \circ f^{l+1}$. Since we need to adjust z^l with importance scores generated from $\nabla_{z^l} f^{[l+1:L]}(z^l)$ and send the adjusted feature $\bar{z}^l$ to $f^{[l+1:L]}$ again for making the final prediction, it will result in computing the second-order derivatives during the training phase. Moreover, in practice, we may apply CIFS into multiple layers, the forward pass will involve at least the second-order gradients (the latter CIFS-modified layer is recursively called). Thus, the back-propagation has to deal with at least the third-order gradients during training. This will aggravate the problem of training instability.

Instead, inspired by the design of auxiliary classifiers in CAS (Bai et al., 2021), CIFS builds a probe network A^l as the surrogate of $f^{[l+1:L]}$ for a making raw prediction $p^l = A^l(z^l)$, so that we can use the gradients of p^l to approximately assess the channels' relevances to the final predic-

tion. The assessment does not involve other CIFS-modified layers. Thus, we can avoid the problem of back-propagation through high-order derivatives. The probe network A^l is parameterized by θ_A^l and $p^l \in \mathbb{R}^K$ represents the vector of prediction scores/logits. We can jointly optimize A^l with the backbone network during the training phase under the supervision of true labels.

2) Relevance Assessment: With the prediction p^l, we can compute the gradients of logits in p^l *w.r.t.* z^l to assess the feature's relevances to each class. We consider the top-k prediction results ($k \geq 2$) for the assessment of channels' relevances. As data from two semantically similar classes (e.g., "dog" and "cat") usually share common features, the prediction for an input often assigns large scores to the classes similar to the true one and the top-k results may include several of these similar classes (Jia et al., 2020). In case the top-1 prediction is wrong, considering the top-k results may help us reliably extract some common relevant features (see Section 4.2 for more evidence).

Let $y^{l,k}$ denote indices of the k largest logits of prediction p^l. Let $\delta \in \mathbb{R}^{n_{\mathrm{C}}^l}$ be the channel-wise perturbation added to z^l, giving the perturbed representation $z_\delta^l = z^l + \delta \cdot \mathbf{1}^\top$. Here $\mathbf{1} \in \mathbb{R}^{n_{\mathrm{F}}^l}$ is the column vector with all elements as one, i.e., the features in the same channel are perturbed by a common value. We calculate the gradients of the sum of the top-k logits *w.r.t.* the channel-wise perturbation δ:

$$g^l = \nabla_\delta \sum_{i \in y^{l,k}} p^l(\delta)_{[i]} \Big|_{\delta=\mathbf{0}} = \nabla_{z_\delta^l} \sum_{i \in y^{l,k}} A^l(z_\delta^l)_{[i]} \Big|_{z_\delta^l = z^l} \cdot \mathbf{1}, \tag{1}$$

where $g^l = (g^l_{[0]}, ..., g^l_{[n_{\mathrm{C}}^l - 1]})$ represents the vector of channels' relevances to the top-k logits. During the training phase, since the true label of z^l is given, we replace the top-1 prediction in $y^{l,k}$ with the true label y and keep other prediction results untouched.

3) Importance Mask Generation: As we want to suppress or promote channels based on their relevances, we need to design proper Importance Mask Generating Functions (IMGFs), which monotonically map relevances to *non-negative* importance scores; of particular importance is to map negative relevances to values close to zero.

Here, we provide several feasible options: To answer the first question on whether suppressing NR channels enhances robustness, one can use the sigmoid function as the IMGF. With a large value of α, the sigmoid function serves as a switch by mapping negative relevances to importance scores close to zero and the positive close to one. To answer the second question concerning aligning channel activations with their relevances, we can use the softplus or softmax function as the IMGF. Both of them can map negative relevacnes to values close zero and map positive relevances monotonically to positive values. These three functions are stated here for ease of reference:

- sigmoid: $m^l_{[i]} = \frac{1}{1+\exp(-\alpha \cdot g^l_{[i]})}, \quad \alpha > 0.$
- softplus: $m^l_{[i]} = \frac{1}{\alpha} \cdot \log(1 + \exp(\alpha \cdot g^l_{[i]})), \quad \alpha > 0.$
- softmax: $m^l_{[i]} = \frac{\exp(g^l_{[i]}/T)}{\sum_j \exp(g^l_{[j]}/T)}, \quad T > 0.$

The usage of these functions will be discussed in detail in Section 3.3. CIFS selects channels by multiplying the importance mask m^l with z^l as follows:

$$\bar{z}^l = z^l \otimes \mathrm{repmat}(m^l, 1, n_L^l), \tag{2}$$

where the "repmat" operation replicates the column vector m^l along the second axis n_L^l times.

Training of CIFS In practice, we may apply the CIFS mechanism into several layers of a CNN. Let I denote the set of indices of these layers, and θ_A^I denote the parameters of all the probes in the CIFS-modified layers. For each input x, the modified model $\bar{f}^{[L]}$ outputs $|I|$ raw predictions and one final prediction $p = \bar{f}^{[L]}(x)$. Given this, we use an adaptive loss function (Bai et al., 2021) for training the model:

$$\ell_\beta(x, y) = \frac{1}{1+\beta} \cdot \ell_{\mathrm{ce}}(p, y) + \frac{\beta}{(1+\beta)|I|} \cdot \sum_{l \in I} \ell_{\mathrm{ce}}(p^l, y), \tag{3}$$

where $\ell_{\mathrm{ce}}(\cdot, \cdot)$ denotes the cross-entropy loss and the coefficient $\beta > 0$ balances the accuracy of raw predictions by CIFS and the final prediction. Since the subsequent decisions closely depend on the channels of features selected by the previous CIFS-modified layers, we should choose a proper value of β to make sure that the raw predictions made by CIFS are reliable. In practice, we set β to be $|I|$, and the effect of β is discussed in the ablation study (see Appendix E). To robustify the CNN model against malicious attacks, we can train $\bar{f}^{[L]}$ in an adversarial manner with a perturbation budget ϵ. Namely, we solve the following optimization problem:

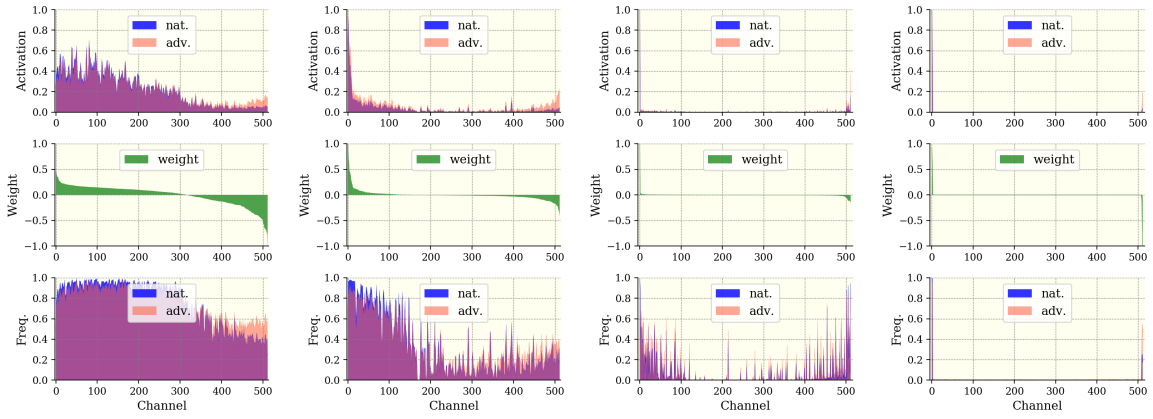
$$\min_{\theta^{[L]}, \theta_A^I} \mathbb{E}_{P_{XY}} \left[\max_{X' \in \mathcal{B}(X, \epsilon, l_\infty)} \ell_\beta(X', Y) \right], \tag{4}$$

where $\mathcal{B}(x, \epsilon, l_\infty) = \{x' \mid \|x' - x\|_{l_\infty} \leq \epsilon\}$.

3.3. Verification of Hypothesis $\mathcal{H}$ on Robustness Enhancement

We verify the hypothesis $\mathcal{H}$ by answering the two questions, Q1 and Q2, in Section 3.1 respectively.

To answer Q1, we applied the sigmoid function to generate the mask m^l from g^l. Setting α to be large enough (here, $\alpha = 10$), we can generate importance scores close to zeros for NR channels and scores close to one for the PR ones, so that we approximately annihilate the NR channels but leave the PR ones as unchanged. We adversarially trained a



(a) non-CIFS (b) CIFS-sigmoid (c) CIFS-softplus (d) CIFS-softmax

Figure 3. The magnitudes of channel-wise activations (top) at the penultimate layer, their activated frequency (bottom), and the weights of the last linear layer (middle) *vs.* channel indices. The activations of natural data and their PGD-20 examples are averaged over all test samples in the "airplane" category. The robust accuracies against PGD-20 (on the whole dataset) are 46.64% for non-CIFS, 49.87% for the CIFS-sigmoid, 50.38% for the CIFS-softplus, and 51.23% for the CIFS-softmax respectively

ResNet-18 model and its CIFS-modified version. As shown in Figure 3a, the NR channels (Channel 300-512) of the vanilla ResNet-18 model are clearly activated (the average activation magnitudes of these channels are larger than 0.1; the activation frequencies are over 0.4). In contrast, the ResNet-18 with CIFS-sigmoid effectively suppresses the activation of NR channels (Channel 100^{th}–512^{th} in Figure 3b). Most of their mean activation magnitudes are smaller than 0.05, and their activation frequencies have clearly decreased. The experimental results show that, under AT, the vanilla ResNet-18 model results in a 46.64% defense rate against the PGD-20 attack while its CIFS-sigmoid modified version achieves 49.87%. More results can be found in Appendix A.2. Thus, we conclude that suppressing NR channels enhances the robustness of CNNs.

To answer Q2, we applied the softplus and softmax functions as IMGFs to generate the mask m^l from g^l respectively. Here the coefficient α in the softplus is set to be 5 and the temperature T in the softmax is set to be 1. From Figures 3c and 3d, we observe that, by generating importance scores positively correlated with the relevances, the model tends to completely focus on few relevant (positive and negative) channels. The channel of the greatest weight (most PR to predictions) is activated with the highest magnitude. Most channels become irrelevant (small absolute values of weights) to the predictions and are activated at a low level. Using softplus as the IMGF (Figure 3c), the irrelevant channels are sparsely activated, and the activation magnitudes are smaller than 5% of the most important channel. In Figure 3d, this phenomenon is enhanced by using softmax as IMGF: most channels become irrelevant to predictions and are usually deactivated. We evaluated the robustness of these two CIFS-modified CNNs against PGD-20 attack. Both of them outperformed the CIFS-sigmoid and the vanilla ResNet-18 classifiers (robust accuracies of CIFS-softplus and CIFS-softmax are 50.38% and 51.23% respectively, *vs.*, 49.87% for CIFS-sigmoid and 46.64% for the vanilla ResNet-18). We also found CIFS can ameliorate the class-wise imbalance of adversarial robustness (*e.g.*, CIFS-softmax increases the PGD-20 accuracy from 16.7% to 22.3% for class "cat"). More details are provided in Appendix A.2. Thus, we conclude that aligning channels activations with their relevances to predictions can further robustify CNNs upon suppressing NR ones.

Given these empirical results, we verified the hypothesis $\mathcal{H}$ and justified that CIFS is an effective mechanism to improve the adversarial robustness of CNNs. In the following section, we conduct extensive experiments to evaluate the robustness enhancement through CIFS and study CIFS in an ablation manner.

4. Experiments

4.1. Robustness Evaluation

We utilize the CIFS to modify CNNs in different architectures to perform classification tasks on benchmark datasets, namely a ResNet-18 and a WideResNet-28-10 on the CIFAR10 (Krizhevsky, 2009) dataset, a ResNet-18 on the SVHN (Netzer et al., 2011) dataset and a ResNet-10 on the Fashion-MNIST (Xiao et al., 2017) dataset. We train the models with the standard PGD adversarial training (AT) (Madry et al., 2018) and its variants, such as FAT (Zhang et al., 2020), to show that CIFS can work under various AT-strategies. We compare CIFS-modified CNNs with the vanilla versions as well as the CAS-modifications, where CAS (Bai et al., 2021) also modifies CNNs by adjusting channels' activations. Here, we report the results on CIFAR10 and SVHN. Results on FMNIST are presented in Appendix D.

Adaptive Attacks As mentioned in Section 3.2, each CIFS-modified layer of CNNs outputs a raw prediction. To generate adversarial examples that are as strong as possible, we follow the strategy used in CAS and attack CNNs via the adaptive loss function ℓ_β in Equation (3) that considers all the raw and final predictions. We let the value of β be chosen by the attacker, i.e., the attacker can try various values of β_{atk} and select one that results in the most harmful perturbations. Our setting is more challenging for defense than CAS where the *same* value of β is used for training and attack. Here, for each adversarial attack, we evaluate the robustness by choosing β_{atk} from $\{0, 0.1, 1, 2, 10, 100, \infty\}$ and report the worst robust accuracy[1]. Setting $\beta_{\text{atk}} = \infty$ means the attacks completely focus on the CIFS-modified layers [2] and only consider the second term in Equation (3).

4.1.1. Robustness Enhancement of CIFS under AT

We adversarially train ResNet-18 and WRN-28-10 models with PGD-10 ($\epsilon = 8/255$) adversarial data. CIFS is applied to the last two residual blocks of each model. The probes for the last and penultimate blocks are a linear layer and a multi-layer perceptron (MLP) respectively. Channels' relevances are assessed based on top-2 results and we use the softmax function with $T = 1$ as the IMGF. Other training details are provided in Appendix B and C.

Defense Results We evaluate the robustness of CNNs against four types of white-box attacks: FGSM (Szegedy et al., 2014), PGD-20 (Madry et al., 2018), C&W (Carlini & Wagner, 2017), and PGD-100. The l_∞-norm of the perturbations are bounded by the value of $\epsilon = 8/255$. Here, we report the robustness evaluated at the last epoch for each model. Detailed attack settings and more defense results (AutoAttack[3] (Croce & Hein, 2020) and the best epochs' results), are present in Appendix B and C.

The defense results on CIFAR10 are reported in Table 1. We observe that, for both of the ResNet-18 and WRN-28-10 architectures, CIFS consistently outperforms the counterparts against various types of adversarial attacks. For example, the CIFS-modified WRN-28-10 can defend the PGD100 attack with a success rate of 48.74%, which exceeds the second best by more than 4 percentage points. In contrast, under the strong adaptive attack, we see that the baseline CAS cannot improve and even worsens the robustness of CNNs. The defense results on SVHN are reported in Table 2. The results also verify the effectiveness of CIFS on improving robustness.

Table 1. Robustness comparison of defense methods on CIFAR10. We report the accuracies (%) for adversarial and natural data. For each model, the results of the strongest attack are marked with an underline.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	84.56	55.11	46.62	45.95	44.72
CAS	86.73	55.99	45.29	44.18	43.22
CIFS	83.86	**58.86**	**51.23**	**50.16**	**48.70**
WRN-28-10	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	87.29	58.50	49.17	48.68	47.08
CAS	88.05	57.94	49.03	47.97	47.25
CIFS	85.56	**61.34**	**53.74**	**53.20**	**51.51**

Table 2. Robustness comparison of defense methods on SVHN. The accuracies (%) for natural and adversarial data are reported.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	93.72	**65.87**	50.35	47.89	45.81
CAS	94.08	65.24	48.47	46.15	43.75
CIFS	93.94	66.24	**52.02**	**50.13**	**47.49**

Training Procedure We train CNN classifiers in an adversarial manner for 120 epochs and adjust the learning rate with a multiplier 0.1 at epoch 75 and epoch 90. We summarize the training procedure by plotting the curves of training losses and the PGD-20 accuracies *w.r.t.* epochs in Figure 4. We observe the best adversarial robustness of the vanilla ResNet-18 (50.64% defense rate) appears around the 75^{th} epoch. After epoch 75, the model starts to overfit to training data, i.e., the training loss continues decreasing, but the robust accuracy drops as well. In contrast, the overfitting problem is ameliorated by the application of CIFS. We can see that the best robust accuracy appears around the 90^{th} epoch; After the best epoch, the training loss continues decreasing, but the robustness is maintained around the peak. This phenomenon may result from the fact that CIFS suppresses redundant channels. The model redundancy can be controlled by selecting few highly relevant channels and deactivating others. In this way, the overfitting in training is ameliorated.

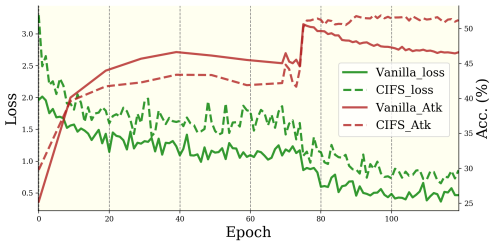


Figure 4. Comparison on the training of a vanilla ResNet-18 and its CIFS-modified version. Training losses and accuracies against PGD-20 attack are plotted.

On the computation overhead, we report the training time and the evaluation time of a ResNet-18 classifier on the

[1] Results of various values of β are present in Appendix B.1. We observe that CAS can improve the robustness of CNNs in most cases but *fail* when attackers completely focus on CAS modules.

[2] For $\beta_{\text{atk}} = \infty$, we consider the cases of attacking both CIFS-modified layers simultaneously and attacking each separately.

[3] AutoAttack consists of both white and black-box attacks.

CIFAR10 dataset for reference. For PGD-10 adversarial training, the vanilla CNN takes 166s for each epoch while the CIFS-modified model takes 172s. For the PGD-20 evaluation, the vanilla CNN takes 53s for the CIFAR10 test set; the CIFS-modified model takes 56s instead. In short, the proposed CIFS does not result in too much extra computation.

4.1.2. CIFS Working in Conjunction with Variants of AT

CIFS improves the adversarial robustness by adjusting channels' activations, which is orthogonal to defense training strategies. Here, we train the CIFS-modified CNNs with variants of AT to examine whether CIFS can work in conjunction with other state-of-the-art training-based defense techniques. We consider the FAT (Zhang et al., 2020) and TRADES (Zhang et al., 2019) strategies. We report the defense results of FAT in Table 3 and provide the results of TRADES in Appendix B.2. We observe that FAT training strategy improves the natural accuracy and robustness of CNNs (compared to the results in Table 1). Under the FAT strategy, CIFS also improves the adversarial robustness of the vanilla CNNs in both ResNet-18 and WRN-28-10 architectures.

Table 3. Robustness comparison of defense methods trained with FAT on CIFAR10. Comparing the accuracies (%) against the strongest attacks, we observe that CIFS clearly robustifies CNNs.

ResNet-18	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	87.16	56.43	47.64	46.01	45.35
CIFS	86.35	**59.47**	**51.68**	**51.84**	**49.52**
WRN-28-10	Natural	FGSM	PGD-20	C&W	PGD-100
Vanilla	88.37	58.81	49.62	48.49	47.58
CIFS	86.74	**60.67**	**51.99**	**52.34**	**49.87**

4.2. Ablation Study

Here, we conduct an ablation study to further understand the robustness properties of CIFS. Specifically, we investigate the effects of the feedback from the top-k predictions. The ablation experiments are conducted on CIFAR10 based on the ResNet-18 model. Besides, in Appendix E, we also study cases in which the CIFS is applied to different layers, the probe networks are in different architectures, various values of β are used for training.

Feedback from Top-k Prediction Results As is well-known, the top-k classification accuracy for $k > 1$ is always not worse than the top-1. For example, in Table 4, we can see that the top-2 accuracy of an adversarially trained ResNet-18 against the PGD-20 attack exceeds the top-1 accuracy by 25 percentage points. This implies that, although adversarial data can usually fool the classifier (i.e., low top-1 accuracy), the prediction confidence of the true class is still high and the corresponding score highly likely lies among the top-2 or 3 logits.

Table 4. Top-k accuracies (%) against adversarial attacks on CIFAR10 of an adversarially trained ResNet-18.

ResNet-18	top-1	top-2	top-3
FGSM	55.11	76.22	85.20
PGD-20	46.62	71.71	81.60

CIFS generates the importance mask from the raw prediction and uses it to suppress or promote channels at the current layer. The final prediction made by subsequent layers strongly depends on the channels selected by CIFS. To ensure the accuracy of final predictions, the logits used for generating importance scores should include the true label's logit for each input so that the truly important channels will be highlighted. According to Table 4, if we use the top-1 logit to assess the importance of channels for PGD-20 adversarial data, the probability of incorrect assessment is over 50%. Instead, if we use the feedback from top-2 or 3 logits, the truly important channels can highly likely be promoted. The following table presents more experiments that justify this argument.

Table 5. Robustness comparison (%) of importance assessment based on top-k results against the PGD-20 attack. The column header $*/\#$ represents the attack point and the output prediction. For example, CIFS/Final means the model outputs the final prediction and the attacker solely focuses on the CIFS's raw prediction.

ResNet-18	Natural/Final	CIFS/CIFS	CFIS/Final	Adap/Final
Vanilla	84.56	-	-	46.62
top-1	87.63	47.24	47.24	47.24
top-2	83.86	48.72	54.96	**51.23**
top-3	83.49	47.59	55.39	49.91

From Table 5, we observe that, for the top-1 case, the defense rate of 'CIFS/CIFS' is the same as that of 'CIFS/Final' and that of 'Adap/Final.' This implies that, once an adversarial example successfully fools the raw prediction of CIFS, the final prediction also will be incorrect. Thus, the attacker only needs to focus on the CIFS's raw predictions to break the model. In contrast, for the top-2 case, the defense rate of 'CIFS/Final' exceeds the 'CIFS/CIFS' by 6 percentage points. This means that nearly 6% adversarial data mislead the CIFS's raw predictions. However, through the channel adjustment via CIFS, these adversarial data are "purified", and more relevant characteristic features are thus transmitted to subsequent layers of CNNs. As such, these adversarial data are finally classified correctly. In this case, the attacker has to exhaustively search for an adaptive loss function to generate attacks, and the CIFS-modified CNNs are safer and more reliable. More discussion on why the top-k assessment performs better and how to choose k is provided in Appendix E.

5. Conclusion

We developed the CIFS mechanism to verify the hypothesis that suppressing NR channels and aligning PR ones with their relevances to predictions benefits adversarial robustness. Empirical results demonstrate the effectiveness of CIFS on enhancing CNNs' robustness.

There are two limitations of our current work: 1) We empirically verify the hypothesis $\mathcal{H}$, but it is still difficult to explicitly, not intuitively, explain why the adjustment of channels improves robustness. 2) Although CIFS ameliorates the overfitting during AT and improves the robustness, it sometimes leads to a bit drop in natural accuracies on certain datasets. In the future, we will attempt to address these two limitations.

Acknowledgements

HY and VYFT are funded by a Singapore National Research Foundation (NRF) Fellowship (R-263-000-D02-281).

JF is supported by the National Research Foundation Singapore under its AI Singapore Programme (Award Number: AISG-100E-2019-035)

JZ, GN, and MS are supported by JST AIP Acceleration Research Grant Number JPMJCR20U3, Japan. MS is also supported by the Institute for AI and Beyond, UTokyo.

References

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In ECCV, 2020.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In ICML 2018. PMLR, 2018.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 2015.

Bai, Y., Zeng, Y., Jiang, Y., Xia, S.-T., Ma, X., and Wang, Y. Improving Adversarial Robustness via Channel-wise Activation Suppressing. In ICLR 2021. OpenReview.net, 2021.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In ICML 2009. ACM, 2009.

Cai, Q., Liu, C., and Song, D. Curriculum adversarial training. In IJCAI 2018. ijcai.org, 2018.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp). IEEE, 2017.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In International Conference on Learning Representations, 2021.

Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y. N., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In ICML 2017. PMLR, 2017.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In ICML 2020. PMLR, 2020.

Dhillon, G. S., Azizzadenesheli, K., Lipton, Z. C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. In ICLR 2018. OpenReview.net, 2018.

Du, J., Yan, H., Tan, V. Y. F., Zhou, J. T., Goh, R. S. M., and Feng, J. RAIN: A Simple Approach for Robust and Accurate Image Classification Networks. arXiv:2004.14798 [cs, eess], 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In ICLR 2015, 2015.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 37(3):362–386, 2020.

Guo, Y., Zhang, C., Zhang, C., and Chen, Y. Sparse dnns with improved adversarial robustness. In NeurIPS 2018, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In CVPR 2016. IEEE, June 2016.

Hsieh, Y.-L., Cheng, M., Juan, D.-C., Wei, W., Hsu, W.-L., and Hsieh, C.-J. On the robustness of self-attentive models. In ACL 2019. Association for Computational Linguistics, 2019.

Hu, D., Liang, J., Hou, Q., Yan, H., Chen, Y., Yan, S., and Feng, J. Panda: Prototypical unsupervised domain adaptation. arXiv preprint arXiv:2003.13274, 2020.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In NeurIPS 2019,, 2019.

Jia, J., Cao, X., Wang, B., and Gong, N. Z. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. 2020.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial Logit Pairing. arXiv:1803.06373 [cs, stat], 2018.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. pp. 60, 2009.

Lamb, A., Verma, V., Kannala, J., and Bengio, Y. Interpolated Adversarial Training: Achieving Robust Neural Networks without Sacrificing Too Much Accuracy. arXiv:1906.06784 [cs, stat], 2019.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In CVPR 2018. IEEE Computer Society, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In ICLR 2018. OpenReview.net, 2018.

Mustafa, A., Khan, S. H., Hayat, M., Goecke, R., Shen, J., and Shao, L. Adversarial defense by restricting the hidden space of deep neural networks. In ICCV 2019. IEEE, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading Digits in Natural Images with Unsupervised Feature Learning. pp. 9, 2011.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. arXiv:1511.04508 [cs, stat], 2016.

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. In ICML 2020. PMLR, 2020.

Sankaranarayanan, S., Jain, A., Chellappa, R., and Lim, S. Regularizing deep networks using efficient layerwise adversarial training. In AAAI 2018. AAAI Press, 2018.

Sanyal, A., Kanade, V., Torr, P. H. S., and Dokania, P. K. Robustness via Deep Low-Rank Representations. arXiv:1804.07090 [cs, stat], 2020.

Shu, C., Yan, H., Lin, K., Lim, C. M., Zheng, W., Feng, J., and Huang, Z. Enhancing in vivo nose cancer detection with rapid fiberoptic raman and deep learning techniques. In Optical Biopsy XVIII: Toward Real-Time Spectroscopic Imaging and Diagnosis. ISOP, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In ICLR 2014, 2014.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On Adaptive Attacks to Adversarial Example Defenses. arXiv:2002.08347 [cs, stat], 2020.

Trebeschi, S., van Griethuysen, J. J., Lambregts, D. M., Lahaye, M. J., Parmar, C., Bakers, F. C., Peters, N. H., Beets-Tan, R. G., and Aerts, H. J. Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric mr. Scientific reports, 7(1): 1–9, 2017.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In ICLR 2019. OpenReview.net, 2019.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Advances in Neural Information Processing Systems, 2018.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In ICML, 2019.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In ICLR 2020. OpenReview.net, 2020.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. NeurIPS, 33, 2020.

Xiao, C., Zhong, P., and Zheng, C. Enhancing adversarial defense by k-winners-take-all. In ICLR 2020, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. L. Mitigating adversarial effects through randomization. In ICLR 2018. OpenReview.net, 2018.

Xie, C., Wu, Y., van der Maaten, L., Yuille, A. L., and He, K. Feature denoising for improving adversarial robustness. In CVPR 2019. Computer Vision Foundation / IEEE, 2019.

Xu, K., Liu, S., Zhang, G., Sun, M., Zhao, P., Fan, Q., Gan, C., and Lin, X. Interpreting Adversarial Examples by Activation Promotion and Suppression. arXiv:1904.02057 [cs], 2019.

Yan, H., Du, J., Tan, V. Y. F., and Feng, J. On robustness of neural ordinary differential equations. In ICLR 2020. OpenReview.net, 2020.

Zeiler, M. D. and Fergus, R. Visualizing and Understanding Convolutional Networks. In ECCV 2014, 2014.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In ICML 2019. PMLR, 2019.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In ICML 2020, 2020.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. Geometry-aware Instance-reweighted Adversarial Training. In ICLR, 2021, 2021.

Zoran, D., Chrzanowski, M., Huang, P., Gowal, S., Mott, A., and Kohli, P. Towards robust image classification using sequential attention models. In CVPR 2020. IEEE, 2020.