Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes

Sravanti Addepalli*, Vivek B.S.*, Arya Baburaj, Gaurang Sriramanan, R.Venkatesh Babu
Video Analytics Lab, Department of Computational and Data Sciences
Indian Institute of Science, Bangalore, India

Abstract

As humans, we inherently perceive images based on their predominant features, and ignore noise embedded within lower bit planes. On the contrary, Deep Neural Networks are known to confidently misclassify images corrupted with meticulously crafted perturbations that are nearly imperceptible to the human eye. In this work, we attempt to address this problem by training networks to form coarse impressions based on the information in higher bit planes, and use the lower bit planes only to refine their prediction. We demonstrate that, by imposing consistency on the representations learned across differently quantized images, the adversarial robustness of networks improves significantly when compared to a normally trained model. Present state-of-the-art defenses against adversarial attacks require the networks to be explicitly trained using adversarial samples that are computationally expensive to generate. While such methods that use adversarial training continue to achieve the best results, this work paves the way towards achieving robustness without having to explicitly train on adversarial samples. The proposed approach is therefore faster, and also closer to the natural learning process in humans.

1. Introduction

Deep Networks are known to be vulnerable to carefully crafted imperceptible noise known as Adversarial Perturbations [22], which could have disastrous implications in critical applications such as autonomous navigation and surveillance systems. The compelling need of securing these systems, coupled with the goal of improving the worst-case robustness of Deep Networks has propelled research in the area of Adversarial Robustness over the last few years. While adversarial training methods [16, 31] have led to significant progress in improving adversarial robustness, these methods are computationally expensive and also non-intuitive when compared to the learning process in humans.

*Equal contribution

Humans perceive images based on features of large magnitude and use finer details only to enhance their impressions [21, 20]. This *background knowledge* of giving higher importance to information present in higher bit planes naturally equips the human visual system to develop resistance towards adversarial perturbations, which are of relatively lower magnitude. On the contrary, these adversarial perturbations can arbitrarily flip the predictions of Deep Networks to completely unrelated classes, suggesting that such *background knowledge* of giving hierarchical importance to different bit planes is missing in these networks. In this work, we propose to equip Deep Networks with such knowledge, and demonstrate that this improves their robustness to adversarial examples.

We propose a novel *Bit Plane Feature Consistency (BPFC)* regularizer, which can significantly improve adversarial robustness of models, without exposure to adversarial samples during training. The proposed method is considerably faster than methods that require multi-step adversarial samples for training [16], and is therefore scalable to large datasets such as ImageNet. Through this work, we hope to pave the path towards training robust Deep Networks without using adversarial samples, similar to the learning process that exists in human beings. Our code is available at: `https://github.com/val-iisc/BPFC`. We refer the reader to our CVPR paper [1] for further details.

2. Related Works

In this section, we discuss existing alternatives to the multi-step, computationally expensive methods of Adversarial Training (AT) such as PGD-AT [16] and TRADES [31]. Early formulations such as FGSM-AT [7] proposed training with single-step adversarial samples, using a first-order linear approximation to the loss function. This was later shown to be ineffective against multi-step attacks by Kurakin *et al*. [12], wherein the effect of gradient masking was identified. Vivek *et al*. [26] introduced a regularisation term to minimize ℓ_2 distance between logits of images perturbed with FGSM and R-FGSM attacks, in order to mitigate the effect of gradient-masking during adversarial training. In the proposed method, we achieve adversarial robustness without using adversarial samples during training, and hence achieve a further reduction in computation time.

Existing attempts of achieving adversarial robustness without using adversarial samples during training have largely been ineffective. Works such as Mixup [30] and Manifold-Mixup [25] encourage the network to behave in a linearized manner between input data points, or between hidden-layers deeper in the network. While these methods resulted in improved performance against single-step FGSM attacks, they were susceptible to stronger multi-step attacks. In the work by Guo *et al*. [8], the effect of various input transformations such as bit-depth reduction and JPEG compression was studied. The robustness from these techniques primarily originated from the non-differentiable pre-processing steps, in order to possibly thwart gradient-based iterative attacks. This method, along with a few others [3, 15, 6, 28, 19], were broken in the work by Athalye *et al*. [2], where it was identified that obfuscated gradients do not provide reliable security against adversaries.

Feature Squeezing, proposed by Xu *et al*. [29], used transformations such as reduction of color bit depth, spatial smoothing with a median filter and a combination of both, for detection of adversarial samples. However, in the work by He *et al*. [10], it was shown that an adaptive attacker cognizant of this defense strategy could fool the model. While we use the concept of quantization to defend against adversarial attacks, we do not introduce any pre-processing blocks that lead to obfuscated or shattered gradients.

3. Preliminaries: Notation and Threat Model

We consider $f(.)$ as the function mapping of a classifier C, from an image x, to its corresponding softmax output $f(x)$. The predicted class label, which is an argmax over the softmax output, is denoted by $c(x)$. The ground truth label corresponding to x is denoted by y. The pre-softmax output of the classifier C is denoted by $g(x)$. We define $\mathcal{A}(x)$ to be the set of all *Adversarial Samples* corresponding to x, where a specific adversarial sample is denoted by x'.

We consider the task of improving the worst-case robustness of Deep Networks. The goal of an adversary is to cause an error in the prediction of the classifier. We define an *Adversarial Sample* x', as one that causes the output of the network to be different from the ground truth label y. We restrict x' to be in the ℓ_∞-ball of radius ε around x. The set of *Adversarial Samples* can be formally defined as:

$$\mathcal{A}(x) = \{x' : c(x') \neq y, \|x - x'\|_\infty \leq \varepsilon\} \tag{1}$$

Therefore, any individual pixel in the image x cannot be perturbed by more than ε. Since the goal of this work is to improve worst-case robustness, we do not impose any restrictions on the access to the adversary. We consider that the adversary has complete knowledge of the model architecture, weights and the defense mechanism employed.

4. Proposed Method

In this section, we first present the motivation behind our proposed method, followed by a detailed discussion of the proposed algorithm. We further describe local properties of networks trained using the proposed regularizer, which lead to improved robustness.

4.1. Hierarchical Importance of Bit Planes

Bit planes of an image are the spatial maps (of the same dimension as the image) corresponding to a given bit position. For an n-bit representation of an image, bit plane $n-1$ corresponds to the most significant bit (MSB), and bit plane 0 corresponds to the least significant bit (LSB). An n-bit image can be considered as the sum of n bit planes weighted by their relative importance. The importance of features embedded within lower bit planes is significantly lower than that of features embedded within higher bit planes, both in terms of pixel value, and information content [18].

The human visual system is known to give higher importance to global information when compared to fine details [20]. Sugase *et al*. [21] demonstrate that global information is used for coarse classification in early parts of the neural response, while information related to fine details is perceived around 51ms later. This demonstrates a hierarchical classification mechanism, where the response to an image containing both coarse and fine information is aligned with that containing only coarse information.

We take motivation from this aspect of the human visual system, and enforce Deep Networks to maintain consistency across decisions based on features in high bit planes alone (quantized image) and all bit planes (normal image). Adversarial examples constrained to the ℓ_∞-ball utilize low bit planes to transmit information which is inconsistent with that of higher bit planes. The fact that Deep Networks are susceptible to such adversarial noise demonstrates the weakness of these networks, which emanates from the lack of consistency between predictions corresponding to coarse information and fine details. Therefore, enforcing feature consistency across bit planes results in a significant improvement in adversarial robustness when compared to conventionally trained networks.

4.2. Proposed Training Algorithm

We present the proposed training method in Algorithm-1. Broadly, each image in the training set is first quantized, and subsequently used to impose local smoothness in the network.

Quantization: The input image x_i is assumed to be represented using n-bit quantization. The intensity of pixels is

Algorithm 1: Bit Plane Feature Consistency

Input: Network f with parameters θ, fixed weight λ, training data $\mathcal{D} = \{(x_i, y_i)\}$ of n-bit images, quantization parameter k, learning rate η, minibatch size M

for minibatch $B \subset \mathcal{D}$ **do**

  Set $L = 0$

  for $i = 1$ **to** M **do**

    $x_{pre} = x_i + \mathcal{U}(-2^{k-2}, 2^{k-2})$ // Add noise

    $x_q = x_{pre} - \left(x_{pre} \mod 2^k\right)$ // Quantization

    $x_q = x_q + 2^{k-1}$ // Range Shift

    $x_q = min(max(x_q, 0), 2^n - 1)$ // Clip

    $L = L + ce(f(x_i), y_i) + \lambda \left\| g(x_i) - g(x_q) \right\|_2^2$

  end for

  $\theta = \theta - \frac{1}{M} \cdot \eta \cdot \nabla_\theta L$ // SGD update

end for

hence assumed to be in the range $[0, 2^n)$. We generate an $n-k+1$ bit image using the quantization process described here. The allowed range of k is between 1 and $n-1$.

Since low magnitude noise does not always reside in low bit planes and can overflow to higher bit planes as well, we introduce pre-quantization noise in our proposed approach. Uniform noise of small magnitude is added to each pixel in the image. Next, each pixel is quantized to $n-k$ bits, by setting the last k bits to 0. Further, the intensity of all pixels is shifted up by a constant, which is half of the quantization step size. This shifts the range of quantization error from $[0, 2^k)$ to $[-2^{k-1}, 2^{k-1})$. Finally, the quantized image is clipped to the original range $[0, 2^n - 1]$.

Bit Plane Feature Consistency Regularizer: The loss function used for training is shown below:

$$L = \frac{1}{M} \sum_{i=1}^{M} ce(f(x_i), y_i) + \lambda \| g(x_i) - g(q(x_i)) \|_2^2 \quad (2)$$

For a given image x_i, the first term of Eq. (2) is the cross-entropy (ce) loss obtained from the softmax output of the network $f(x_i)$, and the corresponding ground truth label y_i. The second term is the squared ℓ_2 distance between the pre-softmax activation of the image x_i, and that of the corresponding quantized image $q(x_i)$ (generated using the process described in Algorithm-1). We call this squared ℓ_2 loss term as the *Bit Plane Feature Consistency* (*BPFC*) regularizer, as it ensures that the network learns consistent feature representations across the original image as well as the coarse quantized image. The loss for each minibatch of size M is an average over all samples in the minibatch.

The cross-entropy term on original images ensures that a combination of coarse and fine features is used to learn the overall function mapping $g(.)$. This helps preserve the accuracy on clean images, while the *BPFC* regularizer helps improve the adversarial robustness of the model.

4.3. Local Properties of BPFC Trained Networks

In this section, we examine local properties of the function $g(.)$ learned using the proposed *BPFC* regularizer.

Let x_i denote an n-bit image sampled from the data distribution $\mathbb{P}_D$ with pixel intensities in the range $[0, 2^n)$, and let $q(x_i)$ denote a quantized image corresponding to x_i. We assume that $q(x_i)$ is not identically equal to x_i. For a fixed value of λ, let $\Theta_{g(\lambda)}$ denote the set of parameters corresponding to a family of functions that lead to the cross-entropy term in Eq. (2) being below a certain threshold. Minimization of *BPFC* loss among the family of functions parameterized by $\Theta_{g(\lambda)}$ is shown in Eq. (3):

$$\min_{\theta_g \in \Theta_{g(\lambda)}} \mathbb{E}_{x_i \sim \mathbb{P}_D} \, \mathbb{E}_{q(x_i)} \, \| g(x_i) - g(q(x_i)) \|_2^2 \quad (3)$$

$$\min_{\theta_g \in \Theta_{g(\lambda)}} \mathbb{E}_{x_i \sim \mathbb{P}_D} \, \mathbb{E}_{q(x_i)} \, \frac{\| g(x_i) - g(q(x_i)) \|_2^2}{\| x_i - q(x_i) \|_2^2} \quad (4)$$

The expression in Eq. (3) can be lower bounded by the expression in Eq. (4), which is equivalent to minimizing the local Lipschitz constant of the network at each sample x_i. Hence, imposing *BPFC* regularizer encourages the network to be locally Lipschitz continuous with a reduced Lipschitz constant. While the *BPFC* regularizer imposes local smoothness, the cross-entropy term in Eq. (2) requires $g(.)$ to be a complex mapping for better accuracy on clean images. The final selection of θ_g would depend on λ, which is selected based on the amount by which clean accuracy can be traded-off for adversarial accuracy [31, 23]. Since the function learned is relatively smooth in the initial epochs, we start with a low value of λ and step it up during training.

Therefore, the *BPFC* formulation leads to functions with improved local properties, which is closely related to adversarial robustness as explained by Szegedy *et al.* [22].

5. Experiments and Analysis

5.1. Preliminaries

We use CIFAR-10 [11], Fashion-MNIST [27] and MNIST [13] datasets for validating our proposed approach. We use ResNet-18 [9] architecture for CIFAR-10, and a modified LeNet [14] architecture with two additional convolutional layers for MNIST and Fashion-MNIST. We train CIFAR-10 models for 100 epochs, MNIST and Fashion-MNIST models for 50 epochs each. The hyperparameters to be selected for training are k (number of bits eliminated during the quantization step in Algorithm-1) and λ (weighting factor for the *BPFC* loss in Eq.2). We refer the reader to Section-5.1 in our CVPR paper [1] for details on datasets, optimizer settings and hyperparameters.

5.2. Overview of Experiments

We compare the proposed approach with Normal Training (NT), FGSM-AT [7], PGD-AT [16] and Regularized

Table 1: **CIFAR-10**: Recognition accuracy (%) of models in a white-box attack setting.

Training method	Clean	FGSM	IFGSM 7 steps	PGD (n-steps) 7	20	1000
FGSM-AT	92.9	96.9	0.8	0.4	0.0	0.0
RSS-AT	82.3	55.0	50.9	50.0	46.2	45.8
PGD-AT	82.7	54.6	51.2	50.4	47.4	47.0
NT	92.3	16.0	0.0	0.0	0.0	0.0
Mixup	90.3	27.4	1.6	0.6	0.1	0.0
BPFC (**Ours**)	82.4	50.1	44.1	41.7	35.7	34.4
Ablations of the proposed approach (BPFC)						
A1: Simple quant	82.6	49.2	41.4	38.8	31.6	30.1
A2: Uniform noise	82.6	48.7	42.3	40.0	33.3	31.9
A3: ℓ_1 norm [1]	92.1	68.3	60.8	57.1	46.8	35.9

Table 2: **White-box setting:** Recognition accuracy (%) of different models on clean samples and adversarial samples generated using PGD-1000 step attack.

Training method	CIFAR-10		F-MNIST		MNIST	
	Clean	PGD	Clean	PGD	Clean	PGD
FGSM-AT	92.9	0.0	93.1	15.1	99.4	3.7
RSS-AT	82.3	45.8	87.7	71.8	99.0	90.4
PGD-AT	82.7	47.0	87.5	79.1	99.3	94.1
NT	92.3	0.0	92.0	0.3	99.2	0.0
Mixup	90.3	0.0	91.0	0.0	99.4	0.0
BPFC (**Ours**)	82.4	34.4	87.2	67.7	99.1	85.7

Table 3: **Black-box setting:** Recognition accuracy (%) of different models on FGSM black-box adversaries. Columns represent the source model used for generating the attack.

Training method	CIFAR-10		Fashion-MNIST		MNIST	
	VGG19	ResNet18	Net-A	M-LeNet	Net-A	M-LeNet
FGSM-AT	78.67	77.58	94.36	90.76	87.99	85.68
RSS-AT	79.80	79.99	84.99	84.16	95.28	95.19
PGD-AT	80.24	80.53	84.99	85.68	95.75	95.36
NT	36.11	15.97	34.71	16.67	29.94	16.60
Mixup	42.67	43.41	54.65	66.31	58.47	69.46
BPFC (**Ours**)	78.92	78.98	81.38	83.46	94.17	94.56

Single-Step Adversarial Training (RSS-AT) [26] across all datasets. For an image with pixel intensities in the range $[0, 1]$, we consider an ε value of $8/255$ for CIFAR-10, 0.3 for MNIST and 0.1 for Fashion-MNIST. We consider ε_{step} to be $2/255$ for CIFAR-10, 0.01 for MNIST and Fashion-MNIST. These restrictions do not apply to the unbounded attacks, DeepFool and C&W. We present the important experimental results and observations below:

- **White-box attacks:** The proposed method achieves a significant improvement over non-adversarial training methods (NT and Mixup) in robustness to single-step and multi-step white-box attacks, despite not being exposed to adversarial samples during training (Tables-1, 2). The proposed method is faster than methods that are robust to multi-step attacks (PGD-AT, RSS-AT).
- **Ablations:** The proposed method (*BPFC*) achieves an improvement over the two ablation experiments of Simple Quantization and addition of Uniform Noise (A1, A2 in Table-1). While results using ℓ_1 norm (A3 in Table-1) show an improvement over the proposed method, the 500-step worst case PGD accuracy goes down from 37.5% to 24.8% with 100 random restarts, indicating that it achieves robustness due to gradient masking. For the proposed approach, the drop in accuracy over multiple random restarts is negligible.
- **Black-box attacks:** Robustness to black-box attacks (Table-3) is significantly better with the proposed approach, when compared to non-adversarial training methods (NT, Mixup). Further, we achieve results that are comparable to adversarial training methods.
- **Sanity checks to verify robustness:** We observe that iterative attacks (PGD and I-FGSM) are stronger than the FGSM attack (Table-1), and white-box attacks are stronger than black-box attacks (Tables-2 and 3).

We present comprehensive results on single-step (FGSM) and multi-step (I-FGSM, PGD) attacks, epsilon-bounded and unbounded attacks (DeepFool [17], Carlini-Wagner [5]), untargeted and targeted attacks, gradient-free attacks (random attacks, SPSA [24]), adaptive attacks and computational complexity in our CVPR paper [1]. We follow the guidelines laid out by Athalye *et al.* [2] and Carlini *et al.* [4] to ascertain the validity of our claim on the achieved robustness, and to prove that the proposed method does not achieve robustness due to gradient masking.

6. Conclusions

We have proposed a novel Bit Plane Feature Consistency (*BPFC*) regularizer, which improves the adversarial robustness of models using a normal training regime. Results obtained using the proposed regularizer are significantly better than existing non-adversarial training methods, and are also comparable to adversarial training methods. Since the proposed method does not utilize adversarial samples, it is faster than adversarial training methods. We demonstrate through extensive experiments that the robustness achieved is indeed not due to gradient masking. Motivated by human vision, the proposed regularizer leads to improved local properties, which results in better adversarial robustness. We hope this work would lead to further improvements on the front of non-adversarial training methods to achieve adversarial robustness in Deep Networks.

[1] A3: The 500-step worst case PGD accuracy goes down from 37.5% to 24.8% with 100 random restarts (over 1000 test samples)

References

[1] Sravanti Addepalli, B S Vivek, Arya Baburaj, Gaurang Sriramanan, and R Venkatesh Babu. Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://arxiv.org/pdf/2004.00306.pdf. 1, 3, 4

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018. 2, 4

[3] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[4] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 4

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017. 4

[6] Guneet S. Dhillon, Kamyar Azizzadenesheli, Jeremy D. Bernstein, Jean Kossaifi, Aran Khanna, Zachary C. Lipton, and Animashree Anandkumar. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3

[8] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[9] 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 (CVPR)*, 2016. 3

[10] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th $USENIX$ Workshop on Offensive Technologies ($WOOT$)*, 2017. 2

[11] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 3

[12] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1

[13] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*. 3

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[15] Xingjun Ma, Bo Li, Yisen Wang, Sarah M. Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Michael E. Houle, Dawn Song, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 1, 3

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[18] Zhao Shan and Wang Hai-Tao. Image retrieval based on bit-plane distribution entropy. In *International Conference on Computer Science and Software Engineering (CSSE)*. IEEE, 2008. 2

[19] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[20] Arun P Sripati and Carl R Olson. Representing the forest before the trees: a global advantage effect in monkey inferotemporal cortex. *Journal of Neuroscience*, 29(24):7788–7796, 2009. 1, 2

[21] Yasuko Sugase, Shigeru Yamane, Shoogo Ueno, and Kenji Kawano. Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, 400(6747):869, 1999. 1, 2

[22] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2013. 1, 3

[23] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 3

[24] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018. 4

[25] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. *arXiv preprint arXiv:1806.05236*, 2018. 2

[26] B. S. Vivek, Arya Baburaj, and R. Venkatesh Babu. Regularizer to mitigate gradient masking effect during single-step adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 1, 4

[27] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. 3

[28] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[29] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017. 2

[30] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2

[31] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019. 1, 3