Does Low Rank Adaptation Lead to Lower Robustness against Training-Time Attacks?

XXXXXXX, XXXXX*,
The Hong Kong Polytechnic University, Hong Kong, China
xxxxxxxx@connect.polyu.hk
xxxxxxx@polyu.edu.hk

Overview

- First theoretical analysis of LoRA's security vulnerabilities during fine-tuning
- Examines robustness against data poisoning and backdoor attacks
- Uses Neural Tangent Kernel and Information Geometry for analysis

Summary of Findings

- **LoRA is more vulnerable than full fine-tuning to untargeted poisoning attacks but demonstrates greater robustness against backdoor attacks**.
- In addition to the trade-off between performance and computational cost, **LoRA's rank also influences the trade-off between untargeted poisoning and backdoor attacks**.
- Besides of the rank, the **initialization variance** of the A matrix in LoRA significantly impacts training-time robustness.
- To improve robustness against backdoor attacks, the rank should be set **as low as possible**, provided that performance requirements are met.
- A **small** scale of initialization variance is recommended to enhance training-time robustness.

Preliminary

• Low Rank Adaptation (LoRA) introduces a mechanism to reduce the number of trainable parameters by freezing the original matrix $W^{(l)}$ and learn a low-rank update $\Delta W^{(l)}$. This update is factorized as the product of two low-rank submatrices,

$$\begin{aligned} W^{(l)\prime} &= W^{(l)} + \Delta W^{(l)} \\ \Delta W^{(l)} &= B^{(l)} A^{(l)}, \end{aligned} \tag{1}$$

where $A^{(l)} \in \mathbb{R}^{r \times n_l}$ and $B^{(l)} \in \mathbb{R}^{n_{l+1} \times r}$ are learnable matrices, and $r \ll \min\{n_l, n_{l+1}\}$.

• Neural Tangent Kernel (NTK) is a special kernel function, which is defined as the inner product of gradients:

$$K_{ntk}(x, x') = \nabla_\theta F(x;\theta)^T \nabla_\theta F(x';\theta). \tag{2}$$

As the width of the neural network approaches infinity, the NTK exhibits the following two key properties:

- The NTK converges to a **deterministic** limiting kernel that depends only on three factors: *i)* the variance of the parameter initialization, *ii)* the neural network structure, and *iii)* the selection of activation functions;
- NTK keeps **constant** through out each training step t.

• Introducing of Poisoning Attacks:

- Untargeted Poisoning Attacks (UPA): adversely reducing the performance of fine-tuned models;
- Backdoor Poisoning Attacks (BPA): to cause the model to misclassify **only** when a particular trigger is present

Modeling Objectives of Different Poisoning Attacks

• Untargeted Poisoning aims to maximize:

$$|\nabla_\theta \mathcal{L}(x_c, \theta)^T \cdot \nabla_\theta \mathcal{L}(x_u, \theta)|, \tag{3}$$

i.e., as adversaries, we aim to align the optimization direction of the poisoned sample x_u as closely as possible with that of the clean training objective, because we aim to maximally influence the model's predictions while injecting only a small fraction of poisoned data.

• Backdoor Poisoning aims to minimize:

$$|\nabla_\theta \mathcal{L}(x_c, \theta)^T \cdot \nabla_\theta \mathcal{L}(x_t, \theta)|, \tag{4}$$

i.e., the adversary aims to ensure that the optimization process driven by $\nabla_\theta \mathcal{L}(x_c, \theta)$ and $\nabla_\theta \mathcal{L}(x_t, \theta)$ occur simultaneously and both significantly influence the training, which aligns with the target of BPA to maintain performance on most inputs while producing significantly altered predictions only when a specific trigger is present.

Theorem (Relationship between NTK and Fisher)

When the width of F_Θ approaches infinity, its Fisher information $\mathcal{I}_\Theta$ under $\tilde{\mathcal{D}}$ is equal to its weighted $\mathcal{M}'(\tilde{\mathcal{D}}, \tilde{\mathcal{D}})$, i.e.,

$$\begin{aligned} I_\theta &= \mathbb{E}_{x \sim \tilde{\mathcal{D}}} \left[\nabla_\theta \mathcal{L}(x, \theta)^T \nabla_\theta \mathcal{L}(x, \theta) \right] \\ &= \mathbb{E}_{\tilde{x}_c \in \tilde{\mathcal{D}}} \left[\nabla_{F_\theta} \mathcal{L}(x, \theta)^T K_{ntk}(x, x) \nabla_{F_\theta} \mathcal{L}(x, \theta) \right]. \end{aligned} \tag{5}$$

Metrics to Measure the LoRA's Resilience

Let $\lambda_1, \lambda_2, ..., \lambda_{n_L}$ denote the n_L eigenvalues of the Fisher information matrix $\mathcal{I}_\Theta$. Then we can quantify the *information bits* (IB) of the model as

$$\mathbf{IB} = \frac{1}{2} \log \det_{\text{pseudo}} \mathcal{I}_\Theta = \frac{1}{2} \sum_{\lambda_i > 0}^{n_L} \lambda_i. \tag{6}$$

We can measure the curvature of the fine-tuning manifold with *Rényi entropy*

$$H_\alpha = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n_L} \lambda_i^\alpha \right), \tag{7}$$

where $\alpha \geq 0$ controls the norm formation on the $\mathcal{I}_\Theta$. Specifically, H_1 corresponds to the Shannon entropy, while $H_\infty = \max\{\lambda_1, \lambda_2, ..., \lambda_{n_L}\}$.
Intuitively, a higher $\mathbf{IB}$ and H_α indicates a more complex fine-tuning manifold of the model, which implicitly demonstrates a higher function fitting ability.

An L-layer ANN's NTK

The NTK of FF can be represented by

$$\begin{aligned} K_{\text{ff}}^{(1,k)}(x, x') &= I_{n_l} \otimes \Sigma^{(1)}(x, x') = x^T \cdot x', \\ K_{\text{ff}}^{(l,k)}(x, x') &= K_{\text{ff}}^{(l-1,k)}(x, x') \dot{\Sigma}^{(l)}(x, x') + \Sigma^{(l)}(x, x'), \end{aligned} \tag{8}$$

where $k = \{0, 1, ..., n_l - 1\}$, $\otimes$ denotes the Kronecker product, and

$$\begin{aligned} &\Sigma^1(X, X') = X^T X', \\ &\Sigma^l(X, X') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{l-1})}[\sigma(f(X))^T \sigma(f(X'))] = \sum_{j=1}^{n_{l-1}} \sigma(y_j^{(l-1)}(X))^T \sigma(y_j^{(l-1)}(X')). \end{aligned} \tag{9}$$

Lemma (NTK of LoRA)

The neural tangent kernel of an l-layer ANN trained with LoRA can be expressed as follows.

$$\begin{aligned} K_{\textit{LoRA}}^{(1,k)}(x, x') &= K_{\textit{ff}}^{(1,k)} \\ K_{\textit{LoRA}}^{(l,k)}(x, x') &= K_{\textit{LoRA}}^{(l-1,k)}(x, x') \dot{\Sigma}^{(l)} + \Sigma_{\textit{LoRA}}^{(l)}(x, x'), \end{aligned} \tag{10}$$

where

$$\Sigma_{\textit{LoRA}}^{(l)}(x, x') = \sigma(y^{(l-1)}(x))^T A^{(l)\ T} A^{(l)} \sigma(y^{(l-1)}(x')),$$

and $W_{\textit{LoRA}}^{(l)} = W_0^{(l)} + B^{(l)} A^{(l)}$ denotes the l-th layer's weight matrix of LoRA.

Theorem (NTK Relationship between FF and LoRA)

For an l-layer ANN with infinite width, the NTK functions of FF and LoRA at the l-th layer are related by the following expression:

$$K_{\textit{LoRA}}^{(l,k)} = K_{\textit{ff}}^{(l,k)} + \Delta_r^{(l)}, \tag{11}$$

where

$$\Delta_r^{(l)} = [\sigma(y^{(l-1)}(x))]^T (A^{(l)\ T} A^{(l)} - I_{n_{l-1} \times n_{l-1}}) [\sigma(y^{(l-1)}(x'))].$$

Theorem ($M_\Delta^{(l)}$'s Negative Semi-Definiteness)

*When the LoRA submatrix $A^{(l)} \in \mathbb{R}^{r \times n_{l-1}}$ is initialized with variance σ_a^2, $\sigma_a^2 < 1/n_{l-1}$, and $r \leq n_{l-1}$, then $M_\Delta^{(l)}$ is a **negative semi-definite** matrix, with r eigenvalues equal to $\sigma_a^2 \cdot n_{l-1}$ and $n_l - r$ eigenvalues equal to 0.*

Corollary (Ideal Full Rank Adaptation)

When $n_{l-1} \to \infty$, the kernel matrix $M_\Delta^{(l)}$ strictly converges to $\mathbf{0}$, i.e., $K_{\textit{LoRA}}^{(l)}(x, x) \equiv K_{\textit{ff}}^{(l)}(x, x)$ if $r = n_{l-1}$ and the initialization variance satisfies $\sigma_a^2 = 1/n_{l-1}$.

Theorem ($\mathbf{IB}_{\textbf{ff}} \geq \mathbf{IB}_{\textbf{LoRA}}$ & $H_{\alpha\textbf{ff}} \geq H_{\alpha\textbf{LoRA}}$)

*The information bits and the Rényi entropy of LoRA are always **smaller** than those of FF if $M_\Delta^{(l)}$ is a negative semi-definite matrix, i.e., $r \leq n_{l-1}$ and $\sigma^2 \leq 1/n_{l-1}$.*