Efficient and Effective Augmentation Strategy for Adversarial Training

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

Abstract

The sample complexity of Adversarial training is known to be significantly higher than standard ERM training. Although complex augmentation techniques have led to large gains in standard training, they have not been successful with Adversarial Training. In this work, we propose Diverse Augmentation based Joint Adversarial Training (DAJAT) that uses a combination of simple and complex augmentations with separate batch normalization layers to handle the conflicting goals of enhancing the diversity of the training dataset, while being close to the test distribution. We further introduce a Jensen-Shannon divergence loss to encourage the joint learning of the diverse augmentations, thereby allowing simple augmentations to guide the learning of complex ones. Lastly, to improve the computational efficiency of the proposed method, we propose and utilize a two-step defense, Ascending Constraint Adversarial Training (ACAT) that uses an increasing epsilon schedule and weight-space smoothing to prevent gradient masking.

1. Introduction

While early Adversarial defenses focused on designing suitable loss functions for training [10, 20], subsequent works [14] observed that adversarial training has a large sample complexity, and further gains require the use of additional training data that is closely related to the original data distribution [2, 6]. The large data requirement, which is impractical to assume, has led to an exploration towards augmentations based on Generative Adversarial Networks and Diffusion based models [7, 11]. However, the use of such generative models incurs an additional training cost, and suffers from limited diversity, specifically in low-data regimes and in datasets with high resolution images.

While standard Empirical Risk Minimization (ERM) based training also benefits from the use of additional data, the most practical and efficient approach for augmenting the training dataset has been the use of a series of random transformations such as Random Crop, Random Rotation, Color Jitter, contrast, sharpness and brightness adjustments [4]. These augmentations can change the images significantly in input space while belonging to the same class as the original image. However, prior works [5, 13, 18] have surprisingly found that such complex augmentations that cause large changes in the input distribution cannot help adversarial training. Thus, the commonly used augmentations in adversarial training are the simple transformations, padding followed by random crop, and horizontal flip [5, 13].

In this work, we show that it is indeed possible to utilize complex augmentations effectively in Adversarial training as well, by jointly training on simple and complex data augmentations using separate batch-normalization layers for each kind of augmentation. While complex augmentations increase the data diversity resulting in better generalization, simple augmentations ensure that the model specializes on the training data distribution as well, leading to direct gains at inference time. We further minimize the Jensen-Shannon divergence between the softmax outputs of various augmentations to enable the simple augmentations to guide the learning of complex ones. In order to improve the computational efficiency of the proposed method, we use two attack steps (instead of 10 steps) during training, while progressively increasing the magnitude of perturbations and performing smoothing in weight space to improve the stability of training. Our contributions are listed below:

- We propose an efficient two-step defense, Ascending Constraint Adversarial Training (ACAT) with a linearly increasing ε schedule, cosine learning rate and weight-space smoothing to prevent gradient masking.
- We propose Diverse Augmentation based Joint Adversarial Training (DAJAT) that effectively combines the benefits of simple and complex augmentations to obtain significant gains in performance.

2. Ascending Constraint Adversarial Training

Prior works [15] have shown that training convergence at large ℓ_∞ norm bounds can be improved by linearly increasing the perturbation radius ε as training progresses. Inspired by this, we propose Ascending Constraint Adversarial Training (ACAT) that utilizes an increasing ε schedule

[*] Equal contribution

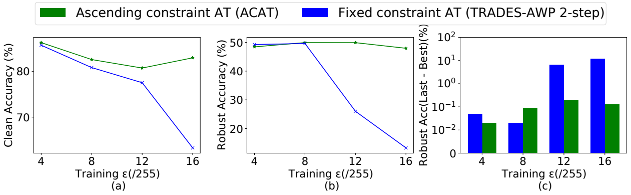

Figure 1. **Comparison of the proposed 2-step defense ACAT against TRADES-AWP [19] 2-step baseline** on the CIFAR-10 with ResNet-18 architecture. ACAT has significantly better performance and stability even at large training ε values. Robust Accuracy is reported against GAMA attack [16] with $\varepsilon = 8/255$

alongside a cosine learning rate schedule with TRADES-AWP [19] loss formulation for improving the stability and convergence of two-step adversarial training. We use a cosine learning rate schedule that decays monotonically over the training epochs, since at large training ε, lower learning rate could further stabilize training. As shown in Fig.1, the performance and stability of the proposed 2-step defense ACAT are significantly better when compared to the TRADES-AWP 2-step baseline, at the same computational cost, specifically at larger perturbations bounds of $12/255$ and $16/255$. The proposed defense maintains a good clean accuracy at all the training ε values considered, and has almost 0 difference between best and last epochs.

3. Diverse Augmentation based Joint Adversarial Training (DAJAT)

The use of augmentations in training can be viewed as a problem of domain generalization, where performance on the source distribution or augmented dataset is crucial towards improving the performance on the target distribution or test set [1]. Since adversarial training is inherently challenging, for limited model capacity it is difficult to obtain good performance on the training data that is transformed using complex augmentations. Moreover, the large distribution shift between augmented data and test data, specifically with respect to low-level statistics, results in poor generalization of robust accuracy to the test set. Specifically, since adversarial attacks perturb images in pixel space, and there is a large difference between the distributions of augmented and test data in input space, it is likely that unless this difference is accounted for, complex augmentations cannot improve the performance of adversarial training. This trend has also been observed empirically by Rebuffi et al. [12], based on which they conclude that the augmentations designed for robustness need to preserve low-level features.

To mitigate these challenges, we propose the combined use of simple and complex augmentations during training so that the model can benefit from the diversity introduced by complex augmentations, while also specializing on the original data distribution that is similar to the simple augmentations. We propose to use separate batch normalization layers for simple and complex augmentations, so as to offset the shift in distribution between the two kinds of augmentations. We additionally minimize the Jensen-Shannon divergence between the softmax outputs of different augmentations, so as to allow the simple augmentations to guide the learning of complex ones.

DAJAT Algorithm: Firstly, the TRADES loss [20] is computed on each of the augmentations of every image x. As shown in the equation below, this loss is a combination of cross-entropy loss on the natural image x and the KL divergence between the softmax predictions of the natural image x and the adversarially perturbed image $\tilde{x}$. The KL divergence term is weighted by a factor β that controls the overall robustness-accuracy trade-off.

$$\mathcal{L}_{\mathrm{T}}(\theta, x, y) = \mathcal{L}_{\mathrm{CE}}(f_\theta(x), y) + \beta \max_{\tilde{x} \in \mathcal{A}(x)} \mathrm{KL}(f_\theta(x) || f_\theta(\tilde{x}))$$

Different from TRADES Adversarial training, we compute $\tilde{x}$ using two attack steps with a step-size of ε. As discussed in Section-2, we use a combination of a linearly increasing schedule of ε alongside a cosine learning rate schedule in order to improve the stability and performance of adversarial training. We additionally use model weight-averaging to improve generalization of the network, as is common in literature [9].

The overall DAJAT loss is a combination of the TRADES 2-step loss on each of the augmentations, x_{base} and $x_{\mathrm{auto(t)}}$ along with an adversarial weight perturbation step [19] on the loss corresponding to the base augmentations alone, to improve training efficiency. For every batch normalization layer, two sets of running statistics and affine parameters are maintained and used for simple and complex augmentations respectively. The computation of the Adversarial Weight Perturbation model $\tilde{\theta}$ within the constraint set $\mathcal{M}(\theta)$ [19] is shown below:

$$\tilde{\theta} = \operatorname*{arg\,max}_{\tilde{\theta} \in \mathcal{M}(\theta)} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{T}}(\tilde{\theta}, x_{i,\mathrm{base}}, y_i)$$

The overall DAJAT loss is shown below:

$$\begin{aligned}\min_{\tilde{\theta}} \frac{1}{N} \sum_{i=1}^{N} \{ & \mathcal{L}_{\mathrm{T}}(\tilde{\theta}, x_{i,\mathrm{base}}, y_i) + \sum_{t=1}^{T} \mathcal{L}_{\mathrm{T}}(\tilde{\theta}, x_{i,\mathrm{auto(t)}}, y_i) \\ & + \mathrm{JSD}(f_{\tilde{\theta}}(x_{i,\mathrm{base}}), f_{\tilde{\theta}}(x_{i,\mathrm{auto(1)}}), \ldots, f_{\tilde{\theta}}(x_{i,\mathrm{auto(T)}})) \}\end{aligned}$$

While in our main algorithm we use AutoAugment [4], that uses Proximal Policy Optimization to find the set of policies that can yield optimal performance on a given dataset for standard training, we find that the proposed approach works well with other augmentations as well. The role of the base augmentations is primarily to learn the batch normalization layers that would be used during inference time, and also to provide better supervision for the training of complex augmentations using the JS divergence term. The role of the complex augmentations is to enhance the diversity of the training dataset. Therefore we use a single base aug-

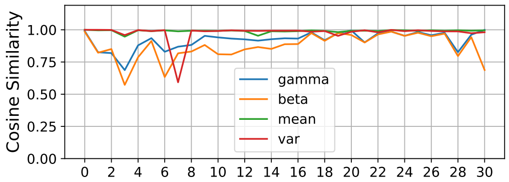

Figure 2. **Cosine Similarity of the two sets of Batch Normalization layer statistics** for a WideResNet-34-10 model trained on CIFAR-10 using the proposed DAJAT defense (Base, 2*AA). Batch normalization layers corresponding to the Base augmentations (Pad+Crop,H-Flip) are compared with those of AutoAugment. Parameters of initial layer (Layer-3) channels are diverse, while those of deeper layers (Layer-25) are similar.

Table 1. Performance of the proposed defenses ACAT and DAJAT when compared to state-of-the-art defenses on **CIFAR-10 dataset**. Robust evaluations are done on GAMA [16] and AutoAttack [3]

		CIFAR-10, ResNet-18				**CIFAR-10, WideResNet-34**		
	Steps	**Clean**	**GAMA**	**AutoAttack**	**Train time/ epoch (sec)**	**Clean**	**GAMA**	**AutoAttack**
NuAT2-WA	2	82.21	50.97	50.75	109	86.32	55.08	54.76
ACAT, Ours (Base, 2step)	2	82.41	50.00	49.80	95	86.71	55.58	55.36
PGD-AT	10	81.12	49.08	48.75	182	86.07	52.70	52.19
TRADES-AWP	10	80.47	50.06	49.87	228	85.19	55.87	55.69
TRADES-AWP (200 epochs)	10	81.99	51.65	51.45	228	85.36	56.35	56.17
TRADES-AWP-WA	10	80.41	49.89	49.67	228	85.10	56.07	55.87
DAJAT, Ours (Base, AA)	2 + 2	85.60	51.27	51.06	160	87.87	56.97	56.68
DAJAT, Ours (Base, 2*AA)	2 + 4	85.99	51.71	51.48	219	**88.90**	57.22	56.96
DAJAT, Ours (Base, 3*AA)	2 + 6	**86.67**	**51.81**	**51.56**	280	88.64	**57.34**	**57.05**

mentation and multiple complex augmentations. The gains in performance saturate with the addition of more complex augmentations, and therefore the use of a single base augmentation and two complex augmentations achieves the best performance-accuracy trade-off. We note from Table-1 that in this setting, the computational complexity of the proposed method is on par with the TRADES-AWP defense which is the current state-of-the-art approach, while achieving considerable performance gains.

Split Batch Normalization Layers for Different Augmentations: The proposed defense DAJAT uses separate batch normalization layers for simple and complex augmentations as discussed above. A Batch Normalization (BN) layer is implemented as follows on a given feature map $g(x_i)$ of the input image x_i: $\hat{g}(x_i) = \frac{g(x_i)-\mu}{\sigma} \cdot \gamma + \beta$ Here, μ and σ denote the mean and standard deviation of the current mini-batch during training. During inference, these are set to the running mean and variance computed during training. γ and β are parameters of the network that are trained. In the proposed approach we maintain two sets of batch normalization statistics μ and σ, and two sets of affine parameters, β and γ for every batch normalization layer.

We plot the cosine similarity between the batch normalization vectors corresponding to the base augmentations and autoaugment of every layer in Fig.2. While the mean and variance of the batch normalization have a high similarity across all layers, we note significant differences in the γ and β values, specifically in the initial layers. This shows that the difference in low-level statistics between the two distributions of images are being offset effectively by incorporating separate batch normalization layers. The network learns more similar parameters in deeper layers since the feature representations of different types of augmentations are expected to be more aligned in these layers.

4. Experiments and Results

We compare the proposed approach against several state-of-the-art defenses in Tables-1 and 2 on CIFAR-10, CIFAR-100 and ImageNette [8]. We integrate model weight averaging with the TRADES-AWP baseline (termed as TRADES-AWP-WA) as well for a fair comparison.

Firstly, we compare the performance of the proposed 2-step defense ACAT with the existing state-of-the-art 2-step defense NuAT-WA [17] in the first partition of Table-1. While we achieve similar performance on ResNet-18, we obtain a marginal boost in both clean and robust accuracy on WideResNet-34 architecture. We note that our proposed defense ACAT can be integrated with the Nuclear Norm training objective as well to obtain improved results. The performance of the proposed ACAT defense is superior when compared to the multi-step training method PGD-AT [13] as well. When compared to the TRADES-AWP 10-step defense [19, 20], we obtain improved clean accuracy with a slight drop in robust accuracy at half the computational cost. On the CIFAR-100 dataset, we obtain substantial gains in both clean and robust accuracy when compared to the 10-step baselines.

We present three variants of the proposed defense DAJAT, by using one, two and three AutoAugment based augmentations for every image. We denote them as DAJAT(Base, AA), DAJAT(Base, 2*AA) and DAJAT(Base, 3*AA) respectively. Using a single AutoAugment based augmentaion (Base, AA), we obtain improved clean and robust accuracy when compared to most of the baselines considered across all datasets and models. By increasing the number of AutoAugment based transformations to 2, we observe consistent gains in robust and clean accuracy in all cases. In this setting, the computational complexity of the proposed approach matches with that of TRADES-AWP [19] as shown in Table-1. With the setting (Base,3*AA), we obtain marginal improvements in performance.

Overall, using the (Base, 2*AA) approach, which has comparable time complexity as the TRADES-AWP 10-step defense, we obtain large gains ranging from 3.8% to 7% on clean accuracy and around 1.8% higher robust accuracy against AutoAttack [3] across most settings. On the Imagenette dataset [8] we obtain 5.48% higher clean accuracy and 5.42% higher robust accuracy, showing that augmentation strategies work best when the amount of training data is less when compared to the complexity of the task.

Table 2. Performance (%) on **CIFAR-100 and ImageNette [8] datasets** against GAMA attack [16] and AutoAttack [3]

		CIFAR-100, ResNet-18			**CIFAR-100, WideResNet-34**			**IN-10, ResNet-18**		
	No. of Steps	**Clean**	**GAMA**	**AutoAttack**	**Clean**	**GAMA**	**AutoAttack**	**Clean**	**GAMA**	**AutoAttack**
TRADES-AWP	10	58.81	25.51	25.30	62.41	29.70	29.54	82.73	57.52	57.40
TRADES-AWP-WA	10	59.88	25.81	25.52	62.73	29.92	29.59	82.03	57.04	56.89
ACAT, Ours (Base, 2step)	2	62.05	26.35	26.10	65.75	30.61	30.23	82.34	57.12	56.96
DAJAT, Ours (Base, AA)	2 + 2	65.75	27.58	27.21	67.82	**31.65**	31.26	85.27	61.50	61.19
DAJAT, Ours (Base, 2*AA)	2 + 4	66.84	27.61	27.32	68.74	31.58	**31.30**	86.01	**62.52**	**62.31**
DAJAT, Ours (Base, 3*AA)	2 + 6	**66.96**	**27.90**	**27.62**	**70.35**	31.15	30.89	**86.92**	62.14	61.89

5. Conclusions

Contrary to prior knowledge, we show that it is possible to use common augmentation strategies that modify the low-level statistics of images, in adversarial training as well. We propose a novel defense Diverse Augmentation based Joint Adversarial Training (DAJAT) that uses a combination of simple and complex augmentations with separate batch normalization layers, in order to benefit from complex augmentations, while also being trained on a distribution that is close to the test set. The use of JS divergence term between network predictions of different augmentations enables the joint learning across various augmentations. We improve the efficiency of the proposed defense by utilizing the proposed Ascending Constraint Adversarial Training (ACAT) that improves the stability and performance of TRADES 2-step adversarial training significantly by using a linearly increasing ε schedule along with a cosine learning rate schedule and weight-space smoothing. We believe this work can open up further possibilities towards finding better data augmentations for adversarial training.

6. Acknowledgements

This work was supported by Uchhatar Avishkar Yojana (UAY) project (IISC_10), MHRD, 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] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 2

[2] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019. 1

[3] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 3, 4

[4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv:1805.09501*, 2018. 1, 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

[6] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy Mann. Improving robustness using generated data. *NeurIPS*, 34, 2021. 1

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 1

[8] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020. 3, 4

[9] Pavel Izmailov, Dmitrii Podoprikhin, T. Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *ArXiv*, abs/1803.05407, 2018. 2

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

[11] 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

[12] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *NeurIPS*, 34, 2021. 2

[13] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 1, 3

[14] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *NeurIPS*, 31, 2018. 1

[15] Amirreza Shaeiri, Rozhin Nobahari, and Mohammad Hossein Rohban. Towards deep learning models resistant to large perturbations. *arXiv preprint arXiv:2003.13370*, 2020. 1

[16] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *NeurIPS*, 2020. 2, 3, 4

[17] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Towards Efficient and Effective Adversarial Training. In *NeurIPS*, 2021. 3

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

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

[20] 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, 3