Avaliação da Sensibilidade de Modelos a Ataques de Inversão de Gradiente no Aprendizado Federado

João Pedro Camargo Batista[1], **Vinicius F. S. Mota**[1], **Rodolfo S. Villaça**[1]

[1]Departamento de Informática (DI)
Universidade Federal do Espírito Santo (Ufes)

joao.c.batista@edu.ufes.br
{vinicius.f.mota, rodolfo.villaca}@ufes.br

Abstract. *Federated Learning (FL) enables distributed model training while keeping local data private. However, FL remains vulnerable to Gradient Inversion Attacks (GIA), in which malicious agents reconstruct clients' private data from shared gradients. This work evaluates the sensitivity of local models to GIA across individual layers and investigates the effectiveness of Differential Privacy (DP) as a mitigation strategy. Experiments conducted on the LFW facial dataset show that convolutional layers are more susceptible to GIA, especially in early training stages, and that relatively low noise levels are sufficient to degrade attack reconstruction quality.*

Resumo. *O Aprendizado Federado (FL) permite o treinamento distribuído de modelos. O FL permanece vulnerável aos Ataques de Inversão de Gradiente (GIA), nos quais agentes maliciosos reconstroem dados dos clientes a partir dos gradientes compartilhados. Este trabalho avalia a sensibilidade de modelos locais aos GIA e investiga a eficácia da Privacidade Diferencial (DP) como estratégia de mitigação. Experimentos preliminares com o dataset LFW demonstram que camadas convolucionais são mais suscetíveis aos GIA, especialmente nas épocas iniciais do treinamento, e que níveis relativamente baixos de ruído são suficientes para degradar a qualidade das reconstruções.*

1. Introdução

Com o crescimento exponencial no volume de dados gerados por dispositivos conectados, as abordagens tradicionais de aprendizado de máquina centralizado enfrentam gargalos severos de infraestrutura e crescentes preocupações com a privacidade dos usuários. Nesse cenário, o Aprendizado Federado (do inglês, *Federated Learning*, FL), proposto por [McMahan et al. 2017], surge como uma alternativa promissora, permitindo o treinamento de modelos de forma colaborativa e distribuída. A principal inovação do FL está no fato de que os dados proprietários permanecem nos dispositivos dos clientes, sendo que estes compartilham apenas as atualizações de parâmetros com um servidor de agregação central, o que mitiga a necessidade de comunicação direta de informações sensíveis.

Contudo, as garantias de privacidade do FL não tornam o sistema livre de vulnerabilidades, visto que o processo federado pode sofrer ataques de agentes internos que põem em risco a integridade do treinamento e a segurança dos dados proprietários [Mammen 2021]. Esses agentes atacantes podem atuar como clientes maliciosos ou como um servidor de agregação honesto-mas-curioso. Dentre essas ameaças, destacam-se os

Ataques de Inversão de Gradiente (do inglês, *Gradient Inversion Attacks*, GIA), que utilizam métodos de otimização e engenharia reversa sobre os gradientes compartilhados pelos clientes para reconstruir os seus dados e rótulos de treinamento originais.

Diante desse cenário, é necessário aplicar uma camada extra de proteção aos dados e modelos locais. Atualmente, a Privacidade Diferencial (do inglês, *Differential Privacy*, DP) é o principal conceito de defesa existente. A sua ideia fundamental é garantir que a inclusão ou a exclusão de uma única amostra de dado não mude de forma significativa os resultados estatísticos do modelo. Na prática, métodos como o DP-SGD [Abadi et al. 2016] e o DP-FedAvg [McMahan et al. 2019] aplicam esse conceito fazendo o corte (*clipping*) dos gradientes e adicionando ruídos controlados antes de enviá-los ao servidor. Essas modificações buscam evitar que o servidor descubra os dados originais, embora impliquem na redução da utilidade do modelo global final. Neste contexto, este trabalho tem como objetivo preliminar avaliar a sensibilidade dos modelos locais dos clientes aos GIA, mesmo em cenários em que há utilização de métodos de defesa, como a DP.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 descreve os experimentos e resultados preliminares; e a Seção 4 apresenta as conclusões e trabalhos futuros.

2. Fundamentação Teórica

2.1. Ataques de Inversão de Gradiente

Os ataques podem ser classificados de acordo com a sua natureza e seu objetivo [Mammen 2021]. Uma dessas classificações engloba os chamados ataques de inversão de gradiente (do inglês, *gradient inversion attacks*, GIA), que são ataques que buscam reconstruir os dados de treinamento dos clientes a partir da observação e engenharia reversa dos gradientes enviados por eles. A base teórica dos GIA está fundamentada sobre o funcionamento da descida estocástica do gradiente (método utilizado para o ajuste de parâmetros em modelos), em especial o passo do *backpropagation*, em que há o cálculo efetivo das derivadas parciais da função de perda em relação aos pesos do modelo para a obtenção das direções de maior crescimento dessa função (vetor gradiente, g_t). Um agente malicioso que obtém acesso ao g_t pode tentar invertê-lo para obter o par (x,y), respectivamente, dado e rótulo, que o originou.

Um tipo de GIA foi proposto por [Zhu et al. 2019] e é chamado *Deep Leakage from Gradients* (DLG), um ataque que fundamenta-se na geração de dados e rótulos sintéticos (x^*, $y*$) que são ajustados iterativamente até que produzam um gradiente g'_t próximo do gradiente real g_t observado. Dessa forma, o agente malicioso utiliza métodos de otimização para minimizar a distância entre os gradientes falsos (g'_t, gerados por x^* e y^*) e os reais (g_t, enviados pelo cliente), conforme a seguinte função objetivo:

$$x^*, y^* = \arg\min_{x^*, y^*} ||g_t - g'_t||^2 \tag{1}$$

Esse método mostrou-se funcional na reconstrução fiel de imagens e textos a partir dos gradientes [Zhu et al. 2019]. Entretanto, o DLG exige a otimização simultânea de duas variáveis (x^* e y^*), o que pode ser desafiador e resultar em eventuais falhas na convergência do ataque. Nesse sentido, [Zhao et al. 2020] propõe um método denominado

Improved Deep Leakage from Gradients (iDLG), que aprimora o DLG ao eliminar a necessidade de otimização do rótulo sintético $y*$. Assim, o iDLG extrai o rótulo verdadeiro ao se beneficiar do fato de que, em modelos que utilizam *Cross-Entropy* como função de perda, o gradiente da perda em relação aos *logits* de saída só apresenta valores negativos para a classe correta [Zhao et al. 2020].

Esse ataque demonstra desempenho experimental superior ao DLG padrão [Zhao et al. 2020]. Ainda assim, o iDLG possui dificuldades com gradientes gerados por *batches* de dados de tamanho grande, uma vez que estes tornam-se uma média de todos os dados do *batch*, o que dificulta a reconstrução fiel. Nesse cenário, alguns trabalhos propõem melhorias ao método, como o trabalho de [Leite et al. 2024], que implementa o *blend* de recuperações intermediárias para facilitar reconstruções futuras, e o trabalho de [Wang et al. 2023], que busca superar essa dificuldade por meio da utilização da distância de Wasserstein para o cálculo da perda durante a fase de otimização do ataque.

2.2. Privacidade Diferencial

Dado o cenário em que é possível efetuar ataques no FL, é necessário aplicar uma camada extra de proteção aos dados e modelos locais dos clientes. Nesse sentido, a DP é o principal conceito envolvido, tendo sido proposto por [Dwork 2006]. Esse conceito provê uma garantia matemática sobre o comportamento de algoritmos de Aprendizado Profundo. A sua ideia principal é garantir que a inclusão ou exclusão de uma amostra em um conjunto de dados não altere de forma significativa os processos estatísticos dos modelos. Formalmente, um mecanismo $\mathcal{M}$ fornece (ϵ-δ)-DP se, para todos os conjuntos de dados adjacentes $\mathcal{D}$ e $\mathcal{D}'$ (que diferem em apenas um registro excluído ou incluído) e todos os subconjuntos de resultados $S \subseteq Range(\mathcal{M})$:

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(D') \in S] + \delta \tag{2}$$

onde ϵ representa o orçamento de privacidade (quanto menor o valor, mais forte é a garantia de privacidade) e δ é a probabilidade de a garantia de privacidade ser violada (idealmente um valor pequeno, menor do que $1/|\mathcal{D}|$). Em outras palavras, a DP diz que se um atacante olhar para dois conjuntos de dados diferentes (e que diferem minimamente), este não conseguirá distingui-los facilmente.

Para aplicar esse conceito na prática e evitar que os modelos memorizem dados sensíveis, [Abadi et al. 2016] propuseram o DP-SGD. Esse método funciona em duas etapas durante o treinamento local: primeiro, faz o corte (*clipping*) dos gradientes para limitá-los a uma norma máxima C, restringindo a influência de cada amostra de dado; em seguida, adiciona ruído gaussiano (controlado por um nível σ) a esses gradientes antes de atualizar os pesos do modelo via *backpropagation*. Assim, os gradientes enviados pelos clientes ao servidor estão modificados de tal forma que não sejam reconhecidos como os originais por ele, embora esse método traga certa degradação para a utilidade do modelo global resultante [McMahan et al. 2019].

2.3. Sensibilidade de Modelos

Como supracitado, o método DP-SGD pode degradar a utilidade dos modelos. Assim, é interessante avaliar se um modelo está sensível ou não aos GIA, para evitar a adição

de ruído desnecessário. Nesse sentido, o estudo de [Novak et al. 2018] introduz uma métrica de sensibilidade que busca avaliar o quanto os gradientes de um modelo variam com alterações nos dados de entrada. Essa métrica é descrita pela seguinte equação:

$$\mathbb{E}_{\mathbf{x}_{\text{test}}}[\|\mathbf{J}(\mathbf{x}_{\text{test}})\|_F] = \mathbb{E}_{\mathbf{x}_{\text{test}}}\left[\sqrt{\sum_{i,j} J_{ij}(\mathbf{x}_{\text{test}})^2}\right] \tag{3}$$

Em que $J_{ij}(\mathbf{x})$ é a matriz jacobiana do modelo $\mathbf{f}$ em relação ao dado de entrada $\mathbf{x}$ e a sensibilidade média esperada $\mathbb{E}_{\mathbf{x}_{\text{test}}}$ para o modelo é resultante da aplicação da norma de Frobenius sobre essa matriz. Em outras palavras, a interpretação dessa métrica vem do entendimento de que a norma de Frobenius determina a magnitude da sensibilidade dos pesos do modelo perto do ponto no espaço que representa o dado de entrada.

Embora essa métrica tenha sido proposta inicialmente para averiguar a relação entre a sensibilidade do modelo aos dados de entrada e a capacidade de generalização no treinamento, o estudo de [Mo et al. 2021] a utilizou para verificar a vulnerabilidade das camadas individuais dos modelos aos GIA, sendo essa a abordagem adotada nos experimentos deste trabalho. Os dados empíricos demonstraram que quanto maior for a sensibilidade dos gradientes em relação aos dados de entrada, mais suscetível aos GIA está o modelo [Mo et al. 2021].

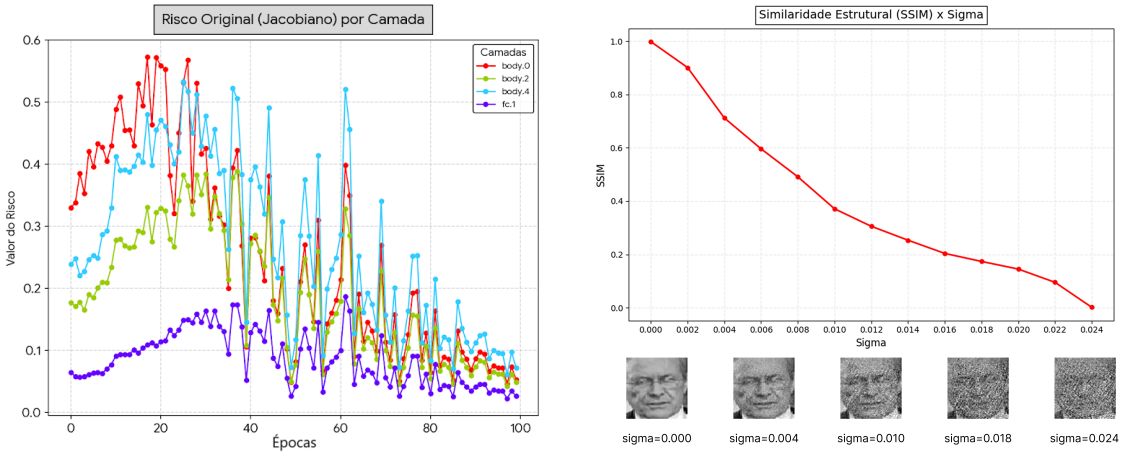
3. Experimentos e Resultados Preliminares

Para avaliar a sensibilidade dos modelos locais dos clientes aos GIA e a eficácia do DP-SGD na mitigação desse tipo de ataque, foram propostos dois experimentos. No primeiro, um modelo convolucional simples, com três camadas de convolução e uma camada linear, foi treinado com o *dataset* Labelled Faces in the Wild[1] (LFW) durante 100 épocas. Esse *dataset* contém imagens de rostos de pessoas reais, simulando um problema de reconhecimento facial relevante ao contexto de privacidade abordado. Além disso, em cada rodada foi computado o valor de sensibilidade de cada camada do modelo, conforme Equação 3, com o objetivo de verificar a sua suscetibilidade aos GIA o longo do treinamento. Ademais, no segundo experimento, o mesmo modelo foi treinado por 10 épocas (suficiente para alcançar a convergência da acurácia) e, em seguida, acrescido de ruído gaussiano modelado com diferentes valores de σ (sem aplicar o *clipping* aos gradientes). Para cada valor de σ, foi aplicado o ataque iDLG e o resultado da reconstrução da imagem alvo foi comparado à imagem original por uma métrica de similaridade estrutural (do inglês, *Structural Similarity Index Metric*, SSIM[2]), que avalia a similaridade de forma mais aproximada da percepção humana. Assim, foi possível verificar a quantidade de ruído necessário para mitigar o ataque.

Para o primeiro experimento, os resultados estão expressos na Figura 1a. É possível perceber que as camadas convolucionais possuem sensibilidade maior aos GIA do que a camada linear final. Além disso, os valores de sensibilidade são maiores nas camadas iniciais, o que é um indicativo de que os filtros convolucionais, principalmente

[1] https://www.kaggle.com/datasets/jessicali9530/lfw-dataset

[2] https://scikit-image.org/docs/0.25.x/auto_examples/transform/plot_ssim.html



(a) Sensibilidade das camadas a cada época

(b) SSIM das reconstruções para diferentes σ

Figura 1. Em (a), sensibilidade das camadas do modelo por época de treinamento, medida pela norma de Frobenius do Jacobiano. Em (b), similaridade estrutural (SSIM) das imagens reconstruídas pelo ataque iDLG em função do nível de ruído gaussiano adicionado ao modelo.

os iniciais, contribuem mais para o vazamento dos dados, ao realizarem operações diretamente sobre eles e carregarem mais informações inerentes. Em adição, é possível perceber que a sensibilidade das camadas vai diminuindo para um patamar pequeno com o passar das rodadas, o que indica que modelos treinados por mais épocas ou que sofreram *overfitting* são menos suscetíveis aos GIA. Esse resultado sugere que, em cenários reais, um atacante obteria maior sucesso ao focar na interceptação dos gradientes nas primeiras camadas, onde a informação sobre os dados brutos está menos abstraída pelo modelo.

Para o segundo experimento, os resultados estão expressos na Figura 1b. Como esperado, as imagens reconstruídas pelo ataque iDLG são menos fiéis quanto maior o nível de ruído adicionado ao modelo. Assim, pode-se pensar em um *threshold* de similaridade para definir o ruído ideal, de forma a mitigar os GIA sem comprometer contundentemente a utilidade do modelo. Além disso, percebe-se que a reconstrução fica degradada com níveis relativamente baixos de ruído. Observa-se que, para $\sigma \geq 0{,}010$, o SSIM das reconstruções cai abaixo de $0{,}4$, tornando as imagens praticamente irreconhecíveis, o que sugere a existência de um *threshold* prático para a aplicação eficaz do DP-SGD sem comprometer excessivamente a utilidade do modelo.

4. Conclusões e Trabalhos Futuros

Este trabalho investigou a sensibilidade de modelos locais aos GIA no contexto de FL. Os experimentos evidenciaram que camadas convolucionais apresentam maior sensibilidade e propensão ao vazamento de dados, embora essa suscetibilidade diminua naturalmente com o avançar do treinamento. Ademais, a inserção de ruído gaussiano mostrou-se eficaz na mitigação dos GIA, degradando severamente a qualidade das reconstruções mesmo com baixos níveis σ, indicando que esses ataques são sensíveis a defesas conhecidas do estado-da-arte. Como trabalhos futuros, avalia-se a aplicação a arquiteturas mais complexas e mais conjuntos de dados, bem como a definição da avaliação de sucesso dos GIA como métrica para busca automática de um valor ótimo de σ.

Agradecimentos

Este trabalho contou com financiamento parcial de: CNPq; Capes; Fapes (2024-9G1WP, 2023-RWXSZ, 2026-B74MN).

Referências

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 308–318, New York, NY, USA. Association for Computing Machinery.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Leite, L., Santo, Y., Dalmazo, B., and Riker, A. (2024). Federated learning under attack: Improving gradient inversion for batch of images. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 794–800, Porto Alegre, RS, Brasil. SBC.

Mammen, P. M. (2021). Federated learning: Opportunities and challenges.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

McMahan, H. B., Andrew, G., Erlingsson, U., Chien, S., Mironov, I., Papernot, N., and Kairouz, P. (2019). A general approach to adding differential privacy to iterative training procedures.

Mo, F., Borovykh, A., Malekzadeh, M., Haddadi, H., and Demetriou, S. (2021). Layer-wise characterization of latent information leakage in federated learning.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. (2018). Sensitivity and generalization in neural networks: an empirical study.

Wang, Z., Peng, C., He, X., and Tan, W. (2023). Wasserstein distance-based deep leakage from gradients. *Entropy*, 25(5).

Zhao, B., Mopuri, K. R., and Bilen, H. (2020). idlg: Improved deep leakage from gradients. *ArXiv*, abs/2001.02610.

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.