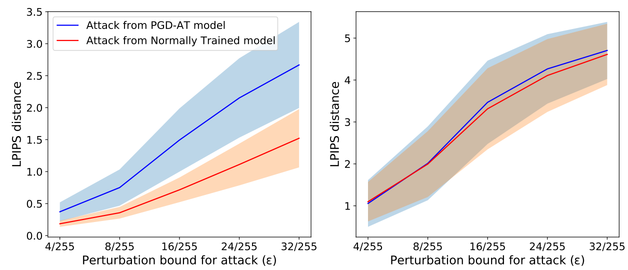
Scaling Adversarial Training to Large Perturbation Bounds

Sravanti Addepalli*, Samyak Jain*, Gaurang Sriramanan, R.Venkatesh Babu
Video Analytics Lab, Department of Computational and Data Sciences
Indian Institute of Science, Bangalore, India

Algorithm 1 Oracle-Aligned Adversarial Training

Input: Deep Neural Network f_θ with parameters θ, Training Data $\{x_i, y_i\}_{i=1}^{M}$, Epochs T, Learning Rate η, Perturbation budget ε_{max}, Adversarial Perturbation function $A(x, y, \ell, \varepsilon)$ which maximises loss ℓ
for epoch $= 1$ **to** T **do**
  $\widetilde{\varepsilon} = \max\{\varepsilon_{max}/4, \varepsilon_{max} \cdot \text{epoch}/T\}$
  for $i = 1$ **to** M **do**
    $\delta_i \sim U(-\min(\widetilde{\varepsilon}, \varepsilon_{max}/4), \min(\widetilde{\varepsilon}, \varepsilon_{max}/4))$
    if $\widetilde{\varepsilon} < 3/4 \cdot \varepsilon_{max}$ **then**
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$, $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
      $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) || f_\theta(x_i)\right)$
    else if $i \,\%\, 2 = 0$ **then**
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i)$, $\widehat{\delta}_i = A(x_i, y_i, \ell, \varepsilon_{ref})$, $\widetilde{\delta}_i = \Pi_\infty(\widehat{\delta}_i, \widetilde{\varepsilon})$
      $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \;||\; \alpha \cdot f_\theta(x_i) + (1 - \alpha) \cdot f_\theta(x_i + \widehat{\delta}_i)\right)$
    else
      $\delta_i \sim U(-\widetilde{\varepsilon}, \widetilde{\varepsilon})$
      $\ell = \ell_{CE}(f_\theta(x_i + \delta_i), y_i) - \text{LPIPS}(x_i, x_i + \delta_i)$, $\widetilde{\delta}_i = A(x_i, y_i, \ell, \widetilde{\varepsilon})$
      $L_{adv} = \text{KL}\left(f_\theta(x_i + \widetilde{\delta}_i) \;||\; f_\theta(x_i)\right)$
    $L = \ell_{CE}(f_\theta(x_i), y_i) + L_{adv}$
    $\theta = \theta - \eta \cdot \nabla_\theta L$



(a) PGD-AT LPIPS model (b) Normally Trained LPIPS model

Figure 2. LPIPS distance between clean and adversarially perturbed images. Attacks generated from PGD-AT [7, 8] model (Oracle-Sensitive) and Normally Trained model (Oracle-Invariant) are considered. (a) PGD-AT ResNet-18 model is used for computation of LPIPS distance (b) Normally Trained AlexNet model is used for computation of LPIPS distance. PGD-AT model based LPIPS distance is useful to distinguish between Oracle-Sensitive and Oracle-Invariant attacks.

1. Related Works

Robustness against imperceptible attacks: Following the discovery of adversarial examples by Szegedy et al., [15], a myriad of adversarial attack and defense methods have been proposed. Adversarial Training has emerged as the most successful defense strategy against ℓ_p norm bound imperceptible attacks. PGD Adversarial Training (PGD-AT) proposed by Madry et al. [7] 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 [8,10,13,16,18,19] that improved accuracy against such imperceptible threat models further.

Zhang et al. [18] 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. [16] 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 [14], 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 [6] is also seen to improve the flatness of loss surface, resulting in a boost in adversarial robustness [3,5]. Recent works attempt to use training techniques such as early stopping [10], optimal weight decay [8], Cutmix data augmentation [9, 17] and label smoothing [9] to

*Equal contribution

Table 1. **CIFAR-10: Standard Adversarial Training using Large-ε:** Performance (%) of various existing Defenses trained using $\varepsilon = 8/255$ or $16/255$ against attacks bound within $\varepsilon = 8/255$ and $16/255$. A large drop in clean accuracy is observed with existing approaches [7, 16, 18, 19] when trained using perturbations with $\varepsilon = 16/255$.

Method	Attack ε (Training)	Clean Acc	GAMA (8/255)	AA (8/255)	GAMA (16/255)	Square (16/255)
TRADES	8/255	80.53	49.63	49.42	19.27	27.82
TRADES	16/255	75.30	35.64	35.12	10.10	18.87
AWP	8/255	80.47	**50.06**	**49.87**	19.66	28.51
AWP	16/255	71.63	40.85	40.55	15.92	24.16
PGD-AT	8/255	81.12	49.03	48.58	15.77	26.47
PGD-AT	16/255	64.93	46.66	46.21	**26.73**	**32.25**
FAT	8/255	**84.36**	48.41	48.14	15.18	25.07
FAT	16/255	75.27	47.68	47.34	22.93	29.47

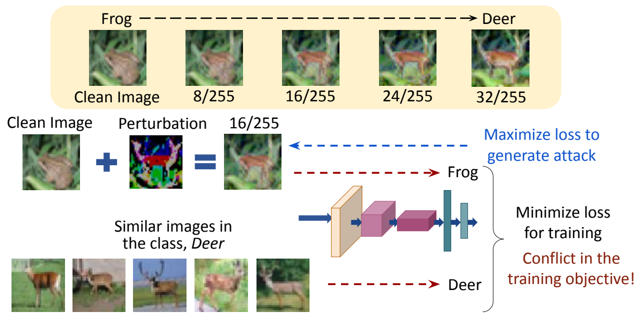


Figure 1. **Issues with Standard Adversarial Training at Large-ε:** An adversarial example generated from the original image of a frog looks partially like a deer at an ℓ_∞ bound of $16/255$, but is trained to predict the true label, Frog. This induces a conflicting objective, leading to a large drop in clean accuracy.

Table 2. **Comparison with existing methods:** Performance (%) of the proposed defense OA-AT when compared to baselines against the attacks, GAMA-PGD100 [13], AutoAttack (AA) [4] and an ensemble of Square [1] and Ray-S [2] attacks (SQ+RS), with different ε bounds. Sorted by AutoAttack (AA) accuracy at $\varepsilon = 8/255$ for CIFAR-10, CIFAR-100 and Imagenette, and $4/255$ for SVHN.

(a) **CIFAR-10, SVHN**

Method	Metrics of interest: Clean	GAMA 8/255	AA 8/255	SQ+RS 16/255	Others: GAMA 16/255	AA 16/255
CIFAR-10 (ResNet-18), 110 epochs						
FAT	**84.36**	48.41	48.14	23.22	15.18	14.22
PGD-AT	79.38	49.28	48.68	25.43	18.18	17.00
AWP	80.32	49.06	48.89	25.99	19.17	18.77
ATES	80.95	49.57	49.12	26.43	18.36	16.30
TRADES	80.53	49.63	49.42	26.20	19.27	18.23
ExAT + PGD	80.68	50.06	49.52	25.13	17.81	19.53
ExAT + AWP	80.18	49.87	49.69	27.04	20.04	16.67
AWP	80.47	50.06	49.87	27.20	19.66	19.23
Ours	80.24	**51.40**	**50.88**	**29.56**	**22.73**	**22.05**
CIFAR-10 (ResNet-34), 110 epochs						
AWP	83.89	52.64	52.44	27.69	20.23	19.69
OA-AT (Ours)	**84.07**	**53.54**	**53.22**	**30.76**	**22.67**	**22.00**
CIFAR-10 (WRN-34-10), 200 epochs						
AWP	85.36	56.34	56.17	30.87	23.74	23.11
OA-AT (Ours)	**85.32**	**58.48**	**58.04**	**35.31**	**26.93**	**26.57**
SVHN (PreActResNet-18), 110 epochs						
Method	**Clean**	**GAMA 4/255**	**AA 4/255**	**SQ+RS 12/255**	**GAMA 12/255**	**AA 12/255**
AWP	91.91	75.92	75.72	35.49	30.70	30.31
OA-AT (Ours)	**94.61**	**78.37**	**77.96**	**39.24**	**34.25**	**33.63**

(b) **CIFAR-100, ImageNette**

Method	Metrics of interest: Clean	GAMA 8/255	AA 8/255	SQ+RS 16/255	Others: GAMA 16/255	AA 16/255
CIFAR-100 (ResNet-18), 110 epochs						
AWP	58.81	25.51	25.30	11.39	8.68	8.29
AWP+	59.88	25.81	25.52	11.85	8.72	8.28
OA-AT (no LS)	60.27	26.41	26.00	13.48	**10.47**	**9.95**
OA-AT (Ours)	**61.70**	**27.09**	**26.77**	**13.87**	10.40	9.91
CIFAR-100 (PreActResNet-18), 200 epochs						
AWP	58.85	25.58	25.18	11.29	8.63	8.19
AWP+	**62.11**	26.21	25.74	12.23	9.21	8.55
OA-AT (Ours)	62.02	**27.45**	**27.14**	**14.52**	**10.64**	**10.10**
CIFAR-100 (WRN-34-10), 110 epochs						
AWP	62.41	29.70	29.54	14.25	11.06	10.63
AWP+	62.73	29.92	29.59	14.96	11.55	11.04
OA-AT (no LS)	65.22	30.75	**30.35**	16.77	12.65	11.95
OA-AT (Ours)	**65.73**	**30.90**	**30.35**	**17.15**	**13.21**	**12.01**
Imagenette (ResNet-18), 110 epochs						
Method	**Clean**	**GAMA 8/255**	**AA 8/255**	**SQ+RS 16/255**	**GAMA 16/255**	**AA 16/255**
AWP	82.73	57.52	57.40	42.52	29.14	28.86
OA-AT (Ours)	**82.98**	**59.51**	**59.31**	**48.01**	**48.66**	**31.78**

achieve enhanced robust performance on base defenses such as PGD-AT [7] and TRADES [18]. We utilize some of these methods in our approach, and also present improved baselines by combining AWP-TRADES [16] with these enhancements.

Robustness against large perturbation attacks:

Table 3. **CIFAR-10, CIFAR-100**: Ablation experiments on ResNet-18 architecture (E1-E7) and WideResNet-34-10 (F1-F2) architecture to highlight the importance of various aspects in the proposed defense OA-AT. Performance (%) against attacks with different ε bounds is reported.

Method	**CIFAR-10** Clean	**CIFAR-10** GAMA (8/255)	**CIFAR-10** GAMA (16/255)	**CIFAR-10** Square (16/255)	**CIFAR-100** Clean	**CIFAR-100** GAMA (8/255)	**CIFAR-100** GAMA (16/255)	**CIFAR-100** Square (16/255)
E1: OA-AT (Ours)	80.24	**51.40**	22.73	31.16	60.27	**26.41**	10.47	14.60
E2: LPIPS weight = 0	78.47	50.60	24.05	31.37	58.47	25.94	10.91	14.66
E3: Alpha = 1	79.29	50.60	23.65	31.23	58.84	26.15	10.97	14.89
E4: Alpha = 1, LPIPS weight = 0	77.16	50.49	**24.93**	**32.01**	57.77	25.92	**11.33**	**15.03**
E5: Using Current model (without WA) for LPIPS	**80.50**	50.75	22.90	30.76	59.54	26.23	10.50	14.86
E6: Without 2*eps perturbations for AWP	79.96	50.50	22.61	30.60	60.18	26.27	10.15	14.20
E7: Maximizing KL div in the AWP step	81.19	49.77	21.17	29.39	59.48	25.03	7.93	13.34
F1: OA-AT (Ours)	**85.32**	**58.48**	26.93	**36.93**	**65.73**	**30.90**	13.21	**18.47**
F2: LPIPS weight = 0	83.47	57.58	**27.21**	36.68	63.16	30.22	**13.59**	18.42

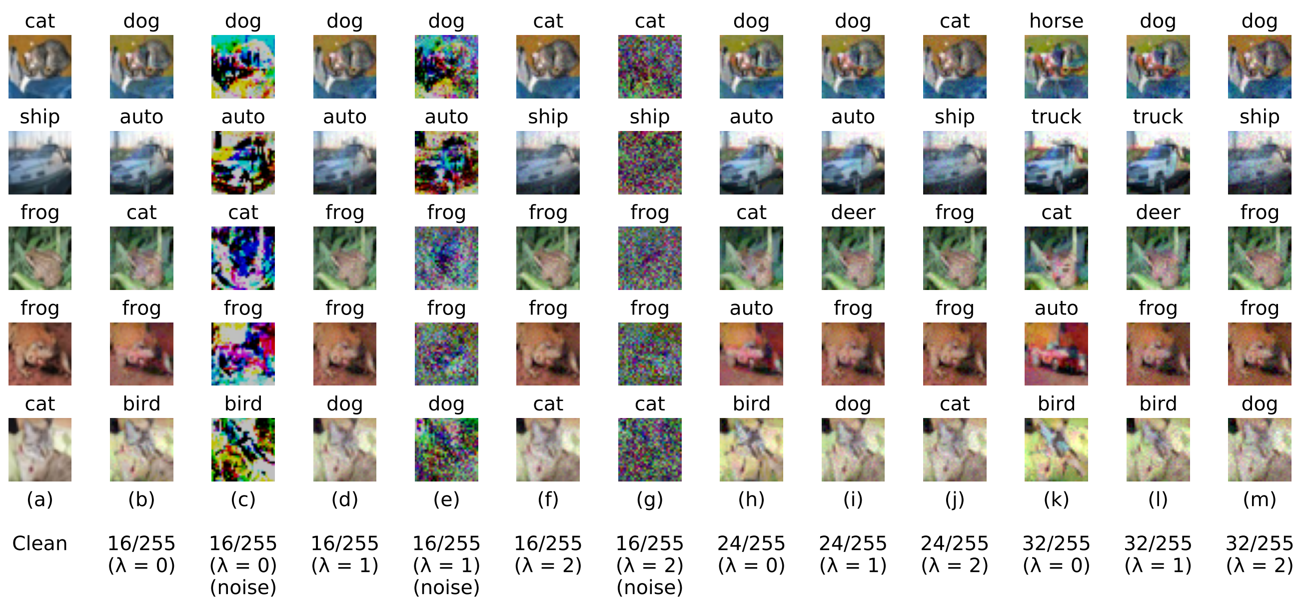


Figure 3. Oracle-Invariant adversarial examples generated using the LPIPS based PGD attack across various perturbation bounds. White-box attacks and predictions on the model trained using the proposed OA-AT defense on the CIFAR-10 dataset with ResNet-18 architecture are shown: (a) Original Unperturbed image, (b, h, k) Adversarial examples generated using the standard PGD 10-step attack, (d, f, i, j, l, m) LPIPS based PGD attack generated within perturbation bounds of $16/255$ (d, f), $24/255$ (i, j) and $32/255$ (l, m) by setting the value of λ_{LPIPS} to 1 and 2, (c, e, g) Perturbations corresponding to (b), (d) and (f) respectively.

Shaeiri et al. [11] 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 just 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) [19] 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. Along similar lines, Sitawarin et al. [12] propose Adversarial Training with Early Stopping (ATES), which performs early stopping of a PGD attack based on the margin (difference between true and maximum probability class softmax outputs) of the perturbed image being greater than a threshold that is increased over epochs.

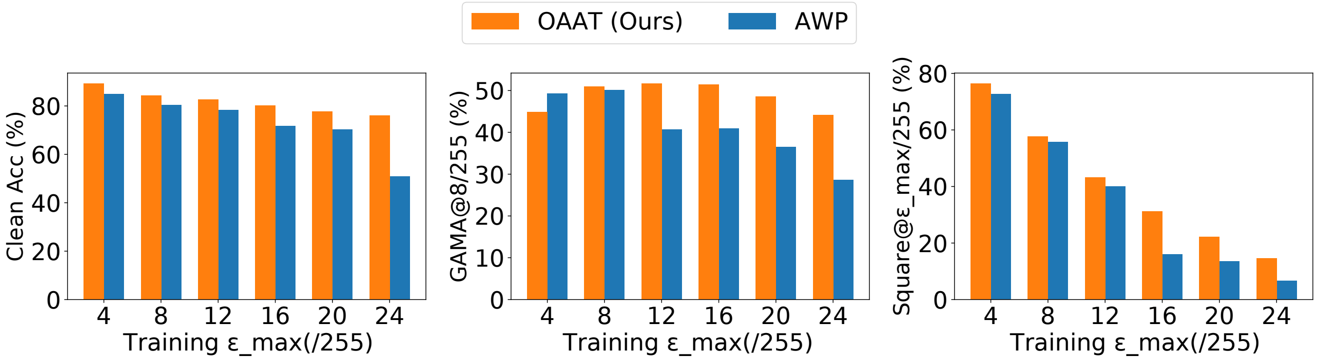


Figure 4. **Results across variation in training ε_{max}:** While the proposed approach works best at moderate-ε bounds such as 16/255 on CIFAR-10, we observe that it outperforms the baseline for various ε_{max} values $\geq 8/255$ as well.

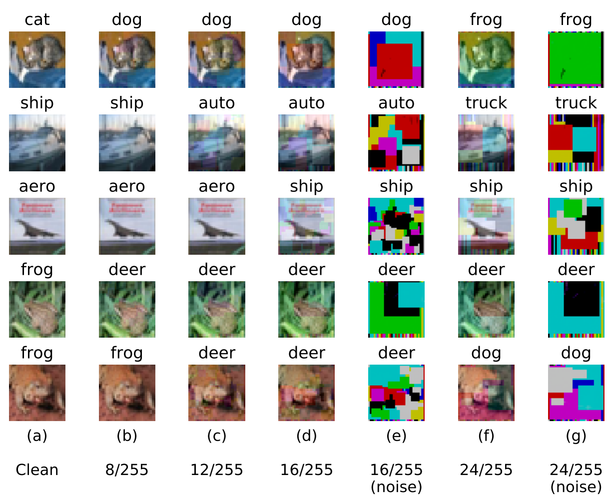


Figure 5. **Square attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various ℓ_∞ bounds generated using the gradient-free random search based attack Square [1]. The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

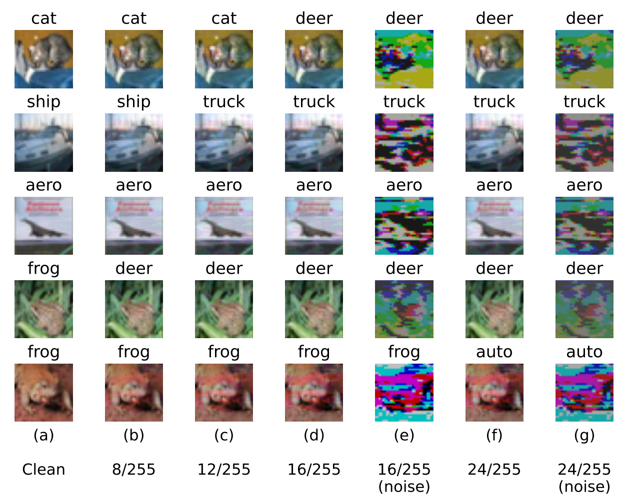


Figure 6. **RayS attack:** Adversarially attacked images (b, c, d, f) and the corresponding perturbations (e, g) for various ℓ_∞ bounds generated using the gradient-free binary search based attack RayS [2]. The clean image is shown in (a). Attacks are generated from a model trained using the proposed Oracle-Aligned Adversarial Training (OA-AT) algorithm on CIFAR-10. Prediction of the same model is printed above each image.

We compare against these methods and improve upon them significantly using our proposed approach.

2. Ablation Study

In order to study the impact of different components of the proposed defense, we present a detailed ablative study using ResNet-18 models in Table-3. We present results on the CIFAR-10 and CIFAR-100 datasets, with E1 representing the proposed approach. First, we study the efficacy of the LPIPS metric in generating Oracle-Invariant attacks. In experiment E2, we train a model without LPIPS by setting its coefficient to zero. While the resulting model achieves a slight boost in robust accuracy at $\varepsilon = 16/255$ due to the use of stronger attacks for training, there is a considerable drop in clean accuracy, and a corresponding drop in robust accuracy at $\varepsilon = 8/255$ as well. We observe a similar trend by setting the value of α to 1 as shown in E3, and by combining E2 and E3 as shown in E4. We note that E4 is similar to standard adversarial training, where the model attempts to learn consistent predictions in the ε ball around every data sample. While this works well for large ε attacks ($\varepsilon = 16/255$), it leads to poor clean accuracy.

As discussed in Sec.3 of the Main paper, we maximize loss on $x_i + 2 \cdot \widetilde{\delta}_i$ (where $\widetilde{\delta}_i$ is the attack) in the additional weight perturbation step. We present results by using the standard ε limit for the weight perturbation step as well, in E6. This leads to a drop across all metrics, indicating the importance of using large magnitude perturbations in the weight perturbation step for producing a flatter loss surface that leads to better generalization to the test set. Different from the standard TRADES formulation, we maximize Cross-Entropy loss for attack generation in the proposed method. From E7, we note that the use of KL divergence leads to a drop in robust accuracy since the KL divergence based attack is weaker. This is consistent with the observation by Gowal et al. [5]. However, on the SVHN dataset, we find that the use of KL divergence based attack results in a significant improvement in clean accuracy, leading to better robust accuracy as well. We therefore utilize the KL divergence loss for attack generation on the SVHN dataset.

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. 2, 4

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

[3] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*, 2020. 1

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

[5] 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. 1, 5

[6] 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. 1

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Tsipras Dimitris, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2

[8] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *ICLR*, 2021. 1

[9] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021. 1

[10] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 1

[11] Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020. 2

[12] Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020. 3

[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

[14] David Stutz, Matthias Hein, and Bernt Schiele. Relating adversarially robust generalization to flat minima. In *ICCV*, 2021. 1

[15] 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

[16] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *NeurIPS*, 2020. 1, 2

[17] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1

[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. 1, 2

[19] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020. 1, 2, 3