CS 6220 - Big Data Systems and Analytics

Georgia Institute of Technology
Assignment 3

Saurab Sirpurkar

October 12, 2021

A critique on Deep Leakage from Gradients

Motivation

When training in a federated or distributed fashion, the involved parties share the locally computed gradients against a global model, and this was believed to be safe. Sure, it is safer than sharing the data to a centralised node that trains a model. But does it mean it is really safe from all kinds of attacks? Once piece of information that an attacker could get it the gradients they share among themselves. Do these gradients reveal any information about the training data?

Previous works have shown that there is indeed some data leakage in these shared gradients. But they neither establish nor rule out the presence of enough leakage to obtain the full training dataset. For example, [Mel+19a] recover partial properties and [HAP17] generate similar samples as in training data. However, both of them assume some prior knowledge about the dataset and its labels. Unlike them, this paper examines the vulnerability that comes from sharing gradients by attempting to reconstruct entire private datasets solely from gradients.

Contributions

1. Stands as the first algorithm to obtain actual training data from the shared gradients. This is an eye-opener on the security vulnerability of Federated Learning to such attacks. Previous attacks could only infer partial properties using gradient attacks.
2. Demonstrates the reconstruction attack in both Computer Vision and NLP domains with clear examples of how the data is revealed in both parameter server based as well as decentralized training architectures with mini-batch sizes upto 8. They achieve close to perfect reconstruction in almost all of their tasks of choice including attacks against a ResNet-20 training on CIFAR with B=8.
3. Demonstrates three defense strategies and their effectiveness against this attack - gradient perturbation, low precision and gradient compression.

Summary of Methods

The core of the paper revolves around the Deep Leakage from Gradients (DLG) algorithm and its modification for non-singular batch-size models. Coming to the details of the algorithm, they first obtain a single leaked gradient $\Delta W_{t,k}$ from one of the participants (k) who is the target. This could be in the form of a honest-but-curious participant within the system in case of decentralized learning or a due to a parameter server attack in case of centralized learning. Next, they initiate dummy inputs and labels and run an iteration on the model to get reconstruction gradient. $\Delta W_i'$. Now, instead of optimizing against a loss, they optimize the input such that this gradient looks more like (using MSE) $\Delta W_{t,k}$ and use the resulting gradient $\frac{\partial ||\Delta W' - \Delta W||^2}{\partial X}$ to make changes to the input. Slowly, the input get closer to the target training image. Look at the figure below for a better understanding of this initialization and reconstruction:

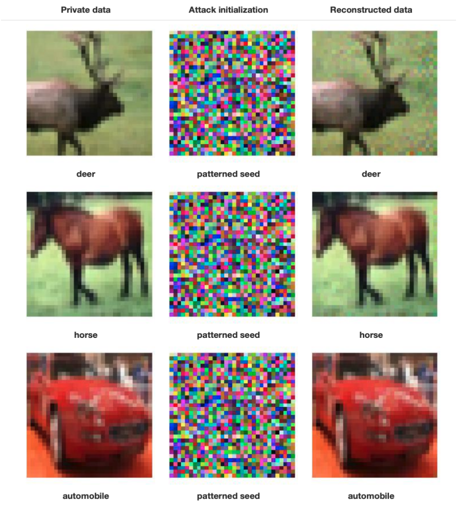

Figure 1: Deep Leakage from Gradients - Demonstration on the CIFAR 10 w. Lenet with 300 iterations
Source: ai-privacy-vlab applet, Distributed Data Intensive Systems Lab, Georgia Tech

The next important aspect of this paper is its demonstration of this attack algorithm on two tasks. First, they consider the image classification task on a modern ResNet-56 against the MNIST, CIFAR-100, LFW and SVHN datasets. After training for a few hundreds of iterations, they show how their attack clearly outperforms other attacks like [Mel+19b] with respect to MSE from the original training images. In fact, the previous attacks were not even aiming to get to accurate reconstruction as they had a different model of attack. The strengths of this paper discuss this aspect more elaborately below.

Next, they move on to BERT-MLM trained on masked-words (15%) prediction. Here, they choose to reconstruct a NeurIPS conference announcement page. The randomly initialized gibberish turns close to token-level-perfectly accurate text within a few dozens of iterations.

So far, they are assuming only a single pair of input and label per batch. Next, they move on to performing the same attack on batched data instead. Here, they simply move on to working the update corresponding to each N'th input data. Although more iterations are required, very clean resconstructions were shown for upto B = 8 on the CIFAR dataset.

Finally, they study the efficiency of the defenses against the DLG attack. They find that noisy gradients cannot meaningfully stop this attack without losing out significant accuracy. Similarly, low precision perturbations and integer quantizations do not stop the attack or lose out too much on accuracy. Next they observe that Gradient Compression techniques. Here they find that pruning ratios as low at 20% can completely stop the convergence of this attack. Finally, they also propose that a larger batch size (>8) and higher resolutions (>64×64) for images can prevent this attack.

Strengths of the Paper

1. This paper formally defines an attack as follows: given a model F(), weights W, can we obtain training data (x,y) by just looking at the Δw's of a given image? This is the purest form of an adversarial attack, and can be considered as the root attack that subsumes other forms of attacks. For example, if the attack reveals the full training dataset, we can always infer the properties on this data (like the majority background color in the images) using the reconstructed data. Hence, it paves the way for future attacks to focus on this attack formality rather that attacking weaker problems. Therefore a strong point about this paper is that the authors attacked the core of privacy directly and it is surprising that they achieved such good results despite being the first paper to work with this stronger attack formality.

2. It clearly establishes the attack and its limitations, and is well received by its successors for this reason. For example, they demonstrate this attack on a multiple datasets including MNIST, CIFAR-100, SVHN and LFW in Computer Vision and on the BERT-MLM model against a conference page in NLP. They also clearly talk about their limitations when it comes to batch size. The future attacks like the [Yin+21], [ZMB20] all cite this paper, discuss these limitations and use it as a clear base and improve on it. However,

some questionable architectural assumptions are not elaborated enough as we see in the weaknesses.

3. Another strong point of this paper is that despite being the pioneer of such attacks, they make room for enough discussion on defenses. For example, they study the trade-off between accuracy and defendability for two noises with different variances and report their findings. This contrasts with some other papers on such attacks like [Gei+20] where such studies are essentially absent. Succeeding papers are clearly helped by such evaluations against defenses. For example, [Sun+20] uses this discussion and builds a defense based on it.

4. The attack does not require any prior knowledge about the training data and its labels unlike previous GAN-based methods like [HAP17] and other attacks [Mel+19b]. Therefore, this is a strictly stronger attack compared to previous works, both in terms of requiring less knowledge as well as its ability to completely reconstruct the training dataset. It can also attack a model that is not fully trained unlike other attacks.

Weaknesses of the Paper

1. One of the key issues I found with the paper is the over-simplistic representations. For example, the attack model was not very clear to me (more on that below). Next, when extending to non-singular batch size, it was confusing what they meant by taking the gradient of each n^{th} input data ($\nabla_{x_{t+1}^{imodN}}$ D). Does it mean that N independent images are trained in lock-step to arrive at N final images? In that case, by a total random choice, if two initial inputs are very similar giving rise to similar gradients, do they not turn into the same final image? Does it leave the attacker with only N-1 final images leaving them to retry randomly until they get the last image? Owing to such questions, I feel like this could have been explained better in the paper.

2. I found the description of attack formality to be a little misleading. First, they claim to reconstruct the entire dataset using only the publicly 'shared gradients'. However, in the diagram showing the threat model informally, the demon has the access to the parameter server in case of distributed training and full access to a node in case of decentralized training. This means that it also has access to that model against which the victim computes the gradient. This is also reflected in the algorithm that uses the model $F(x', W)$ to compute gradients against the current reconstruction x', the $\Delta W'$ as many times as it wishes. This clearly might not translate in real-life unless one of the nodes is hacked. That is, a mere eavesdropper of these gradients would not be able to replicate these attacks. Therefore, a more formal description of the attack would help in reader's understanding.

3. Even though the paper admits to making a handful of assumptions, some of them are notable in terms of applying it in real-life scenarios. For example [Gei+20] goes in depth about their mini-batch size assumption being too 'idealistic' and how it fails to reconstrucrt on very-deep NNs. Similarly, [ZMB20] proposes iDLG and elaborates on the weakness of DLG in recovering the labels accurately in harder datasets. According to their experiments, it the accuracy drops to 79% on LFW and it gets worse as the dataset complexities increase. But this is not talked about in this paper showing that the selection of victim models and datasets was too specific at times without enough justification.

4. The final issue with the paper is that it overstates the importance of its findings in its concluding remarks. It blatantly claims that deep leakage can be prevented only when defenses start degrading accuracy. But this clearly contradicts their findings where gradient compression makes it impossible for the attack to converge without significant accuracy loss. The paper goes on to recommend multi-node systems to rethink the safety of sharing scheme based on this erroneous claim. This is a bad recommendation due to presence of a good defense within the paper and due to its batch-size assumptions which may not translate to most of the real-life scenarios. Instead, they should recommend these systems to consider gradient compression strongly and warn them of imminent attacks in future that can be more applicable to real-life settings.

References

[HAP17] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. "Deep models under the GAN: information leakage from collaborative deep learning". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.* 2017, pp. 603–618.

[Mel+19a] Luca Melis et al. “Exploiting Unintended Feature Leakage in Collaborative Learning”. In: *2019 IEEE Symposium on Security and Privacy (SP)*. 2019, pp. 691–706. DOI: 10.1109/SP.2019.00029.

[Mel+19b] 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.

[Gei+20] Jonas Geiping et al. “Inverting Gradients–How easy is it to break privacy in federated learning?” In: *arXiv preprint arXiv:2003.14053* (2020).

[Sun+20] Jingwei Sun et al. “Provable Defense against Privacy Leakage in Federated Learning from Representation Perspective”. In: *arXiv preprint arXiv:2012.06043* (2020).

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

[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.