Literature Review: Adversarial Training of Deep Neural Networks

Daniel Korth [1]

Abstract

Deep neural networks are susceptible to adversarial examples; small and imperceptible changes made to the input that can completely alter the prediction of the model. To make neural networks more robust to these adversaries, multiple defenses have been proposed. Among those, adversarial training is one of the most promising strategies to make neural networks more robust to adversarially crafted attacks. This paper systematically reviews and evaluates current approaches and phenomenons, and outlines future directions in the field.

1. Introduction

In the past decade, deep learning made significant breakthroughs in various domains, including computer vision (Krizhevsky et al., 2012) and natural language processing (Vaswani et al., 2017). Despite these advancements, neural networks remain vulnerable to adversarial examples - subtle perturbations that deceive the classifier (Szegedy et al., 2014). See Figure 1 for an example. Researchers developed stronger attacks (Kurakin et al., 2017; Carlini & Wagner, 2017; Goodfellow et al., 2015), while others explored defense strategies (Madry et al., 2018; Papernot et al., 2016). Notably, adversarial training has proven to be one of the most effective technique, where models are exposed to adversarial examples during training to enhance robustness.

This review presents both phenomenons of adversarial examples and research directions in adversarial training. Afterwards, future research directions are presented. It focuses on the classical supervised learning approach of independent and identically distributed samples. Adversarial attacks also extend to other domains, such as natural language processing (Jia & Liang, 2017) and graph neural networks (Gosch et al., 2023).

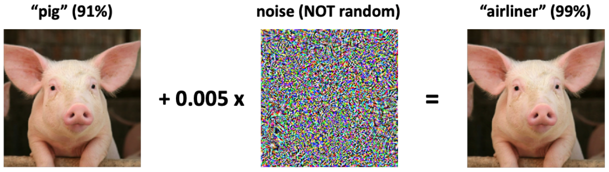

Figure 1. In the left picture, the network confidently predicts ”pig” (91%). When combined with adversarial perturbation, the classifier is fooled, now predicting ”airliner” (99%). Figure taken from Kolter & Madry (2018).

2. Preliminaries

2.1. Adversarial Robustness

Consider a data distribution $\mathcal{D}$ over pairs (x, y) of input $x \in X \subseteq \mathbb{R}^d$ with a corresponding label $y \in Y$. Let $F_\theta : X \rightarrow Y$ denote a deep neural network parameterized by weights θ where for every input x, $F_\theta(x) = \arg\max_i f_\theta^{(i)}(x)$ denotes the classification of the network and $f_\theta^{(i)}(\cdot)$ the i-th logit value. An adversarial example is defined as any $x' = x + \delta$ such that $F_\theta(x) = y$ is the correct prediction, but $F_\theta(x') \neq y$ leads to a misclassification, where $\delta \in \Delta$ is a set of allowed adversarial perturbations (i.e., the threat model). For a ℓ_p-norm perturbation bound of size ϵ, the set of adversarial perturbation is defined as $\Delta_p = \{\delta : ||\delta||_p < \epsilon\}$. A neural network is considered adversarially robust on the input x, if $F_\theta(x + \delta) = y$ holds for all perturbed versions of the input defined by the threat model Δ_p. The majority of work considers the ℓ_2 or ℓ_∞ norm to define the perturbation set.

2.2. Adversarial Attacks

FGSM. One of the simplest forms of generating an adversarial perturbation for the ℓ_∞-ball is the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), where

$$\delta = \epsilon \cdot \text{SIGN}(\nabla_x \mathcal{L}(f_\theta(x), y))$$

It generates an optimal perturbation under the assumption that the network is locally linear in the vicinity of the input.

[1] School of Computation, Information and Technology, Technical University of Munich, Munich, Germany. Correspondence to: Daniel Korth <korth@cs.tum.edu>.

Algorithm 1 FGSM Adversarial Training (Goodfellow et al., 2015)

```
Input: data $X$, epochs $T$, perturbation bound $\epsilon$, learning rate $\tau$, model weights $\theta$
for t $= 1 \dots T$ do
    for batch $B \subset X$ do
        inner maximization
            $\delta \leftarrow \epsilon \cdot \text{SIGN}(\nabla_x \mathcal{L}(f_\theta(x), y))$
        outer minimization
            $\theta \leftarrow \theta - \tau \cdot \nabla_\theta \mathcal{L}(f_\theta(x+\delta), y)$
    end for
end for
```

PGD. Projected Gradient Descent (PGD) (Madry et al., 2018) is one of the most popular methods for generating strong adversarial attacks generated by

$$\delta^{(t+1)} \leftarrow \prod_{\Delta_\infty} (\delta^{(t)} + \alpha \cdot \text{SIGN}(\nabla_x \mathcal{L}(f_\theta(x+\delta^{(t)}), y)))$$

where $\delta^{(0)}$ is randomly sampled from Δ_∞. It is a multi-step attack based on FGSM where after each step for a specific step size α, the negative loss is projected back onto the perturbation set Δ_∞ using the projection operator $\prod$. Madry et al. (2018) hypothesize this attack to be a universal attack amongst all first-order optimization methods and it is regarded as one of the best attacks and basis for many of the most successful training techniques.

2.3. Adversarial Training

Adversarial training is a technique to make models more robust to adversarial attacks. Madry et al. (2018) formulates the optimization procedure as a min-max problem

$$\arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \Delta} \mathcal{L}(f_\theta(x+\delta), y) \right]$$

Intuitively speaking, adversarial training aims to minimize the loss under the worst-case adversarial perturbation for a specific threat model. The *inner maximization* problem is intractable, and one of the main objectives of adversarial training is to find a good approximation through attacks such as FGSM (Goodfellow et al., 2015) or PGD (Madry et al., 2018), which represent a lower bound to the *inner maximization* problem. The *outer minimization* is then using this perturbed version of the input, and tries to minimize the loss, usually cross-entropy. The full adversarial training algorithm for FGSM is shown in Algorithm 1.

3. Phenomenons in Adversarial Training

This section provides some interesting phenomenons, observations, and challenges of adversarial training and adversarial examples in general.

Gradient Masking. The term was first introduced by Papernot et al. (2017) as a defense strategy that prevents the calculation of useful gradients to find a solution to the *inner maximization* and therefore the generation of adversarial perturbation (Papernot et al., 2017). Athalye et al. (2018) define three special cases of gradient masking: (1) shattered gradients, caused by non-differentiable operations or numerical instability, (2) stochastic gradients that depend on test-time randomness, and (3) vanishing/exploding gradients, which are a common problem in deep neural networks that lead to unusable gradients (Athalye et al., 2018). Athalye et al. (2018) claim that defenses that rely on gradient masking give a ”false sense of security”, as they can often be broken by approximations of the gradient, for example by inserting smooth functions that replace the non-differentiable parts of the network. Furthermore, Tramèr et al. (2018) observe that one-step attacks such as FGSM (Goodfellow et al., 2015) used for adversarial training lead to a degenerate solution of the min-max problem, where the model learns to mask the gradients instead of becoming smooth and robust in the vicinity of the input samples.

Robust Overfitting. It is commonly known that neural networks tend to generalize very well despite being trained for very long to overfit on the training data, a phenomenon often referred to as epoch-wise *double descent* (Belkin et al., 2019; Nakkiran et al., 2021). Rice et al. (2020) observe a similar phenomenon for adversarial training which they coin *robust overfitting*. They show that overfitting is a dominant phenomenon in adversarial training, where the robust training loss keeps decreasing when training for longer, while the robust test loss increases again. A typical training curve of standard and adversarial training that display epoch-wise *double decent* and *robust overfitting*, respectively, can be found in Appendix A. To mitigate *robust overfitting*, the authors experiment with several regularization and augmentation techniques, and find simple early stopping to be one of the most effective to prevent the onset of *robust overfitting*.

Robustness and Accuracy Trade-Off. Tsipras et al. (2019) show that there provably exists a trade-off between accuracy on clean images and the robustness with regards to adversarial perturbations. They attribute the reason for the trade-off to be that neural network classifiers assign weight to features that are weakly correlated with the target label, and when perturbed adversarially, these features are correlated with the wrong target class. As a consequence, almost all adversarially trained models have a drop in accuracy compared to standard models. However, this drop has been shown to be a phenomenon of bad choice of perturbation set (Suggala et al., 2019).

Transferability of Adversarial Examples. A very intriguing property of adversarial examples is that they have been shown to transfer across models with different training architectures (Szegedy et al., 2014). An adversarial example generated on model F_A is likely to also fool the classifier of model F_B, despite it being completely different. The transferability allows for black box attacks, where the attacker does not have access to the gradients of the model. Ilyas et al. (2019) hypothesize that one reason for transferability is that inputs usually contain robust and non-robust features, and classifiers make use of non-robust features in their prediction, that can however flip the prediction when small adversarial noise is added.

4. Methods

This section outlines different research directions in adversarial training.

4.1. Regularization

The main motivation behind regularizing adversarial training is to smoothen the loss landscape, as this can promote robustness. Broadly speaking, adversarial regularization approaches add regularization by punishing sharp differences between the perturbed and unperturbed version of the input.

Zhang et al. (2019b) propose TRADES to trade off robustness against accuracy

$$\mathcal{L}_{\text{TRADES}} = \mathcal{L}_{\text{CE}}(f_\theta(x), y) + \lambda \cdot D_{KL}(f_\theta(x), f_\theta(x+\delta))$$

where λ is the hyperparameter for the trade off. The first term optimizes for accuracy via the cross-entropy loss, and the second term encourages a smooth loss landscape by minimizing the KL-Divergence between the logits of the natural example $f(x)$ and the adversarial example $f(x+\delta)$. On top, the authors propose to maximize the KL-Divergence for finding an adversarial perturbation during the *inner maximization*. MART (Wang et al., 2020) extends TRADES by weighting the D_{KL} by $1 - p_y(x)$ where $p_y(x)$ refers to the softmax probability of the correct label. Other approaches for example use a nuclear norm regularizer based on the batched logit differences between clean and perturbed inputs $||f_\theta(X+\delta) - f_\theta(X)||_*$ (Sriramanan et al., 2021) or leverage the first-order Taylor expansion to encourage linearity in the vicinity of the input (Qin et al., 2019).

4.2. Ensembles

Tramèr et al. (2018) are the first to introduce ensembles into adversarial training to enhance the diversity of perturbations. They leverage pre-trained models to generate adversarial examples used to train a different model. This technique allows to decouple the generation and learning step and possibly prevent the onset of gradient masking (Papernot et al., 2017). Motivated by the observation that different adversarial training methods are robust on different examples, Liu et al. (2023) develop Collaborate Adversarial Training (CAT) to enable model and knowledge interaction between multiple classifiers. Croce et al. (2023) propose Model Soups, which are linear interpolations of model parameters trained against different ℓ_p normed attacks to smoothly trade off robustness against a diverse set of adversarial attacks.

4.3. Data Generation & Augmentation

Adversarial training suffers from *robust overfitting* (Rice et al., 2020). Inspired by the work from Schmidt et al. (2018) hinting at the fact that robust generalization requires more training data, there has been an increasing interest in ways to incorporate more natural or generated data into training. Carmon et al. (2019) were among the first to use a semi-supervised learning algorithm for adversarial training by pseudo-labeling data and performing regular adversarial training afterward. Rebuffi et al. (2021) experiment with different data augmentation techniques such as MixUp (Zhang et al., 2018), Cutout (DeVries & Taylor, 2017) or CutMix (Yun et al., 2019). Combined with model weight averaging (Izmailov et al., 2018), an exponential moving average over the model parameters θ with a specific decay rate τ where $\theta' = \tau \cdot \theta' + (1-\tau)\theta$, they significantly enhance adversarial training and mitigate the negative side effects of *robust overfitting*. They apply data augmentation before the adversarial attack, as otherwise, the augmentation will destroy the adversarial perturbation. Other works experimented with pseudo-labeling data generated through generative models (Gowal et al., 2021) and more recently also specifically using diffusion models (Wang et al., 2023).

4.4. Efficient Adversarial Training

The adversarial training computation time is dominated by the multi-step procedure of finding an optimal adversary (e.g. PGD (Madry et al., 2018)). Therefore, a line of research is about making adversarial training more efficient. Shafahi et al. (2019) present a ”free” adversarial training algorithm that does not incur any additional computational cost. The core idea is to calculate the derivate with respect to weights $\nabla_\theta \mathcal{L}$ and input $\nabla_x \mathcal{L}$ in the same backpropagation step. The adversarial noise generated by gradient ascent from the current backpropagation step is then used in the forward pass of the next step. This allows Shafahi et al. (2019) to recycle gradient information and use a single step for weight update and adversarial noise generation. Since the noise is dependend on the input, Shafahi et al. (2019) train on the same batch for multiple hops. The detailed algorithm can be found in Appendix B.

Wong et al. (2020) hypothesize that the main benefit from ”free” adversarial training comes by starting from a non-zero initial input perturbation, and not specifically from using the perturbation from previous steps. Therefore they initialize the noise $\delta = \mathcal{U}(-\epsilon, \epsilon)$ uniformly before performing the conventional FGSM (Goodfellow et al., 2015) adversarial attack. The authors find this technique to perform on par with significantly more time-consuming attacks such as PGD (Madry et al., 2018). Another notable technique to speed up training is YOPO (You Only Propagate Once) (Zhang et al., 2019a) which is motivated by the observation that the adversary update is only coupled with the first network layer. This allows the authors to only rely on the first layer of the neural network for most of the gradient update steps.

4.5. Other Methods

There exist several other promising research directions for adversarial training. Kim et al. (2023) analyze adversarial training from a frequency perspective and find a way to shift the inputs into the low-frequency region which leads to faster convergence, smoother predictions and ultimately improved robustness. Xie et al. (2020) try to promote the gradient quality to generate stronger adversarial attacks by finding smooth alternatives to the ReLU function such as the GELU (Hendrycks & Gimpel, 2016). Another line of research uses an instance adaptive perturbation bound ϵ that can change during training and depend on the distance to other points on the training manifold (Balaji et al., 2019).

5. Review

In recent years, adversarial training has received a lot of attention in the research community, leading to numerous approaches being explored. This section reviews the methods presented in this paper and takes a broader look at adversarial robustness in general. Benchmark results of the methods presented in this paper can be found in Appendix C.

Regularization has been shown to be an effective technique to improve the generalization of adversarial training, both by preventing *robust overfitting* as well as enforcing a smoother loss landscape in the vicinity of the input to counteract gradient masking. It does usually not incur any additional training cost and is an effective tool to enhance adversarial training. At the same time, performance gains solely from regularization are not solving the issue of adversarial examples.

Ensembles improve machine learning performance in a lot of different tasks, including adversarial training. They are great to diversify adversarial examples and mitigate issues such as *robust overfitting*. However, ensemble techniques can lead to significant computational overhead that is currently not justified in performance gains.

Data generation and augmentation methods are highly effective in enhancing robustness and generalization. Still, they come at the cost of increased training time, especially with multi-step attacks for the *inner maximization*.

Lastly, efficient adversarial training methods are of high importance, as it can enhance all other methods by finding good approximations of the optimal adversarial perturbation quickly. One of the biggest challenges for efficient methods is to ensure the diversity of adversarial examples while using as little as possible training overhead, i.e. steps for the *inner maximization*.

Analyzing the RobustBench benchmark for adversarial robustness (Croce et al., 2020) reveal following trends: (1) All leading solutions heavily rely on large amounts of extra (synthetic) data. (2) The top-performing approaches are based in adversarial training. (3) There hasn't been any significant breakthroughs in adversarial robustness in the recent past.

6. Conclusion

Adversarial Training remains to be the most effective technique to enhance the robustness of neural networks. Overcoming challenges like gradient masking and robust overfitting is achieved through diverse regularization, ensemble, and augmentation techniques. The emergence of diffusion models opens up the opportunity to train with an abundance of data. This calls for the development of efficient methods capable of generating strong adversarial examples without the need for multiple gradient ascent steps during the *inner maximization*.

7. Future Work

Future research should focus on efficient adversarial training methods that can handle large amounts of synthetic data. Furthermore, new data augmentation techniques could ensure less overfitting, better sample efficiency, and better generalization. Adversarial training also lacks theoretical understanding. Recently, Latorre et al. (2023) has disproven a corollary of Danskin's Theorem that has long been taken for granted (Madry et al., 2018): a solution to the *inner maximization* problem yields a descent direction for the robust loss. PGD (Madry et al., 2018) remains the gold standard of strong first-order adversary methods, but this cannot be explained by theory anymore. More generally, a better understanding on the theoretical side could give rise to novel and creative solutions.

A direction that could complement adversarial training is certified training (Gowal et al., 2019; Shi et al., 2021), which tries to find provable upper bounds on the worst case perturbation. This can for example be achieved by linear relaxations (Gowal et al., 2019). In a broader context, one could start to rethink how to define adversarial examples, as fundamental principles such as imperceptibility and optimization-based attacks do not transfer well to other domains such as Natural Language Processing (Carlini et al., 2023).

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018.

Balaji, Y., Goldstein, T., and Hoffman, J. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on Artificial Intelligence and Security*, 2017.

Carlini, N., Nasr, M., Choquette-Choo, C. A., Jagielski, M., Gao, I., Awadalla, A., Koh, P. W., Ippolito, D., Lee, K., Tramer, F., et al. Are aligned neural networks adversarially aligned? *arXiv preprint arXiv:2306.15447*, 2023.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 2019.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Croce, F., Rebuffi, S.-A., Shelhamer, E., and Gowal, S. Seasoning model soups for robustness to adversarial and natural distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Gosch, L., Sturm, D., Geisler, S., and Günnemann, S. Revisiting robustness in graph machine learning. In *International Conference on Learning Representation*, 2023.

Gowal, S., Dvijotham, K. D., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 2021.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in Neural Information Processing Systems*, 2019.

Izmailov, P., Wilson, A., Podoprikhin, D., Vetrov, D., and Garipov, T. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Jia, R. and Liang, P. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

Kim, Y., Kim, S., Seo, I., and Shin, B. Phase-shifted adversarial training. *arXiv preprint arXiv:2301.04785*, 2023.

Kolter, Z. and Madry, A. Adversarial robustness: Theory and practice. *Advances in Neural Information Processing Systems Tutorial Session*, 2018. URL https://www.youtube.com/watch?v=TwP-gKBQyic.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

Latorre, F., Krawczuk, I., Dadi, L. T., Pethick, T. M., and Cevher, V. Finding actual descent directions for adversarial training. In *International Conference on Learning Representations*, 2023.

Liu, X., Kuang, H., Lin, X., Wu, Y., and Ji, R. Cat: Collaborative adversarial training. *arXiv preprint arXiv:2303.14922*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, 2016.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 2019.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 2021.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 2018.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! *Advances in Neural Information Processing Systems*, 2019.

Shi, Z., Wang, Y., Zhang, H., Yi, J., and Hsieh, C.-J. Fast certified robust training with short warmup. *Advances in Neural Information Processing Systems*, 2021.

Sriramanan, G., Addepalli, S., Baburaj, A., et al. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems*, 2021.

Suggala, A. S., Prasad, A., Nagarajan, V., and Ravikumar, P. Revisiting adversarial risk. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

Wang, Z., Pang, T., Du, C., Lin, M., Liu, W., and Yan, S. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

Xie, C., Tan, M., Gong, B., Yuille, A., and Le, Q. V. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in Neural Information Processing Systems*, 2019a.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019b.

A. Robust Overfitting

Robust overfitting is a phenomenon first observed by Rice et al. (2020). It describes that overfitting can occur during adversarial training, and that it can have negative consequences on the adversarial robustness. A visualization of robust overfitting can be found below (Figure 2).

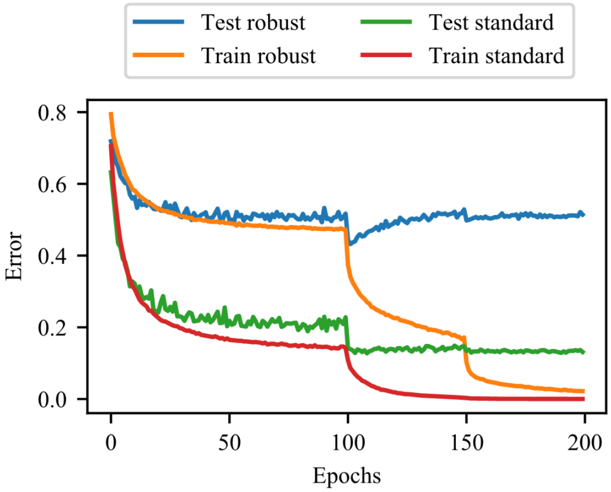

Figure 2. The learning curves for both a robust and standard trained model replicating the experiment done by (Madry et al., 2018) on CIFAR-10. The curves demonstrate “robust overfitting”; shortly after the first learning rate decay the model momentarily attains 43.2% robust error, and is actually more robust than the model at the end of training, which only attains 51.4% robust test error against a 10-step PGD adversary. Figure and Caption taken from Rice et al. (2020).

B. ”Free” Adversarial Training

Free Adversarial Training (Shafahi et al., 2019) is a training technique that incurs almost no overhead compared to standard training. The *inner maximization* and *outer minimization* are combined into one update step. The backpropagation step calculates both the gradient with respect to θ as well as the gradient with respect to x at the same time. To ensure that the adversarial noise is used on the same example, the algorithm trains on the same batch for multiple hops m.

Algorithm 2 ”Free” Adversarial Training (Shafahi et al., 2019)

```
Input: data $X$, epochs $T$, perturbation bound $\epsilon$, learning rate $\tau$, hop steps $m$, model weights $\theta$
$\delta \leftarrow 0$
for $t = 1 \ldots T/m$ do
    for batch $B \subset X$ do
        for step = 1...m do
            Compute gradients $\nabla_\theta$ and $\nabla_x$ simultaneously
                $\nabla_\theta, \nabla_x \leftarrow \nabla L(f_\theta(x + \delta), y)$
            Update model weights $\theta$
                $\theta \leftarrow \theta - \tau \cdot \nabla_\theta$
            Clip perturbation $\delta$
                $\delta \leftarrow \delta + \epsilon \cdot \textsc{sign}(\nabla_x)$
                $\delta \leftarrow \max(\min(\delta, \epsilon), -\epsilon)$
        end for
    end for
end for
```

C. Benchmark

Table 1 shows benchmark results of all methods presented in the main body that have an entry on the RobustBench Leaderboard (Croce et al., 2020). It is important to note that the methods are only ordered and classified by the main contribution of the paper. A method presented in the regularization section can still use generated data or efficient training techniques. The reader is encouraged to check out the RobustBench leaderboard (Croce et al., 2020) for a more elaborate leaderboard over more methods, different datasets and threat models.

Table 1. Clean and Robust accuracy for different methods on CIFAR-10. Robust accuracy is calculated via Auto-Attack (Croce & Hein, 2020) with an l_∞ perturbation budget of $\epsilon = 8/255$. Entries are sorted by highest robust accuracy within each method type. Numbers taken from (Croce et al., 2020).

Method	Standard accuracy	Robust accuracy	Additional Data
Regularization			
(Wang et al., 2020)	87.50	56.29	√
(Zhang et al., 2019b)	84.92	53.08	×
(Qin et al., 2019)	86.28	52.84	×
Data Generation & Augmentation			
(Wang et al., 2023)	93.25	70.69	√
(Rebuffi et al., 2021)	92.23	66.58	×
(Gowal et al., 2021)	88.74	66.11	√
(Carmon et al., 2019)	89.69	59.53	√
Efficient Adversarial Training			
(Zhang et al., 2019a)	87.20	44.83	×
(Wong et al., 2020)	83.34	43.21	×
(Shafahi et al., 2019)	86.11	41.47	×
Baselines			
(Madry et al., 2018)	87.14	44.04	×
Plain	94.78	0.0	×