Towards Achieving Adversarial Robustness Beyond Perceptual Limits

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, Shivangi Khare, R.Venkatesh Babu
Video Analytics Lab, Indian Institute of Science, Bangalore, India

Abstract

The vulnerability of Deep Neural Networks to Adversarial Attacks has fuelled research towards building robust models. While most existing Adversarial Training algorithms aim towards defending against imperceptible attacks, real-world adversaries are not limited by such constraints. In this work, we aim to achieve adversarial robustness at larger epsilon bounds. We first discuss the ideal goals of an adversarial defense algorithm beyond perceptual limits, and further highlight the shortcomings of naively extending existing training algorithms to higher perturbation bounds. In order to overcome these shortcomings, we propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT), that attempts to align the predictions of the network with that of an Oracle during adversarial training. The proposed approach achieves state-of-the-art performance at large epsilon bounds (ℓ_∞ bound of $16/255$) while outperforming adversarial training algorithms such as AWP, TRADES and PGD-AT at standard perturbation bounds (ℓ_∞ bound of $8/255$) as well.

1. Introduction

Deep Neural Networks are known to be vulnerable to Adversarial Attacks, which are perturbations crafted with an intention to fool the network [20]. In a classification setting, adversarially perturbed images cause the network prediction to flip to unrelated classes, while causing no change in a human's prediction (Oracle label). The definition of adversarial attacks involves the presence of an Oracle, and this makes it challenging to formalize threat models for the training and verification of adversarial defenses. The widely accepted convention used in practice is the ℓ_p norm based threat model [3] with low-magnitude bounds to ensure imperceptibility [8]. For example, attacks constrained within an ℓ_∞ norm of $8/255$ on the CIFAR-10 dataset are imperceptible to the human eye as shown in Fig.1(b), ensuring that the Oracle label is unchanged.

While low-magnitude ℓ_p norm based threat models form a crucial subset of the widely accepted definition of adversarial attacks [7], they are not sufficient, as there exist valid attacks at higher ε-bounds as well, as shown in Fig.1(g). However, the challenge at large perturbation bounds is the existence of attacks that can flip Oracle labels as well [21], as shown in Fig.1(c-e). This makes it difficult to naively scale existing Adversarial Training algorithms to large ε bounds. In this work, we aim to improve robustness at larger epsilon bounds, such as an ℓ_∞ norm bound of $16/255$. We define this as a moderate-magnitude bound, and discuss the ideal goals for achieving robustness under this threat model in Sec.3. We further propose a novel defense Oracle-Aligned Adversarial Training (OA-AT), which attempts to align the predictions of the network with that of an Oracle, rather than enforcing all samples within the constraint set to have the same label as the unperturbed image.

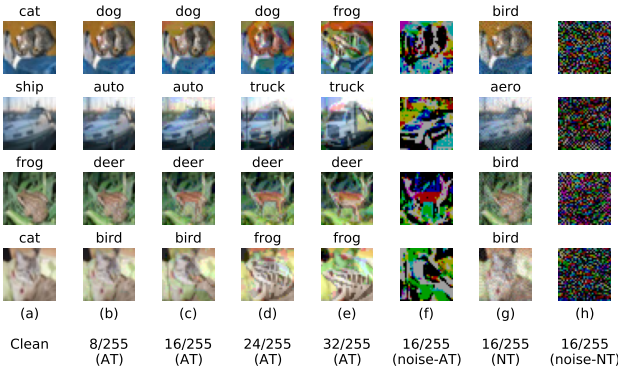

Figure 1. Adversarially attacked images (b-e, g) and perturbations (f, h) for various ℓ_∞ bounds. Attacks are generated from a PGD Adversarially Trained model (AT) [13, 14] or a Normally Trained model (NT). Original unperturbed image is shown in (a). Prediction of the attack source model is printed above each image.

Our contributions have been summarized below:

- We define the ideal goals for a moderate-ε threat model (ℓ_∞ radius of $16/255$) and construct our goals as a feasible subset of the same.
- We propose methods for generating Oracle-Aligned adversaries, which can be used for adversarial training.
- We propose Oracle-Aligned Adversarial Training (OA-AT) to improve robustness within the defined moderate-ε threat model.
- We demonstrate superior performance when com-

*Equal contribution.

Correspondence to: Sravanti Addepalli, sravantia@iisc.ac.in

pared to state-of-the-art methods such as AWP [23], TRADES [24] and PGD-AT [13, 14] at $\varepsilon = 16/255$ while also performing better at $\varepsilon = 8/255$.

- We achieve improvements over the baselines even at larger model capacities such as ResNet-34 and WideResNet-34-10.

2. Related Works

Robustness against imperceptible attacks: Adversarial Training has emerged as the most successful defense strategy against ℓ_p norm bound imperceptible attacks. PGD Adversarial Training (PGD-AT) [13] constructs multi-step adversarial attacks by maximizing Cross-Entropy loss within the considered threat model and subsequently minimizes the same for training. This was followed by several adversarial training methods [24, 1, 15, 23, 18, 14] that improved accuracy against such imperceptible threat models further. Zhang *et al*. [24] proposed the TRADES defense, which maximizes the Kullback-Leibler (KL) divergence between the softmax outputs of adversarial and clean samples for attack generation, and minimizes the same in addition to the Cross-Entropy loss on clean samples for training.

Improving Robustness of base defenses: Wu *et al*. [23] proposed an additional step of *Adversarial Weight Perturbation* (AWP) to maximize the training loss, and further train the perturbed model to minimize the same. This generates a flatter loss surface [19], thereby improving robust generalization. While this can be integrated with any defense, AWP-TRADES is the state-of-the-art adversarial defense today. On similar lines, the use of stochastic weight averaging of model weights [10] is also seen to improve the flatness of loss surface, resulting in a boost in robustness [9, 4]. Recent works [14, 15, 9] attempt to find the best training techniques such as early stopping, use of optimal weight decay and weight averaging to achieve enhanced robust performance on base defenses such as PGD-AT [13] and TRADES [24].

Robustness against large perturbation attacks: Shaeiri *et al*. [16] demonstrate that the standard formulation of adversarial training is not well-suited for achieving robustness at large perturbations, as the loss saturates very early. The authors propose Extended Adversarial Training (ExAT), where a model trained on low-magnitude perturbations ($\varepsilon = 8/255$) is fine-tuned with large magnitude perturbations ($\varepsilon = 16/255$) for merely 5 training epochs, to achieve improved robustness at large perturbations. The authors also discuss the use of a varying epsilon schedule to improve training convergence. Friendly Adversarial Training (FAT) [1] performs early-stopping of an adversarial attack by thresholding the number of times the model misclassifies the image during attack generation. The threshold is increased over training epochs to increase the strength of the attack over training. On similar lines, Sitawarin *et al*. [17] propose Adversarial Training with Early Stopping (ATES), which performs early stopping of a PGD attack based on the margin of the perturbed image being greater than a threshold that is increased over epochs. We improve upon these methods significantly using our proposed approach (Sec.4).

3. Preliminaries and Threat Model

3.1. Notation

We consider an N-class image classification problem with access to a labelled training dataset $\mathcal{D}$. The input images are denoted by $x \in \mathcal{X}$ and their corresponding labels are denoted as $y \in \{1, ..., N\}$. The function represented by the Deep Neural Network is denoted by f_θ where $\theta \in \Theta$ denotes the set of trained network parameters. The N-dimensional softmax output of the input image x is denoted as $f_\theta(x)$. Adversarial examples are defined to be images that are crafted specifically to fool a model into making an incorrect prediction [7]. An adversarial image corresponding to a clean image x would be denoted as $\widetilde{x}$. The set of all images within an ℓ_p norm ball of radius ε, $\mathcal{S}(x)$ is defined as, $\mathcal{S}(x) = \{\hat{x} : ||\hat{x} - x||_p < \varepsilon\}$. The set of all ℓ_p norm bound adversarial examples, $\mathcal{A}(x)$ is defined as, $\mathcal{A}(x) = \{\widetilde{x} : f_\theta(\widetilde{x}) \neq y, \widetilde{x} \in \mathcal{S}(x)\}$. In this work, we specifically consider robustness to ℓ_∞ norm bound adversarial examples. We define the Oracle prediction of a sample x as the label that a human is likely to assign to the image, and denote it as $O(x)$. For a clean image, $O(x)$ would correspond to the true label y, while for a perturbed image it could differ from the original label.

3.2. Nomenclature of Adversarial Attacks

Tramer *et al*. [21] discuss the existence of two types of adversarial examples: Sensitivity-based examples, where the model prediction changes, but the Oracle prediction remains the same as the unperturbed image, and Invariance-based examples, where the Oracle prediction changes, while the model prediction remains unchanged. Models trained using standard empirical risk minimization are susceptible to sensitivity-based adversarial examples, while models which are overly robust to large perturbation bounds could be susceptible to invariance-based examples. Since these definitions are dependent on the model being considered, we define a nomenclature which only depends on the input image and the threat model considered, as below:

- Oracle-Invariant set $OI(x)$, is defined as the set of all images within the bound $\mathcal{S}(x)$, which preserve the Oracle label. The Oracle is invariant to such perturbations: $OI(x) := \{\hat{x} : O(\hat{x}) = O(x), \hat{x} \in \mathcal{S}(x)\}$
- Oracle-Sensitive set $OS(x)$, is defined as the set of all images within the bound $\mathcal{S}(x)$, which flip the Oracle label. The Oracle is sensitive to such perturbations: $OS(x) := \{\hat{x} : O(\hat{x}) \neq O(x), \hat{x} \in \mathcal{S}(x)\}$

3.3. Objectives of the Proposed Defense

Defenses based on the conventional ℓ_p norm threat model defined in Sec.3.1 attempt to train models which are invariant to all samples within $\mathcal{S}(x)$. This is an ideal requirement for low ε-bound perturbations, where the added noise is imperceptible, and hence all samples within the threat model are Oracle-Invariant. An example of a low ε threat model is the constraint set defined by $\varepsilon = 8/255$ for the CIFAR-10 dataset, which produces adversarial examples that are perceptually similar to the corresponding clean images, as shown in Fig.1(b).

As we move to larger ε bounds, Oracle-labels begin to change, as shown in Fig.1(c, d, e). For a very high perturbation bound such as $32/255$, the changes produced by an attack are clearly perceptible and cause a change in the Oracle label in many cases. Hence, robustness at such large bounds may not be of much practical relevance. The focus of this work is to achieve robustness within a moderate-magnitude ℓ_p norm bound threat model, where some perturbations look partially modified (Fig.1(c)), while others look unchanged (Fig.1(g)), as is the case with $\varepsilon = 16/255$ for CIFAR-10. The existence of attacks that do not significantly change the perception of the image necessitates the requirement of robustness within such bounds, while the existence of partially Oracle-Sensitive samples makes it difficult to use standard adversarial training methods on the same. The ideal goals for training defenses under this moderate-magnitude threat model are described below:

- Robustness against samples which belong to $OI(x)$
- Sensitivity towards samples which belong to $OS(x)$, with model's prediction matching the Oracle label
- No specification on Out-of-Distribution (OOD) images

We incorporate these goals in the training objective of our proposed defense, which is discussed in Sec.4. Given the practical difficulty in assigning Oracle labels, we consider the following criteria for our defense evaluations:

- Robustness-Accuracy trade-off, measured using accuracy on clean samples and robustness against valid attacks within the threat model (discussed below)
- Robustness against all attacks within $\varepsilon = 8/255$, measured using strong white-box attacks [5, 18]
- Robustness to Oracle-Invariant samples within $\varepsilon = 16/255$, measured using gradient-free attacks [2]

We do not explicitly define goals for white-box attacks within the moderate ε bound of $16/255$ since the existence of Oracle-Sensitive samples within this bound is image specific. We note from Fig.1(c) and Fig.A3(b) (Suppl.) that most adversarial examples look partially modified at $\varepsilon = 16/255$.

4. Proposed Method

In order to achieve the goals discussed in Sec.3.3, we require to generate Oracle-Sensitive and Oracle-Invariant samples and impose specific training losses on each of them individually. Since labeling adversarial samples as Oracle-Invariant or Oracle-Sensitive is expensive and cannot be done while training networks, we propose to use attacks which ensure a given type of perturbation (OI or OS) by construction, and hence do not require explicit annotation.

Generation of Oracle-Sensitive examples: Robust models are known to have perceptually aligned gradients [22]. Adversarial examples generated using a robust model tend to start looking like the target (other) class images at large perturbation bounds, as seen in Fig.1(c, d, e). We therefore use large ε white-box adversarial examples generated from the model being trained as Oracle-Sensitive samples, and the model prediction as a proxy to the Oracle prediction.

Generation of Oracle-Invariant examples: While the strongest Oracle-Invariant examples are generated using the gradient-free Square attack [2] (Fig.A1, Suppl.), it uses 5000 queries, which is computationally expensive for use in adversarial training. Reducing the number of queries weakens the attack significantly. The most efficient attack that is widely used for adversarial training is the PGD 10-step attack. However, it cannot be used for the generation of Oracle-Invariant samples as gradient-based attacks generated from adversarially trained models produce Oracle-Sensitive samples. We propose to use the Learned Perceptual Image Patch Similarity (LPIPS) measure for the generation of Oracle-Invariant attacks, as it is known to match well with perceptual similarity [25, 12]. As shown in Fig.A2 (Suppl.), while the standard AlexNet model used in prior work [12] fails to distinguish between Oracle-Invariant and Oracle-Sensitive samples, an adversarially trained model is able to distinguish between the two effectively. We therefore propose to minimize the LPIPS distance between natural and perturbed images, in addition to the maximization of Cross-Entropy loss for attack generation: $\mathcal{L}_{CE}(x, y) - \lambda \cdot \text{LPIPS}(x, \hat{x})$. We choose λ as the minimum value that transforms the attack from Oracle-Sensitive to Oracle-Invariant (OI), to generate strong OI attacks (Fig.A3, Suppl.). This is fine-tuned during training to achieve the optimal robustness-accuracy trade-off.

Oracle-Aligned Adversarial Training (OA-AT): The training algorithm for the proposed defense, Oracle-Aligned Adversarial Training (OA-AT) is presented in Algorithm-A1 (Suppl.). We use the Trades-AWP formulation [24, 23] as the base implementation, with Cross-Entropy loss instead of KL-divergence loss for attack generation, as it results in stronger attacks [9]. We maximize loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) in the additional weight perturbation step, as it results in improved robust generalization. We use cosine learning rate schedule.

We start with an initial ε value of $4/255$ upto one-fourth the training epochs, and ramp up this value linearly to a value of $16/255$ at the last epoch. We use 5 attack steps

Table 1. **CIFAR-10, CIFAR-100**: Performance (%) of the proposed defense OA-AT compared to baselines, against attacks with different ε bounds. Sorted by AutoAttack (AA) [5] accuracy at $\varepsilon = 8/255$

Method	Clean	GAMA (8/255)	AA (8/255)	GAMA (12/255)	Square (12/255)	GAMA (16/255)	Square (16/255)
CIFAR-10 (ResNet-18)							
FAT [1]	**84.36**	48.41	48.14	29.39	39.48	15.18	25.07
PGD-AT [13]	79.38	49.28	48.68	32.40	41.46	18.18	28.29
AWP [23]	80.32	49.06	48.89	32.88	40.27	19.17	27.56
ATES [17]	80.95	49.57	49.12	32.44	42.21	18.36	29.07
TRADES [24]	80.53	49.63	49.42	33.32	40.94	19.27	27.82
ExAT-PGD [16]	80.68	50.06	49.52	32.47	41.10	17.81	27.23
ExAT + AWP	80.18	49.87	49.69	33.51	41.04	20.04	28.40
AWP [23]	80.47	50.06	49.87	33.47	41.05	19.66	28.51
OA-AT (Ours)	80.24	**51.40**	**50.88**	**36.01**	**43.20**	**22.73**	**31.16**
Gain w.r.t. AWP	−0.23	+1.34	+1.01	+2.54	+2.15	+3.07	+2.65
CIFAR-100 (ResNet-18)							
AWP [23]	59.88	25.81	25.52	14.80	20.24	8.72	12.80
OA-AT (Ours)	**60.27**	**26.41**	**26.00**	**16.28**	**21.47**	**10.47**	**14.60**
Gain w.r.t. AWP	+0.39	+0.60	+0.48	+1.48	+1.23	+1.75	+1.80

when $\varepsilon = 4/255$ and 10 attack steps later. We perform standard adversarial training upto $\varepsilon = 12/255$ as the attacks in this range are imperceptible. Beyond this, we start incorporating separate training losses for Oracle-Invariant and Oracle-Sensitive samples in alternate training iterations. Oracle-Sensitive samples are generated by maximizing Cross-Entropy loss in a PGD attack formulation. Rather than enforcing the predictions of such attacks to be similar to the original image, we allow the network to be partially sensitive to such attacks by training them to be similar to a convex combination of predictions on the clean image and perturbed samples at larger $(1.5 \cdot \varepsilon_{max})$ bounds as shown: $L_{adv} = KL\big(f_\theta(x_i + \widetilde{\delta}_i) \,||\, \alpha\, f_\theta(x_i) + (1-\alpha)\, f_\theta(x_i + \widehat{\delta}_i)\big)$ Here $\widetilde{\delta}_i$ is the perturbation at the varying epsilon value $\widetilde{\varepsilon}$, and $\widehat{\delta}_i$ is the perturbation at $24/255$. This results in better robustness-accuracy trade-off as shown in Table-A1 (Suppl.). In the other alternate iteration, we use the LPIPS metric to generate strong and efficient Oracle-Invariant attacks during training. We perform exponential weight-averaging of the network being trained and use this for computing the LPIPS metric for improved and stable results (Table-A1, Suppl.). We increase α and λ over training, as the nature of attacks changes with varying ε. The use of both Oracle-Invariant (OI) and Oracle-Sensitive (OS) samples ensures robustness to OI samples while allowing sensitivity to partially OS samples.

5. Experiments and Results

We compare performance of the proposed approach with the existing defenses discussed in Sec.2 on the CIFAR-10 [11] dataset in Table-1. We train all models on ResNet-18 architecture for 110 epochs. For each baseline, we find the best set of hyperparameters to achieve clean accuracy of around 80% to ensure a fair comparison across all methods. We also perform baseline training across various ε values and report the best baselines in Table-1. We observe that baseline defenses do not perform well when trained using large ε bounds such as $16/255$ (Table-A2, Suppl.). We report adversarial robustness against the strongest known attacks, AutoAttack (AA) [5] and GAMA PGD-100 (GAMA) [18] for $\varepsilon = 8/255$ in order to obtain the worst-case robust accuracy. For larger bounds such as $12/255$ and $16/255$, we primarily aim for robustness against the Square attack [2], as it is the strongest known Oracle-Invariant attack. We compare the proposed approach against the strongest baseline AWP-TRADES [23] on CIFAR-100 in Table-1 (ref. Table-A3, Suppl. for detailed results), and on CIFAR-10 with larger capacity models in Table-2. We observe significant gains with the use of AutoAugment [6, 19] on CIFAR-100, and additionally with Model Weight Averaging (WA) [10, 9, 4] at larger model capacities. To ensure a fair comparison, we consider these for the AWP baseline as well.

Table 2. **CIFAR-10**: Performance (%) of the proposed defense OA-AT (Ours) compared to the strongest baseline, AWP-TRADES (AWP) [23] against attacks with different ε bounds

Method	Model	Clean	AA (8/255)	Square (12/255)	AA (12/255)	Square (16/255)	AA (16/255)
AWP	RN-18	80.47	49.87	41.05	33.19	28.51	19.23
Ours	RN-18	80.24	50.88	43.20	35.39	31.16	22.00
AWP	RN-34	83.89	52.44	42.84	34.61	29.22	19.69
Ours	RN-34	84.07	53.22	45.03	36.31	32.47	22.00
AWP	WRN-34	85.19	55.69	46.48	38.05	32.68	23.46
Ours	WRN-34	85.54	55.67	48.15	38.13	35.20	22.92
AWP + WA	WRN-34	85.10	55.87	46.52	37.97	32.50	23.27
Ours + WA	WRN-34	**85.67**	**55.93**	**48.79**	**39.06**	**35.76**	**24.05**

Results: The proposed defense achieves consistent gains across all metrics considered in Sec.3.3 (AutoAttack [5] at $\varepsilon = 8/255$ and Square attack [2] at larger ε bounds). Although we train the model for achieving robustness at larger ε bounds, we achieve an improvement in the robustness at $\varepsilon = 8/255$ as well, which is not observed in any of the existing methods (Table-A2, Suppl.). We evaluate the proposed defense against diverse attacks (Table-A4, Suppl.) and sanity checks (Sec.A5, Suppl.) to ensure the absence of gradient masking.

6. Conclusions

We explore the idea of robustness beyond perceptual limits in an ℓ_p norm based threat model. We first discuss the ideal goals of an adversarial defense at larger perturbation bounds, and further propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT) that aims to align model predictions with that of an Oracle during training. The key aspects of the defense include the use of LPIPS metric for generating Oracle-Invariant attacks during training, and the use of a convex combination of clean and adversarial image predictions as targets for Oracle-Sensitive samples. We achieve significant gains in robustness at low and moderate perturbation bounds, and a better robustness-accuracy trade-off.

7. Acknowledgements

This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC_10), MHRD, Govt. of India. Sravanti Addepalli is supported by a Google PhD Fellowship in Machine Learning. We are thankful for the support.

References

[1] Attacks which do not kill training make adversarial learning stronger. In *International Conference on Machine Learning (ICML)*, 2020.

[2] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *The European Conference on Computer Vision (ECCV)*, 2020.

[3] 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] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2020.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[7] Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it? Blog post on Feb 15, 2017.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[9] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

[10] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[11] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[12] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *International Conference on Learning Representations (ICLR)*, 2021.

[13] 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.

[14] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *International Conference on Learning Representations (ICLR)*, 2021.

[15] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning (ICML)*, 2020.

[16] Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020.

[17] Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020.

[18] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[19] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. *arXiv preprint arXiv:2104.04448*, 2021.

[20] 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.

[21] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning (ICML)*, 2020.

[22] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations (ICLR)*, 2019.

[23] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.