Less is More: **Data Pruning** for Faster Adversarial Training

Yize Li[1], Pu Zhao[1], Xue Lin[1], Bhavya Kailkhura[2], Ryan Goldhahn [2]

[1]Northeastern University, [2]Lawrence Livermore National Laboratory

Northeastern University

SafeAI

01 Introduction

Deep Neural Networks (DNNs) are vulnerable to adversarial attacks

Attack

FGSM(Fast Gradient Sign Method, one-step):

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y; \theta))$$

PGD(Project Gradient Descent, multi-step):

$$x'_{t+1} = \text{Clip}_{x,\epsilon}(x'_t + \alpha \cdot \text{sign}(\nabla_x L(x'_t, y; \theta))$$

Adversarial Training (Defense)

FGSM-based: **Fast** but fail on stronger attacks.

PGD-based: Effective but **slow**.

01 Introduction

Efficient *Adversarial Training (Defense)*

At the **generation** level (Make **FGSM-based** Great Again):

Free AT

Fast AT:random start, gradient alignment regularization, suitable step size

01 Introduction

Efficient Training

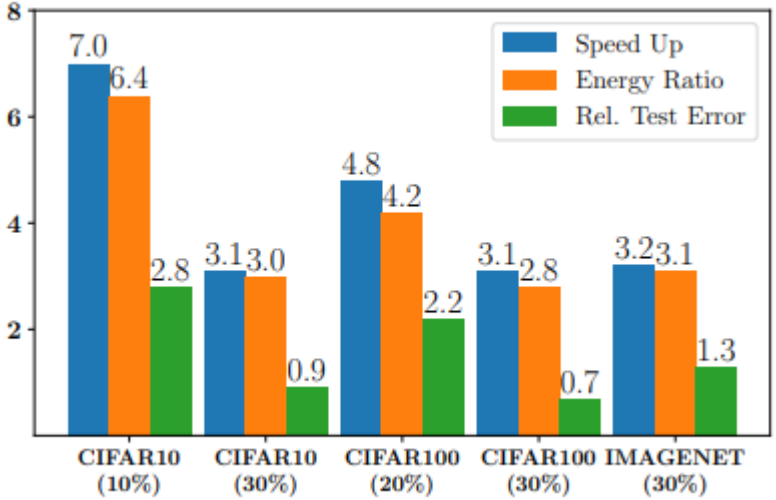
At the **data** level:

Find **subsets** of the data

Approximate certain desirable **characteristics** of the full data

Efficient AT by **less data**

Efficiency on the clean data



Data selection Flowchart

02 Data Pruning Based Adversarial Training

- **Adversarial Training**

$$\min_{\theta} \frac{1}{|D|} \sum_{(x,y)\in\mathcal{D}} \left[\max_{\delta\in\triangle} \mathcal{L}(\theta; x+\delta, y) \right]$$

$$x^{t+1} = \mathrm{Proj}_{\triangle}\left(x^t + \alpha\,\mathrm{sign}\left(\nabla_{x^t}\mathcal{L}\left(\theta; x^t, y\right)\right)\right)$$

- **Adversarial Loss**

TRADES

$$\min_{f} \mathbb{E}\Big\{ \underbrace{\phi(f(\boldsymbol{X})Y)}_{\text{for accuracy}} + \underbrace{\max_{\boldsymbol{X}'\in\mathbb{B}(\boldsymbol{X},\epsilon)} \phi(f(\boldsymbol{X})f(\boldsymbol{X}')/\lambda)}_{\text{regularization for robustness}} \Big\}$$

MART

$$\mathcal{L}^{\mathrm{MART}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i, y_i, \boldsymbol{\theta}),$$

$$\ell(\mathbf{x}_i, y_i, \boldsymbol{\theta}) := \mathrm{BCE}\big(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i\big) + \lambda \cdot \mathrm{KL}\big(\mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}) \| \mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta})\big) \cdot \big(1 - \mathbf{p}_{y_i}(\mathbf{x}_i, \boldsymbol{\theta})\big)$$

02 Data Pruning Based Adversarial Training

- **Adversarial Data Pruning**

$$\min_{\theta} \frac{1}{k} \sum_{(x,y)\in\mathcal{S}} \left[\max_{\delta\in\triangle} \mathcal{L}(\theta; x+\delta, y) \right]$$

$$\min_{\mathcal{S}\subseteq\mathcal{D},|\mathcal{S}|=k} G(\mathcal{S})$$

- **Adv-GLISTER**

$$G(\mathcal{S}) = \sum_{(x_V,y_V)\in\mathcal{V}} L_V(\theta_S; x_V + \delta_V^*, y_V)$$

- **Adv-GRAD-MATCH**

$$G(\mathcal{S}) = \| \sum_{(x_S,y_S)\in\mathcal{S}} w\nabla_\theta \mathcal{L}_\mathcal{S}(\theta; x_S + \delta_S^*, y_S) - \nabla_\theta \sum_{(x_D,y_D)\in\mathcal{D}} \mathcal{L}_\mathcal{D}(\theta; x_D + \delta_D^*, y_D) \|$$

03 Results

- **Set up**
  - **Adv Loss:** TRADES, MART
  - **Dataset:** CIFAR-10, CIFAR-100
  - **Model:** ResNet-18
  - **Data Pruning:** subset size [30%, 50%], selection interval 20, 100/200 epochs
  - **Evaluation:** PGD(eps=4,8,16/255, 50 steps), AutoAttack
  - **Baselines:** TRADES, MART, Bullet-Train

03 Results

Table 1: TRADES results where data pruning methods use only 30% data points on CIFAR-10 and 50% data points on CIFAR-100 for 100 epochs of training.

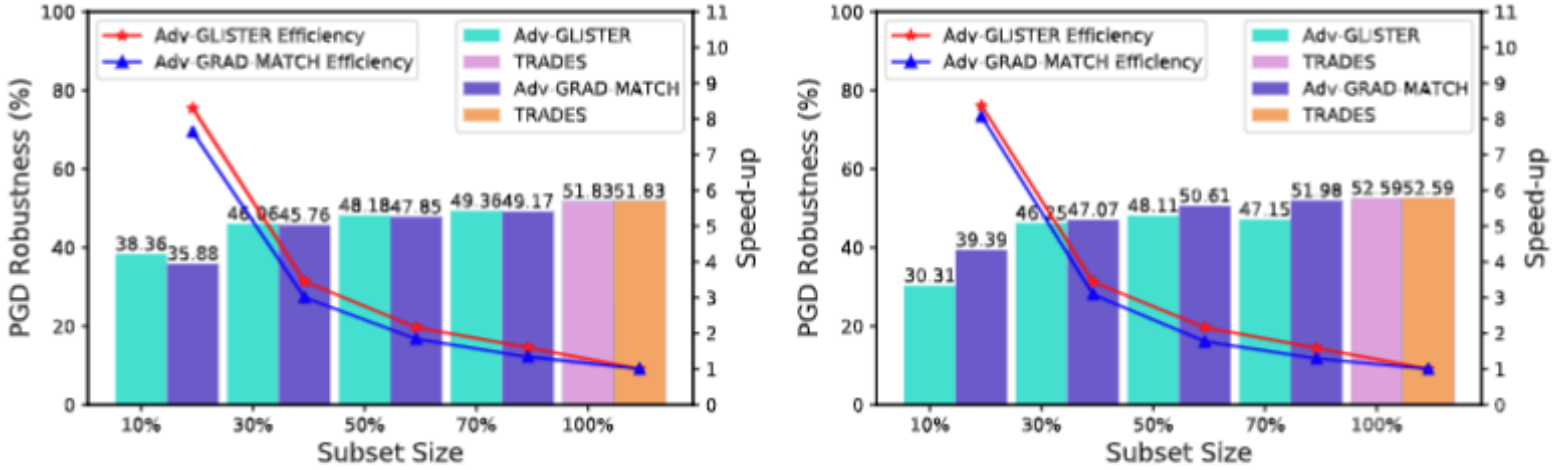
Dataset	Method	Clean	PGD 4/255	PGD 8/255	PGD 16/255	AutoAttack	Time/epoch (Speed-up)
CIFAR-10	TRADES [49]	82.73	69.17	51.83	19.43	49.06	416.20 (-)
	Bullet [21]	84.60	70.24	50.82	16.05	47.93	193.06 (2.16×)
	Adv-GLISTER (Ours)	77.62	63.06	46.06	16.52	41.61	128.00 (3.25×)
	Adv-GRAD-MATCH (Ours)	75.67	61.85	45.96	17.49	42.19	148.76 (2.80×)
	Adv-GLISTER&Bullet (Ours)	79.21	63.02	44.52	13.33	40.77	**68.51 (6.08×)**
	Adv-GRAD-MATCH&Bullet (Ours)	77.57	62.00	45.13	14.65	41.94	77.51 (5.37×)
CIFAR-100	TRADES [49]	55.85	40.31	27.35	10.71	23.39	387.72 (-)
	Bullet [21]	59.43	42.23	28.08	9.40	23.85	173.59 (2.23×)
	Adv-GLISTER (Ours)	51.26	37.16	24.78	9.49	20.57	202.7 (1.91×)
	Adv-GRAD-MATCH (Ours)	51.03	37.17	24.60	9.70	20.42	206.05 (1.88×)
	Adv-GLISTER&Bullet (Ours)	53.54	37.24	23.91	7.69	20.02	**105.66 (3.67×)**
	Adv-GRAD-MATCH&Bullet (Ours)	52.98	36.92	24.24	8.01	20.17	105.61 (3.67×)

03 Results

Table 2: MART results where data pruning methods use only 30% data points on CIFAR-10 and 50% data points on CIFAR-100 for 100 epochs of training.

Dataset	Method	Clean	PGD 4/255	PGD 8/255	PGD 16/255	AutoAttack	Time/epoch (Speed-up)
CIFAR-10	MART [20]	80.96	68.21	52.59	19.52	46.94	329.54 (-)
	Bullet [21]	85.29	70.92	50.64	13.33	43.77	199.42 (1.65×)
	Adv-GLISTER (Ours)	71.97	60.13	46.25	16.59	39.86	96.72 (3.41×)
	Adv-GRAD-MATCH (Ours)	73.67	61.35	47.07	18.16	40.98	114.24 (2.88×)
	Adv-GLISTER&Bullet (Ours)	73.87	58.89	41.01	10.20	35.99	**67.97 (4.85×)**
	Adv-GRAD-MATCH&Bullet (Ours)	78.78	64.42	46.72	13.50	39.53	76.50 (4.31×)
CIFAR-100	MART [20]	54.85	39.24	25.08	8.59	22.66	307.43 (-)
	Bullet [21]	57.44	39.22	24.14	6.66	21.55	187.73 (1.64×)
	Adv-GLISTER (Ours)	46.36	34.37	24.01	9.20	19.79	152.11 (2.02×)
	Adv-GRAD-MATCH (Ours)	48.07	36.19	26.11	10.79	21.24	153.86 (2.00×)
	Adv-GLISTER&Bullet (Ours)	52.13	35.07	20.67	5.64	18.21	**100.22 (3.07×)**
	Adv-GRAD-MATCH&Bullet (Ours)	52.46	35.81	22.20	6.48	18.68	113.03 (2.72×)

03 Results



(a) TRADES. (b) MART.

Figure 2: PGD evaluation ($\epsilon = 8/255$) with the corresponding speed-up under different subset sizes for 100 epoch CIFAR-10 training. Note that when the size is 100%, data pruning methods are not applied and the speed-up is compared with the baselines (TRADES or MART).

03 Results

- Epoch

Table 3: 100 *v.s.* 200 epoch TRADES CIFAR-10 results with ResNet-18 when using 30% data points with robustness regularization factor to be 1.

Method	Epoch	Clean	PGD			AutoAttack
			4/255	8/255	16/255	
Adv-GLISTER	100	77.62	63.06	46.06	16.52	41.61
Adv-GRAD-MATCH	100	75.61	60.81	45.76	17.49	42.19
Adv-GLISTER	200	78.76	64.15	46.11	16.92	42.43
Adv-GRAD-MATCH	200	75.75	61.24	46.49	18.55	43.63

- Number of selection rounds

Table 4: TRADES results on CIFAR-10 with ResNet-18 using 30% data samples under different selection counts for 200 epoch training.

Method	Number of selections	Clean	PGD			AutoAttack	Speed-up
			4/255	8/255	16/255		
TRADES	-	83.32	68.91	49.64	17.31	47.53	-
Adv-GLISTER	4	75.80	60.48	44.62	16.07	40.44	3.15×
Adv-GRAD-MATCH	4	73.80	60.43	46.06	18.33	43.03	2.83×
Adv-GLISTER	9	78.76	64.15	46.11	16.92	42.43	2.93×
Adv-GRAD-MATCH	9	75.75	61.24	46.49	18.55	43.63	2.75×

04 Conclusion

- Explore efficient AT from the lens of data pruning, where the acceleration is achieved by only focusing on the representative subset of the data.
- Propose two data pruning algorithms, Adv-GRAD-MATCH and Adv-GLISTER, and perform a comprehensive experimental study.
- Combining our efficient AT framework with the existing Bullet-Train approach achieves state-of-the-art performance in training cost.