Supplementary Material for "Prior-Guided Adversarial Initialization for Fast Adversarial Training"

Xiaojun Jia[1,2,†], Yong Zhang[3,⋆] Xingxing Wei[4], Baoyuan Wu[5], Ke Ma[6], Jue Wang[3], Xiaochun Cao[7,⋆]

[1]SKLOIS, Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3] Tencent, AI Lab, Shenzhen, China
[4] Institute of Artificial Intelligence, Beihang University, Beijing, China.
[5]School of Data Science, Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China
[6] School of Computer Science and Technology, UCAS, Beijing, China
[7] School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China
jiaxiaojun@iie.ac.cn; zhangyong201303@gmail.com; xxwei@buaa.edu.cn; wubaoyuan@cuhk.edu.cn; make@ucas.ac.cn; arphid@gmail.com; caoxiaochun@mail.sysu.edu.cn

This supplementary material contains the following contents:

- The difference between the training processes of FAT and SAT methods on the more datasets. (see Sec. 1).
- Detailed algorithms of FGSM-BP, FGSM-EP and FGSM-MEP in Sec. 3.3 of the manuscript (see Sec. 2).
- Proof of **Proposition 1** in Sec. 3.4 of the manuscript.(see Sec. 3).
- Detailed hyper-parameter settings. (see Sec. 4).
- Experiments with a larger model as the backbone (see Sec. 5).
- More comparative experiments with using a cyclic learning rate strategy (see Sec. 6).

1 Difference between the Training Processes of FAT and SAT Methods

In Fig.2 of the manuscript (Sec.3.1), we compared the intermediate results of the training processes of FAT and SAT on the CIFAR-10 dataset. Here, we also reinvestigate catastrophic overftting on more datasets, *i.e.,* CIFAR-100 [6] and Tiny ImageNet [2]. We adopt ResNet18 [3] as the backbone on CIFAR-100 and PreActResNet18 [4] on Tiny ImageNet. The training setting is presented in the Sec.4.1 of the manuscript.

We observe a similar overfitting phenomenon on CIFAR-100 and Tiny ImageNet. There is also a distinct difference between the attack success rates (ASRs)

⋆ Correspondence to: Yong Zhang (zhangyong201303@gmail.com) and Xiaochun Cao (caoxiaochun@mail.sysu.edu.cn).† Work done in an internship at Tencent AI Lab.

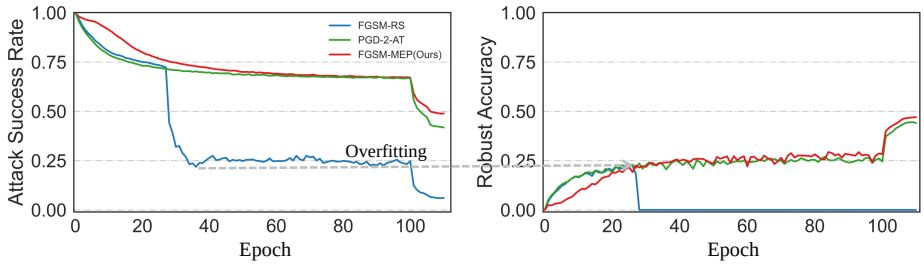

Fig. 1. The difference between the training processes of FAT and SAT methods on the CIFAR-100 dataset. Left: the attack success rate of generated AEs. Right: the PGD-10 robust accuracy of the target model.

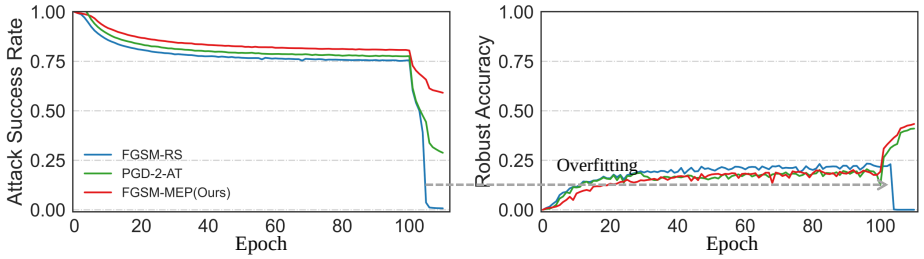

Fig. 2. The difference between the training processes of FAT and SAT methods on the Tiny ImageNet dataset. Left: the attack success rate of generated AEs. Right: the PGD-10 robust accuracy of the target model.

of AEs used in FAT and SAT in the late training stage. The results on CIFAR-100 are shown in Fig. 1. It can be observed that the ASRs of FGSM-RS drop sharply at the 26-th epoch, resulting in the dramatic decreases of the robust accuracy. But the ASRs of PGD-2-AT do not drop sharply during the whole training. PGD-2-AT does not suffer from catastrophic overftting. The results on Tiny ImageNet are shown in Fig. 2. It can be observed that the ASRs of FGSM-RS drop sharply at the 105-th epoch.

2 Detailed Algorithms of FGSM-BP, FGSM-EP and FGSM-MEP

FGSM-BP uses the adversarial perturbations from the previous batch onto clean images and then conduct FGSM based on the perturbed examples. The FGSM-MEP algorithm is shown in Algorithm 1. FGSM-EP uses the adversarial perturbations from the previous epoch onto clean images and then conduct FGSM based on the perturbed examples. The FGSM-EP algorithm is shown in Algorithm 2. And The FGSM-MEP algorithm is shown in Algorithm 3.

3 Proof of Proposition 1

Proof. The first part of the desired result is true by the Jensen's inequality as

$$\left(\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\left\|\hat{\boldsymbol{\delta}}_{adv}\right\|_2\right]\right)^2 \leq \mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\left\|\hat{\boldsymbol{\delta}}_{adv}\right\|_2^2\right]. \tag{1}$$

In the following we focus on $\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2^2\right]$. Denote

$$\boldsymbol{\nabla} = \nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{pgi}};\mathbf{w}),\mathbf{y}), \tag{2}$$

we have

$$\begin{aligned}
&\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2^2\right] \\
=\ &\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}}\left[\left\|\Pi_{\boldsymbol{\Omega}}\left[\boldsymbol{\delta}_{\mathbf{pgi}}+\alpha\cdot\operatorname{sign}(\boldsymbol{\nabla})\right]\right\|_2^2\right] \\
=\ &\sum_{i=1}^{d}\mathbb{E}_{\delta_{pgi}(i)}\left[\Pi_{\boldsymbol{\Omega}}\left[\delta_{pgi}(i)+\alpha\cdot\operatorname{sign}(\nabla_i)\right]^2\right] \\
\leq\ &d\cdot\mathbb{E}_{\delta_{pgi}(i)}\left[\mathbf{min}\left\{\frac{\epsilon^2}{d},\left(\delta_{pgi}(i)+\alpha\cdot\operatorname{sign}(\nabla_i)\right)^2\right\}\right] \\
=\ &d\cdot\mathbb{E}_{r_i}\left[\mathbb{E}_{\delta_{pgi}(i)}\left[\mathbf{min}\left\{\frac{\epsilon^2}{d},\left(\delta_{pgi}(i)+\alpha\cdot\operatorname{sign}(\nabla_i)\right)^2\right\}\right]\ \middle|\ \operatorname{sign}(\nabla_i)=r_i\right],
\end{aligned} \tag{3}$$

where the last setp follows the low of total expectation as $r_i := \operatorname{sign}(\nabla_i)$ is also a random variable depending on $\delta_t(i)$.

As r_i is a binary random variable, d is the feature dimension, and $\alpha < \epsilon$, it holds that

$$\begin{aligned}
-\epsilon d^{-1/2} &> -\epsilon+\alpha \\
\epsilon d^{-1/2} &< \epsilon+\alpha,
\end{aligned}$$

and we could separate the procedure into the following two cases:

(i) $r_i = 1$, the inner conditional expectation has the form:

$$\begin{aligned}
&\int_{-\epsilon}^{\epsilon} \mathbf{min}\left\{\frac{\epsilon^2}{d}, (\delta_{pgi}(i)+\alpha)^2\right\}\frac{1}{2\epsilon}d\delta_{pgi}(i) \\
=\ & \frac{1}{2\epsilon}\int_{-\epsilon+\alpha}^{\epsilon+\alpha} \mathbf{min}\left\{\frac{\epsilon^2}{d}, x^2\right\}dx \\
=\ & \frac{1}{2\epsilon}\left(\int_{\epsilon d^{-1/2}}^{\epsilon+\alpha}\frac{\epsilon^2}{d}dx + \int_{-\epsilon d^{-1/2}}^{\epsilon d^{-1/2}} x^2 dx + \int_{-\epsilon+\alpha}^{-\epsilon d^{-1/2}}\frac{\epsilon^2}{d}dx\right) \\
=\ & \frac{\epsilon}{2d}\left(2\epsilon - 2\epsilon d^{-1/2}\right) + \frac{1}{3}\epsilon^2 d^{3/2} \\
=\ & \frac{\epsilon^2}{2d} - \frac{\epsilon^2}{d^{3/2}} + \frac{1}{3}\epsilon^2 d^{-3/2} \\
\leq\ & \frac{\epsilon^2}{d}.
\end{aligned} \tag{4}$$

(ii) $r_i = -1$, the inner conditional expectation will be:

$$\begin{aligned}
&\int_{-\epsilon}^{\epsilon} \mathbf{min}\left\{\frac{\epsilon^2}{d}, (\delta_{pgi}(i)-\alpha)^2\right\}\frac{1}{2\epsilon}d\delta_{t}(i) \\
=\ & \frac{1}{2\epsilon}\int_{-\epsilon-\alpha}^{\epsilon-\alpha} \mathbf{min}\left\{\frac{\epsilon^2}{d}, x^2\right\}dx \\
=\ & \frac{1}{2\epsilon}\left(\int_{\epsilon d^{-1/2}}^{\epsilon-\alpha}\frac{\epsilon^2}{d}dx + \int_{-\epsilon d^{-1/2}}^{\epsilon d^{-1/2}} x^2 dx + \int_{-\epsilon-\alpha}^{-\epsilon d^{-1/2}}\frac{\epsilon^2}{d}dx\right) \\
=\ & \frac{\epsilon}{2d}\left(2\epsilon - 2\epsilon d^{-1/2}\right) + \frac{1}{3}\epsilon^2 d^{3/2} \\
=\ & \frac{\epsilon^2}{2d} - \frac{\epsilon^2}{d^{3/2}} + \frac{1}{3}\epsilon^2 d^{-3/2} \\
\leq\ & \frac{\epsilon^2}{d}.
\end{aligned} \tag{5}$$

Combining (4) and (5) together with (3), we obtain

$$\begin{aligned}
\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2\right] &\leq \sqrt{\mathbb{E}_{\hat{\boldsymbol{\delta}}_{adv}\sim\boldsymbol{\Omega}}\left[\|\hat{\boldsymbol{\delta}}_{adv}\|_2^2\right]} \\
&\leq \sqrt{\frac{1}{d}}\cdot\epsilon.
\end{aligned}$$

Remark. If $\boldsymbol{\Omega}$ is a bounded set like $\boldsymbol{\Omega} = \left\{\hat{\boldsymbol{\delta}}_{adv} \ :\ \|\hat{\boldsymbol{\delta}}_{adv} - \boldsymbol{\delta}_{pgi}\|_2^2 \leq \epsilon^2\right\}$, we can obtain the upper bound of the proposed method which is $\sqrt{\frac{1}{d}}\cdot\epsilon$. It is less than

Algorithm 1 FGSM-BP

Require: The epoch N, the maximal perturbation ϵ, the maximal label perturbation $\epsilon_{\mathbf{y}}$, the step size α, the dataset $\mathcal{D}$ including the benign sample batch $\mathbf{x}_B$ and the label $\mathbf{y}_B$, the dataset batch number M_B, the network $f(\cdot, \mathbf{w})$ with parameters $\mathbf{w}$, the decay factor μ, the hyper-parameter λ, the adversarial initialization set $\mathcal{D}^{\boldsymbol{\delta}}$.

for $n = 1, ..., N$ **do**
  for $i = 1, ..., M_B$ **do**
    if $i == 1$ **then**
      $\boldsymbol{\delta}_{pgi} = \mathbf{U}(-\epsilon, \epsilon)$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\boldsymbol{\delta}_{adv} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}^{\boldsymbol{\delta}} = \boldsymbol{\delta}_{adv}$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    else
      $\boldsymbol{\delta}_{pgi} = \mathcal{D}^{\boldsymbol{\delta}}$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\boldsymbol{\delta}_{adv} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}^{\boldsymbol{\delta}} = \boldsymbol{\delta}_{adv}$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    end if
  end for
end for

the bound $\sqrt{\frac{d}{3}} \cdot \epsilon$ of FGSM-RS provided in [1] when $d > \sqrt{3}$. It requires that the prior-guided adversarial perturbation $\boldsymbol{\delta}_{pgi}$ is not far from the current adversarial perturbation $\hat{\boldsymbol{\delta}}_{adv}$. Moreover, d represents the dimension of image data whose value is much larger than 4.

4 Detailed Hyper-parameter Settings

In this section, we present the detailed hyper-parameter settings of the proposed FGSM-MEP. There are two hyper-parameters in the proposed FGSM-MEP: the decay factor μ and the lambda λ. To evaluate the influence of the decay factor on FGSM-MEP, we perform an experiment on the CIFAR-10 database. The metric is the robust accuracy under PGD-50, C&W and AA attack. The results are shown in Fig 3. It is observed that when the decay factor μ is set to 0.3, FGSM-MEP achieves the best performance under all attack scenarios.

To evaluate the influence of the lambda λ, we conduct an experiment by using FGSM-MEP with different lambda values λ. The metric is also the robust accuracy unde rPGD-50, C&W and AA attack. The results are shown in Fig 3. It can be observed that when the lambda λ is set to 10, FGSM-MEP achieves the best performance under all attack scenarios. Hence, we set the the decay factor μ to 0.3 and the lambda λ to 8 for FGSM-MEP to conduct adversarial training.

Algorithm 2 FGSM-EP

Require: The epoch N, the maximal perturbation ϵ, the maximal label perturbation $\epsilon_{\mathbf{y}}$, the step size α, the dataset $\mathcal{D}$ including the benign sample $\mathbf{x}$ and the label $\mathbf{y}$, the dataset size M, the network $f(\cdot, \mathbf{w})$ with parameters $\mathbf{w}$, the decay factor μ, the hyper-parameter λ, the adversarial initialization set $\mathcal{D}^{\boldsymbol{\delta}}$.

for $n = 1, ..., N$ **do**
  for $i = 1, ..., M$ **do**
    if $n == 1$ **then**
      $\boldsymbol{\delta}_{pgi} = \mathbf{U}(-\epsilon, \epsilon)$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\boldsymbol{\delta}_{\mathbf{adv}} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}_i^{\boldsymbol{\delta}} = \boldsymbol{\delta}_{adv}$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    else
      $\boldsymbol{\delta}_{pgi} = \mathcal{D}_i^{\boldsymbol{\delta}}$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\boldsymbol{\delta}_{adv} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}_i^{\boldsymbol{\delta}} = \boldsymbol{\delta}_{adv}$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    end if
  end for
end for

5 Experiments with a Larger Model as the Backbone

As for using a larger architecture, we conduct an experiment on CIFAR-10 with WideResNet34-10 [13] as the backbone. The results are shown in Table 1. It can be observed that compared with other FAT methods, the proposed FGSM-MEP can achieve the best robustness performance under all attack scenarios. In terms of training efficiency, the proposed FGSM-MEP requires a bit more calculation cost than FGSM-RS, but much less time than other FAT variants.

6 Experiments with a Cyclic Learning Rate Strategy

In the manuscript, we conducted all experiments using the multi-step learning rate strategy. Here, we also conduct comparative experiments using a cyclic learning rate strategy [9] on CIFAR-10 and CIFAR-100. Following [1,5], we set the maximum learning rate of FGSM-GA [1] and FGSM-CKPT [5] to 0.3. Following [12], we set the maximum learning rate of FGSM-RS [12], Free [8], GAT [10] and NuAT [11], and the proposed method to 0.2.

The results are shown in Table 2 and Table 3. We can observe the similar phenomenons as the models trained using a multi-step learning rate strategy. Specifically, compared with other fast AT methods, the proposed FGSM-MEP can achieve the best adversarial robustness among the fast AT methods under all adversarial attack scenarios. As [1] discovered, using the cyclic learning rate strategy can improve the robustness against adversarial examples. But it also

Algorithm 3 FGSM-MEP

Require: The epoch N, the maximal perturbation ϵ, the maximal label perturbation $\epsilon_{\mathbf{y}}$, the step size α, the dataset $\mathcal{D}$ including the benign sample $\mathbf{x}$ and the label $\mathbf{y}$, the dataset size M, the network $f(\cdot, \mathbf{w})$ with parameters $\mathbf{w}$, the decay factor μ, the hyper-parameter λ, the adversarial initialization set $\mathcal{D}^{\boldsymbol{\delta}}$ and the historical model gradient $\mathcal{D}^m$.

for $n = 1, ..., N$ **do**
  for $i = 1, ..., M$ **do**
    if $n == 1$ **then**
      $\boldsymbol{\delta}_{pgi} = \mathbf{U}(-\epsilon, \epsilon)$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\mathcal{D}_i^m = \mathbf{g}_c$
      $\boldsymbol{\delta}_{\mathbf{adv}} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}_i^{\boldsymbol{\delta}} = \boldsymbol{\delta}_{adv}$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    else
      $\boldsymbol{\delta}_{pgi} = \mathcal{D}_i^{\boldsymbol{\delta}}$
      $\mathbf{g}_c = \text{sign}\left(\nabla_{\mathbf{x}_i} \mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{pgi}; \mathbf{w}), \mathbf{y}_i)\right)$
      $\mathcal{D}_i^m = \mu \cdot \mathcal{D}_i^m + \mathbf{g}_c$
      $\boldsymbol{\delta}_{adv} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \mathbf{g}_c]$
      $\mathcal{D}_i^{\boldsymbol{\delta}} = \Pi_{[-\epsilon,\epsilon]}[\boldsymbol{\delta}_{pgi} + \alpha \cdot \text{sign}(\mathcal{D}_i^m)]$
      $\mathbf{w} \leftarrow \mathbf{w} - \nabla_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}_i + \boldsymbol{\delta}_{adv}; \mathbf{w}), \mathbf{y}_i) + \lambda \cdot \|f(\mathbf{x} + \boldsymbol{\delta}_{adv}; \mathbf{w}) - f(\mathbf{x} + \boldsymbol{\delta}_{pgi}; \mathbf{w})\|_2^2]$
    end if
  end for
end for

reduces the accuracy of clean images. Compared with FGSM-GA, the proposed FGSM-MEP achieves a higher robustness accuracy. For example, the FGSM-GA achieves the accuracy of about 43.06% under AA attack while the proposed FGSM-MEP achieves the accuracy of about 46.65% under AA attack. In terms of efficiency, our training process is about 3 times faster than Free-AT, 2.5 times faster than FGSM-GA, and 1.4 times faster than GAT and NuAT.

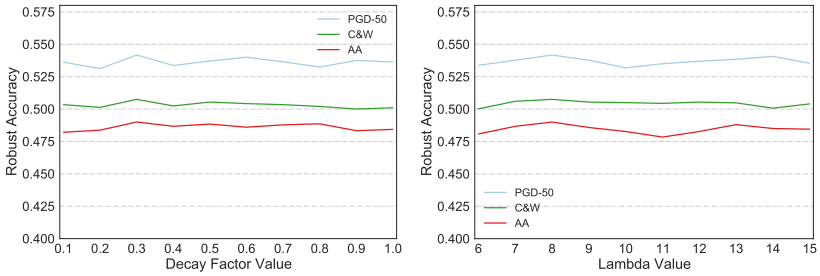

Fig. 3. Selection of the hyper-parameters decay factor μ and lambda λ uesd in the proposed method. Left: robust accuracy (%) of different decay factor values on the CIFAR10 database using ResNet18. Right: robust accuracy (%) of different lambda values on the CIFAR10 database using ResNet18.

Table 1. Comparisons of clean and robust accuracy (%) and training time (hour) with WideResNet34-10 on the CIFAR-10 database. Number in bold indicates the best of the fast AT methods.

CIFAR-10	Clean	PGD-10	PGD-20	PGD-50	AA	Time(h)
PGD-AT [7]	85.17	56.1	55.07	54.87	51.67	31.9h
FGSM-RS [12]	74.3	42.3	41.2	40.9	38.4	5.8h
FGSM-CKPT [5]	**91.8**	44.7	42.6	42.2	40.4	8.7h
NuAT [11]	85.30	55.8	54.68	53.75	50.06	11.8h
GAT [10]	85.17	56.3	55.23	54.97	50.01	12.9h
FGSM-GA [1]	82.1	48.9	47.1	46.9	45.7	20.3h
Free-AT [8]	80.1	47.9	46.7	46.3	43.9	23.7h
FGSM-MEP(ours)	85.09	**57.72**	**56.86**	**56.4**	**50.11**	8.3h

Table 2. Comparisons of clean and robust accuracy (%) and training time (minute) with ResNet18 on the CIFAR-10 database. Number in bold indicates the best of the fast AT methods. **All models are trained using a cyclic learning rate strategy.**

CIFAR-10		Clean	PGD-10	PGD-20	PGD-50	AA	Time(min)
FGSM-RS [12]	Best	83.75	48.05	46.47	46.11	42.92	15
	Last	83.75	48.05	46.47	46.11	42.92	
FGSM-CKPT [5]	Best	**89.08**	40.47	38.2	37.69	35.66	23
	Last	**89.08**	40.47	38.2	37.69	35.66	
NuAT [11]	Best	76.23	51.52	50.81	50.64	46.33	30
	Last	76.23	51.52	50.81	50.64	46.33	
GAT [10]	Best	81.91	50.43	49.82	49.62	45.24	33
	Last	81.91	50.43	49.82	49.62	45.24	
FGSM-GA [1]	Best	80.83	48.76	47.83	47.54	43.06	53
	Last	80.83	48.76	47.83	47.54	43.06	
Free-AT [8]	Best	75.22	44.67	43.97	43.72	40.30	58
	Last	75.22	44.67	43.97	43.72	40.30	
FGSM-MEP(ours)	Best	80.68	**52.48**	**51.69**	**51.5**	**46.65**	22
	Last	80.68	**52.48**	**51.69**	**51.5**	**46.65**	

Table 3. Comparisons of clean and robust accuracy (%) and training time (minute) with ResNet18 on the CIFAR-100 database. Number in bold indicates the best of the fast AT methods. **All models are trained using a cyclic learning rate strategy.**

CIFAR-100		Clean	PGD-10	PGD-20	PGD-50	AA	Time(min)
FGSM-RS [12]	Best	57.71	24.82	23.91	23.64	20.66	19
	Last	57.71	24.82	23.91	23.64	20.66	
FGSM-CKPT [5]	Best	**70.75**	10.85	7.86	6.32	2.07	25
	Last	**70.75**	10.85	7.86	6.32	2.07	
NuAT [11]	Best	59.52	27.17	22.52	20.29	11.45	31
	Last	59.52	27.17	22.52	20.29	11.45	
GAT [10]	Best	59.88	21.54	20.91	20.56	17.68	34
	Last	59.88	21.54	20.91	20.56	17.68	
FGSM-GA [1]	Best	55.44	27.14	26.43	26.19	22.08	55
	Last	55.44	27.14	26.43	26.19	22.08	
Free-AT [8]	Best	47.12	23.05	22.7	22.66	18.90	62
	Last	47.12	23.05	22.7	22.66	18.90	
FGSM-MEP(ours)	Best	56.69	**29.34**	**28.74**	**28.53**	**23.00**	24
	Last	56.69	**29.34**	**28.74**	**28.53**	**23.00**	

References

1. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. In: Annual Conference on Neural Information Processing Systems (2020)
2. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255. IEEE Computer Society (2009). https://doi.org/10.1109/CVPR.2009.5206848, https://doi.org/10.1109/CVPR.2009.5206848
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90, https://doi.org/10.1109/CVPR.2016.90
4. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 9908, pp. 630–645. Springer (2016). https://doi.org/10.1007/978-3-319-46493-0_38, https://doi.org/10.1007/978-3-319-46493-0_38
5. Kim, H., Lee, W., Lee, J.: Understanding catastrophic overfitting in single-step adversarial training (2020)
6. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
7. Rice, L., Wong, E., Kolter, J.Z.: Overfitting in adversarially robust deep learning. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 8093–8104. PMLR (2020), http://proceedings.mlr.press/v119/rice20a.html
8. Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! (2019)
9. Smith, L.N.: Cyclical learning rates for training neural networks. In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017. pp. 464–472. IEEE Computer Society (2017). https://doi.org/10.1109/WACV.2017.58, https://doi.org/10.1109/WACV.2017.58
10. Sriramanan, G., Addepalli, S., Baburaj, A., et al.: Guided adversarial attack for evaluating and enhancing adversarial defenses. Advances in Neural Information Processing Systems **33**, 20297–20308 (2020)
11. Sriramanan, G., Addepalli, S., Baburaj, A., et al.: Towards efficient and effective adversarial training. Advances in Neural Information Processing Systems **34** (2021)
12. Wong, E., Rice, L., Kolter, J.Z.: Fast is better than free: Revisiting adversarial training (2020)
13. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Wilson, R.C., Hancock, E.R., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016. BMVA Press (2016), http://www.bmva.org/bmvc/2016/papers/paper087/index.html