Efficient Adversarial Training with Transferable Adversarial Examples

Haizhong Zheng   Ziqi Zhang   Juncheng Gu   Honglak Lee   Atul Prakash
University of Michigan, Ann Arbor
{hzzheng, ziqizh, jcgu, honglak, aprakash}@umich.edu

Abstract

Adversarial training is an effective defense method to protect classification models against adversarial attacks. However, one limitation of this approach is that it can require orders of magnitude additional training time due to high cost of generating strong adversarial examples during training. In this paper, we first show that there is high transferability between models from neighboring epochs in the same training process, i.e., adversarial examples from one epoch continue to be adversarial in subsequent epochs. Leveraging this property, we propose a novel method, Adversarial Training with Transferable Adversarial Examples (ATTA), that can enhance the robustness of trained models and greatly improve the training efficiency by accumulating adversarial perturbations through epochs. Compared to state-of-the-art adversarial training methods, ATTA enhances adversarial accuracy by up to 7.2% on CIFAR10 and requires $12 \sim 14\times$ less training time on MNIST and CIFAR10 datasets with comparable model robustness.

1. Introduction

State-of-the-art deep learning models for computer vision tasks have been found to be vulnerable to adversarial examples [27]. Even a small perturbation on the image can fool a well-trained model. Recent works [7, 25, 26] also show that adversarial examples can be physically realized, which can lead to serious safety issues. Design of robust models, which correctly classifies adversarial examples, is an active research area [5, 9, 14, 21, 30], with *adversarial training* [18] being one of the most effective methods. It formulates training as a game between adversarial attacks and the model: the stronger the adversarial examples generated to train the model are, the more robust the model is.

To generate strong adversarial examples, iterative attacks [13], which use multiple attack iterations to generate adversarial examples, are widely adopted in various adversarial training methods [4, 18, 28, 34]. Since adversarial perturbations are usually bounded by a constrained space $\mathcal{S}$ and attack perturbations outside $\mathcal{S}$ need to be projected back to

$\mathcal{S}$, k-step projected gradient descent method [13, 18] (PGD-k) has been widely adopted to generate adversarial examples. Typically, using more attack iterations (higher value of k) produces stronger adversarial examples [18]. However, each attack iteration needs to compute the gradient on the input, which causes a large computational overhead. As shown in Table 1, the training time of adversarial training can be close to 100 times larger than natural training.

Recent works [17, 19, 20] show that adversarial examples can be transferred between models: adversarial examples generated for one model can still stay adversarial to another model. The key insight in our work, which we experimentally verified, is that, because of high transferability between models (i.e., checkpoints) from neighboring training epochs, attack strength can be accumulated across epochs by repeatedly reusing the adversarial perturbation from the previous epoch.

We take advantage of this insight in coming up with a novel adversarial training method called ATTA (Adversarial Training with Transferable Adversarial examples) that can be significantly faster than state-of-the-art methods while achieving similar model robustness. In traditional adversarial training, when a new epoch begins, the attack algorithm generates adversarial examples from the original input images, which ignores the fact that these perturbations can be reused effectively. In contrast, we show that it can be advantageous to reuse these adversarial perturbations through epochs. Even using only *one attack iteration* to generate adversarial examples, ATTA-1 can still achieve comparable robustness with respect to traditional PGD-40.

We apply our technique on Madry's Adversarial Training method (MAT) [18] and TRADES [34] and evaluate the performance on both MNIST and CIFAR10 dataset. Compared with traditional PGD attack, our method improves training efficiency by up to $12.2\times$ ($14.1\times$) on MNIST (CIFAR10) with comparable model robustness. Trained with ATTA, the adversarial accuracy of MAT can be improved by up to 7.2%. Noticeably, for MNIST, with one attack iteration, our method can achieve 96.31% adversarial accuracy within 5 minutes (compared to 65 minutes with MAT). For CIFAR10, compared to MAT whose training time is

more than one day, our method achieves comparable adversarial accuracy in about 2 hours (compared to 33 hours with MAT). We also verify that our method is practical for large dataset: we evaluate our method with ImageNet, which can achieve comparative robustness within 97 hours with just one 2080 TI GPU. We also compare our method with two fast adversarial training methods, YOPO [32] and Free [23]. Evaluation results show that our method is both more effective and efficient than these two methods.

Contribution. To the best of our knowledge, our work is the first to enhance the efficiency and effectiveness of adversarial training by taking advantage of high transferability between models from different epochs. In summary, we make the following contributions:

- We are the first to reveal the high transferability between models of neighboring epochs in adversarial training. With this property, we verify that the attack strength can be accumulated across epochs by reusing adversarial perturbations from the previous epoch.

- We propose a novel method (ATTA) for iterative attack based adversarial training with the objectives of both efficiency and effectiveness. It can generate the similar (or even stronger) adversarial examples with much fewer attack iterations via accumulating adversarial perturbations through epochs.

- Evaluation result shows that, with comparable model robustness, ATTA is $12.2\times$ ($14.1\times$) faster than traditional adversarial methods on MNIST (CIFAR10). ATTA can also enhance model adversarial accuracy by up to 7.2% for MAT on CIFAR10.

2. Adversarial training and transferability

In this section, we introduce relevant background on adversarial training and transferability of adversarial examples. We also discuss the trade-off between training time and model robustness.

2.1. Adversarial Training

Adversarial training is an effective defense method to train robust models against adversarial attacks. By using adversarial attacks as a data augmentation method, a model trained with adversarial examples achieves considerable robustness. Recently, lots of works [2, 10, 11, 18, 22, 33, 34] focus on analyzing and improving adversarial machine learning. Madry *et al*. [18] first formulate adversarial training as a min-max optimization problem:

$$\min_{f \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} [\max_{\delta \in \mathcal{S}} L(f(x + \delta), y)] \quad (1)$$

where $\mathcal{H}$ is the hypothesis space, $\mathcal{D}$ is the distribution of the training dataset, L is a loss function, and $\mathcal{S}$ is the allowed perturbation space that is usually selected as an L-p

norm ball around x. The basic strategy in adversarial training is, given a original image x, to find a perturbed image $x + \delta$ that maximizes the loss with respect to correct classification. The model is then trained on generated adversarial examples. In this work, We consider adversarial examples with a high loss to have *high attack strength*.

PGD-k attack based adversarial training: Unfortunately, solving the inner maximization problem is hard. Iterative attack [13] is commonly used to generate strong adversarial examples as an approximate solution for the inner maximization problem of Equation 1. Since adversarial perturbations are usually bounded by the allowed perturbation space $\mathcal{S}$, PGD-k (k-step projected gradient descent [13]) is adopted to conduct iterative attack [18, 23, 32, 34]. PGD-k adversarial attack is the multi-step projected gradient descent on a negative loss function:

$$x_{t+1} = \Pi_{x+\mathcal{S}}(x_t + \alpha \operatorname{sgn}(\Delta_{x_t} L(\theta, x_t, y)))$$

In the above, x_t is the adversarial example in the t-th attack iteration, α is the attack step size, and Π is the projection function to project adversarial examples back to the allowed perturbation space $\mathcal{S}$.

With a higher value of k (more attack iterations), PGD-k can generate adversarial examples with higher loss [18]. However, there is a trade-off between training time and model robustness in adversarial training. On one hand, since each attack iteration needs to calculate the gradient for the input, using more attack iterations requires more time to generate adversarial examples, thus causing a large computational overhead for adversarial training. As shown in Table 1, compared to natural training, adversarial training may need close to 100x more training time until the model converges. Most of the training time is consumed in generating adversarial examples (attack time). On the other hand, reducing the number of attack iterations k can reduce the training time, but that negatively impacts robustness of the trained model (Table 2).

Dataset	Natural Training	Adversarial training		
		Training	Attack	Total
MNIST	39.7 sec	210 sec	3723 sec	3933 sec
CIAFR10	55min	214 min	1813 min	2027 min

Table 1: Training time of natural training and adversarial training. *Attack* column shows the time consumed in adversarial example generation.

2.2. Transferability of Adversarial Examples.

Szegedy *et al*. [27] show that adversarial examples generated for one model can stay adversarial for other models. This property is named as *transferability*. This property is usually leveraged to perform a black-box attack

Defense	PGD-1	PGD-3	PGD-5	PGD-10
Nat. Acc.	93.18%	89.95%	85.24%	87.49%
Adv. Acc.	22.3%	41.38%	43.55%	47.07%
Time (min)	435	785	1140	2027

Table 2: The relation between the number of attack iterations and adversarial accuracy against PGD-20 attack on CIFAR10 dataset.

[16, 17, 19, 20]. To attack a targeted model f_t, the attacker generates transferable adversarial examples from the source model f_s. The higher the transferability between f_s and f_t is, the higher the success rate the attack has.

Substitute model training is a commonly used method to train a source model f_s. Rather than the benchmark label y, f_s is trained with $f_t(x)$ which is the prediction result of the targeted model [20] to achieve a higher black-box attack success rate. While our work does not use black-box attacks, we do rely on a similar intuition as behind substitute model training, namely, two models with similar behavior and decision boundaries are likely to have higher transferability between each other. We use this intuition to show high transferability between models from neighboring training epochs, as discussed in the next section.

3. Attack strength accumulation

In this section, we first conduct a study and find that models from neighboring epochs show very high transferability and are naturally good substitute models to each other. Based on this observation, we design an accumulative PGD-k attack that accumulates attack strength by reusing adversarial perturbations from one epoch to the next. Compared to traditional PGD-k attack, accumulative PGD-k attack achieves much higher attack strength with fewer number of attack iterations in each epoch.

3.1. High transferability between epochs

Transferability between models in different training epochs of the same training program has not been studied. Because fluctuations of model parameters between different epochs are very small, we think that they are likely to have similar behavior and similar decision boundaries, which should lead to high transferability between these models.

To evaluate the transferability between models from training epochs, we adversarially train a model T as the targeted model, while saving intermediate models at the end of three immediately prior epochs as T_1, T_2, and T_3, with T_3 being the model from the epoch immediately prior to T. We measure the transferability of adversarial examples of each of T_i with T. For comparison, we also train three additional models S_1, S_2, and S_3 that are trained by exactly the same
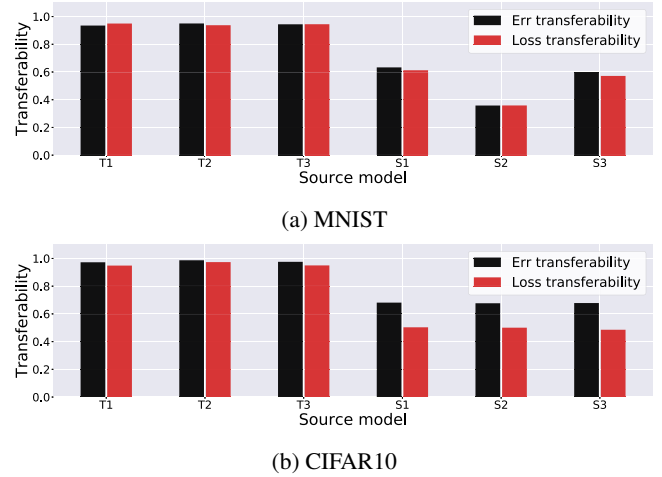

(a) MNIST

(b) CIFAR10

Figure 1: Error rate transferability and loss transferability with the different source models.

training method as T but with different random seeds. And we measure the transferability between each of S_i and T.

To measure transferability from the source model to the targeted model, we use two metrics. The first metric is *error rate transferability* used in [1, 19], which is the ratio of the number of adversarial examples misclassified by source model to that of the targeted model. The other metric is *loss transferability*, which is the ratio of the loss value caused by adversarial examples on the source model to the loss value caused by the same examples on the targeted model.

We conduct experiments on both MNIST and CIFAR10 dataset, and the results are shown on Figure 1. We find that, compared to the baseline models S_i, the models T_i from neighboring epochs of T have higher transferability for both transferability metrics (the transferability metrics for all models T_i are larger than 0.93). This provides strong empirical evidence that adversarial examples generated in one epoch still retain some strength in subsequent epochs.

Inspired by the above result, we state the following hypothesis. **Hypothesis:** *Repeatedly reusing perturbations from the previous epoch can accumulate attack strength epoch by epoch. Compared to current methods that iterate from original images in each epoch, this can allow us to use few attack iterations to generate the same strong adversarial examples.*

3.2. Accumulative PGD-k attack

To validate the aforementioned hypothesis, we design an *accumulative PGD-k attack*. As shown in Figure 2b, we longitudinally connect models in each epoch by directly reusing the attack perturbation of the previous epoch. Accumulative PGD-k attack Figure 2b generates adversarial examples for f_{n-m+1} first. Then, for the following epochs, the attack is performed based on the accumulated perturba-
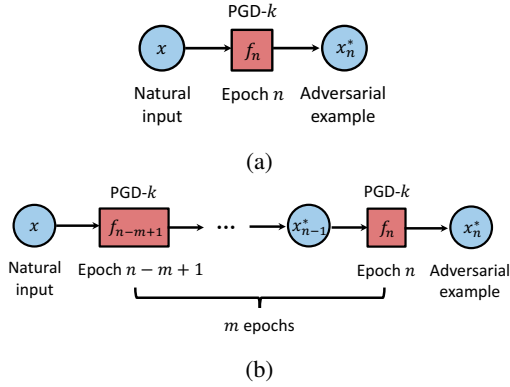
tions from previous epochs.

(a)

(b)

Figure 2: The traditional PGD-k attack (a) and the accumulative PGD-k attack through m epochs (b). The PGD-k attack is performed on the model in the red rectangle to get adversarial examples.

To compare the attack strength of two attacks, we use Madry's method [18] to adversarially train two models on MNIST and CIFAR10 and evaluate the loss value $\mathcal{L}(f_n(x_n^*), y)$ of adversarial examples generated by two attacks. Figure 3 summarises the evaluation result. We can find that, with more epochs m involved in the attack, accumulative PGD-k attack can achieve a higher loss value with the same number of attack iterations k.

Especially, when adversarial examples are transferred through a large number of epochs, even accumulative PGD-1 attack can cause high attack loss. For MNIST, accumulative PGD-1 attack can achieve the same attack loss as traditional PGD-35 attack when $m = 40$. For CIFAR10, accumulative PGD-1 attack can achieve the same attack loss as traditional PGD-12 attack when $m = 10$.

This result indicates that, with high transferability between epochs, adversarial perturbations can be reused effectively, which allows us to use fewer attack iterations to generate the same or stronger adversarial examples. Reuse of perturbations across epochs can help us reduce the number of attack iterations k in PGD-k, leading to more efficient adversarial training. Next section describes our proposed algorithm, **ATTA** (Adversarial Training with Transferable Adversarial examples), based on this property.

4. Adversarial training with transferable adversarial examples

The discussion on transferability in Section 3 suggests that adversarial examples can retain attack strength in subsequent training epochs. The results of accumulative attack in Section 3 suggest that stronger adversarial examples can be generated by accumulating attack strength. This property inspires us to *link* adversarial examples between adjacent
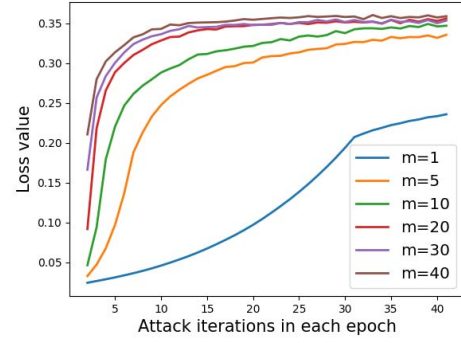
(a) MNIST
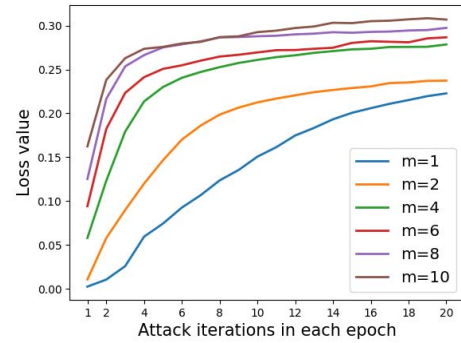
(b) CIFAR10

Figure 3: Given the number of epochs m, the relationship between the number of attack iterations k and the attack loss. $m = 1$ stands for the traditional PGD-k attack.

epochs as shown in Figure 4. Unlike starting from a natural image x to generate an adversarial example in each epoch as shown in Figure 4(a), we start from a previously saved adversarial example x_i^* from the previous epoch to generate an adversarial example (Figure 4(b)). To improve the transferability between epochs, we use a connection function C to link adversarial examples, which transforms x_i^* to a start point x_{i+1} for the next epoch. During the training, with repeatedly reusing adversarial example between epochs, attack strength can be accumulated epoch by epoch:

$$x_{i+1}^* = \mathcal{A}(f_{i+1}, x, y, C(x_i^*))$$

where $\mathcal{A}$ is the attack algorithm, C is a connection function (described in the next section). f_i is the model in the i-th epoch, x_i^* is the adversarial examples generated in the i-th epoch and x, y are natural image and benchmark label [1].

As shown in the previous section, adversarial examples can achieve high attack strength as they are transferred across epochs via the above linking process, rather than

---
[1] Note that the x_{i+1}^* is still bounded by the natural image x, rather than $C(x_i^*)$.
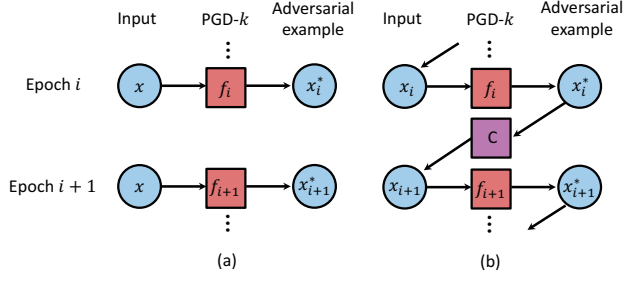
Figure 4: Traditional adversarial training (PGD-k) (a) and ATTA-k (b). C is the connection function which improves the transferability between epochs.

Figure 5: The workflow of inversed data augmentation.

starting from natural images. This, in turn, should allow us to train a more robust model with fewer attack iterations.

4.1. Connection function design

Designing connection function can help us overcome two challenges that we encountered in achieving high transferability of adversarial examples from one epoch to the next:

1. *Data augmentation problem:* Data augmentation is a commonly used technique to improve the performance of deep learning. It applies randomly transformation on original images so that models can be trained with various images in different epochs. This difference can cause a mismatch of images between epochs. Since the perturbation is calculated with the gradient of image, if we directly reuse the perturbation, the mismatch between reused perturbation and new augmented image can cause a decrease in attack strength. Simply removing data augmentation also hurts the robustness. We experimentally show this negative influence of these two alternatives in Section 5.3.

2. *Drastic model parameter change problem:* As discussed in Section 3, similar parameters between models tends to cause a similar decision boundary and thus high transferability. Unfortunately, model parameters tend to change drastically at the early stages of training. Thus, adversarial perturbations in early epochs tend to be useless for subsequent epochs.

Overcoming challenges: Inverse data augmentation. To address the first issue, we propose a technique called *inverse data augmentation* so that adversarial examples retain a high transferability between training epochs despite data augmentation. Figure 5 shows the workflow with inverse data augmentation. Some transformations (like cropping and rotation) pad augmented images with background pixels. To transfer a perturbation on background pixels, our method transfers the padded image x_i rather than standard
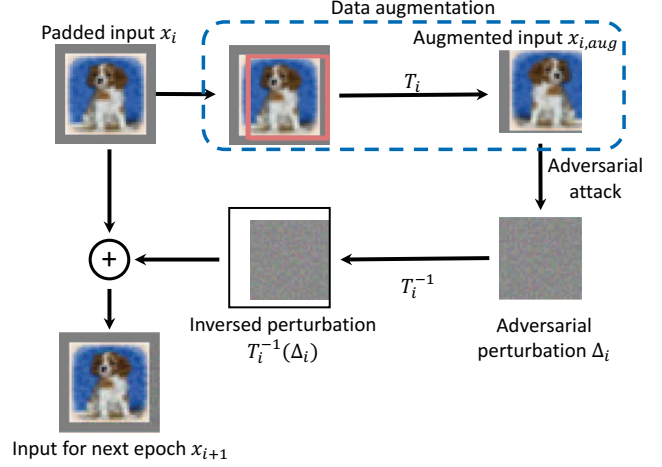
images so that the background pixels used by data augmentation can be covered by these paddings.

After the adversarial perturbation of augmented image is generated, we can perform the inverse transformation[2] $T^{-1}(\cdot)$ to calculate the inversed perturbation $T^{-1}(\Delta_i)$ on the padded image x_i. By adding $T^{-1}(\Delta_i)$ to x_i, we can store and transfer all perturbation information in x_{i+1} to next epoch. (Note that, when we perform the adversarial attack on the augmented image $x_{i,aug}$, the perturbation is still bounded by the natural image rather than $x_{i,aug}$.)

Periodically reset perturbation. To solve the second issue, we propose a straightforward but effective solution: our method resets the perturbation and lets adversarial perturbations be accumulated from the beginning periodically, which mitigates the impact caused by early perturbations.

4.2. Attack Loss

Various attack loss functions are used in different adversarial training algorithms. For example, TRADES [34] uses robustness loss (Equation 2) as the loss to generate the adversarial examples.

$$\mathcal{L}(f(x^*), f(x)) \tag{2}$$

where $\mathcal{L}$ is the loss function, f is the model and x, x^* are natural and adversarial example respectively. This loss represents how much the adversarial example diverges from the natural image. Zhang *et al*. [34] shows that this loss has a better performance in the TRADES algorithm.

In our method, we use the following loss function:

$$\mathcal{L}(f(x^*), y) \tag{3}$$

It represents how much the adversarial examples diverges to the benchmark label.

---

[2]In adversarial training, most data augmentation methods used are linear transformations which are easy to be inversed.

1182

The objection ($f(x)$) of Equation 2 for adversarial attack varies across epochs, which may weaken the transferability between epochs. Equation 3 applies a fixed objection (y), which doesn't have this concern. In addition, Equation 3 has a smaller computational graph, which can reduce computational overhead during training.

The overall training method is described in Algorithm 1.

Algorithm 1 Adversarial Training with Transferable Adversarial Examples (ATTA)

1: **Input**: Padded training dataset D_{nat}, model f_θ, attack algorithm $\mathcal{A}$, perturbation bound ϵ, the number of epochs to reset perturbation $reset$
2: Initialize θ
3: Initialize D by cloning D_{nat}
4: **for** $epoch = 1 \cdots N$ **do**
5:     **for** x_{nat}, y in D_{nat} and corresponding $x \in D$ **do**
6:         **if** $epoch \% reset = 0$ **then**
7:             $\beta \leftarrow$ a small random perturbation
8:             $x \leftarrow x_{nat} + \beta$
9:         **end if**
10:         Store the transformation T_{aug} for the inverse augmentation:
11:             $x_{aug}, x_{nat,aug}, T_{aug} \leftarrow \text{data\_aug}(x, x_{adv})$
12:             $x^* \leftarrow \mathcal{A}(f_\theta, x_{nat,aug}, y, x_{aug}, \epsilon)$
13:             $\theta \leftarrow \theta - \nabla_{f_\theta} \frac{\partial \mathcal{L}(f_\theta, x^*, y)}{\partial \theta}$
14:             $x \leftarrow \text{inverse\_aug}(x, x^*, T_{aug})$
15:     **end for**
16: **end for**

---

5. Evaluation

In this section, we integrate ATTA with two popular adversarial training methods: Madry's Adversarial Training (MAT) [18] and TRADES [34]. By evaluating the training time and robustness, we show that ATTA can provide a better trade-off than other adversarial training methods. To understand the contribution of each component to robustness, we conduct an ablation study.

5.1. Setup

Following the literature [18, 32, 34], we use both MNIST [14] and CIFAR10 dataset [12] to evaluate ATTA.

For the MNIST dataset, the model has four convolutional layers followed by three full-connected layers which is same architecture as used in [18,34]. The adversarial perturbation is bounded by l_∞ ball with size $\epsilon = 0.3$.

For the CIFAR10 dataset, we use the wide residual network WRN-34-10 [31] which is same as [18, 34]. The perturbation is bounded by l_∞ ball with size $\epsilon = 0.031$.

5.2. Efficiency and effectiveness of ATTA

In this part, we evaluate the training efficiency and the robustness of ATTA, comparing it to state-of-the-art adversarial training algorithms. To better verify effectiveness of ATTA, we also evaluate ATTA under various attacks.

5.2.1 Training efficiency

We select four state-of-the-art adversarial training methods as baselines: MAT [18], TRADES [34], YOPO [32] and Free [23]. For MNIST, the model trained with ATTA can achieve a comparable robustness with up to 12.2 times training efficiency and, for CIFAR10, our method can achieve a comparable robustness with up to 14.1 times training efficiency. Compared to MAT trained with PGD, our method improve the accuracy up to 7.2% with 3.7 times training efficiency.

Defense / Attack		Natural	PGD-40	Time (sec)
MAT	PGD-1	99.52%	15.82%	226
	PGD-40	**99.37%**	**96.21%**	**3933**
	YOPO-5-10	99.15%	93.69%	789
	ATTA-1	**99.45%**	**96.31%**	**297**
	ATTA-40	99.23%	97.28%	4650
TRADES	PGD-1	99.41%	39.53%	583
	PGD-40	**98.89%**	**96.54%**	**6544**
	ATTA-1	**99.03%**	**96.10%**	**460**
	ATTA-40	98.21%	96.03%	4660

Table 3: The result of different attacks on MNIST dataset.

MNIST. The experiment results of MNIST are summarised in Table 3. For MAT, to achieve comparable robustness, ATTA is about 12.2 times faster than the traditional PGD training method. Even with one attack iteration, model trained with ATTA-1 achieves 96.31% adversarial accuracy within 297 seconds. For TRADES, we get a similar result. With one attack iteration, ATTA-1 achieves 96.1% adversarial accuracy within 460 seconds, which is about 13 times faster than TRADES(PGD-40). Compared to another fast adversarial training method YOPO, ATTA-1 is about 2.7 times faster and achieves higher robustness (96.1% versus YOPO's 93.69%).

CIFAR10. We summarise the experiment results of CIFAR10 in Table 4. For MAT, compared to PGD-10, ATTA-1 achieves 3.89% higher adversarial accuracy with about 15.1 times training efficiency, and ATTA-10 improves adversarial accuracy by 7.26% with 2.9 times training efficiency when the model is trained with 10 attack iterations.

Attack / Defense	Natural	PGD-20	Time (min)
MAT			
PGD-1	93.18%	22.3%	435
PGD-3	89.95%	41.38%	785
PGD-10	**87.49%**	**47.07%**	**2027**
Free($m = 8$)	85.54%	47.68%	640
YOPO-5-3	86.43%	48.24%	335
ATTA-1	**85.71%**	**50.96%**	**134**
ATTA-3	85.44%	52.56%	267
ATTA-10	83.80%	54.33%	690
TRADES			
PGD-1	93.58%	35.52%	592
PGD-3	88.52%	54.20%	861
PGD-10	**84.13%**	**56.6%**	**2028**
YOPO-3-4 [3]	87.82%	46.13%	-
ATTA-1	85.04%	54.50%	199
ATTA-3	**84.23%**	**56.36%**	**320**
ATTA-10	83.67%	57.34%	752

Table 4: The result of different attacks on CIFAR10 dataset.

For TRADES, ATTA-3 achieves comparable adversarial accuracy to PGD-10 with 5.3 times faster training efficiency. By comparing the experiment results with YOPO and Free, for MAT, our method is 3.78 (1.5) times faster than Free (YOPO) with 3.28% (2.72%) better adversarial accuracy.

To better understand the performance of ATTA, we present the trade-off between training time and robustness of different methods in Figure 6. We find that our method (indicated by the solid markers in the left-top corner) gives a better trade-off on efficiency and effectiveness on adversarial training.

5.2.2 Defense under other attacks

To verify the robustness of our method, we evaluate models in the previous section with other attacks: PGD-100 [13], FGSM [8], CW-20 [3].

As shown in Table 5, models trained with ATTA are still robust to other attacks. Compared to baselines, our method still achieves a comparable or better robustness under other attacks. We find that, although ATTA-40 has a similar robustness to PGD-40 under PGD-20 attack, with a stronger attack (e.g. PGD-100), ATTA-40 shows a better robustness

---

³The author-implemented YOPO-3-4 can't converge in our experiment. We pick the accuracy data from YOPO paper.
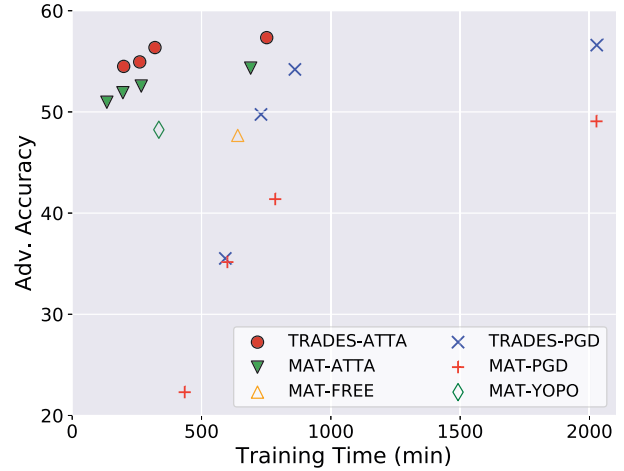

Figure 6: The scatter plot presents adversarial accuracy against PGD-20 attack and training time of different adversarial training methods on CIFAR10.

Defense	PGD100	FGSM	CW20
MNIST			
M-PGD-40	94.69%	97.37%	99.06%
M-ATTA-40	96.85%	98.55%	98.02%
CIFAR10			
M-PGD-10	46.77%	63.5%	56.7%
M-ATTA-10	52.6%	63.49%	75.37%
T-PGD-10	55.36%	63.02%	79.4%
T-ATTA-10	56.39%	64.1%	82.01%

Table 5: The robustness comparison between ATTA and PGD under other attacks. The first 'M' and 'T' stand for MAT and TRADES, respectively.

5.2.3 ATTA with ImageNet

Table 6 shows the results for training an ImageNet Classifier [6] with ATTA for $l_\infty (\epsilon = 2/255)$ attack with the loss used in MAT. We also show previously reported results of Free ($m = 4$) from [23]. Shafahi *et al.* [23] report requiring about 50 hours with four P100s to adversarially train their model. ATTA required 97 hours with one 2080Ti.

Defense	Natural	PGD10	PGD50	PGD100
Free($m = 4$)	64.44%	43.52%	43.39%	43.40%
ATTA-2	60.70%	44.57%	43.57%	43.51%

Table 6: Evaluation of ATTA on ImageNet

5.3. Ablation study

To study the contribution of each component to robustness, we do an ablation study on inverse data augmentation and different attack loss functions.

Inverse data augmentation. To study the robustness gain of inverse data augmentation(i.d.a.), we use ATTA-1 to adversarially train models by reusing the adversarial perturbation directly. As shown in Table 7, for both MAT (ATTA-1) and TRADES (ATTA-1), models trained with inverse data augmentation achieve about 6% higher accuracy, which means that inverse data augmentation does help improve the transferability between training epochs. As discussed in 4.1, another alternative is to remove data augmentation. However, Table 7 shows that removal of data augmentation hurts both natural accuracy and robustness.

Attack Defense	Natural	PGD-20
MAT(w/o d.a., w/o i.d.a.)	82.53%	41.64%
MAT(w/ d.a., w/o i.d.a.)	91.65%	42.55%
MAT(w/ d.a., w/ i.d.a.)	85.71%	50.96%
TRADES(w/o d.a., w/o i.d.a.)	81.87%	41.65%
TRADES(w/ d.a., w/o i.d.a.)	90.46%	48.38%
TRADES(w/ d.a., w/ i.d.a.)	85.04%	54.50%

Table 7: The accuracy under PGD attack of models trained by ATTA-1 with or without d.a.(data augmentation) and i.d.a. (inverse data augmentation).

Attack loss. Zhang *et al.* [34] show that, for TRADES, using Equation 2 leads to a better robustness. However, in this attack loss, we noticed that both inputs to the loss function $\mathcal{L}$ are related to the model f. Since the model f is updated every epoch, compared to Equation 3 whose y is fixed, the instability of $f(x^*)$ and $f(x)$ may have a larger influence on transferability. To analyze the performance difference between these two attack losses in ATTA, we train two TRADES(ATTA-1) models with different attack losses. In Table 8, we find that Equation 3 leads to 2.48% higher accuracy against PGD-20 attack. This result suggests that the higher stability of Equation 3 helps ATTA increase transferability between training epochs.

6. Related work

Several works [2, 4, 10, 15, 18, 23, 24, 28, 29, 32] focus on enhancing either the efficiency or effectiveness of the adversarial training that was first proposed in [13]. We highlight most relevant work here. YOPO [32]. TRADES [34] improves the robustness of an adversarially trained model by adding a robustness regularizer in the loss function.

Attack Defense	Natural	PGD-20
TRADES loss	85.95%	52.08%
MAT loss	85.04%	54.50%

Table 8: The accuracy under PGD attack of models trained by TRADES(ATTA-1) with MAT loss and TRADES loss.

BAT [29] uses perturbed labels (besides perturbed images) to generate strong attacks with fewer iterations than Madry *et al.* [18]. ATTA does not perturb labels and is based on a completely different insight of transferability.

Free [23] speeds up adversarial training by recycling gradients calculated during training when generating adversarial examples. In contrast, ATTA speedups adversarial training with a completely different insight based on transferability. While recycling gradients, Free does reuse perturbations through iterations on the top of the same batch of images (inner-batch). In contrast, ATTA reuses perturbations *across* epochs (inter-epoch). Inter-epoch reuse is a non-trivial problem involving new challenges discussed in Section 4.1, requiring use of connection functions to overcome them. Free did not discuss transferability, but our analysis of transferability potentially provides an alternative way of understanding Free. For many training scenarios, ATTA both outperforms existing techniques as well as achieves better robustness.

7. Conclusion

ATTA is a new method for iterative attack based adversarial training that significantly speeds up training time and improves model robustness. The key insight behind ATTA is high transferability between models from neighboring epochs, which is firstly revealed in this paper. Based on this property, the attack strength in ATTA can be accumulated across epochs by repeatedly reusing adversarial perturbations from the previous epoch. This allows ATTA to generate the same (or even stronger) adversarial examples with fewer attack iterations. We evaluate ATTA and compare it with state-of-the-art adversarial training methods. It greatly shortens the training time with comparable or even better model robustness. More importantly, ATTA is a generic method and can be applied to enhance the performance of other iterative attack based adversarial training methods.

8. Acknowledgments

References

[1] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *AAAI*, pages 2687–2695, 2018.

[2] Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2018.

[3] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017.

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

[5] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2014.

[9] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.

[10] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.

[11] Yunseok Jang, Tianchen Zhao, Seunghoon Hong, and Honglak Lee. Adversarial defense via learning to generate diverse attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2740–2749, 2019.

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[13] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2017.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Hyeungill Lee, Sungyeob Han, and Jungwoo Lee. Generative adversarial trainer: Defense to adversarial perturbations with gan. *arXiv preprint arXiv:1705.03387*, 2017.

[16] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *AAAI*, 2020.

[17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[19] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[20] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519. ACM, 2017.

[21] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] 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.

[23] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Neural Information Processing Systems (NeurIPS)*, 2019.

[24] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. *arXiv preprint arXiv:1811.11304*, 2018.

[25] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540. ACM, 2016.

[26] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX Workshop on Offensive Technologies (WOOT 18)*, 2018.

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[28] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.

[29] Jianyu Wang and Haichao Zhang. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6629–6638, 2019.

[30] Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope.

In *International Conference on Machine Learning (ICML)*, 2018.

[31] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[32] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[33] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations (ICLR)*, 2019.

[34] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.