Improving Cost Function For iDLG

Yixuan Zeng, Yifan Wang
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
zengy5@rpi.edu, wangy77@rpi.edu

Abstract

Efficiency in neural network operations is pivotal for practical deployments, especially when dealing with large-scale data and complex models. Traditional techniques for reconstructing inputs from network gradients, while effective, often incur substantial computational costs. This study proposes a novel method integrating cosine similarity and gradient matching with ground truth label extraction, aiming to refine the reconstruction process and, consequently, enhance efficiency. By aligning gradient vectors' directions and magnitudes more precisely, we hypothesize that the proposed method will reduce the iterations needed for accurate reconstruction. Furthermore, by leveraging known label information, the method is designed to converge on the correct input representation faster than existing approaches. Preliminary experiments suggest that this integrated approach not only maintains the fidelity of reconstructions but also significantly reduces computational overhead, marking a promising advancement in the field of neural network interpretability and security. This report details the methodology, experimental setup, and comparative analysis against baseline models, highlighting the efficiency gains of the proposed method.

1 Introduction

The fundamental assumption in distributed learning frameworks, like Collaborative Learning and Federated Learning, is that privacy can be preserved through the sharing of gradients rather than raw data. While it was initially believed that sharing gradients would disclose minimal information about a client's private data, innovative attacks have demonstrated the capability to reverse engineer sensitive data from these gradients, thereby posing a significant threat to privacy. Recent research [Zhu et al., 2019; Zhao et al., 2020; Geiping et al., 2020] has introduced the concept of a "gradient inversion attack." This technique allows an attacker, who is eavesdropping on a client's communications with a server, to begin reconstructing the client's private data. We proposed a unified framework that combines cosine similarity for gradient matching with an effective ground truth label extraction technique, aiming to enhance both the fidelity and security of reconstructions.

Cosine similarity offers a measure of orientation alignment between two non-zero vectors, making it an apt choice for refining loss functions that compare gradient vectors. By integrating cosine similarity into the loss function, our methodology aims to more accurately align the dummy data gradients with the shared gradients, thereby improving the quality of the reconstructed data.

Concurrently, the extraction of ground truth labels from the gradients enables a more precise selection of inputs. By identifying the labels directly from the gradient information, we could ensure that the inputs selected for model updates are those that contribute to a more accurate and truthful representation of the data distribution.

Preprint. Under review.

The ensuing sections of this report detail the theoretical underpinnings of our integrated approach, outline the experimental setup, present our findings, and discuss the broader implications of our method on the privacy and also, efficiency of distributed learning systems.

2 Methodology

[Zhao et al., 2020] proposed an enhanced version of the DLG by incorporating a label prediction process using the Euclidean cost function. [Geiping et al., 2020] suggested that employing a cost function based on cosine similarity could expedite convergence. While [Geiping et al., 2020] have conducted experiments showing that their approach does accelerate convergence, their setup differed significantly and was based on several substantial assumptions. First, private labels are generally not disclosed in Federated Learning; however, knowing these labels could enhance the effectiveness of an attack. [Zhao et al., 2020] demonstrated that the label information for a single private image could be deduced from its gradient. Building on this, [Geiping et al., 2020] postulate that an attacker has knowledge of private labels, as noted in the remark at the end of Section 4 in their paper. Nonetheless, this assumption might not be valid in scenarios where multiple images in a batch share the same label.

We also observe that [Geiping et al., 2020] implemented their method under the assumption that BatchNorm statistics (i.e., mean(x^*), var(x^*)) of the private batch are provided alongside the gradient. Possessing these BatchNorm statistics allows the attacker to apply the same batch normalization that was used in the private batch to their reconstructed batch, resulting in improved reconstruction accuracy. Consequently, their results do not offer a direct comparison and it is unclear how much the observed speedup is attributable to the new cost function. Addressing this concern, our approach combines the strengths of cosine similarity with ground truth label extraction to enhance the loss function used in the reconstruction of inputs from gradients.

The proposed method combines and involves two main components: Extracting ground truth labels before the reconstruction process and optimizing gradient using a cost function based on cosine similarity.

2.1 Cosine Similarity for Gradient Matching

The cosine similarity between the gradients of the actual input x and the reconstructed input x' is maximized by defining the following loss function:

$$l(x, x') = 1 - \frac{\langle \nabla_\theta L_\theta(x, y), \nabla_\theta L_\theta(x', y') \rangle}{\|\nabla_\theta L_\theta(x, y)\| \|\nabla_\theta L_\theta(x', y')\|} \tag{1}$$

where L_θ is the neural network loss function parameterized by θ, and y and y' are the true and reconstructed labels, respectively.

2.2 Ground Truth Label Extraction

The ground truth label y is determined by analyzing the gradients corresponding to each class output. Specifically, the label is identified as the class index i for which the gradient with respect to the output y_i is most negative, indicating the greatest decrease in loss with respect to an increase in that class's output. This is mathematically defined as:

$$y = \arg\min_i \left\{ g_i : g_i = \frac{\partial L}{\partial y_i} \text{ and } g_i \text{ is the most negative} \right\} \tag{2}$$

2.3 Integrated Optimization Procedure

The reconstruction of x' and y' is performed by jointly minimizing the combined loss:

$$\min_{x', y'} \left(\alpha \cdot l(x, x') + \beta \cdot l(y, y') \right) \tag{3}$$

where α and β balance the importance of gradient similarity and label accuracy.

Algorithm 1 Enhanced iDLG with Cosine Similarity

Input: $F(x;W)$: Differentiable learning model, W: Model parameters, ∇W: Gradients produced by private training datum (x, c), N: maximum number of iterations, η: learning rate.
Output: (x', c'): Dummy datum and label.
Initialize: $x' \sim \mathcal{N}(0, 1)$
$c' \leftarrow i$ s.t. $\nabla W_i^T \cdot \nabla W_j \leq 0, \forall j \neq i$ ▷ Extract the ground-truth label
for $i = 1$ **to** N **do**
  $\nabla W' \leftarrow \partial_x F(x'; W, c')$ ▷ Calculate the dummy gradients
  $L_G \leftarrow \|\nabla W' - \nabla W\|_2^2$ ▷ Calculate the Euclidean loss
  $L_C \leftarrow 1 - \frac{\langle \nabla W', \nabla W \rangle}{\|\nabla W'\| \|\nabla W\|}$ ▷ Calculate the cosine similarity loss
  $x' \leftarrow x' - \eta(\nabla_x L_G + \nabla_x L_C)$ ▷ Update the dummy datum
end for

3 Empirical Results

In this section, we examined the empirical performance of the proposed algorithms and the experimental setup that yielded these results. We compared the performance of the traditional DLG and iDLG with our proposed method to ensure fair comparisons. All three methods were configured identically, using the LeNet model, the CIFAR-100 dataset, a learning rate of 0.1, 300 iterations, and the LBFGS optimizer. Our code can be found at: *https://github.com/wyfbw07/iDLG-with-cosine-similarity*

We began by running the traditional DLG and iDLG methods, with the results presented in Figures 1 and 2. We observed that the random initialization of the inputs and labels introduced variability in the convergence process, particularly evident in the significantly different initial losses. To facilitate a more straightforward comparison, we selectively chose two sets of results that exhibit similar initial losses. Generally, both methods exhibit comparable convergence rates, though iDLG tends to converge slightly faster. However, during our experiments, we noted that the traditional DLG method has a higher failure rate and often requires multiple restarts of the reconstruction process.

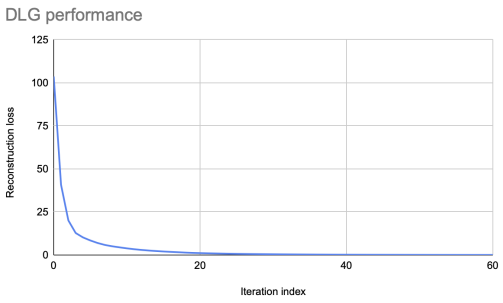

Figure 1: DLG performance

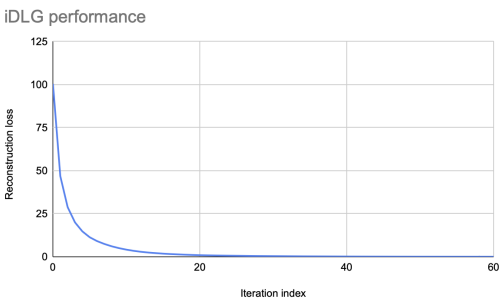

Figure 2: iDLG performance

We continued by assessing whether incorporating a cost function based on cosine similarity enhances the convergence speed using our method. As shown in Figure 3, by including both predicted labels and a cosine similarity-based cost function, our method achieved a significantly faster convergence rate, as indicated by a steeper curve. We also compared the Mean Squared Error (MSE) across the different methods. The MSE measures the cumulative squared error between the compressed and the original image. In our comparisons, iDLG consistently outperforms DLG, beginning to yield meaningful reconstructions earlier. Our enhanced iDLG model, although initially struggling, converges to zero more rapidly and ultimately surpasses the other two methods after approximate 45 iterations in performance. It is also the first to produce meaningful reconstruction results.

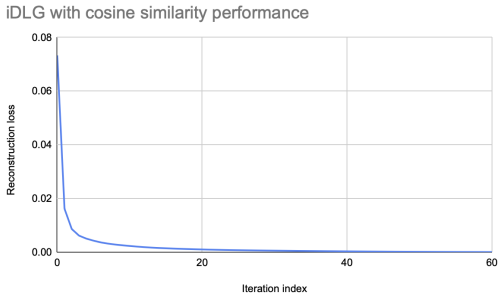

Figure 3: iDLG with cosine similarity

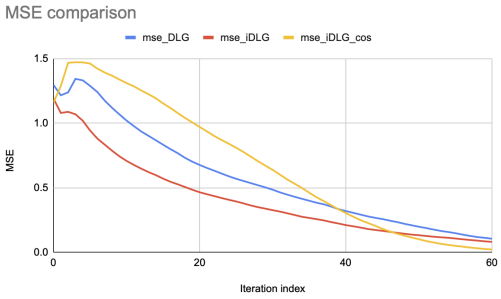

Figure 4: Reconstruction MSE comparison

4 Conclusion and Future work

This paper, while not unveiling ground-breaking discoveries, it reflects our commitment to understanding the full scope of academic writing through a comprehensive project that goes beyond basic theory and experimentation. We aim to maintain a high standard of quality, which is particularly essential as first-time undergraduate students. Looking ahead, we intend to expand our testing to include more models and datasets, with an emphasis on higher resolution images. Due to time constraints, we were unable to assess performance in a distributed setting in this round, but we plan to address this in future work. Additionally, we noted that all three methods experienced initial loss surges; we aim to further investigate these occurrences to better understand and address the issues

References

[1] Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020, September 11). *Inverting gradients – How easy is it to break privacy in Federated Learning?*. arXiv.org. Retrieved from `https://arxiv.org/abs/2003.14053`.

[2] Zhao, B., Mopuri, K. R., & Bilen, H. (2020, January 8). *IDLG: Improved deep leakage from gradients*. arXiv.org. Retrieved from `https://arxiv.org/abs/2001.02610`.

[3] Ligeng Zhu, Zhijian Liu, & Song Han. (2019, December 19). *Deep Leakage from Gradients*. arXiv.org. Retrieved from `https://arxiv.org/abs/1906.08935`.

Appendix

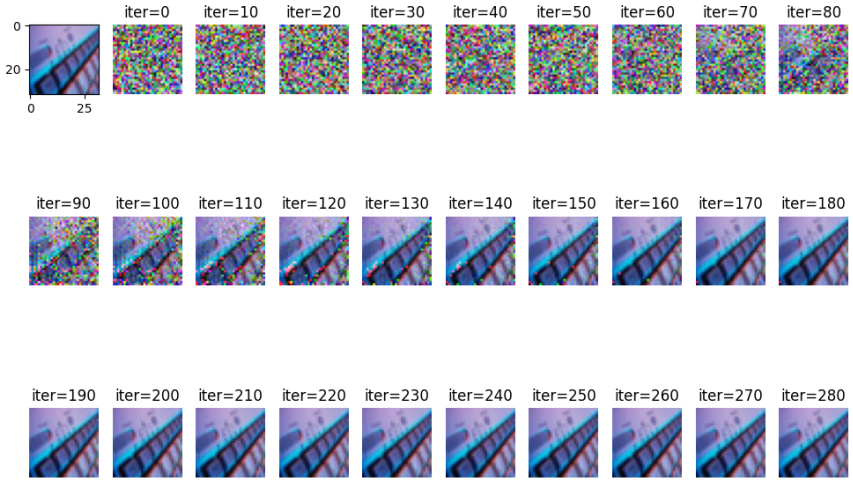

Figure 5: DLG reconstruction

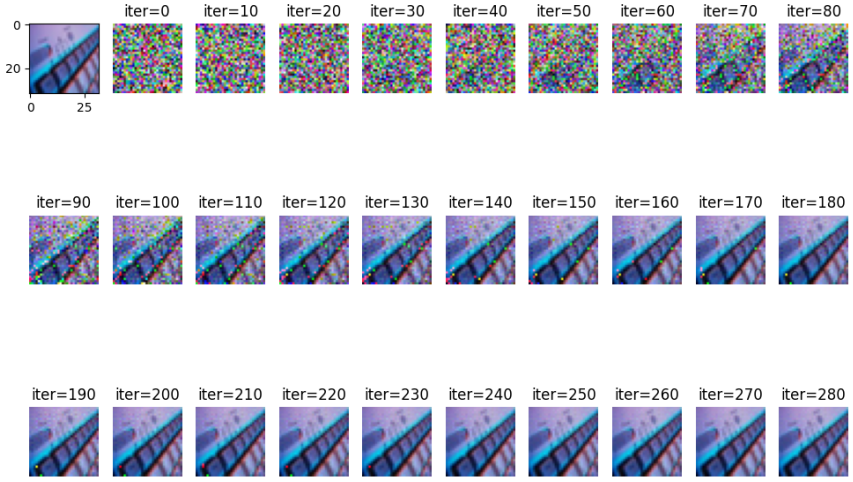

Figure 6: iDLG reconstruction

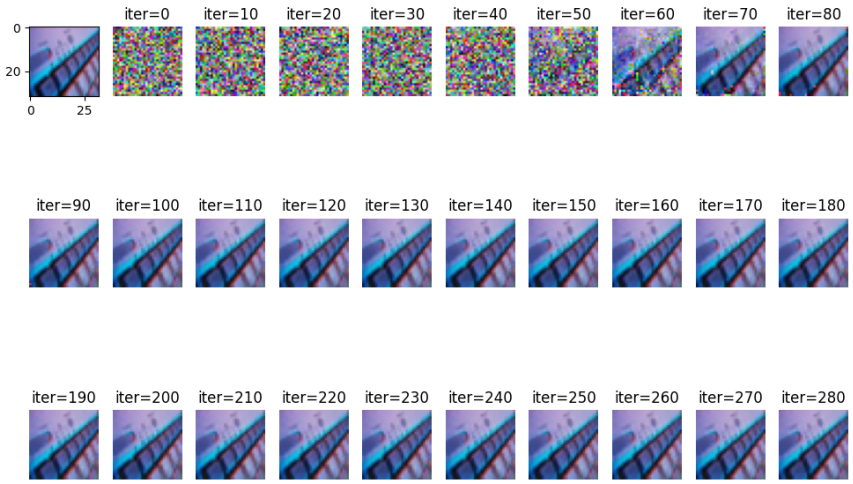

Figure 7: iDLG with cosine similarity reconstruction