A Simple Fine-tuning Is All You Need: Towards Robust Deep Learning Via Adversarial Fine-tuning

Ahmadreza Jeddi, Mohammad Javad Shafiee, Alexander Wong
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada
{a2jeddi, mjshafiee, alexander.wong}@uwaterloo.ca

Abstract

Adversarial Training (AT) with Projected Gradient Descent (PGD) is an effective approach for improving the robustness of the deep neural networks. However, PGD AT has been shown to suffer from two main limitations: i) high computational cost, and ii) extreme overfitting during training that leads to reduction in model generalization. While the effect of factors such as model capacity and scale of training data on adversarial robustness have been extensively studied, little attention has been paid to the effect of a very important parameter in every network optimization on adversarial robustness: the learning rate. In particular, we hypothesize that effective learning rate scheduling during adversarial training can significantly reduce the overfitting issue, to a degree where one does not even need to adversarially train a model from scratch but can instead simply adversarially fine-tune a pre-trained model. Motivated by this hypothesis, we propose a simple yet very effective adversarial fine-tuning approach based on a 'slow start, fast decay' learning rate scheduling strategy which not only significantly decreases computational cost required, but also greatly improves the accuracy and robustness of a deep neural network. Experimental results show that the proposed adversarial fine-tuning approach outperforms the state-of-the-art methods on CIFAR-10, CIFAR-100 and ImageNet datasets in both test accuracy and the robustness, while reducing the computational cost by 8–10×. Furthermore, a very important benefit of the proposed adversarial fine-tuning approach is that it enables the ability to improve the robustness of any pre-trained deep neural network without needing to train the model from scratch, which to the best of the authors' knowledge has not been previous demonstrated in research literature.

1. Introduction

The simple adversarial training (AT) approach remains the most popular and effective adversarial defense mechanism; especially, since Madry *et al.* [3] introduced PGD adversarial attack and empirically illustrated that PGD is the universal first-order adversary (i.e. no other adversarial algorithm that uses first-order gradients can be more effective than PGD in fooling DNNs), AT with a PGD adversary has been the *de facto* adversarial defense mechanism. This is mainly due to the robustness guarantee that PGD AT can provide, such that if a model is robust against PGD, then it is robust against all the other first-order adversaries as well. As a result of this certified robustness, almost all the recent state-of-the-art methods [1, 2, 11] take advantage of PGD AT as a part of their algorithms.

In this work, we demonstrate how by taking a different view at the AT approach, it is possible to reduce the training time and improve the scalability of AT by a large degree. Using the proposed algorithm not only faces no loss on the accuracy and the robustness of the model, but it can also significantly improve the robust generalization of the model at the same time. Therefore, our adversarial fine-tuning approach mitigates the existing trade-off between the training time of AT and the model accuracy and robustness.

Motivated by the finding of Schmidt *et al.* [5] that during PGD AT a model highly overfits on the training data, we hypothesize that this issue is partially related to learning rate scheduling at the training stage, and effective learning rate scheduling can mitigate the overfitting issue significantly. Experimental results show that the proposed approach is able to improve the robust generalization of the DNN models, achieving state-of-the-art performance on many well-known datasets. Furthermore, one of the main benefits of the proposed adversarial fine-tuning algorithm is that it can be applied to any pre-trained model to increase its robustness. This is specially important when dealing with AT of models with very large training data sizes (e.g. ImageNet) or in scenarios where a model has been trained using special techniques which may not be reproducible, such as when the model is trained by using a pipeline of transfer learning, or when weak or semi supervision is applied on billion scale datasets [9]. In such scenarios, the usual PGD AT will not be practical due to its very high computational overhead, whereas adversarial fine-tuning not only is very computationally feasible and scalable, but it can also improve the

adversarial robustness by decreasing the overfitting.

2. Adversarial Training (AT)

Generally, any adversarial attack algorithm can be incorporated in the inner maximization of an adversarial training process. However, multi-step attacks and specially PGD are usually more powerful in providing effective perturbations. Especially, since Madry *et al.* [3] experimentally showed that PGD is the universal first-order adversary, this adversary has been wildly popular both for the adversarial training and for evaluating the ultimate robustness of deep models. An iterative PGD-k (PGD with k iterations) crafts the following adversarial example for a given natural sample x:

$$x_{t+1} = \Pi_{x+S}\Big(x_t + \alpha \; sign(\nabla_x \; L(\theta, x, y)\Big) \qquad (1)$$

where $\Pi(\cdot)$ is the projection function forcing the generated adversarial example remain within the boundary S and x_t is the adversarial example at step t, resulting from taking the ascent step of size α.

Intuitively speaking, the PGD AT approach tries to train the model to associate not only a single location in the space to the corresponding label of the sample x, but to associate a l_∞-ball around the example x to the same class label.

While increasing the number of steps k would result in more powerful adversarial examples with higher loss, one important limitation is the high computational overhead of PGD. Computing the model's gradient for the input data in each step is the main bottleneck of this approach and as such increasing the number of steps k, increases the training time significantly. PGD AT with its default setup is on average $\sim$8–10$\times$ more computationally complex than a model being trained only on natural samples.

Furthermore, PGD AT highly overfits on the training data resulting in drops in both the generalization of the model on natural samples and even the model robustness on test data. To this end, we analyze the relation between the model overfitting and learning rate scheduling in Figure 1. We hypothesis that it is possible to reduce the effect of overfitting for a given model and dataset and consequently improve the adversarial robustness and generalization of the model by utilizing a smart learning rate schedule.

A very intriguing pattern that we observe in our experiments is that long plateaus in the learning rate scheduling of the model during training further contributes to the overfitting problem. For a gradient-based optimizer, we consider the learning rate scheduler formulated as *step-LR*(i, γ), where i and γ show the number of epochs for each plateau and the step scale, respectively. The step-LR(i,γ) schedules decrease the learning rate at each e where $e\%i = 0$ by the factor of γ. For example, if $i = 5$ and $\gamma = 0.5$ the learning rate is multiplied by 0.5 after each 5 epochs.

Figure 1 compares the effect of using 6 different learning rate schedulers for fine-tuning a pre-trained PreAct ResNet18 on CIFAR-10 dataset. The only difference between different runs in the Figure is the size of the plateau, and the step scale is 0.5 for all, so, all of these learning rate schedulers can be formulated as *step-LR*$(i, 0.5)$. For the sake of a fair comparison, all trials use the exact same pre-trained model as their initialization. As seen in Figure 1, as i increases, the model gets more time to thoroughly learn the bubble around each sample and therefore, the overfitting on the training data increases, resulting in a drop in both the accuracy and the robustness of model on the test data. On the other hand, for smaller values of i the exact opposite trend happens causing a decrease in the train data robustness and an increase in both the accuracy and the robustness of model on the test data meaning less overfitting.

Motivated by the illustrated experiment and our observations regarding the sample complexity and the learning rate scheduling, we hypothesize that a simple adversarial *fine-tuning* approach can mitigate the overfitting issue, and achieve great robustness generalization. It is worth mentioning that, although other factors such as the model capacity have important effects on the robust generalization of the trained models, in this work, we only study the effects of the learning rate scheduling and the sample complexity of the training data on the model's robustness.

2.1. Adversarial Fine-tuning

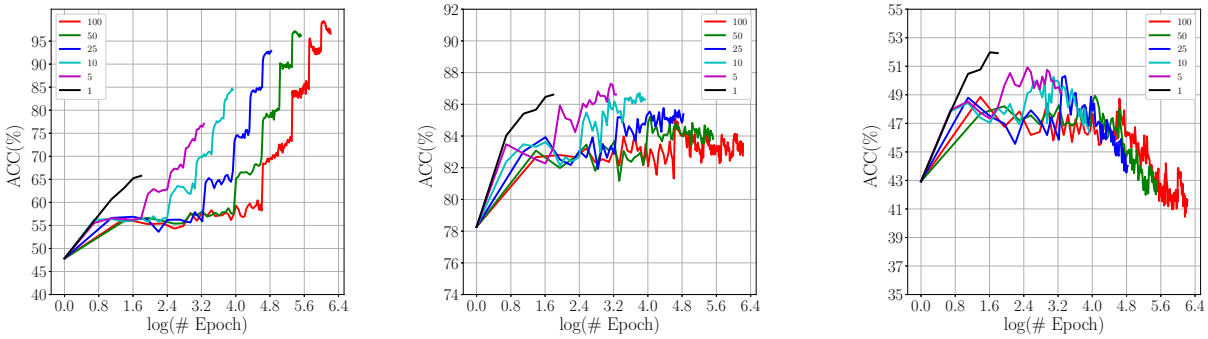
Motivated by the empirical evidence on the significant impact of learning rate scheduling on adversarial robustness, we propose a simple yet effective adversarial fine-tuning (AFT) technique for not only reducing training time (and hence computational cost) but also improving the robustness of a deep neural network. More specifically, the proposed AFT approach comprises of two main aspects:

- **Model pre-training**: A model is trained regularly using natural samples without consideration of adversarial perturbations for stronger initial generalization.
- **'Slow Start, Fast Decay' fine-tuning**: The pre-trained model is fine-tuned using adversarial perturbations following a 'slow start, fast decay' learning rate schedule for a small number of epochs for stronger adversarial robustness while preserving generalization.

This proposed technique is contrary to previously proposed AT methods that involve training models with adversarial perturbations in an end-to-end manner from scratch, which is significantly more computationally costly and lead to reduced model generalization. Details of the two main aspects of the proposed AFT technique are described below.

2.1.1 Step 1: Model Pre-training

The first step of the proposed AFT strategy involves pre-training a model which is performed regularly with natural



(a) Performance on Training Data (b) Performance on Clean Test Data (c) Performance on Adversarial Test Data

Figure 1: The effect of learning rate scheduling on model generalization and robustness; PGD AT with more number of epochs improves the model's robustness on the training data. However, bigger number of training iterations causes some drops in the model's accuracy and robustness on test data as evident in (b) and (c). As seen, training a model with less number of epochs can result in better reducing the overfitting issue and improves the adversarially robust generalization.

samples. Our experiments suggest that having a good pre-trained model is of high value, and we empirically find that the more data the pre-trained model is exposed to during its training, the better initialization it would be for the fine-tuning step. Experimental results validate this hypothesis on the CIFAR-10 and ImageNet datasets. This observation is particularly exciting since one can take advantage of already pre-trained models that have been trained on a very large set of data. This is especially important in many classification problems that leverage semi- or weak-supervision techniques to enrich their training data where an additional set of samples are used to improve the classification performance. SWSL [9] is an example of such approach where billion-sample scale data [4, 7] is used to achieve state-of-the-art performance on the ImageNet dataset. In our experiments, we show that adversarially fine-tuning such pre-trained models only on the main training data can improve the robustness and test accuracy by 7-8%. It is worth mentioning that due to the high computational overhead of PGD AT, conducting PGD AT on the dataset augmented by weak or semi-supervised method is not feasible. As such, we are motivated to introduce a 'slow start, fast decay' finetuning strategy.

2.1.2 Step 2: 'Slow Start, Fast Decay' Fine-tuning

The second step of the proposed AFT strategy involves fine-tuning the pre-trained model using adversarial perturbations via a 'slow start, fast decay' learning rate schedule. Given the overfitting issue explained before and the tendency of neural network to catastrophically forget their previously learned distributions when exposed to new samples, it is very crucial that the selected learning rate scheduling helps the model learn the new distribution of adversarial samples without sacrificing the previously learnt knowledge (natural data examples). Therefore, it is important that the learning rate is slow at first, so that the model gradually learns the new distribution.

The proposed 'slow start scheduling strategy follows a linear increase of the learning rate, and the 'fast decay' is followed by an exponential decrease to avoid the overfitting. The learning rate scheduler is formulated as, $LR = 0.0001 \times e$ for $1 \leq e \leq 5$ and $LR = \frac{0.0005}{2^{e-5}}$ for $6 \leq e \leq 10$; where e represents the epoch number. This approach helps the model learn the distribution of the adversarial examples without forgetting the distribution of the natural samples. After these first few epochs, the learning rate is reduced very fast so that model performance converges to a steady state, without having too much time to overfit on the training data.

Table 1: Evaluation results on CIFAR-10 dataset; the proposed algorithm is compared with the state-of-the-art methods which have been proposed in the recent years to improve the efficiency and the performance of (PGD) AT. The competing methods aim to provide an efficient approach in AT while reducing the computational complexity compared to original PGD AT (PGD AT). As seen, the proposed fine-tuning algorithm can result to higher accuracy on clean data while outperforms others significantly in robustness against PGD attack. Result (AFT (+500K)) shows that a model with better initialization can offer higher robustness after performing adversarial fine-tuning algorithm.

Method	Architecture	Clean	PGD	Time (min)
Natural	WideRes-32x10	95.01	00.00	780
PGD AT [3]	WideRes-32x10	87.25	45.84	5418
Free AT [6]	WideRes-32x10	85.96	46.82	785
Fast AT [8]	PreAct ResNet18	83.81	46.06	12
YOPO [10]	WideRes-34x10	86.70	47.98	476
ATTA [12]	WideRes-34x10	85.71	50.96	134
AFT	WideRes-28x10	**88.15**	**51.7**	486
AFT (+500K)	WideRes-28x10	**88.42**	**52.8**	486

3. Experimental Results & Discussion

We evaluate the proposed adversarial fine-tuning (AFT) method on three well-known classification datasets of CIFAR-10, CIFAR-100, and ImageNet, and compare the results with state-of-the-art techniques.

Table 2: CIFAR-100 experimental results; the accuracy and PGD robustness of the proposed method and the state-of-the-art methods are compared against PGD adversarial attack. Two different ϵ (AFT ($\epsilon = \frac{8}{255}$) and AFT ($\epsilon = \frac{10}{255}$)) are used in the proposed fine-tuning technique to illustrate the effect of PGD adversarial training in model robust generalization. YOPO has not been evaluated on CIFAR-100.

Method	Architecture	Clean	PGD	Time (min)
Natural	WideRes-32x10	80.00	00.00	817
Natural	WideRes-28x10	82.00	00.00	$\sim$ 750
PGD AT [3]	WideRes-32x10	60.00	22.50	5157
PGD AT [3]	WideRes-28x10	62.00	20.50	$\sim$ 5000
Free AT [6]	WideRes-32x10	62.13	25.88	780
AFT ($\epsilon = \frac{8}{255}$)	WideRes-28x10	**68.15**	23.29	486
AFT ($\epsilon = \frac{10}{255}$)	WideRes-28x10	66.57	**25.12**	486

3.1. Results

As the first experiment, the proposed method and the competing algorithms are compared via CIFAR-10 dataset, and the robustness of the model are evaluated against a PGD adversary with $\epsilon = \frac{8}{255}$. As seen in Table 1, the proposed fine-tuning algorithm can provide models with both highest generalization on natural images (accuracy on clean data) and greatest robustness against adversarial attack. Results show that using data augmentation and taking advantage of 500K additional data samples to augment the CIFAR-10 dataset improves the robustness of the model against adversarial attack significantly as well. It is important to note that this additional data samples are not used in adversarial fine-tuning step but only in the training of the model on natural images. As such, the result confirms the hypothesis that a model with a higher generalization can offer better robustness against adversarial attacks when trained properly.

To confirm the effectiveness of the proposed algorithm as the second experiment, it is evaluated via CIFAR-100 dataset. A same setup as CIFAR-10 experiment is used, where a PGD adversary with 20 iterations and $\epsilon = \frac{8}{255}$ is utilized to evaluate the robustness of the competing methods. Results reported in Table 2 further illustrates the effectiveness of the proposed algorithm in providing robust DNN models while does not sacrifice the model's generalization on neutral images. To better analyze the effect of PGD adversarial training in the proposed fine-tuning technique, two different ϵ values (AFT ($\epsilon = \frac{8}{255}$) and AFT ($\epsilon = \frac{10}{255}$)) have been used to trained the model. As seen, while using perturbed images with stronger attack can improve the robustness of the model, it resulted a drop in the accuracy of the model against natural data samples which further validates the overfitting issue explained in Section **??**. Higher value of ϵ means bigger l_∞-ball around the samples and this forces the model to use more complex decision boundary to fit on the data. The proposed fine-tuning techniques is more than $10\times$ faster than the conventional PGD adversarial training method (PGD AT) and is it even $\sim 2\times$ faster compared to Free AT algorithm in training the final robust model.

Table 3: Comparison results on ImageNet dataset; The proposed method outperforms competing algorithms on clean data samples (natural images) while provide comparable robustness performance as evident by ResNet50 results. The reported result for ResNet50-SWSL architecture shows the significant effect of pre-training and the generalization of the model on robustness. As seen, the model offers $\sim 7\%$ robustness improvement. YOPO has not been evaluated on ImageNet.

Method	Architecture	Clean	PGD	Time (hours)
Natural	ResNet50	76.04	0.13	-
PGD AT [3]	ResNet50	68.00	45.0	$\sim$280
Free AT [6]	ResNet50	64.50	43.5	52
Fast AT [8]	ResNet50	61.00	43.5	12
ATTA [12]	ResNet50	60.70	44.5	-
AFT	ResNet50	69.5	43.0	32
AFT	ResNet50-SWSL	**74.5**	**50.5**	32

As the last experiment, the proposed algorithm and the competing methods are compared against ImageNet dataset. To evaluate the model $\epsilon = \frac{2}{255}$ is chosen for the PGD adversarial attack. As seen in Table 3, while the proposed fine-tuning technique outperforms the competing methods in clean accuracy which shows the generalization of the DNN model on natural images, it provides comparable robustness against adversarial attack. This is evident by the reported result for ResNet50 network architecture. The reported result for ResNet50-SWSL demonstrates the significant effect of pre-training and the effect of the model generalization on the robustness result. The ResNet50-SWSL architecture is further trained via a semi-supervised technique. As seen, this further training can result a significant boost in both the generalization of the model and model accuracy on clean data and robustness of the model against adversarial attack. Results show that the robustness of the model can improve by more than 7% and outperforms competing methods significantly while it can provide the final model in reasonable time-frame.

4. Conclusion

Here, we further illustrated the severe overfitting issue with adversarial training and we argued why this phenomena takes place. Motivated by the finding and experimental results, we proposed simple yet effective fine-tuning approach to improve the robustness of deep neural network models against adversarial attacks without sacrificing the generalization of the model on natural data samples. The proposed fine-tuning framework can reduce the training run-time by $10\times$ while outperforms state-of-the-art algorithms in adversarial training. One important benefit of the proposed method is that it can be easily applied on any pre-trained model without requiring to trained the model from scratch. This is very crucial when the model is trained via customized training frameworks which it is impracticable to train the model again while it is important to improve the robustness of that against adversarial attacks.

References

[1] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11192–11203, 2019. 1

[2] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–597, 2019. 1

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3, 4

[4] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018. 3

[5] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018. 1

[6] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pages 3358–3369, 2019. 3, 4

[7] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[8] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020. 3, 4

[9] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019. 1, 3

[10] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, pages 227–238, 2019. 3

[11] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, pages 1831–1841, 2019. 1

[12] Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1181–1190, 2020. 3, 4