ICLR

NC STATE UNIVERSITY

Fine-Grained Iterative Adversarial Attacks with Limited Computation Budget

Zhichao Hou, Weizhi Gao, Xiaorui Liu

Attacks Are Essential Yet Expensive

Attack perspective:

- Iterative attack such as PGD (Madry et al., 2017) are widely adopted as strong oracles to benchmark the robustness.
- PGD with 20, 50, or even 100 steps is typically used, which incurs roughly **40–200×** the cost of a natural inference.

Defense perspective:

- The most effective defense: adversarial training (Madry et al., 2017; Zhang et al., 2019; Wang et al., 2019) critically depends on strong iterative attacks to generate challenging adversaries during training.
- The standard PGD-10 procedure already costs around **10×** more than natural training.

Critical Challenge: *Given a **prescribed computation budget**, how can we **maximize the strength** of iterative adversarial attacks ?*

Expensive Computation in Attacks

Algorithm 1 Gradient-based Adversarial Attack

Require: clean input $\boldsymbol{x}$, label y, adversarial loss $\mathcal{L}$, attack budget ϵ, step size α, iteration T.

Ensure: adversarial example $\tilde{\boldsymbol{x}}$

1: $\boldsymbol{x}_1 \leftarrow \boldsymbol{x}$ ▷ initialization

2: **for** $t = 1, \ldots, T$ **do**

3: $\quad \boldsymbol{x}_t \rightarrow \boldsymbol{o}_t^{(1)} \cdots \boldsymbol{o}_t^{(l)} \cdots \rightarrow \boldsymbol{o}_t^{(L)} \rightarrow \mathcal{L}$

4: $\quad \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_t} \leftarrow \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t^{(1)}} \cdots \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t^{(l)}} \cdots \leftarrow \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t^{(L)}} \leftarrow \mathcal{L}$

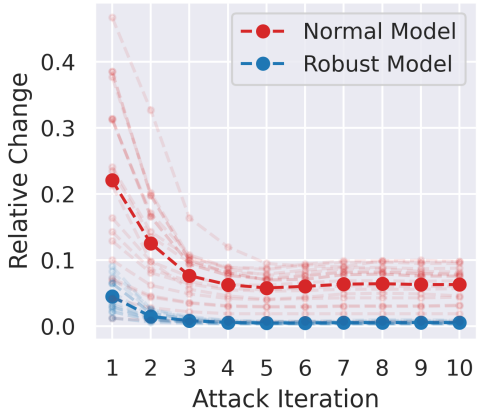
5: $\quad \boldsymbol{x}_{t+1} \leftarrow \Pi_{\mathcal{B}_\epsilon(\boldsymbol{x})}\left(\boldsymbol{x}_t + \alpha \cdot \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_t}\right)$

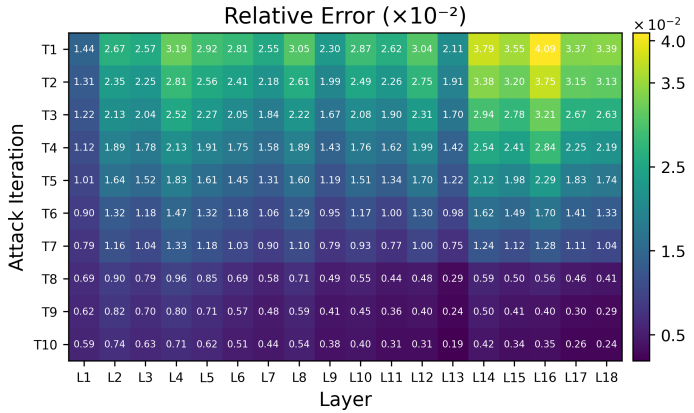
6: **end for**

7: **return** $\tilde{\boldsymbol{x}} \leftarrow \boldsymbol{x}_{T+1}$

Gradient-based iterative attacks are notoriously expensive because it requires **multiple iterations**, and each step **requires both a forward and backward pass**.

Computation Redundancy in Attacks

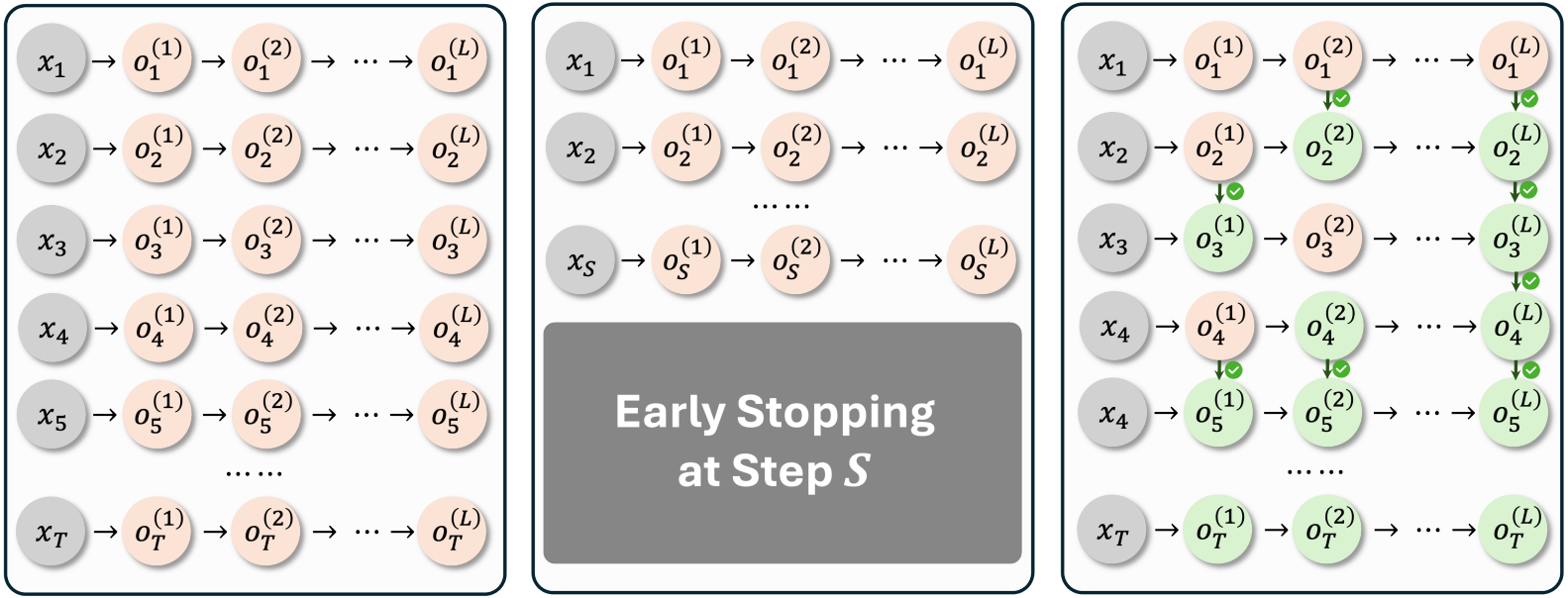




Activation relative change $||\boldsymbol{a}_t - \boldsymbol{a}_{t-1}||/||\boldsymbol{a}_t||$ for ResNet-18 on CIFAR-10.

- Figure (Left) shows that the activations become highly similar after a small number of iterations. This indicates substantial redundancy in repeated computations across attack iterations. Additionally, the "Robust Model" exhibits markedly smaller activation changes than the "Normal Model";
- Figure (Right) shows that while all layers follow the same overall decreasing trend, the decay rate of relative activation change varies across layer.

Spiking Iterative Attack



(a) Vanilla Iterative Attack (b) Coarse-Grained Attack (c) Fine-Grained Attack

Attack as Combinatorial Optimization

Let $\Delta = (\delta_{t,l}) \in \{0,1\}^{T \times L}$ be a binary mask where $\delta_{t,l} = 1$ indicates that layer l is fully computed at attack iteration t, while $\delta_{t,l} = 0$ indicates reuse of previously computed activations. The corresponding *fine-grained* optimization is the following:

$$\max_{\Delta \in \{0,1\}^{T \times L}} \quad \mathcal{L}\big(\boldsymbol{x}_{T+1}(\Delta), y\big) \quad \text{s.t.} \quad \sum_{t=1}^{T} \sum_{l=1}^{L} C_{t,l}\, \delta_{t,l} \leq C_{\text{total}}, \tag{2}$$

where $\boldsymbol{x}_{T+1}(\Delta)$ denotes the final perturbed input after T attack steps executed under mask Δ.

Challenges:

(1) Vast search space: $\mathcal{O}(2^{T \times L})$;

(2) Expensive objective evaluations;

(3) Broken gradient flow;

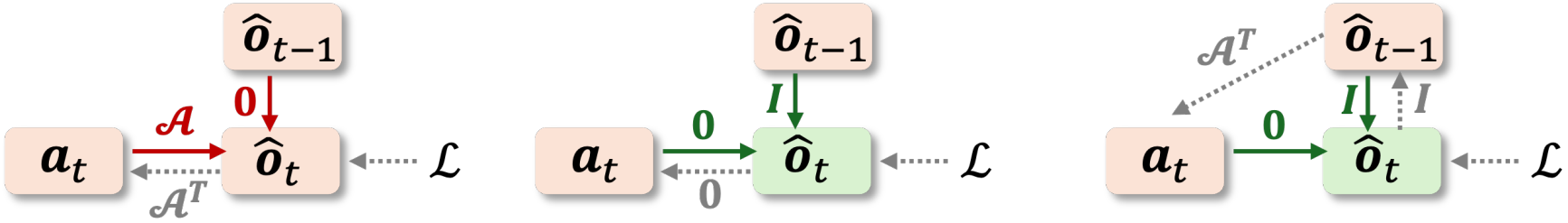
Spiking Forward Computation

$$
\begin{cases}
\boldsymbol{o}_1^{(l)} = \mathcal{A}^{(l)}(\boldsymbol{a}_1^{(l)}) \\
\boldsymbol{o}_2^{(l)} = \mathcal{A}^{(l)}(\boldsymbol{a}_2^{(l)}) = \mathcal{A}^{(l)}(\boldsymbol{a}_2^{(l)} - \boldsymbol{a}_1^{(l)}) + \boldsymbol{o}_1^{(l)} \\
\quad \cdots \\
\boldsymbol{o}_t^{(l)} = \mathcal{A}^{(l)}(\boldsymbol{a}_t^{(l)}) = \mathcal{A}^{(l)}(\boldsymbol{a}_t^{(l)} - \boldsymbol{a}_{t-1}^{(l)}) + \boldsymbol{o}_{t-1}^{(l)} \\
\quad \cdots \\
\boldsymbol{o}_T^{(l)} = \mathcal{A}^{(l)}(\boldsymbol{a}_T^{(l)}) = \mathcal{A}^{(l)}(\boldsymbol{a}_T^{(l)} - \boldsymbol{a}_{T-1}^{(l)}) + \boldsymbol{o}_{T-1}^{(l)}
\end{cases}
\Rightarrow
\begin{cases}
\boldsymbol{o}_1^{(l)} = \hat{\boldsymbol{o}}_1^{(l)} = \mathcal{A}^{(l)}(\boldsymbol{a}_1^{(l)}) \\
\boldsymbol{o}_2^{(l)} \approx \hat{\boldsymbol{o}}_2^{(l)} = \mathcal{A}^{(l)}(\mathcal{S}_\rho(\boldsymbol{a}_2^{(l)}, \boldsymbol{a}_1^{(l)})) + \hat{\boldsymbol{o}}_1^{(l)} \\
\quad \cdots \\
\boldsymbol{o}_t^{(l)} \approx \hat{\boldsymbol{o}}_t^{(l)} = \mathcal{A}^{(l)}(\mathcal{S}_\rho(\boldsymbol{a}_t^{(l)}, \boldsymbol{a}_{t-1}^{(l)})) + \hat{\boldsymbol{o}}_{t-1}^{(l)} \\
\quad \cdots \\
\boldsymbol{o}_T^{(l)} \approx \hat{\boldsymbol{o}}_T^{(l)} = \mathcal{A}^{(l)}(\mathcal{S}_\rho(\boldsymbol{a}_T^{(l)}, \boldsymbol{a}_{T-1}^{(l)})) + \hat{\boldsymbol{o}}_{T-1}^{(l)}
\end{cases}
$$

$$
S_\rho(\boldsymbol{a}_t, \boldsymbol{a}_{t-1}) = \begin{cases}
\mathbf{1} \cdot (\boldsymbol{a}_t - \boldsymbol{a}_{t-1}), & \|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| / \|\boldsymbol{a}_t\| \geq \rho, \\
\mathbf{0} \cdot (\boldsymbol{a}_t - \boldsymbol{a}_{t-1}), & \|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| / \|\boldsymbol{a}_t\| < \rho,
\end{cases}
$$

Virtual Backward Computation

register_hook

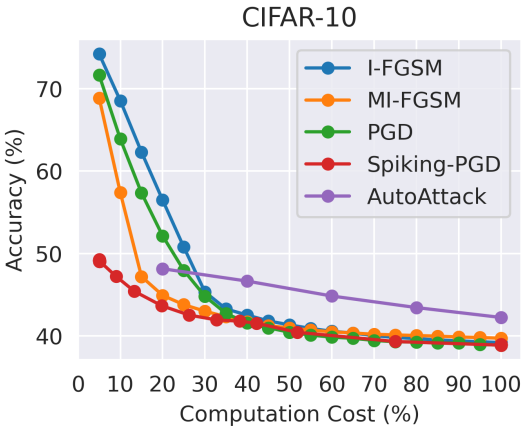


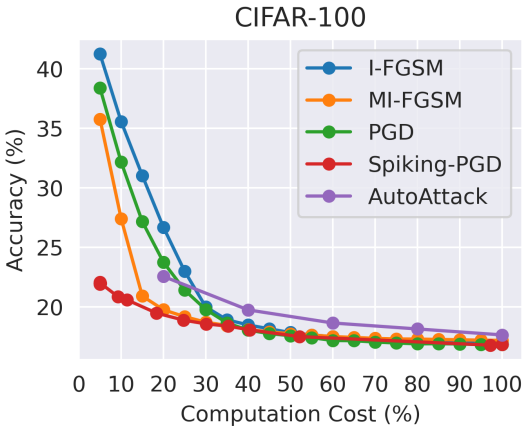
(a) Exact Gradient (b) Vanishing Gradient (c) Virtual Gradient

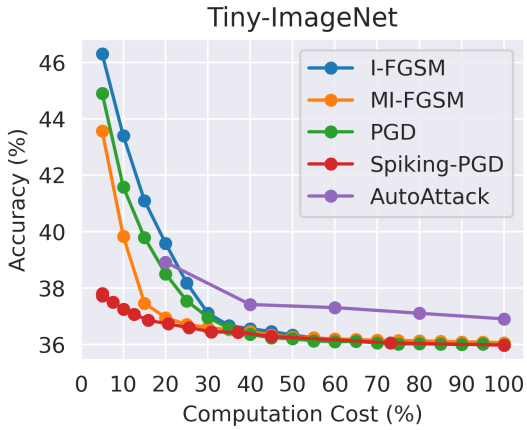
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}_t^{(l)}} = \mathcal{S}'_\rho \cdot \mathcal{A}^{(l)^\top}\left(\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{o}}_t^{(l)}}\right) = \begin{cases} \mathcal{A}^{(l)^\top}\left(\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{o}}_t^{(l)}}\right), & \|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|/\|\boldsymbol{a}_t\| \geq \rho, \Rightarrow \text{ Figure 4 } (a) \\ \mathbf{0}, & \|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\|/\|\boldsymbol{a}_t\| < \rho, \Rightarrow \text{ Figure 4 } (b) \end{cases}$$

Adversarial Attack Strength

Pixel Perturbation on Images

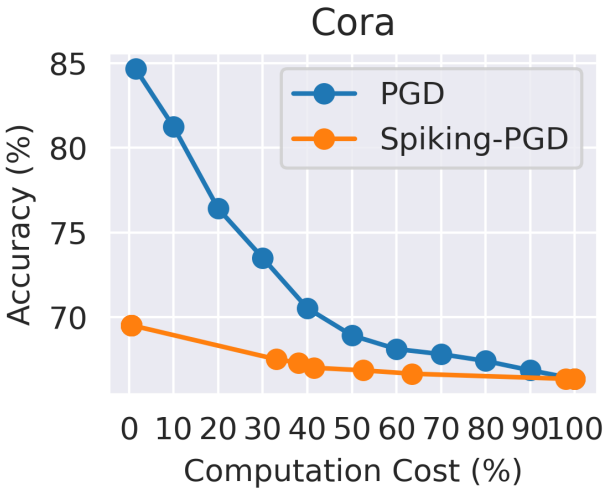


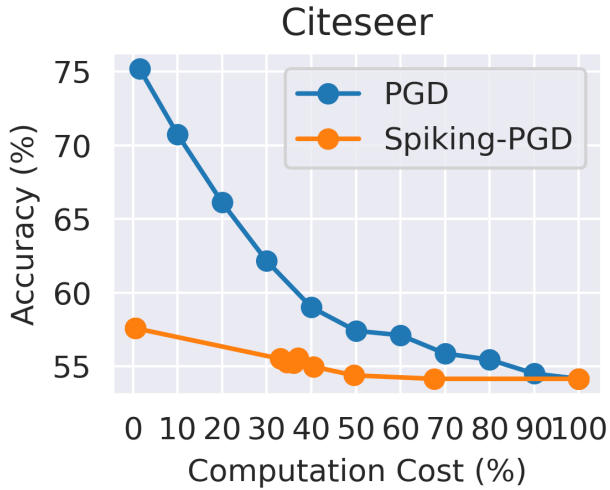




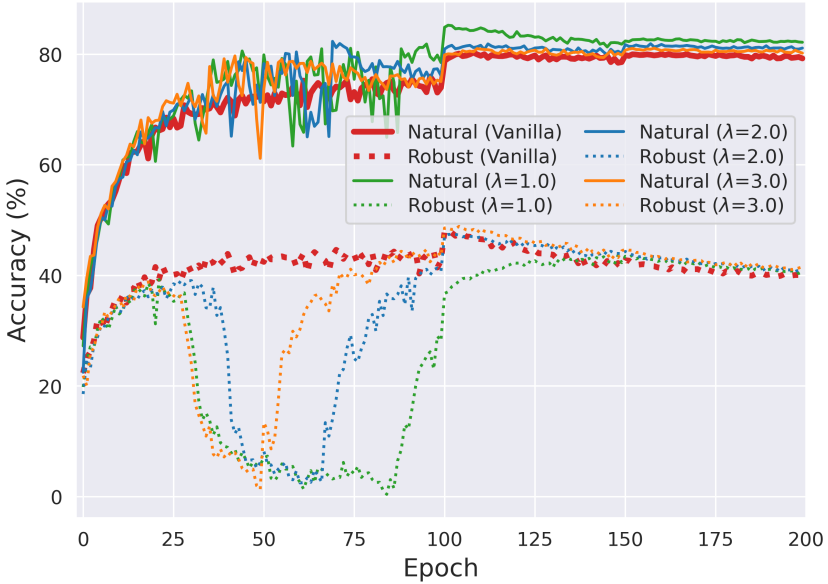
Adversarial Attack Strength

Structure Attack on Graphs

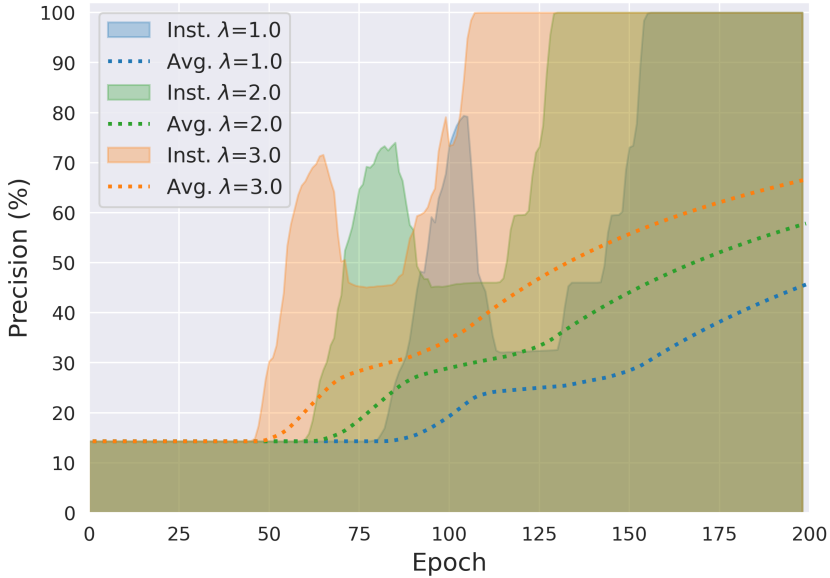




Efficient Adversarial Training

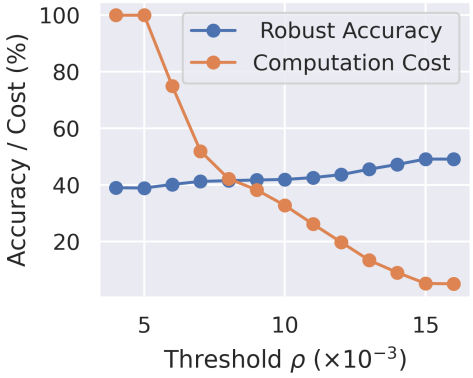


Accuracy curves of adversarial training on CIFAR10.

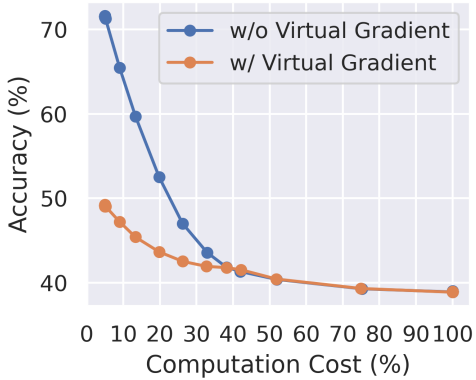


Precision curves of adversarial training on CIFAR-10.

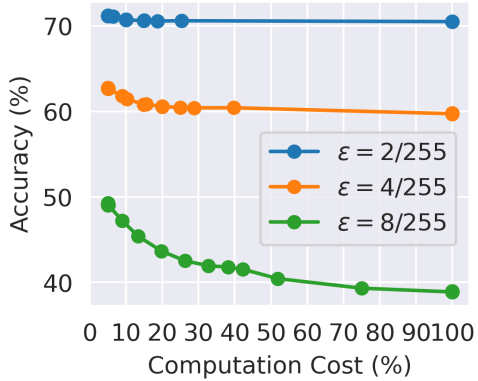
Ablation Study



(a) Ablation on threshold ρ



(b) Ablation on virtual gradient.



(c) Ablation on attack radius.

ICLR

NC STATE UNIVERSITY

Thanks for Listening

Zhichao Hou, Weizhi Gao, Xiaorui Liu