CS 6220 - Big Data Systems and Analytics

Georgia Institute of Technology
Assignment 3

Saurab Sirpurkar

October 10, 2021

A critique on
Inverting Gradients - How easy is it to break privacy in federated learning?

Motivation

Federated learning has multiple advantages such as training efficiency, increased training data availability and participant privacy. But most of the privacy guarantees are based on assumptions such as "Privacy is enhanced by the ephemeral and focused nature of the Federated Learning updates." [Bon+19] or just based on that fact that it is better than sharing data for training. However, this is not true. There have been studies that observe attacks on privacy such as [Pho+17], [Mel+19], [ZH20] in federated settings. But in most cases they either consider simple neural networks with insufficient depth and smooth activations or have unrealistic limits on the batch size. Despite their success in recreating the training data from participants' gradients [ZH20], there is a scope to improve the attacks on more realistic federated learning settings.

This paper looks into whether such attacks can be recreated in more realistic settings such as using larger batch size and multi-epoch averaging in deeper networks. It is crucial for the participants involved in Federated Learning to understand the possible reconstruction efficacy of such attacks, as the scenarios often includes medical centers, personal mobiles etc. all of which have private data and aim to avoid any reconstruction risk.

Contributions

1. Inspects more realistic architectures like deeper networks (ResNet-152), with higher batch sizes (upto N=100) and with variable widths (ResNet-18|w=128), mostly dealing with the Computer Vision domain - a common application domain of Federated Learning.
2. Launches successful attacks in such realistic scenarios by reconstruction training samples from CIFAR-10, ImageNet etc. for different hyperparameter choices (explained in summary). Most of these recreated attacks look very similar to the ground truth images, and in cases when it is not, a human can easily interpret the key details and make inferences about the supposed original image such as the main subject in the image.
3. Provides a novel cost function and argue, with observations, why it captures the similarity between attacker's current gradient and the target gradient more accurately. They also propose a better alternative to optimise such loss functions based on Adam.
4. Proves that a fully-connected layer in a neural network can lead to accurate input reconstruction if the attackers get access to its gradient. Extends the same to a stack of fully connected layers and shows that the inputs to this stack can be reconstructed if all intermediate gradients are known and are non-zero.

Summary of Methods

The most important contribution of the paper is the Numerical Reconstruction Method of training images and its applications for Single and Multi-Image Recovery. First, by experimenting with the gradients from DNNs in different stages of training, the authors infer that the magnitude of a gradient has more to do with the stage of the training and less to do with input training sample's features. Instead they argue that the direction captures more

information about the original image's effect on model updates. Therefore, for their loss function, they propose the use of cosine similarity to compare attacker's gradient and the target gradient, independent of their magnitudes. They also propose a better way to optimize this loss function using only signs of the gradient along with Adam optimization.

In case of single image reconstruction, i.e. B = 1 and gradients correspond to one particular image, by running 24000 iterations, this method outperforms others in reconstruction task on untrained ResNets and comes far ahead of previous attacks in case of trained ResNets. In case of multiple image reconstruction, while there are studies that assume $B \leq 8$ [ZH20], this paper considers a very large B = 100 on CIFAR-100 and shows that the attack efficacy has a high variance across classes. However, for the most vulnerable classes, they come very close to the original images, such that a human could guess the subject of the image.

This paper considers many other aspects of gradient leakage attacks such as trained vs untrained networks, invariant CNNs, effect of Depth and Width of a DNN, variable number of epochs etc. They claim that their attack performs better than others on trained networks as their loss is magnitude independent. Also, they observe that paddings in inputs make their attack reconstruction less effective and hence recommend their usage. In respect to the width of the network, they observe that higher width leads to higher variance in the accuracy of reconstruction. Therefore, given multiple attempts, wider networks are more susceptible to such attacks. They also show that averaging gradient updates across multiple epochs does not deter the attacker either. Also, depth hardly matters from their observations - they successfully reconstruct training images that trained a ResNet-152.

Strengths of the Paper

1. It claims to be the first paper to launch successful recreation attacks on industrially realistic DNNs in the computer vision domain. This has far reaching implications as many federated learning settings clearly value privacy that now comes under question. The paper clearly establishes this by showing high PSNR attacks on multiple datasets like the CIFAR-10, CIFAR-100 and ImageNet using realistic ResNets. Given that [ZH20] has seen success across both NLP and CV with a single attack, there is scope for more studies handling the same attack in NLP domain. Therefore, all Federated Learning systems must be careful of these impending attacks given that most of these attacks have open source codes such as this paper's code, DLG and CPL.

2. Establishes the legitimacy of attacks by honestly reporting the variance of attack efficiency across different settings as seen in the summary above. Therefore, its message for Federated Learning setups to reconsider their defenses comes across as authentic. For example, Turbo-aggregate [SGA21] one of the SOTA secure aggregation techniques for Federated Learning talks about this attack and focuses on prevention of such gradient attacks. This, combined with its exhaustive study on hyper-parameter and architectural effects, makes it a strong paper for practical understanding of these attacks in real-life settings compared to its peers such as [ZH20], [Wan+19], [ZMB20] which consider multiple variations of their approach but in idealistic settings such as limited batch sizes or shallower models.

3. Clearly defines the threat model and makes an original theoretical contribution to optimize attackers estimate of the data by maximizing gradient similarity between local gradient and captured gradient. Here, they propose that the gradient direction matters more than its magnitude, and prove it on actual DNNs. This new loss function has a lot of scope to be used in upcoming SOTA attacks (already looked at in [Yin+21] and [Pan+20]). This singular change (and the proposed Adam-based optimizer) changed the scope of gradient attacks from few-batch shallow networks to mini-batch deep CNNs and improved their PSNR.

Weaknesses of the Paper

1. Towards the beginning of the paper, there is a proof that shows that inputs to a fully connected layer can be entirely reconstructed using its gradient information. Although interesting, this feels out of place with the rest of paper. For example, the immediate next section starts by saying we would not consider in the first layers to be FC which breaks the main assumption in this proof. Also, it is does not serve as a precursor to any other technique in the following sections of the paper. This slows down the flow of the paper. Instead, it could be added as an in-line proof in the third section followed by its applications to reconstruct labels, penultimate layers' outputs and intermediate visual representations [DB16] that are vulnerable to reconstruction.

2. Despite mentioning that they assume industrially realistic settings, their focus on attacking using a single-image gradient on single-batch-size networks is disproportionately high. For example, after all the demonstrations on single-image and small batch networks (something the previous studies have done successfully), they only allocate a single experiment to study the attack on CIFAR-100 (B=100) where they have questionable reconstruction. Since it differentiates itself from previous attacks following the principle of realism, having small batch sizes, that too for image classification task is not very realistic. Therefore, its direct application to attack real Federated Learning systems left room for more work. One recent paper dealing with this issue is [Yin+21] also cites this paper mentioning the same issue and its mitigation.

3. At places, I felt like their focus on comparing with other papers (both discriminative and associative) eclipsed the main point they are trying to communicate. For example, when introducing their new loss function, they reconsider the previous work and explicitly argue that they are all suboptimal. This comparison affects the next sentences and delays their delivery without adding anything significant. Instead, they could have focused more on their original contribution that could speak for itself by demonstrating it on more datasets and tasks with other standard architectures. For example, a more thorough study is [QH20] which takes larger batch size and inspects their work on 3 different tasks across 6 datasets.

   Also, in related work, they talk about other similar tasks such as input attribute inferences, model inversion and inverting visual representations. Here, they try to associate with the closest related tasks. This does not contribute much and takes away from the overall flow of the paper.

4. Similar to last weakness, when establishing the accuracy of 'Proposed Loss' function compared to 'Euclidean' losses, the paper forgets to include ground truth images, but instead focuses on relative performance against another model. This confuses the reader and prevents them from understanding the attack's accuracy in absolute terms. Also, during such comparisons, they give themselves a generous advantage of 16 restarts by claiming "This approach can often fail due to a bad initialization.." without proper theoretical justification on why that is the case. Even in case this is justified, other attacks must also receive 16 restarts, getting to choose their best attack among these attempts.

References

[DB16] Alexey Dosovitskiy and Thomas Brox. "Inverting visual representations with convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4829–4837.

[Pho+17] Le Trieu Phong et al. "Privacy-preserving deep learning: Revisited and enhanced". In: *International Conference on Applications and Techniques in Information Security*. Springer. 2017, pp. 100–110.

[Bon+19] Keith Bonawitz et al. "Towards federated learning at scale: System design". In: *arXiv preprint arXiv:1902.01046* (2019).

[Mel+19] Luca Melis et al. "Exploiting unintended feature leakage in collaborative learning". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 691–706.

[Wan+19] Zhibo Wang et al. "Beyond inferring class representatives: User-level privacy leakage from federated learning". In: *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE. 2019, pp. 2512–2520.

[Pan+20] Xudong Pan et al. "Theory-Oriented Deep Leakage from Gradients via Linear Equation Solver". In: *arXiv preprint arXiv:2010.13356* (2020).

[QH20] Jia Qian and Lars Kai Hansen. "What can we learn from gradients?" In: (2020).

[ZMB20] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. "idlg: Improved deep leakage from gradients". In: *arXiv preprint arXiv:2001.02610* (2020).

[ZH20] Ligeng Zhu and Song Han. "Deep leakage from gradients". In: *Federated learning*. Springer, 2020, pp. 17–31.

[SGA21] Jinhyun So, Başak Güler, and A. Salman Avestimehr. "Turbo-Aggregate: Breaking the Quadratic Aggregation Barrier in Secure Federated Learning". In: *IEEE Journal on Selected Areas in Information Theory* 2.1 (2021), pp. 479–489. DOI: 10.1109/JSAIT.2021.3054610.

[Yin+21] Hongxu Yin et al. "See through Gradients: Image Batch Recovery via GradInversion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16337–16346.