Scaling Limits of the Dropout Inversion Attack: An Empirical Analysis Under Increasing Input Dimensionality

Eduardo Silva[1,3], **Rodrigo Silva**[2,3]

[1]Programa de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

[2]Departamento de Computação, Universidade Federal de Ouro Preto
Ouro Preto - MG - Brasil

[3]Laboratório de Computação de Sistemas Inteligentes - CSI Lab
Ouro Preto - MG - Brasil

eduardo.passos@aluno.ufop.edu.br

Abstract. *Gradient inversion attacks exploit shared gradients in federated learning to reconstruct private training samples. The Dropout Inversion Attack (DIA) bypasses dropout-based defenses by jointly optimizing the synthetic input, class labels, and dropout masks. This work characterizes the scalability limits of DIA through a factorial experiment on CIFAR-10 upscaled from 32×32 to 224×224 pixels across four architectures and three dropout rates. For the MLP, reconstruction quality decreases monotonically with resolution (56.1 ± 15.0 dB at 32×32 to 27.2 ± 6.3 dB at 224×224) while wall-clock time grows superlinearly. LeNet maintains high quality across the entire range without monotonic decay. ResNets fail regardless of resolution, an effect attributed to batch normalization. DIA scalability is architecture-dependent: for MLP-type architectures the practical barrier at scale is computational rather than qualitative.*

Resumo. *Ataques de inversão de gradientes exploram gradientes compartilhados no aprendizado federado para reconstruir amostras privadas de treinamento. O Dropout Inversion Attack (DIA) contorna defesas baseadas em dropout ao otimizar conjuntamente a entrada sintética, os rótulos e as máscaras de dropout. Aqui, caracterizamos os limites de escalabilidade do DIA por meio de um experimento fatorial com CIFAR-10, quatro arquiteturas e três taxas de dropout. No MLP, a qualidade de reconstrução decresce monotonicamente com a resolução (56,1 ± 15,0 dB a 32×32 para 27,2 ± 6,3 dB a 224×224) enquanto o tempo de execução cresce de forma supralinear. A LeNet mantém qualidade alta em toda a faixa testada sem decaimento monotônico. As ResNets falham independentemente da resolução, efeito atribuído ao batch normalization. A escalabilidade do DIA é dependente de arquitetura: para MLPs a barreira prática em alta resolução é computacional, não de qualidade.*

1. Introdução

O Aprendizado Federado (FL) permite treinar modelos de aprendizado de máquina sem que os clientes compartilhem seus dados brutos, transmitindo ao servidor apenas os gradientes calculados localmente [McMahan et al. 2017]. Contudo, a privacidade desta abor-

dagem não é garantida: ataques de inversão de gradiente (*gradient inversion attacks*) demonstraram que é possível reconstruir os dados de treinamento originais a partir apenas dos gradientes observados pelo servidor [Zhu et al. 2019, Geiping et al. 2020].

O *dropout* [Srivastava et al. 2014] é frequentemente considerado como mecanismo de defesa em FL [Zheng 2021, Zhang et al. 2022, Sotthiwat et al. 2026], pois, além de já estar embutido no processo de treinamento da maioria dos modelos de deep learning, introduz estocasticidade nas ativações durante o treinamento e aparentemente degrada a qualidade dos ataques de inversão clássicos. Scheliga et al. [Scheliga et al. 2023] demonstraram, contudo, que essa sensação de segurança é ilusória: a degradação observada resulta do *desalinhamento estocástico* entre a máscara de dropout aplicada pelo cliente (Ψ_C) e aquela assumida pelo atacante (Ψ_A), não de uma proteção estrutural. Os autores propuseram o *Dropout Inversion Attack* (DIA), que supera essa limitação ao otimizar conjuntamente a entrada sintética $\hat{x}$, o rótulo estimado $\hat{y}$ e as máscaras de dropout do atacante Ψ_A. Para $p = 0$, o DIA degenera no ataque clássico de Geiping et al. [Geiping et al. 2020].

Uma limitação explícita do trabalho de Scheliga et al. é que os experimentos foram conduzidos predominantemente em baixa resolução. Permanece em aberto como a viabilidade do DIA se comporta quando a dimensão de entrada $X = C \cdot H \cdot W$ cresce. Ao aumentar a resolução, três fatores degradantes atuam simultaneamente: (i) a dimensionalidade de $\hat{x}$ cresce com $H \cdot W$, expandindo o espaço de busca do otimizador; (ii) o número de variáveis de máscara Ψ_A a otimizar aumenta; e (iii) a "assinatura" de cada amostra no gradiente agregado torna-se proporcionalmente menor.

Este trabalho investiga empiricamente a fronteira de viabilidade do DIA em função da escala de entrada, por meio de um experimento fatorial controlado com CIFAR-10 reescalado de 32×32 a 224×224 pixels, quatro arquiteturas (MLP, LeNet, ResNet18, ResNet50) e três taxas de dropout ($p \in \{0{,}0, 0{,}25, 0{,}5\}$). As hipóteses de trabalho (H1, H2) são formalizadas na Seção 3.

As contribuições deste artigo são:

- Caracterização empírica da degradação de qualidade (PSNR, SSIM) e do crescimento de custo computacional (tempo de parede, iterações) do DIA em função da resolução de entrada;
- Evidência de que, para o MLP, a barreira prática à escala é *computacional*, não de qualidade — o ataque mantém PSNR estruturalmente coerente (27,2 ± 6,3 dB) até 224×224 em entradas *upscaled*, com ressalvas sobre validade externa (Seção 3.5);
- Demonstração de que a penalidade de escala é *dependente de arquitetura*: CNN sem BatchNorm (LeNet) não apresenta tendência monotônica de PSNR na faixa testada, enquanto MLP com $p = 0$ decresce monotonicamente.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta o contexto teórico e os trabalhos relacionados; a Seção 3 detalha a metodologia experimental; a Seção 4 analisa os resultados; e a Seção 5 apresenta as conclusões.

2. Fundamentação e Trabalhos Relacionados

2.1. Aprendizado Federado e Privacidade de Gradientes

No paradigma de FL proposto por McMahan et al. [McMahan et al. 2017], um conjunto de K clientes treina colaborativamente um modelo global $\mathbf{w}$ sem compartilhar dados brutos. Em cada rodada t, o servidor distribui os parâmetros correntes $\mathbf{w}^t$ a um subconjunto

de clientes; cada cliente k computa uma atualização local minimizando sua perda empírica $\mathcal{L}_k$ sobre o conjunto de dados privado $\mathcal{D}_k$:

$$\mathbf{g}_k^t = \nabla_{\mathbf{w}} \mathcal{L}_k(\mathbf{w}^t;\, \mathcal{D}_k), \tag{1}$$

e transmite $\mathbf{g}_k^t$ ao servidor, que agrega as contribuições — tipicamente pela média ponderada do FedAvg — para produzir $\mathbf{w}^{t+1}$. Os dados $\mathcal{D}_k$ nunca saem do dispositivo local; apenas $\mathbf{g}_k^t$ trafega pela rede.

Essa separação entre dados e gradientes é frequentemente apresentada como uma garantia de privacidade, mas essa leitura é imprecisa. O vetor $\mathbf{g}_k^t$ é calculado sobre $\mathcal{D}_k$ e, portanto, carrega informação sobre as amostras que o originaram. Um atacante com acesso a $\mathbf{g}_k^t$ e aos parâmetros $\mathbf{w}^t$ pode formular a reconstrução como um problema de otimização inversa: encontrar uma entrada sintética $\hat{x}$ e um rótulo $\hat{y}$ cujos gradientes coincidam com os gradientes observados. Zhu et al. [Zhu et al. 2019] propuseram minimizar a distância ℓ_2 entre gradientes:

$$\min_{\hat{x},\,\hat{y}} \left\| \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^t;\, \hat{x}, \hat{y}) - \mathbf{g}_k^t \right\|_2^2, \tag{2}$$

demonstrando que $\mathcal{D}_k$ pode ser reconstruído com alta fidelidade a partir de $\mathbf{g}_k^t$, sem acesso direto aos dados. Geiping et al. [Geiping et al. 2020] refinaram a formulação substituindo a distância ℓ_2 por uma função de perda baseada em similaridade de cosseno e adicionando um regularizador de imagem $\Omega(\hat{x})$:

$$\min_{\hat{x},\,\hat{y}} \; 1 - \frac{\langle \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^t;\, \hat{x}, \hat{y}),\, \mathbf{g}_k^t \rangle}{\|\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^t;\, \hat{x}, \hat{y})\| \, \|\mathbf{g}_k^t\|} + \lambda\, \Omega(\hat{x}), \tag{3}$$

obtendo reconstruções de alta qualidade em redes convolucionais profundas (*Inverting Gradients*, IG). As Equações 2 e 3 estabelecem a estrutura geral dos ataques de inversão: dado $\mathbf{g}_k^t$, o atacante resolve um problema de otimização no espaço de entradas para recuperar $\mathcal{D}_k$. O DIA estende essa formulação para o caso em que o cliente emprega dropout durante o treinamento.

2.2. Dropout como Defesa e suas Limitações

O dropout [Srivastava et al. 2014], técnica de regularização amplamente utilizada, introduz estocasticidade nas ativações ao zerá-las aleatoriamente com probabilidade p durante o treinamento. Em FL, isso cria uma divergência entre o modelo realizado pelo cliente (máscara Ψ_C) e o assumido pelo atacante (Ψ_A): quando $\Psi_A \neq \Psi_C$, a similaridade de cosseno entre os gradientes reais e os sintéticos é degradada, e o ataque IG aparentemente falha.

Scheliga et al. [Scheliga et al. 2023] formalizaram esse fenômeno e demonstraram que a proteção do dropout é *ilusória*. O atacante pode modelar explicitamente a incerteza introduzida pelo dropout, expandindo o espaço de otimização para incluir as máscaras:

$$\min_{\hat{x},\,\hat{y},\,\Psi_A} \; 1 - \frac{\langle \nabla_C,\, \nabla(\hat{x}, \hat{y}; \Psi_A) \rangle}{\|\nabla_C\| \, \|\nabla(\hat{x}, \hat{y}; \Psi_A)\|} \; + \; \lambda\, \Omega(\hat{x}), \tag{4}$$

onde ∇_C é o gradiente observado do cliente, Ω é um regularizador de imagem e as máscaras Ψ_A são relaxadas para valores contínuos em $[0, 1]$ para permitir gradiente estável. Esta formulação define o **DIA**, que recupera a qualidade de reconstrução independentemente da taxa de dropout. Para $p = 0$ (sem dropout), a Equação 4 degenera no ataque IG.

2.3. Escalabilidade de Ataques de Inversão

A literatura sobre ataques de inversão concentra-se predominantemente em regimes de baixa resolução [Guo et al. 2025, Carletti et al. 2025]. Por exemplo, os trabalhos seminais de Zhu et al. [Zhu et al. 2019] e Geiping et al. [Geiping et al. 2020] validam seus métodos em CIFAR-10 (32×32) e conjuntos de dados similares. Estudos em resolução maior, como ImageNet (224×224), reportam degradação significativa de qualidade [Geiping et al. 2020], mas sem análise sistemática da relação entre dimensão de entrada e viabilidade do ataque. Geiping et al. reportam PSNR substancialmente inferior em ImageNet nativo em comparação com CIFAR-10 (32×32), o que sugere que conteúdo espectral de alta frequência e escala absoluta de parâmetros modulam a viabilidade do IG; este trabalho *não* replica esse baseline em imagens nativas de alta resolução, medindo em vez disso o efeito isolado da dimensionalidade do espaço de busca via *upscaling* de CIFAR-10 (Seção 3.5).

Trabalhos de defesa baseados em compressão ou perturbação de gradientes [Sun et al. 2021, Zhu et al. 2021] reduzem indiretamente a informação disponível ao atacante, sendo uma estratégia ortogonal ao dropout. O efeito da dimensionalidade sobre ataques de inversão com otimização conjunta de máscaras, como o DIA, permanece, até onde é de conhecimento dos autores, não investigado de forma sistemática — lacuna que este trabalho busca preencher.

3. Metodologia

3.1. Configuração Experimental

Os experimentos estendem a implementação pública de Scheliga et al. [Scheliga et al. 2023], adicionando protocolo de varredura fatorial controlado com múltiplas sementes fixas por execução, *logging* incremental em CSV, lógica de *early-stopping* e suporte a múltiplas resoluções.

3.1.1. Conjunto de dados

O CIFAR-10 (32×32, RGB) foi utilizado como base. Para avaliar o efeito da dimensionalidade, 128 imagens do conjunto original foram reescaladas por interpolação bilinear para 64×64, 128×128 e 224×224 pixels, preservando média e desvio-padrão de normalização do CIFAR-10. Cada resolução constitui uma condição experimental independente. Nota metodológica: as imagens de alta resolução são *upscalings* de originais 32×32 e não imagens nativas de alta frequência; o experimento mede, portanto, o efeito da *dimensionalidade do espaço de busca*, não do conteúdo espectral da imagem. Como o reescalamento bilinear preserva estrutura de baixa frequência, valores elevados de SSIM, como os observados na LeNet (Tabela 2, SSIM $\geq 0{,}94$), podem superestimar a fidelidade percebida em comparação com reconstruções sobre dados nativos de alta resolução; interpretamos SSIM aqui como indicador *interno* de consistência estrutural entre original upscaled e reconstrução, não como evidência de viabilidade em ImageNet ou dados reais.

3.1.2. Arquiteturas

Quatro arquiteturas informam a questão de escalabilidade, organizadas em três papéis experimentais: MLP (camada totalmente conectada, cuja dimensão de entrada cresce com $H \cdot W$), LeNet (CNN sem *batch normalization*) e, como *controles negativos* para o efeito de *batch normalization*, ResNet18 e ResNet50 (ambas com BN). Incluir ResNets no fatorial completo confunde dropout e BN — uma limitação de desenho reconhecida a priori (Seção 3.5); os resultados das ResNets servem apenas para confirmar que BN impede inversão, não para inferir tendências de escala.

3.1.3. Fatores e níveis.

O experimento segue um projeto fatorial completo com os fatores: resolução (32, 64, 128, 224 px), arquitetura (MLP, LeNet, ResNet18, ResNet50) e taxa de dropout ($p \in \{0{,}0, 0{,}25, 0{,}5\}$), totalizando $4 \times 4 \times 3 = 48$ células experimentais. A análise principal de escalabilidade restringe-se a MLP e LeNet (24 células); as ResNets compõem bloco de controle separado.

3.1.4. Ambiente de execução.

Os experimentos foram executados com PyTorch 2.11 e Python 3.14 sobre computador com processador Ryzen 9 7900X, 16 GB de RAM, GPU **RTX 5060 Ti 8 GB**, em `cuda:0`. O tempo de parede (`wall_clock_s`) é medido de forma *end-to-end* por célula, incluindo forward/backward e atualização de $\hat{x}$, $\hat{y}$ e Ψ_A. Valores absolutos de tempo dependem do hardware; neste estudo são usados exclusivamente para comparação *entre células* sob o mesmo ambiente (RTX 5060 Ti/CUDA, sementes 42–51, lote $b = 2$, processo *single-thread* por célula).

3.2. Hipóteses de Trabalho

Com base na lacuna identificada na Seção 2, formulamos:

H1 (escala $\Rightarrow$ qualidade e custo). Para o MLP com $p = 0$ (p conforme definido na Seção 2), a qualidade de reconstrução (PSNR/SSIM) decresce monotonicamente com o aumento da resolução, e o custo (iterações e tempo de parede) cresce de forma supralinear — tornando o ataque progressivamente menos prático em escala.

H2 (interação dropout $\times$ resolução). A interação entre taxa de dropout p e resolução difere do regime de baixa resolução: para $p > 0$, o comportamento do DIA em qualidade e custo não coincide com o de $p = 0$ (IG puro), especialmente em resoluções intermediárias e altas.

Estas hipóteses são testadas empiricamente na Seção 4; conclusões sobre monotonicidade de PSNR restringem-se às colunas $p = 0$ e $p = 0{,}25$ da Tabela 3.

3.3. Protocolo de Ataque

Em cada célula experimental, o atacante observa o gradiente ∇_C produzido por um lote de $b = 2$ imagens e otimiza a Equação 4 usando o Adam com taxa de aprendizado inicial $\eta =$

$0{,}1$ e escalonamento adaptativo via *ReduceLROnPlateau*. A otimização é interrompida quando a perda cai abaixo de 10^{-5} (*early-stopping*), após 1 500 iterações sem melhora, ou após 20 000 iterações (teto máximo). Os experimentos utilizam sementes 42–51, uma por execução independente, controlando o sorteio das imagens vítimas, a inicialização dos pesos e o *dummy* inicial de cada ataque. O *batch normalization* é mantido em modo `eval` nos dois lados (vítima e atacante) para que avaliem a mesma função; o dropout permanece em `train`. Para $p = 0$, o DIA degenera no ataque IG [Geiping et al. 2020].

3.4. Métricas

Quatro métricas são registradas por célula:

- **PSNR** (dB): razão sinal-ruído de pico entre imagem reconstruída e original, calculada após desnormalização;
- **SSIM**: índice de similaridade estrutural;
- **Iterações necessárias** (`needed_iterations`): número de passos até *early-stopping* ou teto;
- **Tempo de parede** (`wall_clock_s`): tempo total de execução em segundos.

3.5. Limitações

Cinco limitações condicionam a interpretação dos resultados. Primeiro, as execuções repetidas permitiram estimar média e desvio-padrão; contudo, a variância entre execuções é elevada (6–21 dB no MLP/LeNet), de modo que diferenças menores que um desvio (incluindo não-monotonicidades de PSNR nas colunas $p > 0$ da Tabela 3) devem ser tratadas como sugestivas. Segundo, durante o experimento, o teto de 20 000 iterações foi atingido em menos de 4% das execuções, exclusivamente no MLP; em tais casos a otimização foi interrompida sem atingir convergência. Monotonicidade de H1 possui folga larga (27,23 dB a 224×224 não depende desses casos). Terceiro, resoluções acima de 224×224 não foram cobertas, de modo que o regime $X \to \infty$ permanece em aberto. Quarto, o uso de *upscaling* em vez de imagens nativas de alta resolução limita a validade externa: SSIM elevado pode refletir baixo conteúdo espectral upscaled, não robustez do DIA em ImageNet ou dados reais. Quinto, incluir ResNet18/ResNet50 no fatorial confunde dropout e *batch normalization*; idealmente, variantes ResNet sem BN deveriam compor o desenho principal. Além disso, os dados de ResNet reportados neste trabalho cobrem apenas a resolução 32×32 (Figura 1); a extensão da falha das ResNets a outras resoluções não foi verificada diretamente e deve ser lida como hipótese plausível, não como resultado testado.

4. Resultados

4.1. Efeito da Resolução no MLP

A Tabela 1 resume os resultados do MLP, a arquitetura cujo número de parâmetros de entrada cresce linearmente com $H \cdot W$, tornando-a o caso mais revelador para o estudo de escala.

O PSNR decresce monotonicamente com a resolução *para* $p = 0$ *e* $p = 0{,}25$ (56,1 dB a 32×32 até 27,2 dB a 224×224 em $p = 0$), apoiando H1 na dimensão qualidade sob essas condições. Na coluna $p = 0{,}5$ da Tabela 3, a monotonicidade *não* se sustenta, o que é compatível com H2. A degradação não é uniforme: há uma queda abrupta entre

Tabela 1. PSNR, SSIM, iterações e tempo de parede do DIA no MLP ($p = 0$) em função da resolução. Média ± desvio.

Resolução	$X = 3 \cdot \text{res}^2$	PSNR (dB)	SSIM	Iter.	Tempo (s)
32×32	3 072	56,09 ± 15,00	0,877 ± 0,379	2 687 ± 1 366	23,6 ± 14,1
64×64	12 288	42,15 ± 14,60	0,876 ± 0,379	5 713 ± 2 165	87,7 ± 40,6
128×128	49 152	28,09 ± 9,77	0,756 ± 0,482	10 312 ± 4 202	196,8 ± 80,2
224×224	150 528	27,23 ± 6,33	0,872 ± 0,292	14 600 ± 3 990	719,1 ± 196,4

Tabela 2. PSNR (dB) e SSIM da LeNet ($p = 0$) por resolução. Média ± desvio.

Resolução	$X = 3 \cdot \text{res}^2$	PSNR (dB)	SSIM	Tempo (s)
32×32	3 072	50,81 ± 11,90	0,9925 ± 0,0200	63,2 ± 16,7
64×64	12 288	51,60 ± 10,27	0,9977 ± 0,0047	85,4 ± 31,1
128×128	49 152	43,42 ± 9,94	0,9946 ± 0,0079	81,4 ± 69,1
224×224	150 528	32,34 ± 9,51	0,9447 ± 0,1120	441,3 ± 187,5

64 e 128 px ($\Delta \approx -14$ dB), após a qual o decaimento desacelera (128 → 224 px: $\Delta \approx -0{,}9$ dB), sugerindo que o ataque se aproxima de um patamar em vez de colapsar. A 224×224, PSNR de $27{,}2 \pm 6{,}3$ dB e SSIM de $0{,}872 \pm 0{,}292$ indicam reconstrução estruturalmente coerente *no regime upscaled* (Seção 3.5), distante do piso de ruído (9,6–16,5 dB observado nas ResNets a 32×32).

4.2. Custo Computacional e a Fronteira Prática

O número de iterações necessárias cresce de forma moderada com a resolução (2 687 a 32×32 para 14 600 a 224×224), verdadeiro apenas para $p = 0$; com dropout, a 224 px o número de iterações se aproxima do teto de 20 000 (19 840 para $p = 0{,}25$; 19 999 para $p = 0{,}5$). O tempo de parede cresce de forma supralinear: de 128 para 224 px, os pixels aumentam ≈3,06×, mas o tempo cresce ≈3,65× (de 196,8 s para 719,1 s).

Esses resultados sugerem que, para o MLP com $p = 0$, a barreira efetiva à escala é computacional, não de qualidade de reconstrução: o ataque permanece viável em PSNR/SSIM dentro da faixa upscaled testada, mas o custo de tempo cresce de forma supralinear, apoiando H1 na dimensão custo. Tempos absolutos são comparativos intra-estudo (Seção 3).

4.3. Dependência Arquitetural da Penalidade de Escala

A Tabela 2 resume o PSNR da LeNet ($p = 0$) ao longo das quatro resoluções; a Figura 1 compara todas as arquiteturas *somente* a 32×32 (baseline de referência).

A LeNet (CNN sem *batch normalization*) mantém PSNR entre 32 e 52 dB de 32×32 a 224×224, *sem tendência monotônica* (Tabela 2): o valor a 64×64 (51,60 dB) supera o de 32×32 (50,81 dB). Esse comportamento é consistente com a estrutura da rede: os pesos convolucionais são compartilhados espacialmente, de modo que o número de parâmetros quase não cresce com a resolução, preservando a relação sinal/gradiente. O MLP, em contraste, expande a camada de entrada linearmente com $H \cdot W$, diluindo a assinatura de cada amostra no gradiente (Tabela 1, $p = 0$).

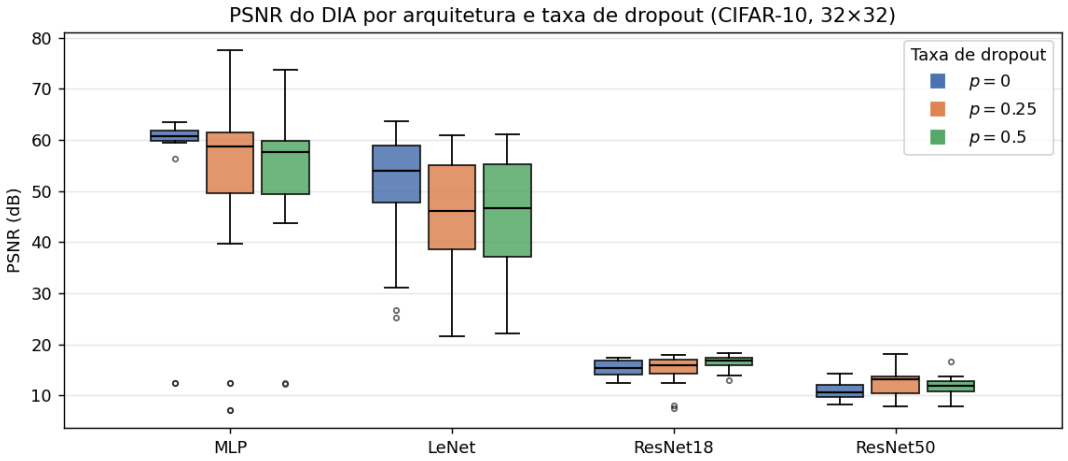


Figura 1. PSNR por arquitetura e taxa de dropout a 32×32 apenas (baseline; não representa variação ao longo das resoluções). Caixas mostram mediana, quartis (Q_1–Q_3) e amplitude dentro de $1{,}5\,\mathrm{IQR}$; pontos são *outliers*.

Tabela 3. PSNR (dB) do MLP por taxa de dropout e resolução. Média ± desvio. Monotonicidade de PSNR com resolução observa-se apenas na coluna $p = 0{,}0$.

Resolução	$p = 0{,}0$	$p = 0{,}25$	$p = 0{,}5$
32×32	$56{,}09 \pm 15{,}00$	$49{,}36 \pm 21{,}57$	$52{,}73 \pm 15{,}40$
64×64	$42{,}15 \pm 14{,}60$	$40{,}19 \pm 16{,}92$	$32{,}57 \pm 15{,}39$
128×128	$28{,}09 \pm 9{,}77$	$29{,}97 \pm 7{,}39$	$19{,}19 \pm 10{,}58$
224×224	$27{,}23 \pm 6{,}33$	$27{,}37 \pm 7{,}54$	$27{,}85 \pm 8{,}35$

As ResNets (ResNet18 e ResNet50) falham em todas as resoluções e taxas de dropout, com PSNR $9{,}6$–$16{,}5$ dB e SSIM $0{,}05$–$0{,}65$ (Figura 1, baseline 32×32). A falha é independente da resolução e do dropout e é atribuída ao *batch normalization*, que introduz acoplamento entre amostras do lote e dificulta a inversão. Esses modelos funcionam como *controles negativos* (Seção 3.5), não como evidência sobre escalabilidade.

4.4. Interação entre Dropout e Resolução

A Tabela 3 apresenta o PSNR do MLP por taxa de dropout e resolução, e a Figura 2 mostra a curva de degradação.

O efeito do dropout sobre a qualidade de reconstrução é não-monotônico e dependente da resolução. A 32×32, as três taxas produzem PSNR entre 49 e 56 dB, e a diferença fica dentro de um desvio. A 64×64, $p > 0$ prejudica a reconstrução em relação a $p = 0$, de forma desigual entre taxas: -2 dB para $p = 0{,}25$ e $-9{,}6$ dB para $p = 0{,}5$ ($-5{,}8$ dB em média).

O efeito do dropout sobre o **custo** é, em contraste, consistente e robusto: o dropout *sempre* aumenta o número de iterações e o tempo de parede em todas as resoluções. A 224×224, por exemplo, as iterações crescem de $14\,600$ ($p = 0$) para $19\,840$ ($p = 0{,}25$) e $19\,999$ ($p = 0{,}5$); o tempo de parede passa de 719 s para $1\,049$ s e $1\,258$ s, respectivamente (Figura 3). Esse padrão apoia parcialmente H2: a interação dropout–resolução difere

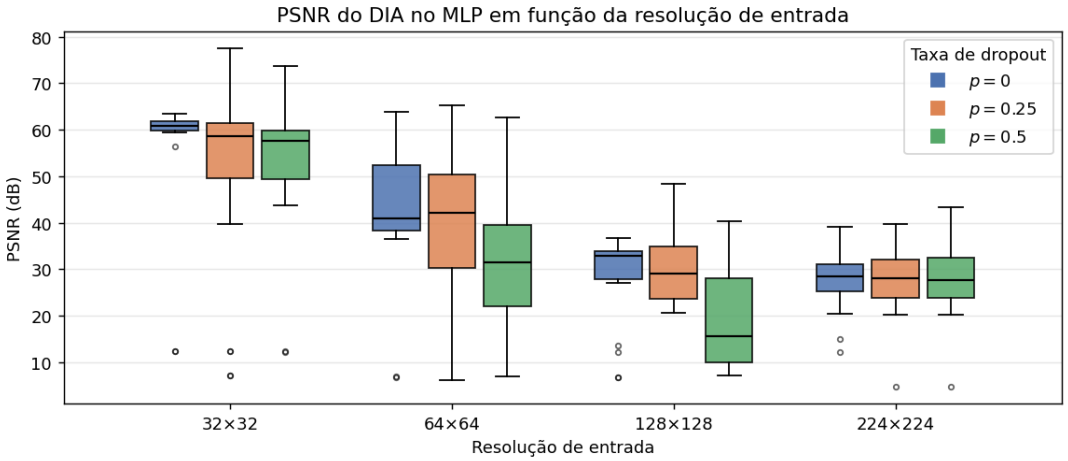


Figura 2. PSNR do MLP em função da resolução (32×32 a 224×224) para cada taxa de dropout. Caixas mostram mediana, quartis (Q_1–Q_3) e amplitude dentro de $1{,}5\,\mathrm{IQR}$; pontos são *outliers*.

do regime de baixa resolução, sobretudo na dimensão custo; a dimensão qualidade não admite generalização monotônica entre colunas de p.

4.5. Validação de Sanidade

É importante ressaltar que antes da varredura fatorial, a implementação foi validada em dois cenários do artigo de referência. Para MLP/$p{=}0$ a 32×32, o DIA converge com *early-stopping* em $2\,687 \pm 1\,366$ iterações (média ± desvio sobre 10 execuções), equivalente ao IG. Para ViT/$p{=}0{,}5$, o IG estaciona (perda $\approx 0{,}336$) enquanto o DIA converge, reproduzindo o fenômeno demonstrado por Scheliga et al.

5. Conclusões

Este trabalho caracterizou empiricamente os limites de escalabilidade do Dropout Inversion Attack (DIA) em função da dimensão de entrada, estendendo os experimentos de Scheliga et al. [Scheliga et al. 2023] para resoluções de 32×32 a 224×224 pixels.

Os resultados mostram que, para o MLP com $p = 0$, a qualidade de reconstrução decresce monotonicamente com a resolução — de 56,1 dB a 32×32 para 27,2 dB a 224×224 — enquanto o custo computacional cresce de forma supralinear. Para $p > 0$, a monotonicidade de PSNR não se sustenta (Tabela 3). A barreira efetiva ao ataque nesse regime é, portanto, **computacional**: o DIA preserva fidelidade estrutural (SSIM $0{,}756$–$0{,}877$) na faixa upscaled testada, sem colapso de qualidade, com ressalva sobre validade externa (Seção 3.5).

Para a LeNet (CNN sem *batch normalization*), PSNR permanece entre 32 e 52 dB de 32 a 224 px (Tabela 2), *sem* decaimento monotônico mensurável, revelando que a penalidade de escala é **dependente de arquitetura**. A intuição de que "maior resolução implica ataque menos eficaz" vale para modelos com camadas FC de entrada (MLP, $p = 0$), mas não se generaliza a CNNs com pesos compartilhados na faixa de resoluções testada. As ResNets falham independentemente de resolução e dropout (controles de BN), efeito atribuído ao *batch normalization*, não à dimensionalidade da entrada.

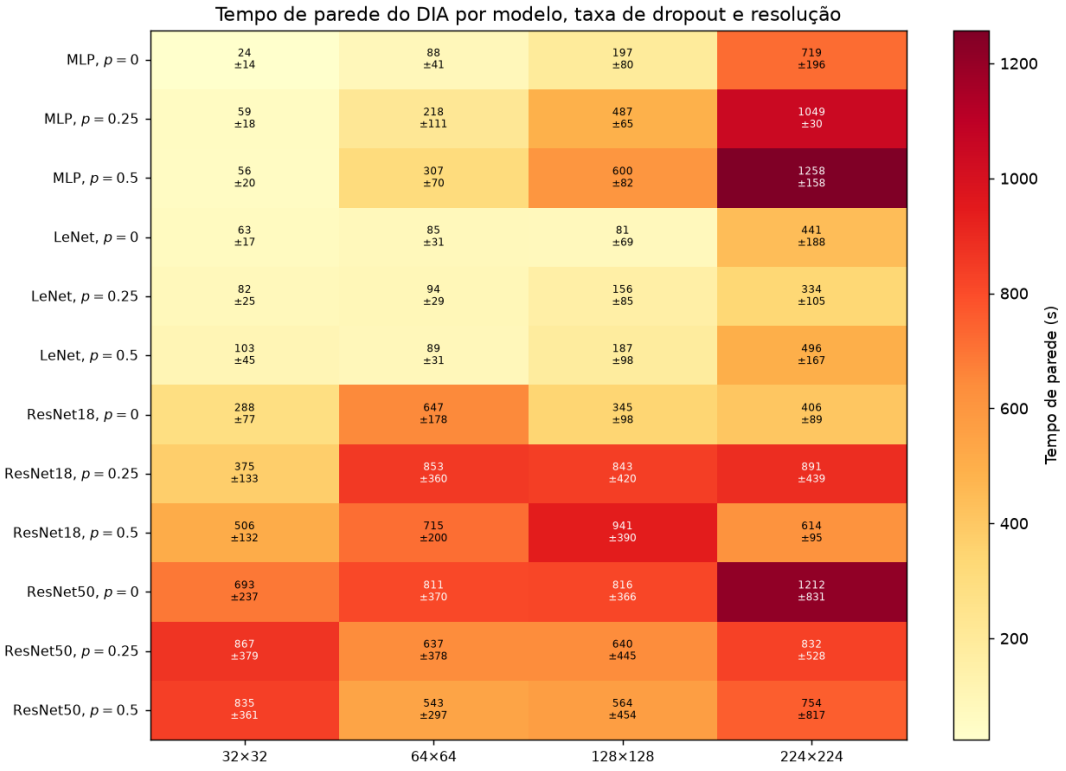


Figura 3. Tempo de parede (s) por modelo, dropout e resolução (resoluções de 32×32 a 224×224). Média sobre 10 execuções.

O dropout eleva consistentemente o custo do ataque em todas as resoluções, sem impedir a reconstrução. Seu efeito sobre a qualidade de saída é não-monotônico e, dado que as diferenças entre taxas de dropout ficam frequentemente dentro de um desvio-padrão entre execuções (Seção 3.5), não conclusivo.

Quatro direções naturais decorrem das limitações deste trabalho. A mais imediata é repetir o protocolo, mantendo o desenho de múltiplas sementes já empregado aqui, sobre imagens nativas de alta resolução (ImageNet ou CelebA-HQ), para separar o efeito da dimensionalidade do espaço de busca do efeito do conteúdo espectral, que o upscaling bilinear não reproduz. Uma segunda direção é incluir variantes ResNet sem batch normalization no desenho fatorial; sem esse controle, não é possível distinguir o efeito do BN do efeito da escala nessas arquiteturas. A terceira é estender a faixa de resolução além de 224×224 pixels: os dados sugerem que o MLP se aproxima de um patamar entre 128 e 224 px, mas não é possível afirmar, com os experimentos atuais, se esse patamar persiste ou se o PSNR colapsa em resoluções maiores. Por fim, a combinação de dropout com compressão de gradiente ou perturbação diferencial permanece sem avaliação sistemática; resta saber se mecanismos ortogonais produzem proteção composta contra o DIA ou se o ataque pode ser adaptado para contornar cada combinação.

Agradecimentos

Os autores agradecem ao CNPq, à FAPEMIG e à PROPPI/UFOP pelo suporte financeiro ao projeto de pesquisa.

Referências

Carletti, V., Foggia, P., Mazzocca, C., Parrella, G., and Vento, M. (2025). {SoK}: Gradient inversion attacks in federated learning. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 6439–6459.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting Gradients – How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, pages 16937–16947, Vancouver, Canada.

Guo, P., Wang, R., Zeng, S., Zhu, J., Jiang, H., Wang, Y., Zhou, Y., Wang, F., Xiong, H., and Qu, L. (2025). Exploring the vulnerabilities of federated learning: A deep dive into gradient inversion attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, Fort Lauderdale, FL, USA.

Scheliga, D., Mäder, P., and Seeland, M. (2023). Dropout Is NOT All You Need to Prevent Gradient Leakage. In *Proceedings of the National Conference on Artificial Intelligence*, pages 9934–9942, Washington, DC, USA.

Sotthiwat, E., Zhang, C., Xiao, X., and Zhen, L. (2026). Safeguarding federated learning from data reconstruction attacks via gradient dropout. *IEEE Transactions on Information Forensics and Security*, 21:1874–1888.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Sun, L., Qian, J., and Chen, X. (2021). Soteria: Provable Defense Against Privacy Leakage in Federated Learning from Representation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9311–9319, Nashville, TN, USA.

Zhang, R., Guo, S., Wang, J., Xie, X., and Tao, D. (2022). A survey on gradient inversion: Attacks, defenses and future directions. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5678–5685. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Zheng, Y. (2021). Dropout against deep leakage from gradients. *arXiv preprint arXiv:2108.11106*.

Zhu, L., Liu, Z., and Han, S. (2019). Deep Leakage from Gradients. In *Advances in Neural Information Processing Systems*, pages 14774–14784, Vancouver, Canada.

Zhu, R., Cao, D., Zheng, W., and Cao, J. (2021). R-GAP: Recursive Gradient Attack on Privacy. In *Proceedings of the International Conference on Learning Representations*, Virtual.