Supplementary Material for: Unveiling Privacy Risks in Stochastic Neural Networks Training: Effective Image Reconstruction from Gradients

Yiming Chen[1,2] ✉, Xiangyu Yang[1,2], and Nikos Deligiannis[1,2]

[1] Department of Electronics and Informatics, Vrije Universiteit Brussel (VUB), B-1050 Brussels, Belgium
[2] Interuniversitair Micro-Elektronica Centrum, B-3001 Leuven, Belgium
{cyiming, xyanga, ndeligia}@etrovub.de

A Proof of the Label Extraction Strategy

In this section, we present a comprehensive proof of the proposition outlined in Section 4.4. The proposition is as follows:

Proposition 1. *Consider the model is generally trained with cross-entropy loss using one-hot labels, and the vector* ∇W^i_{pred} *represents the gradient of weights* W^i_{pred} *that is connected to the* i^{th} *logit in the prediction (last) layer. The ground-truth label* y *can be analytically identified as checking the sign of* ∇W^i_{pred} *in single image training, as follows:*

$$y = i, \quad s.t. \quad {\nabla W^i_{pred}}^T \cdot \nabla W^j_{pred} \leq 0, \quad \forall j \neq i. \tag{1}$$

The authors of [10] established Prop. 1 for traditional (deterministic) neural networks, which is based on the proposition [10] below:

Proposition 2. *For any model with parameters* θ *that is generally trained with cross-entropy loss using one-hot labels, let* $\mathcal{L}(\mathcal{F}(\theta; x), y)$ *represent the model's computed loss, and* o_i *the logit output for the* i^{th} *class, given training images* x *and labels* y*. Then, the gradient of the loss with respect to each output is:*

$$g_i = \frac{\partial \mathcal{L}(\mathcal{F}(\theta; x), y)}{\partial o_i} = \begin{cases} -1 + \frac{e^{o_i}}{\sum_j e^{o_j}}, & \text{if } i = y \\ \frac{e^{o_i}}{\sum_j e^{o_j}}. & \text{otherwise} \end{cases} \tag{2}$$

As $\frac{e^{o_i}}{\sum_j e^{o_j}} \in (0, 1)$*, we have* $g_i \in (-1, 0)$ *when* $i = y$ *and* $g_i \in (0, 1)$ *when* $i \neq y$*. Therefore, the ground-truth label* y *can be identified as the index of the output* o_i *that has a negative gradient.*

We demonstrate that Proposition 1 is still applicable to Bayesian Layer-based SNNs, particularly when the last layer is a Bayesian Layer. The proof proceeds as follows:

Proof. Let W_μ^i and W_σ^i represent the mean and standard deviation, respectively, standing for the posterior distribution of the final layer's weights associated with the i^{th} logit. Following Eq. (7) in Sec. 3 of our paper, the weights linked to the i^{th} logit, W_i, are sampled from $\mathcal{N}(W_\mu^i, {W_\sigma^i}^2\mathbf{I})$ via reparameterization trick [1], as follows:

$$W_i = W_\mu^i + W_\sigma^i \odot \varepsilon_i, \tag{3}$$

where $\varepsilon_i \sim \mathcal{N}(0, \mathbf{I})$. For an input vector $\mathbf{a}$, the output logit o_i of Bayesian layer is given by:

$$o_i = W_i^T \cdot \mathbf{a} = (W_\mu^i + W_\sigma^i \odot \varepsilon_i)^T \cdot \mathbf{a}. \tag{4}$$

Note that the bias is omitted for brevity, as it does not affect the proof. Based on Eq.2 from Prop. 2, the gradients of W_μ^i can be derived as follows:

$$\begin{aligned} \nabla W_\mu^i = \frac{\partial \mathcal{L}(\mathcal{F}(\theta; x), y)}{\partial W_\mu^i} &= \frac{\partial \mathcal{L}(\mathcal{F}(\theta; x), y)}{\partial o_i} \cdot \frac{\partial o_i}{\partial W_\mu^i} \\ &= g_i \cdot \frac{\partial [(W_\mu^i + W_\sigma^i \odot \varepsilon_i)^T \cdot \mathbf{a}]}{\partial W_\mu^i} \\ &= g_i \cdot \mathbf{a}. \end{aligned} \tag{5}$$

Since $\mathbf{a}$ remains constant for all i. Therefore, the ground-truth label y can be analytically determined by checking the sign of ∇W_μ^i, which is different from that of the others. Overall, this suggests that Prop. 1 remains valid for networks with a Bayesian final layer. Specifically, when extracting the true label from the gradients of the last (Bayesian) layer, the gradient vector ∇W_{pred}^i in Prop. 1 refers to the gradient of the "weights" (i.e., the mean of the posterior distribution, W_μ^i), which is linked to the i^{th} logit in the last layer.

B Derivation of Intermediate Noise Regularization

This section provides a full derivation of Eq. (15) from Sec. 4.2, originally established by [5], to aid readers unfamiliar with the Kullback-Leibler Divergence loss. Here, μ and σ are the mean and standard deviation of ε, respectively, calculated by $\texttt{mean}(\cdot)$ and $\texttt{std}(\cdot)$ functions (see Sec. 4.2):

$$
\begin{aligned}
\mathcal{R}_{\texttt{kl}}(\varepsilon) &= \alpha_{\texttt{kl}} \cdot \mathcal{D}_{\texttt{KL}}(\mathcal{N}(\mu, \sigma^2)||\mathcal{N}(0,1)) \\
&= \alpha_{\texttt{kl}} \cdot \int_z \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)} dz \\
&= \alpha_{\texttt{kl}} \int_z [\frac{-(z-\mu)^2}{2\sigma^2} + \frac{z^2}{2} - \log\sigma] \mathcal{N}(\mu, \sigma^2) dz \\
&= \alpha_{\texttt{kl}} [-\int_z \frac{(z-\mu)^2}{2\sigma^2} \mathcal{N}(\mu, \sigma^2) dz + \int_z \frac{z^2}{2} \mathcal{N}(\mu, \sigma^2) dz - \int_z \log\sigma \mathcal{N}(\mu, \sigma^2) dz] \\
&= \alpha_{\texttt{kl}} (-\frac{\mathbb{E}[(z-\mu)^2]}{2\sigma^2} + \frac{\mathbb{E}[z^2]}{2} - \log\sigma) \\
&= \alpha_{\texttt{kl}} \frac{1}{2} (-\log\sigma^2 + \sigma^2 + \mu^2 - 1). \qquad (6)
\end{aligned}
$$

C Additional Reconstructed Images

Additional reconstructed images from CIFAR-10 using gradients of VB-ResNet18 and from CelebA using gradients of BL-ConvNet are depicted in Fig. 1 and 2, respectively.

Fig. 1: Visualization of different attack results of VB-ResNet18 on the CIFAR-10 classification FL tasks. The quality of image reconstruction is quantitatively assessed and reported in Tab. 1 of Sec. 5.2 in our main paper.

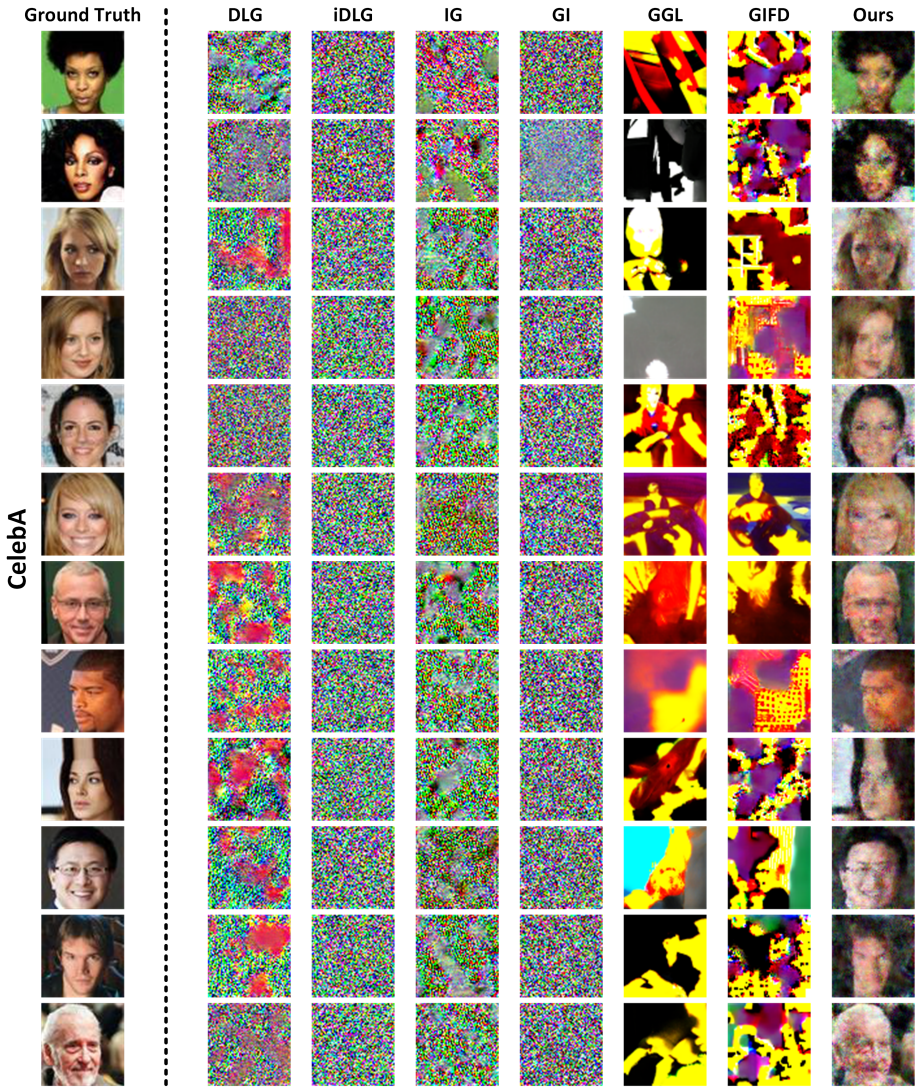


Fig. 2: Visualization of different attack results of Bayesian ConvNet on the gender classification FL tasks. The quality of image reconstruction is quantitatively assessed and reported in Tab. 1 of Sec. 5.2 in our main paper.

D Duration Analysis of Various Attack Algorithms for Image Reconstruction

Configuration. In this section, we explore the time required for our proposed attack approach to complete the reconstruction process from gradients. To ensure a fair comparison, the optimization iterations for IG [3], GI [8], and our proposed attack are uniformly set at 13,000, utilizing the Adam optimizer [4]. For DLG [11] and iDLG [10], the iterations are limited to 1,200, employing the L-BFGS optimizer [7]. GGL [6] and GIFD [2] adopted the pre-trained BigGAN for a deep image prior. Given GGL's rapid convergence as reported in [6], we adhered to the original experimental setup, allocating a total of 2,500 optimization iterations. For GIFD, the approach involves searching the intermediate features of BigGAN's first 13 layers, assigning 1,000 iterations per layer, resulting in a cumulative total of 13,000 iterations.

Results. Tables 1 and 2 present the duration of reconstruction (in seconds) for various attack approaches applied to the VB-ResNet18 and BL-ConvNet models, respectively. The observations reveal that incorporating an approximation of stochasticity does not substantially extend the duration of the reconstruction process. For instance, the time increase is only 17 seconds compared to IG (which took 472.05 seconds) on VB-ResNet18. On the contrary, DLG and iDLG require a significant amount of time to complete the attack task due to the use of the L-BFGS optimizer.

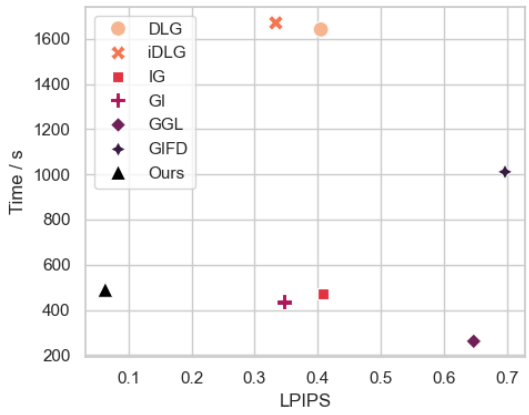
Trade-off between the attack duration and the quality of reconstruction. We report the relationship between the averaged duration of a reconstruction attack and the averaged LPIPS [9] scores of the training data reconstructed from the gradients in Fig. 3. The closer the points in the legend are to the origin, the better the overall performance of the algorithm. The figures illustrate that our proposed approach is effective in terms of both speed and the quality of the reconstructed images.

Table 1: The duration of different gradient inversion attack algorithms to complete a CIFAR-10 image reconstruction from the gradients of VB-ResNet18.

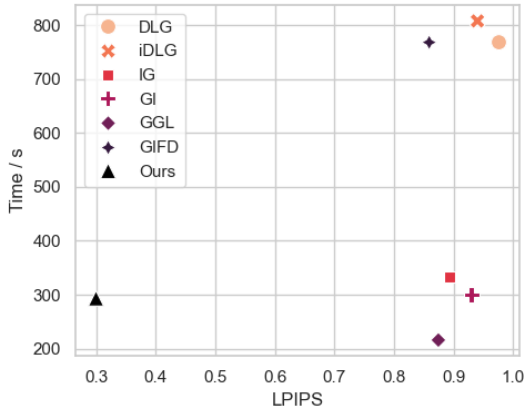
	DLG [11]	iDLG [10]	IG [3]	GI [8]	GGL [6]	GIFD [2]	Ours
Time/s	1642.21	1671.51	472.05	434.76	260.87	1011.52	489.04

Table 2: The duration of different gradient inversion attack algorithms to complete a CelebA image reconstruction from the gradients of BL-ConvNet.

	DLG [11]	iDLG [10]	IG [3]	GI [8]	GGL [6]	GIFD [2]	Ours
Time/s	767.98	807.49	332.90	299.75	215.96	767.80	293.19



(a)



(b)

Fig. 3: Illustration of the relationship between the duration of a single reconstruction attack and the LPIPS scores of the training data reconstructed from the gradients of (a) VB-ResNet18 and (b) BL-ConvNet, respectively

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=HyxQzBceg
2. Fang, H., Chen, B., Wang, X., Wang, Z., Xia, S.T.: Gifd: A generative gradient inversion method with feature domain optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4967–4976 (2023)
3. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients-how easy is it to break privacy in federated learning? Advances in Neural Information Processing Systems **33**, 16937–16947 (2020)
4. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR). San Diega, CA, USA (2015)
5. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), available at: http://arxiv.org/pdf/1312.6114v10
6. Li, Z., Zhang, J., Liu, L., Liu, J.: Auditing privacy defenses in federated learning via generative gradient leakage. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10132–10142 (2022)
7. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical Programming **45**, 503–528 (1989), https://api.semanticscholar.org/CorpusID:5681609
8. Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See through gradients: Image batch recovery via gradinversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16337–16346 (2021)
9. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
10. Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)
11. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Advances in neural information processing systems **32** (2019)