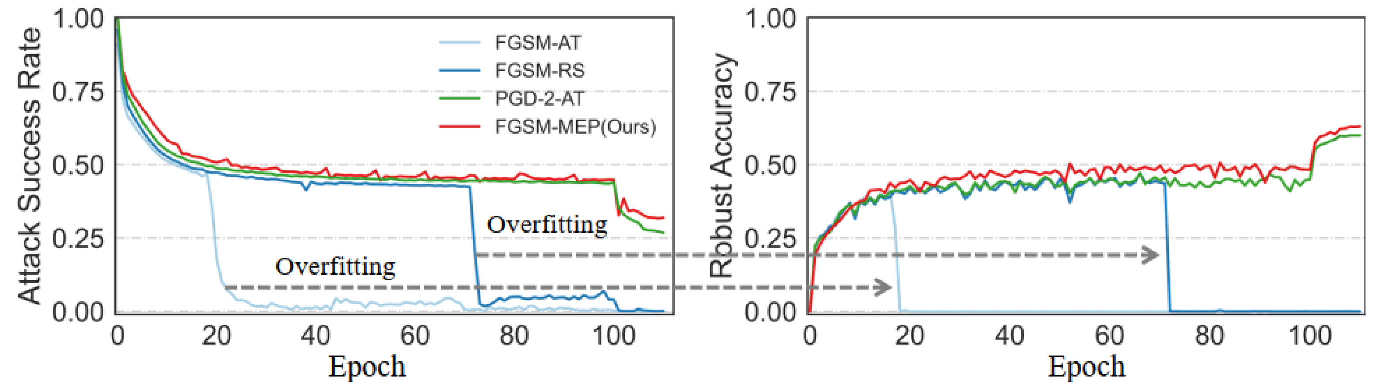
Prior-Guided Adversarial Initialization for Fast Adversarial Training

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

1. SKLOIS, Institute of Information Engineering, CAS 2. School of Cybers Security, University of Chinese Academy of Sciences 3. Tencent, AI Lab
4. Institute of Artificial Intelligence, Beihang University 5. School of Data Science, Secure Computing Lab of Big Data, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong
6. School of Computer Science and Technology, University of Chinese Academy of Sciences 7. School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen

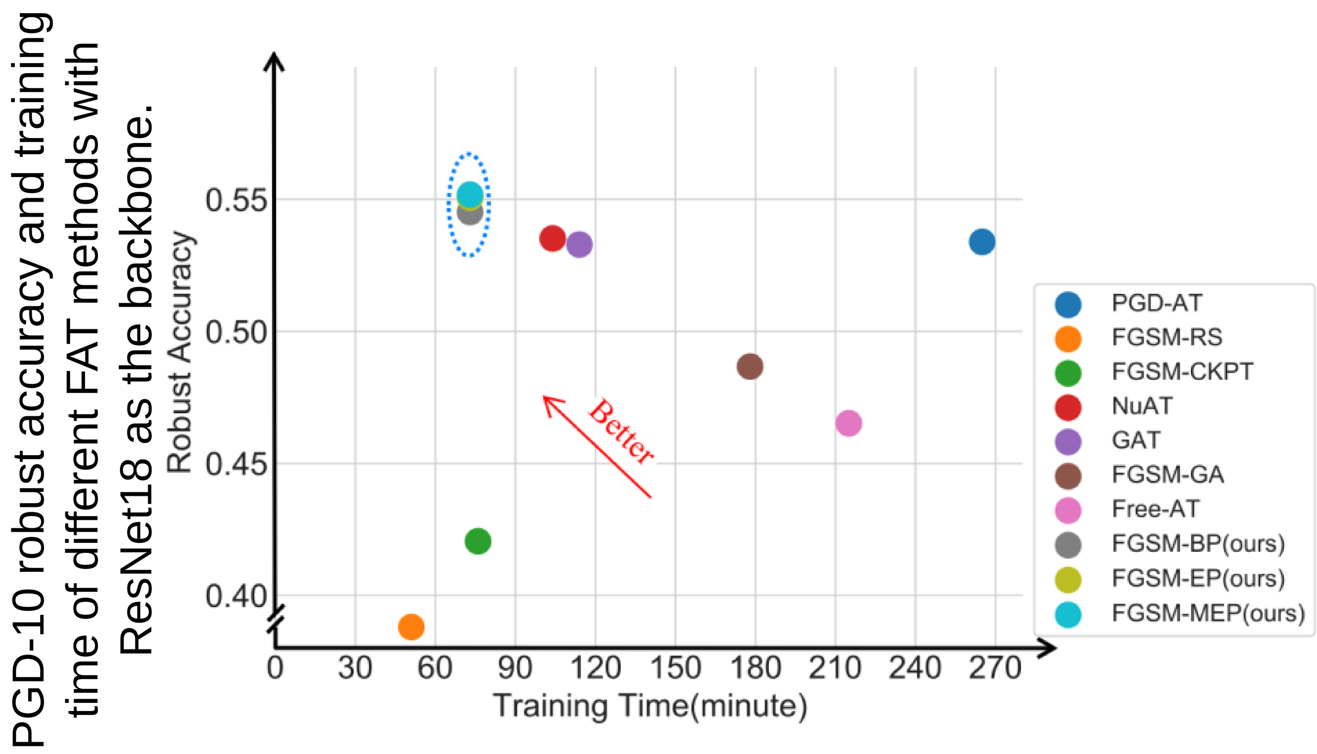
Motivation & Contribution



Motivation: We explore the difference between the training processes of standard adversarial training and fast adversarial training and observe that the attack success rate of adversarial examples (AEs) of fast adversarial training gets worse gradually in the late training stage, resulting in overfitting.

Contribution:

- We propose a prior-guided adversarial initialization to prevent overfitting after investigating several initialization strategies.
- We also propose a regularizer to guide the model learning for better robustness by considering both the currently generated perturbation and the prior-guided initialization.
- Extensive experiments on four datasets demonstrate that the proposed method can outperform state-of-the-art FAT methods in terms of both efficiency and robustness.



Comparisons of clean and robust accuracy (%) and training time (minute) on the CIFAR-10 dataset.

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
FGSM-BP	Best	**83.15**	54.59	53.55	53.2	50.24	47.47	73
	Last	**83.09**	54.52	53.5	53.33	50.12	47.17	
FGSM-EP	Best	82.75	54.8	53.62	53.27	49.86	47.94	73
	Last	81.27	55.07	54.04	53.63	50.12	46.83	
FGSM-MEP	Best	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	73
	Last	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	

Methods

- Prior From the Previous Batch (FGSM-BP): The adversarial perturbation can be defined as:

$$\boldsymbol{\delta}_{B_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\delta}_{B_t} + \alpha \cdot \mathrm{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta}_{B_t};\mathbf{w}),\mathbf{y})\right)\right],$$

- Prior From the Previous Epoch (FGSM-EP): The adversarial perturbation can be defined as:

$$\boldsymbol{\delta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\delta}_{E_t} + \alpha \cdot \mathrm{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta}_{E_t};\mathbf{w}),\mathbf{y})\right)\right],$$

- Prior From the Momentum of All Previous Epochs (FGSM-MEP): The adversarial perturbation can be defined as:

$$\mathbf{g}_c = \mathrm{sign}\left(\nabla_{\mathbf{x}}\mathcal{L}(f(\mathbf{x}+\boldsymbol{\eta}_{E_t};\mathbf{w}),\mathbf{y})\right),$$
$$\mathbf{g}_{E_{t+1}} = \mu \cdot \mathbf{g}_{E_t} + \mathbf{g}_c,$$
$$\boldsymbol{\delta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\eta}_{E_t} + \alpha \cdot \mathbf{g}_c\right],$$
$$\boldsymbol{\eta}_{E_{t+1}} = \Pi_{[-\epsilon,\epsilon]}\left[\boldsymbol{\eta}_{E_t} + \alpha \cdot \mathrm{sign}(\mathbf{g}_{E_{t+1}})\right].$$

The proposed regularization term can be added into the training loss to update the model parameters:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}}[\mathcal{L}(f(\mathbf{x}+\boldsymbol{\delta}_{adv};\mathbf{w}),\mathbf{y}) + \lambda \cdot \|f(\mathbf{x}+\boldsymbol{\delta}_{adv};\mathbf{w}) - f(\mathbf{x}+\boldsymbol{\delta}_{pgi};\mathbf{w})\|_2^2],$$

Detailed algorithms of FGSM-MEP:

Algorithm 3 FGSM-MEP

Require: The epoch N, the maximal perturbation ϵ, the maximal label perturbation ϵ_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 $\mathcal{D}^{\delta}$ and the historical model gradient $\mathcal{D}^{m}$.

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

Experiments & Results

Comparisons on CIFAR-10

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	82.32	53.76	52.83	52.6	51.08	48.68	265
	Last	82.65	53.39	52.52	52.27	51.28	48.93	
FGSM-RS [49]	Best	73.81	42.31	41.55	41.26	39.84	37.07	51
	Last	83.82	00.09	00.04	00.02	0.00	0.00	
FGSM-CKPT [25]	Best	**90.29**	41.96	39.84	39.15	41.13	37.15	76
	Last	**90.29**	41.96	39.84	39.15	41.13	37.15	
NuAT [42]	Best	81.58	53.96	52.9	52.61	**51.3**	**49.09**	104
	Last	81.38	53.52	52.65	52.48	50.63	48.70	
GAT [41]	Best	79.79	54.18	53.55	53.42	49.04	47.53	114
	Last	80.41	53.29	52.06	51.76	49.07	46.56	
FGSM-GA [2]	Best	83.96	49.23	47.57	46.89	47.46	43.45	178
	Last	84.43	48.67	46.66	46.08	46.75	42.63	
Free-AT(m=8) [39]	Best	80.38	47.1	45.85	45.62	44.42	42.17	215
	Last	80.75	45.82	44.82	44.48	43.73	41.17	
FGSM-BP (ours)	Best	83.15	54.59	53.55	53.2	50.24	47.47	73
	Last	83.09	54.52	53.5	53.33	50.12	47.17	
FGSM-EP (ours)	Best	82.75	54.8	53.62	53.27	49.86	47.94	73
	Last	81.27	55.07	54.04	53.63	50.12	46.83	
FGSM-MEP (ours)	Best	81.72	**55.18**	**54.36**	**54.17**	50.75	49.00	73
	Last	81.72	**55.18**	**54.36**	**54.17**	**50.75**	**49.00**	

Comparisons on CIFAR-100

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	57.52	29.6	28.99	28.87	28.85	25.48	284
	Last	57.5	29.54	29.00	28.90	27.6	25.48	
FGSM-RS [49]	Best	49.85	22.47	22.01	21.82	20.55	18.29	70
	Last	60.55	00.45	00.25	00.19	00.25	0.00	
FGSM-CKPT [25]	Best	**60.93**	16.58	15.47	15.19	16.4	14.17	96
	Last	**60.93**	16.69	15.61	15.24	16.6	14.34	
NuAT [41]	Best	59.71	27.54	23.02	20.18	22.07	11.32	115
	Last	59.62	27.07	22.72	20.09	21.59	11.55	
GAT [42]	Best	57.01	24.55	23.8	23.55	22.02	19.60	119
	Last	56.07	23.92	23.18	23.0	21.93	19.51	
FGSM-GA [2]	Best	54.35	22.93	22.36	22.2	21.2	18.88	187
	Last	55.1	20.04	19.13	18.84	18.96	16.45	
Free-AT(m=8) [39]	Best	52.49	24.07	23.52	23.36	21.66	19.47	229
	Last	52.63	22.86	22.32	22.16	20.68	18.57	
FGSM-BP (ours)	Best	57.58	30.78	30.01	28.99	26.40	23.63	83
	Last	83.82	30.56	29.96	28.82	26.32	23.43	
FGSM-EP (ours)	Best	57.74	31.01	30.17	29.93	27.37	24.39	83
	Last	57.74	31.01	30.17	29.93	27.37	24.39	
FGSM-MEP (ours)	Best	58.78	**31.88**	**31.26**	**31.14**	**28.06**	**25.67**	83
	Last	58.81	**31.6**	**31.03**	**30.88**	**27.72**	**25.42**	

Comparisons on Tiny ImageNet

Method		Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Time(min)
PGD-AT [37]	Best	43.6	20.2	19.9	19.86	17.5	16.00	1833
	Last	45.28	16.12	15.6	15.4	14.28	12.84	
FGSM-RS [49]	Best	44.98	17.72	17.46	17.36	15.84	14.08	339
	Last	45.18	0.00	0.00	0.00	0.00	0.00	
FGSM-CKPT [25]	Best	**49.98**	9.20	9.20	8.68	9.24	8.10	464
	Last	**49.98**	9.20	9.20	8.68	9.24	8.10	
NuAT [42]	Best	42.9	15.12	14.6	14.44	12.02	10.28	660
	Last	42.42	13.78	13.34	13.2	11.32	9.56	
GAT [41]	Best	42.16	15.02	14.5	14.44	11.78	10.26	663
	Last	41.84	14.44	13.98	13.8	11.48	9.74	
FGSM-GA [2]	Best	43.44	18.86	18.44	18.36	16.2	14.28	1054
	Last	43.44	18.86	18.44	18.36	16.2	14.28	
Free-AT(m=8) [39]	Best	38.9	11.62	11.24	11.02	11.00	9.28	1375
	Last	40.06	8.84	8.32	8.2	8.08	7.34	
FGSM-BP (ours)	Best	45.01	21.67	21.47	21.43	17.89	15.36	458
	Last	47.16	20.62	20.16	20.07	15.68	14.15	
FGSM-EP (ours)	Best	45.01	21.67	21.47	21.43	17.89	15.36	458
	Last	46.00	20.77	20.39	20.28	16.65	14.93	
FGSM-MEP (ours)	Best	43.32	**23.8**	**23.4**	**23.38**	**19.28**	**17.56**	458
	Last	45.88	**22.02**	**21.7**	**21.6**	**17.44**	**15.50**	

Comparisons with WideResNet34-10

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

Ablation study

CIFAR-10		Clean	PGD-50	C&W	AA	Time(min)
FGSM-RS	Best	73.81	41.26	39.84	37.07	51
	Last	83.82	00.02	0.00	0.00	
FGSM-BP w/o regularizer	Best	**86.51**	45.77	44.8	43.30	51
	Last	86.57	44.39	43.82	42.08	
FGSM-EP w/o regularizer	Best	85.97	45.97	44.6	43.39	51
	Last	86.3	44.97	43.8	42.84	
FGSM-MEP w/o regularizer	Best	86.33	46.71	45.5	43.99	51
	Last	**86.61**	45.69	44.8	43.26	
FGSM-RS with regularizer	Best	84.41	50.63	48.76	46.80	73
	Last	84.41	50.63	48.76	46.80	
FGSM-BP with regularizer	Best	83.15	53.2	50.24	47.47	73
	Last	83.09	53.33	50.12	47.17	
FGSM-EP with regularizer	Best	82.75	53.27	49.86	47.94	73
	Last	81.27	53.63	50.12	46.83	
FGSM-MEP with regularizer	Best	81.72	**54.17**	**50.75**	**49.00**	73
	Last	81.72	**54.17**	**50.75**	**49.00**	

Comparisons on ImageNet

ImageNet	Epsilon	Clean	PGD-10	PGD-50	Time (hour)
Free-AT(m=4)[39]	$\epsilon=2$	68.37	48.31	48.28	127.7
	$\epsilon=4$	63.42	33.22	33.08	
	$\epsilon=8$	52.09	19.46	12.92	
FGSM-RS [49]	$\epsilon=2$	67.65	48.78	48.67	44.5
	$\epsilon=4$	63.65	35.01	32.66	
	$\epsilon=8$	53.89	0.00	0.00	
FGSM-BP (ours)	$\epsilon=2$	**68.41**	**49.11**	**49.10**	63.7
	$\epsilon=4$	**64.32**	**36.24**	**34.93**	
	$\epsilon=8$	**53.96**	**21.76**	**14.33**	

Conclusion

- **Prior-guided adversarial initialization :** we . propose to adopt historically generated adversarial perturbations to initialize adversarial examples.
- **A simple yet effective regularizer:** we also propose a simple yet effective regularizer to further improve model robustness, which prevents the current perturbation deviating too much from the prior-guided initialization. .
- **Superiority:** extensive experimental evaluations are performed on three benchmark databases to demonstrate the superiority of the proposed method.
- The code is released at *https://github.com/jiaxiaojunQAQ/FGSM-PGI*.