Scaling Adversarial Training to Large Perturbation Bounds

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, R.Venkatesh Babu
Video Analytics Lab, Department of Computational and Data Sciences
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 Adversarial Training algorithms aim at defending attacks constrained within low magnitude Lp norm bounds, real-world adversaries are not limited by such constraints. In this work, we aim to achieve adversarial robustness within larger bounds, against perturbations that may be perceptible, but do not change human (or Oracle) prediction. The presence of images that flip Oracle predictions and those that do not, makes this a challenging setting for adversarial robustness. We 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), 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 (such as an L-inf bound of 16/255) while outperforming existing defenses at standard bounds (8/255) as well. The proposed approach generalizes to attacks that are unseen during training as well, such as other Lp norm bound attacks, common corruptions and recolor attacks.

1. Introduction

Deep Neural Networks are vulnerable to Adversarial Attacks, which are perturbations crafted with an intention to fool the network [14]. In a classification setting, Adversarial attacks can flip the prediction of a network to even unrelated classes, while causing no change in a human's prediction (Oracle label). The definition of adversarial attacks involves the prediction of an Oracle, making it challenging to formalize threat models for the training and verification of adversarial defenses. The widely used convention that overcomes this challenge is the ℓ_p norm based threat model with low-magnitude bounds to ensure imperceptibility [2,7]. 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. The goal of Adversarial Training within such a threat model is to ensure that the prediction of the model is consistent within the considered perturbation radius ε, and matches the label associated with the unperturbed image.

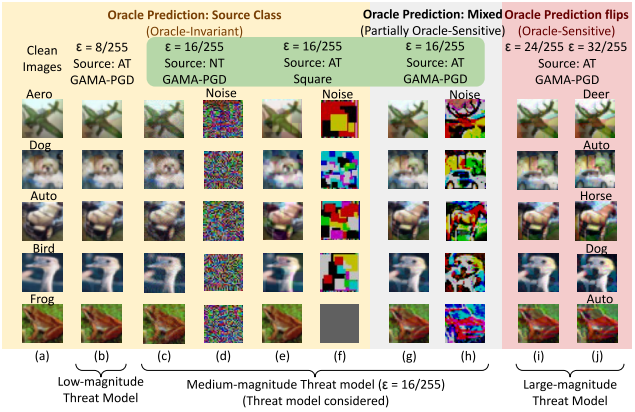

Figure 1. **Perturbations within different threat models:** Adversarial images (b, c, e, g, i, j) and perturbations (d, f, h) along with the corresponding clean image (a) for various ℓ_∞ norm bounds on CIFAR-10. Attacks are generated from an Adversarially Trained model (AT) or a Normally Trained model (NT) using the gradient-based attack GAMA-PGD [13] or the Random-search based attack Square [1]. The medium-magnitude threat model is challenging since it consists of attacks which are Oracle-Invariant and partially Oracle-Sensitive.

While low-magnitude ℓ_p norm based threat models form a crucial subset of the widely accepted definition of adversarial attacks [6], they are not sufficient, as there exist valid attacks at higher perturbation bounds as well, as shown in Fig.1(c) and (e). However, the challenge at large perturbation bounds is the existence of attacks that can flip Oracle labels as well [15], as shown in Fig.1(g), (i) and (j). Naively scaling existing Adversarial Training algorithms to large perturbation bounds would enforce consistent labels on images that flip the Oracle prediction as well, leading to a conflict in the training objective as shown in Fig.1 of the Supplementary. This results in a large drop in clean accu-

*Equal contribution

racy, as shown in Table-1 of the Supplementary. This has triggered interest towards developing perceptually aligned threat models, and defenses that are robust under these settings [10]. However, finding a perceptually aligned metric is as challenging as building a network that can replicate oracle predictions [15]. Thus, it is crucial to investigate adversarial robustness using the well-defined ℓ_p norm metric under larger perturbation bounds.

In this work, we aim to improve robustness at larger epsilon bounds, such as an ℓ_∞ norm bound of $16/255$ on the CIFAR-10 and CIFAR-100 datasets [8]. We define this as a moderate-magnitude bound, and discuss the ideal goals for achieving robustness under this threat model in Sec.2.2. 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 original image. We demonstrate superior performance when compared to state-of-the-art methods [11, 17, 18] at $\varepsilon = 16/255$ while also performing better at $\varepsilon = 8/255$. We achieve improvements even at larger model capacities such as WideResNet-34-10, and outperform existing methods on the RobustBench leaderboard. The proposed approach generalizes remarkably well to attacks that are unseen during training as well, such as other ℓ_p norm bound perturbations, common corruptions and recolor attacks. Our code is available here: https://github.com/val-iisc/OAAT.

2. Preliminaries

2.1. Nomenclature of Adversarial Attacks

Tramer et al. [15] discuss the existence of two types of adversarial examples: Sensitivity-based examples, where the model prediction changes while 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 attacks, while models which are overly robust to large perturbation bounds could be susceptible to invariance-based attacks. Since these definitions are model-specific, we define a different nomenclature which only depends on the input image and the threat model considered:

- Oracle-Invariant set $OI(x)$ is defined as the set of all images within the bound $\mathcal{S}(x)$, that preserve Oracle label. Oracle is invariant to such attacks:

  $OI(x) := \{\hat{x} : O(\hat{x}) = O(x), \hat{x} \in \mathcal{S}(x)\}$

- Oracle-Sensitive set $OS(x)$ is defined as the set of all images within the bound $\mathcal{S}(x)$, that flip the Oracle label. Oracle is sensitive to such attacks:

  $OS(x) := \{\hat{x} : O(\hat{x}) \neq O(x), \hat{x} \in \mathcal{S}(x)\}$

2.2. Objectives of the Proposed Defense

Defenses based on the conventional ℓ_p norm threat model 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-bound constraint set is the ℓ_∞ threat model with $\varepsilon = 8/255$ 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(g, i, j). For a very high perturbation bound such as $32/255$, the changes produced by an attack are clearly perceptible and in many cases flip the Oracle label as well. Hence, robustness at such large bounds is not of practical relevance. The focus of this work is to achieve robustness within a moderate-magnitude ℓ_p norm bound, where some perturbations look partially modified (Fig.1(g)), while others look unchanged (Fig.1(c, e)), 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 i) Robustness against samples which belong to $OI(x)$, ii) Sensitivity towards samples which belong to $OS(x)$, with model's prediction matching the Oracle label, iii) No specification on samples which cannot be assigned an Oracle label. Given the practical difficulty in assigning Oracle labels during training and evaluation, we consider the following subset of these ideal goals in this work:

- Robustness-Accuracy trade-off, measured using accuracy on clean samples and robustness against valid attacks within the threat model
- Robustness against all attacks within an imperceptible radius ($\varepsilon = 8/255$ for CIFAR-10), measured using strong white-box attacks [5, 13]
- Robustness to Oracle-Invariant samples within a larger radius ($\varepsilon = 16/255$ for CIFAR-10), measured using gradient-free attacks [1, 3]

3. Proposed Method

In order to achieve the goals discussed in Sec.2.2, 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

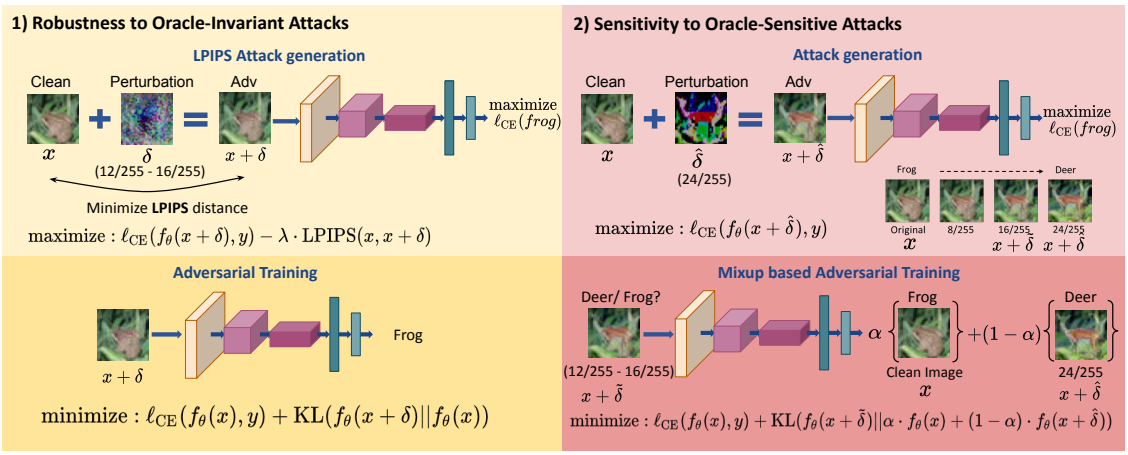

Figure 2. **Oracle-Aligned Adversarial Training:** The proposed defense OA-AT involves alternate training on Oracle-Invariant and Oracle-Sensitive samples. 1) Oracle-Invariant samples are generated by minimizing the LPIPS distance between the clean and perturbed images in addition to the maximization of the Classification Loss. 2) Oracle-Sensitive samples are trained using a convex combination of the predictions of the clean image and the perturbed image at a larger perturbation bound as reference in the KL divergence loss.

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 [16]. Adversarial examples generated using a robust model tend to look like the target (other) class images at large perturbation bounds, as seen in Fig.1(g, i, j). We therefore use large ε-bound 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 attacks Square [1] and Ray-S [3], they require a large number of queries (5000 to 10000), which is computationally expensive for use in adversarial training. Furthermore, 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 based on a study involving human annotators [10, 19]. Further, we observe that while the standard AlexNet model used in prior work [10] fails to distinguish between Oracle-Invariant and Oracle-Sensitive samples, an adversarially trained model is able to distinguish between the two effectively (Ref: Fig.2 of the Supplementary). 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: $\mathcal{L}_{CE}(x, y) - \lambda \cdot \text{LPIPS}(x, \hat{x})$. The ideal setting of λ is the minimum value that transforms attacks from Oracle-Sensitive to Oracle-Invariant (OI) for majority of the images. This results in the generation of strong Oracle-Invariant (OI) attacks. We present Oracle-Invariant examples for visual inspection in Fig.3 of the Supplementary.

Table 1. **Comparison with RobustBench Leaderboard [4] Results:** Performance (%) of the proposed method (OA-AT) when compared to AWP [17], which is the state-of-the-art amongst methods that do not use additional training data/ synthetic data on the RobustBench Leaderboard.

Method	Clean Acc	ℓ_∞ (AA) 8/255	ℓ_∞ (OI) 16/255	ℓ_2 (AA) $\varepsilon=0.5$	ℓ_2 (AA) $\varepsilon=1$	ℓ_1 (AA) $\varepsilon=5$	ℓ_0 (PGD$_0$) $\varepsilon=7$	Comm Corr	ReColor	ReColour+δ
CIFAR-10 (WRN-34-10)										
AWP	**85.36**	56.17	30.87	60.68	28.86	37.29	39.09	75.83	58.80	25.60
Ours	85.32	**58.04**	**35.31**	**64.08**	**34.54**	**45.72**	**44.40**	**76.78**	**70.50**	**39.90**
CIFAR-100 (WRN-34-10)										
AWP	62.73	29.59	14.96	36.62	17.05	21.88	17.40	50.73	37.60	12.40
Ours	**65.73**	**30.35**	**17.15**	**37.21**	**17.41**	**25.75**	**29.20**	**54.88**	**40.40**	**20.60**

Oracle-Aligned Adversarial Training (OA-AT): The training algorithm for the proposed defense, Oracle-Aligned Adversarial Training (OA-AT) is presented in Algorithm-1 of the Supplementary and illustrated in Fig.2. The maximum perturbation bound used for attack generation during training is denoted as ε_{max}. We use the AWP-TRADES formulation [17, 18] as the base implementation, with 10 steps of optimization for attack generation and one additional weight perturbation step. Classification loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) is maximized in the additional weight perturbation step (instead of $x_i + \widetilde{\delta}_i$ [17]), in order to achieve better smoothness in the loss surface. Initially, attacks constrained within a perturbation bound of $\varepsilon_{max}/4$ upto one-fourth the training epochs (Alg.1 of the Supplementary, L6-L8). The perturbation bound is increased linearly to ε_{max} at the last epoch alongside a co-

sine learning rate schedule. The use of a fixed epsilon initially helps in improving the adversarial robustness faster, while the use of an increasing epsilon schedule later results in better training stability [12]. We use 5 attack steps upto $\varepsilon_{max}/4$ to reduce computation, and 10 attack steps later.

Standard adversarial training is implemented upto a perturbation bound of $3/4 \cdot \varepsilon_{max}$, as the attacks in this range are imperceptible, based on the chosen moderate-magnitude threat model discussed in Sec.2.2. Beyond this, separate training losses are incorporated for Oracle-Invariant and Oracle-Sensitive samples in alternate training iterations (Alg.1 of the Supplementary, L9-L15), as shown in Fig.2. Oracle-Sensitive samples are generated by maximizing the classification 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 constrained within a bound of ε_{ref}, which is chosen to be greater than or equal to ε_{max} (Alg.1 of the Supplementary, L10). This component of the overall training loss is shown below:

$$KL\big(f_\theta(x_i + \widetilde{\delta}_i) \,||\, \alpha\, f_\theta(x_i) + (1-\alpha)\, f_\theta(x_i + \widehat{\delta}_i)\big) \quad (1)$$

Here $\widetilde{\delta}_i$ is the perturbation at the varying epsilon value $\widetilde{\varepsilon}$, and $\widehat{\delta}_i$ is the perturbation at ε_{ref}. This loss formulation results in better robustness-accuracy trade-off as shown in E1 versus E3 in Table-3 of the Supplementary. In the alternate iteration, we use the LPIPS metric to efficiently generate strong Oracle-Invariant attacks during training (Alg.1 of the Supplementary, L14). We perform exponential weight-averaging of the network being trained and use this for computing the LPIPS metric for improved and stable results (E1 versus E2 and F1 versus F2 in Table-3 of the Supplementary). We therefore do not need additional training or computation time for training this model. We increase α and λ over training, as the nature of attacks changes with varying $\widetilde{\varepsilon}$. The use of both Oracle-Invariant (OI) and Oracle-Sensitive (OS) samples ensures robustness to Oracle-Invariant samples while allowing sensitivity to partially Oracle-Sensitive samples.

4. Experiments and Results

We present a detailed comparison with respect to prior works on the CIFAR-10, CIFAR-100, SVHN and Imagenette datasets in Tables-2(a) and (b) of the Supplementary. We report adversarial robustness against the strongest known attacks, AutoAttack (AA) [5] and GAMA PGD-100 (GAMA) [13] 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 an ensemble of the Square [1] and Ray-S [3] attacks, as they generate strong Oracle-Invariant examples. We observe that the proposed defense achieves significant and consistent gains across all metrics specified in Sec.2.2. The proposed approach outperforms existing defenses by a significant margin on all four datasets, over different network architectures.

RobustBench Leaderboard Comparisons: As shown in Table-1, using the proposed method, we obtain a significant improvement over state-of-the-art results reported on the RobustBench Leaderboard (AWP) without the use of additional/ synthetic data on both CIFAR-10 and CIFAR-100 datasets. We observe that the proposed approach achieves significant gains against ℓ_∞ norm bound attacks at $\varepsilon = 8/255$ and $16/255$ that were used for training, as well as other ℓ_p norm bound attacks and common corruptions on both datasets. We also observe significant gains on functional threat models like ReColor [9], where the pixel values of an image are modified using a single function applied on all the pixels of the image. Further, combining ℓ_∞ attack with ReColor [9] (Recolor + δ) yields a stronger attack, and the proposed defense OA-AT achieves better robustness against this attack as well.

The training time of OA-AT is comparable with that of AWP [17]. On CIFAR-10, OA-AT takes 7 hours 16 minutes, while AWP takes 7 hours 27 minutes for 110 epochs of training on ResNet-18 using a single V100 GPU.

5. Conclusions

In this paper, we investigate robustness at large perturbation bounds in an ℓ_p norm based threat model. We discuss the ideal goals of an adversarial defense at large perturbation bounds, identify deficiencies of prior works in this setting and further propose a novel defense, Oracle-Aligned Adversarial Training (OA-AT) that aligns 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 state-of-the-art robustness at low and moderate perturbation bounds, and a better robustness-accuracy trade-off. We show the practical applicability of adversarial training at larger perturbation bounds by demonstrating significant improvements against common corruptions, other ℓ_p norm bound attacks (ℓ_2, ℓ_1, ℓ_0) unseen during training, and recolor attacks.

6. Acknowledgements

This work was supported by a research grant (CRG/2021/005925) from SERB, DST, Govt. of India. Sravanti Addepalli is supported by Google PhD Fellowship and CII-SERB Prime Minister's Fellowship for Doctoral Research. We are thankful for the support.

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 1, 2, 3, 4

[2] 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. 1

[3] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *KDD*, 2020. 2, 3, 4

[4] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS Datasets and Benchmarks Track (Round 2)*, 2021. 3

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 2, 4

[6] Ian Goodfellow and Nicolas Papernot. Is attacking machine learning easier than defending it? Blog post on Feb 15, 2017. 1

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1

[8] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 2

[9] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 4

[10] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *ICLR*, 2021. 2, 3

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2

[12] Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020. 4

[13] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *NeurIPS*, 2020. 1, 2, 4

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2013. 1

[15] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *ICML*, 2020. 1, 2

[16] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2019. 3

[17] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *NeurIPS*, 2020. 2, 3, 4

[18] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 2, 3

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3