FL-WBC: Enhancing Robustness against Model Poisoning Attacks in Federated Learning from a Client Perspective

Yuri Dimitre Dias de Faria

y218172@dac.unicamp.br

18 de Junho 2024

Table of Contents

Table of Contents

Introdução
Funcionamento do Aprendizado Federado

- Definição
  - Abordagem de aprendizado de máquina distribuída.
  - Permite a colaboração de múltiplos dispositivos de borda no treinamento de um modelo global.
- Funcionamento
  - Treinamento Local: Cada dispositivo treina um modelo localmente usando seus próprios dados.
  - Envio de Atualizações: Dispositivos enviam apenas atualizações de modelo (gradientes ou parâmetros) para um servidor central.
  - Agregação Centralizada: O servidor central agrega as atualizações recebidas para criar um modelo global atualizado.
  - Distribuição do Modelo: O modelo global atualizado é enviado de volta aos dispositivos.
  - Ciclo Iterativo: O processo de treinamento local, envio de atualizações e agregação é repetido várias vezes.

Introdução
Problemas de segurança em FL

- Alta vulnerabilidade a ataques de envenenamento (*poisoning attacks*) em sistema FL.

Introdução
Objetivo

- Melhorar a robustez contra esse tipo de ataques por mecanismo de defesa *client-based*.

Table of Contents

Background
Model Poisoning Attacks

Tipos de ataques em modelos em FL.

- Não-direcionado (untargeted): Ataca o modelo global indiscriminadamente para que ele tenha uma alta taxa de erro.
- Direcionado (targeted): Ataca o modelo global para gerar classificações incorretas desejadas pelo invasor.

Background
Defesas da literatura

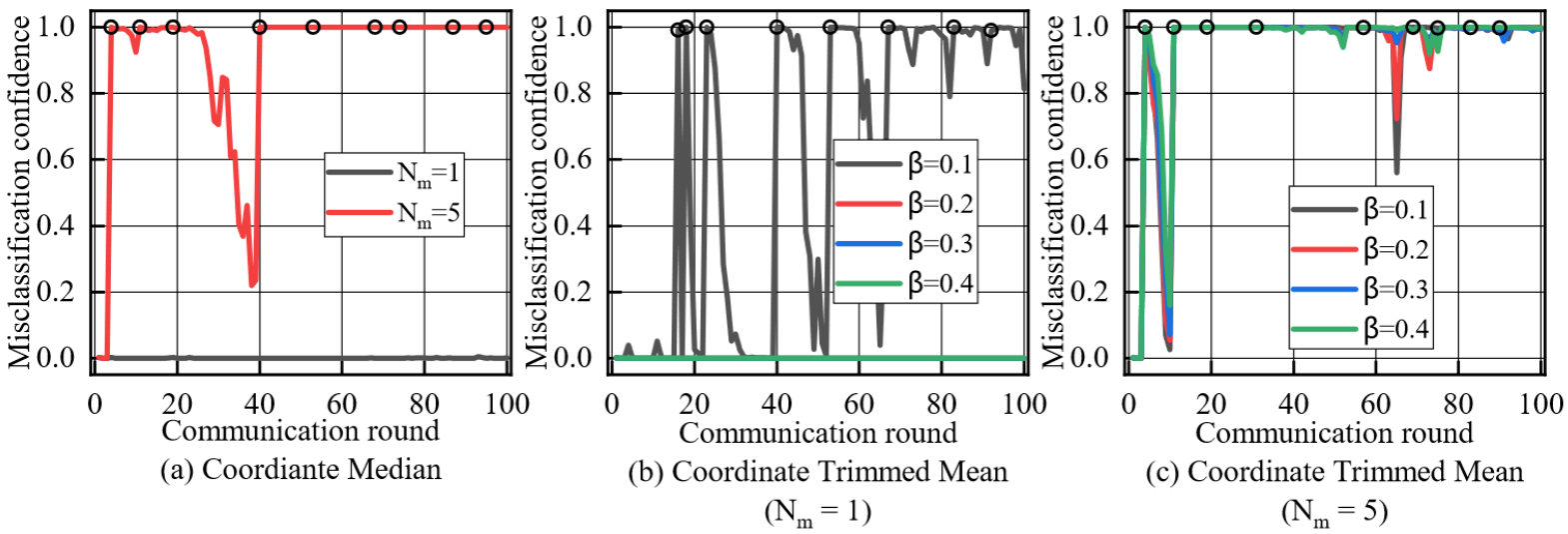

Figure: Desempenho de *robust aggregation* para ataques em grande escala

Table of Contents

Contribuições

Attack Effect on Parameter (AEP)

AEP δ_t é uma métrica do acúmulo das mudanças do modelo global até a rodada t devido a ataques conduzidos.

$$W_t(\mathbb{S}_i/\mathbb{M}) \leftarrow \frac{N}{K} \sum_{k \in \mathbb{S}_t} p^k W_{t,I}^k(\alpha = 1)$$

$$\delta_t \triangleq W_t(\mathbb{S}/M) - W_t$$

Onde $\mathbb{M}$ é um conjunto de atacantes, $W_t(\mathbb{S}_i/\mathbb{M})$ representa os pesos do modelo global na rodada t quando todos os dispositivos maliciosos em $S_i(i \leq t)$ não performaram um ataque em i rodadas, N representa o número de dispositivos e K é o número de dispositivos selecionados para agregação na rodada.

Contribuições

Attack Effect on Parameter (AEP)

- Por que pode ser difícil eliminar um envenenamento do modelo global?
- Razão: A transmissão do AEP δ_t para o modelo global é determinada por $H_{t,i}^k$ (matriz hessiana), inacessível pelo servidor central.

Contribuições
Mecanismo de defesa FL-WBC

- Proposto um mecanismo de pertubação em matrizes hessianas em dispositivos benignos para combater um envenenamento,

Table of Contents

FL-WBC
Objetivos

- Ideia central: Perturbar o espaço de parâmetros no treinamento local para mitigar efeitos de envenenamento.
- Objetivo 1: Para manter a desempenho da tarefa benigna, o *loss* da mesma deve ser minimizado.
- Objetivo 2: Para prevenir que AEP se esconda no *kernel* das matrizes hessianas em dispositivos benignos, o *kernel* $H^k_{l+1,i}$ deve ser perturbado.

Para perturbar $H^k_{l,i}$, foi considerado a segunda derivada parcial da função de perda, onde a diagonal descreve a mudança do gradiente pelas iterações.

FL-WBC

Formulação matematica

$$W^{\hat{k}}_{t,i+1} = W^k_{t,i} - \eta_{t,i} \bigtriangledown F^k(W^k_{t,i}, \xi^k_{t,i})$$

$$W^k_{t,i+1} = W^{\hat{k}}_{t,i} + \eta_{t,i} \Upsilon^k_{t,i} \odot M^k_{t,i}$$

Onde $\Upsilon^k_{t,i_{r,c}}$ é uma matriz ruído de mesma dimensão de W, $M^k_{t,i}$ é uma máscara binaria, na qual os elementos são determinados por:

$$M^k_{t,i_{r,c}} = \begin{cases} 1, & |(W^{\hat{k}}_{t,i+1} - W^k_{t,i}) - \Delta W^k_{t,i}|_{r,c}/\eta_{t,i}| \leq |\Upsilon^k_{t,i_{r,c}}| \\ 0, & |(W^{\hat{k}}_{t,i+1} - W^k_{t,i}) - \Delta W^k_{t,i}|_{r,c}/\eta_{t,i}| > |\Upsilon^k_{t,i_{r,c}}| \end{cases}$$

Nesse trabalho, foi escolhido $\Upsilon^k_{t,i_{r,c}}$ como sendo um ruído laplaciano com média zero e desvio padrão s.

Algorithm 1 Local training process applying FL-WBC on a benign device in round t.

Input: Local training data $\mathbb{D} \in \mathbb{R}^{L \times P \times Q}$; Local objective function $F : \mathbb{R}^{P \times Q} \to \mathbb{R}$; Local model parameters $\boldsymbol{W} \in \mathbb{R}^{M \times N}$; The number of local training iterations I; Learning rates $\eta_{t,i}$ for $i \in [I]$; Standard deviation of Laplace noise s.

Output: Learnt model parameter $\boldsymbol{W}$ with our defense.

Initialize $\boldsymbol{W}_{-1}, \boldsymbol{W}_{-2}$;
$i \leftarrow 0$;
for batch $\mathcal{B}$ in $\mathbb{D}$ **do**
    Randomly generate a Laplace noise matrix $\Upsilon \in \mathbb{R}^{M \times N}$ with $mean = 0$ and $std = s$;
    $\boldsymbol{W}_{-1} \leftarrow \boldsymbol{W}$;
    $\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_{t,i} \nabla F(\boldsymbol{W}, \mathcal{B})$;
    if this is not the first training batch **then**
        $\boldsymbol{W}^* \leftarrow (\boldsymbol{W} - \boldsymbol{W}_{-1}) - (\boldsymbol{W}_{-1} - \boldsymbol{W}_{-2})$;
        Find the set $\mathbb{S}$ which contains the indices of elements in $|\boldsymbol{W}^*| - \eta_{t,i}|\Upsilon|$ which are less than or equal to 0;
        for $j, k \in \mathbb{S}$ **do**
            $\boldsymbol{W}_{j,k} \leftarrow \boldsymbol{W}_{j,k} + \eta_{t,i} \Upsilon_{j,k}$;
        end for
    end if
    $\boldsymbol{W}_{-2} \leftarrow \boldsymbol{W}_{-1}$;
    $i \leftarrow i + 1$
end for

Figure: Algoritmo FL-WBC

Table of Contents

Procedimentos Experimentais

Configurações e datasets

- Datasets: FashionMNIST e CIFAR10.
- Configurações IID e non-IID.
- Dataset D_m malicioso com samples de uma e múltiplas imagens.
- Comparação com métodos de agregação robusta (CMA e CTMA) e *differential privacy* (DP).
- DP: *Central Differential Privacy* (CDP) e *Local Differential Privacy* (LPD).
- Rodadas contam com 10 dispositivos, onde 5 deles são maliciosos caso a rodada for maligna.
- Probabilidade de 10% de ser uma rodada adversária.

Procedimentos Experimentais

Resultados Robust Aggregation com 1 imagem

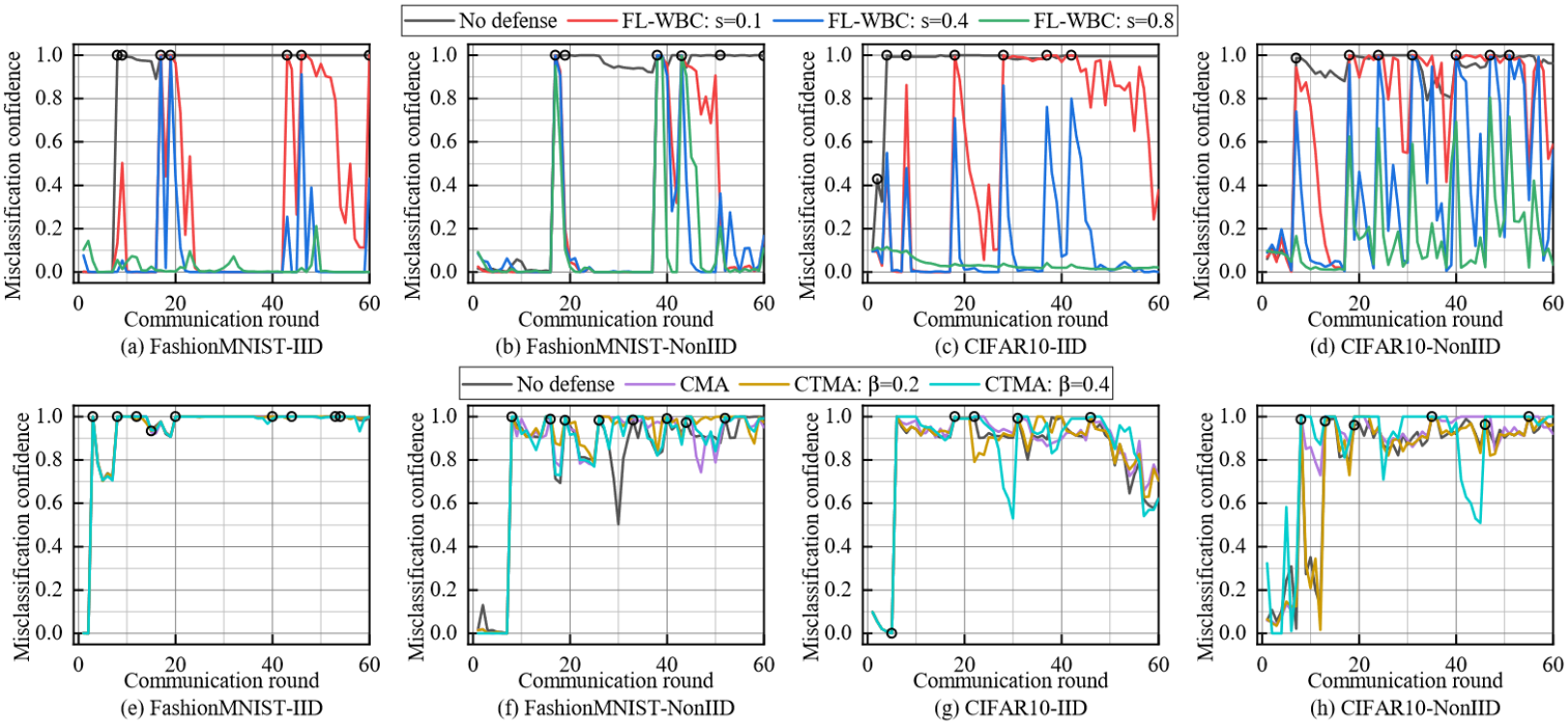

Figure: Comparação da confidência na classificação incorreta

Procedimentos Experimentais

Resultados DP 1 imagem

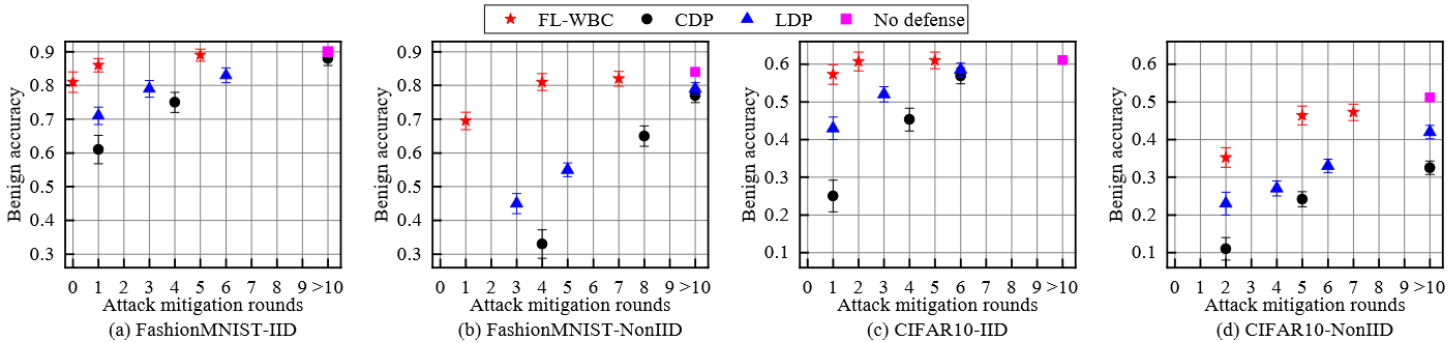

Figure: Comparação da acurácia benigna e rodadas de mitigação de ataques

Procedimentos Experimentais

Resultados Robust Aggregation multiplas imagem

Defense	Fashion-MNIST (IID)	Fashion-MNIST (non-IID)	CIFAR10 (IID)	CIFAR10 (non-IID)
CTMA ($\beta = 0.1$)	7	9	8	>10
CTMA ($\beta = 0.2$)	7	8	8	9
CTMA ($\beta = 0.4$)	6	8	7	9
CMA	5	7	6	8

Figure: Comparação de rodadas de mitigação de ataques

Procedimentos Experimentais

Resultados DP multiplas imagem

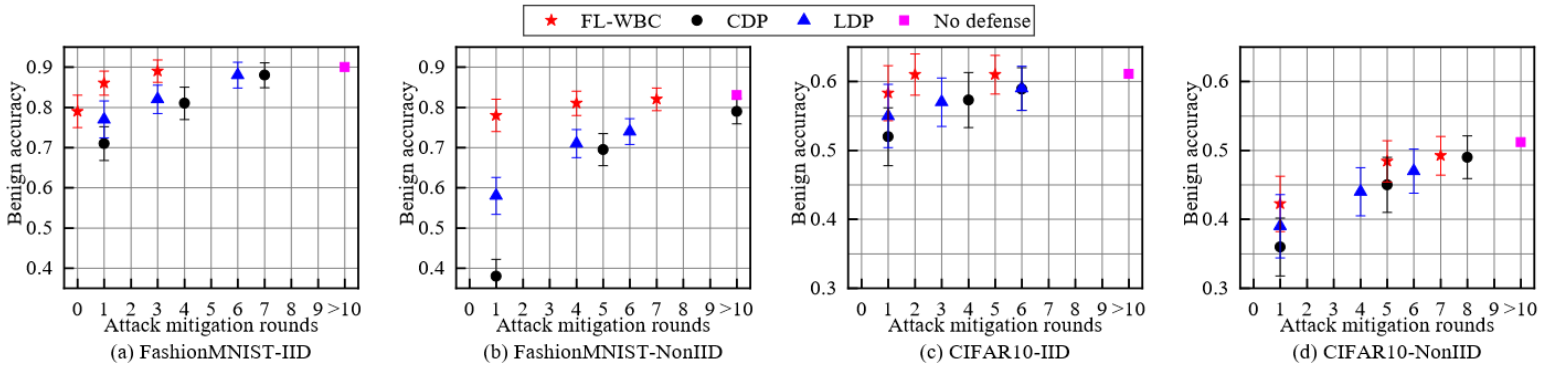

Figure: Comparação da acurácia benigna e rodadas de mitigação de ataques

Procedimentos Experimentais

Integração do FL-WBC com Robust Aggregation

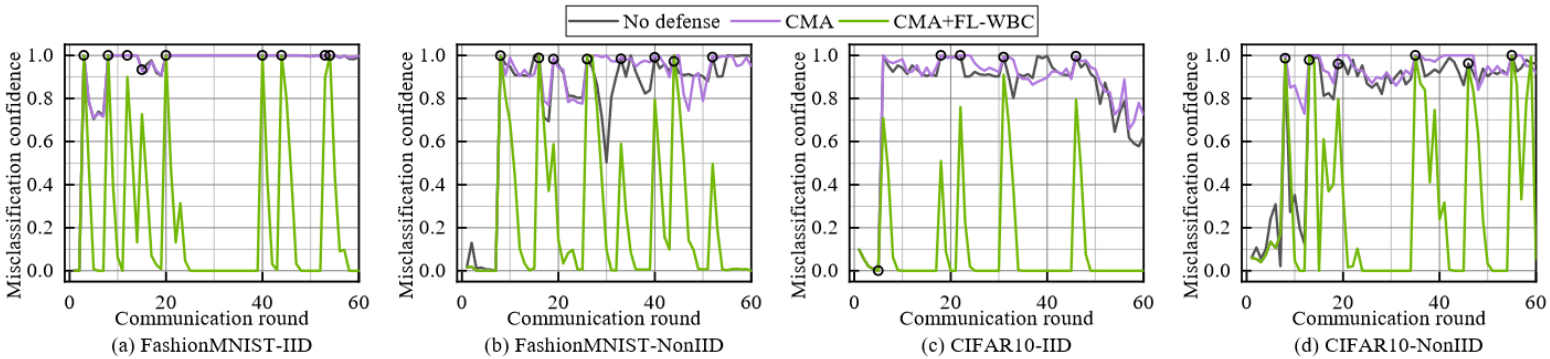

Figure: Comparação da confidência da classificação incorreta e rodadas de comunicação

Table of Contents

Conclusões

- Defesa contra ataque de envenenamento de modelo baseada no cliente .
- Resultados demonstram que o sistema supera as linhas de base de mitigação do efeito do ataque.
- A defesa consegue proteger contra o ataque em menos rodadas de comunicação que outros tipos de defesas com uma menor degradação da utilidade do modelo.
- Pode ser estendido para outros tipos de ataques de envenenamento, como backdoor.