

Deep Leakage from Gradients

Ligeng Zhu, Zhijian Liu, Song Han
Massachusetts Institute of Technology.

Publishing info:
NeurIPS 2019

Presenter:

Xin Fan

2020/10/07

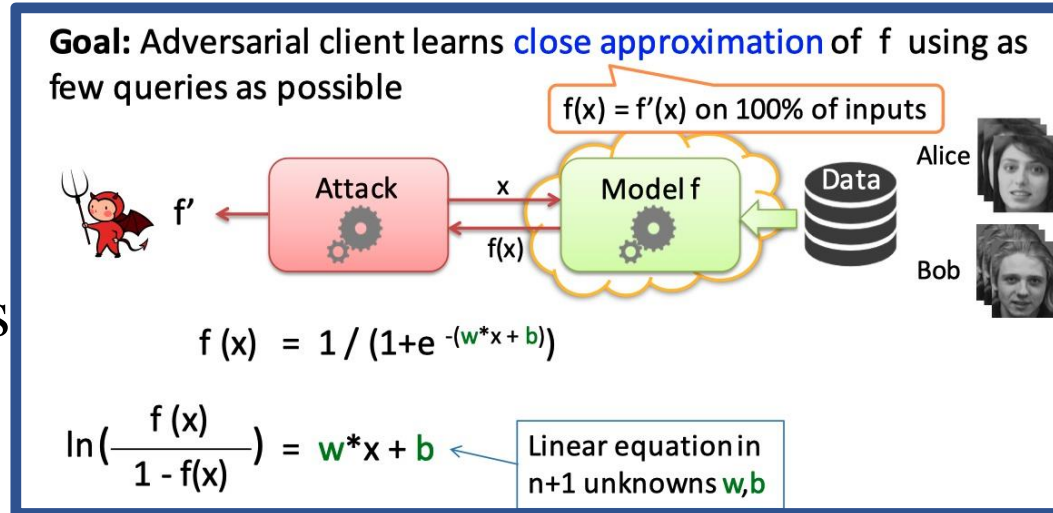
Outline

- ◆ Related works and motivation
- ◆ Deep leakage scenarios
- ◆ Methods
- ◆ Experiments
- ◆ Defense strategies
- ◆ Conclusion
- ◆ Key reference

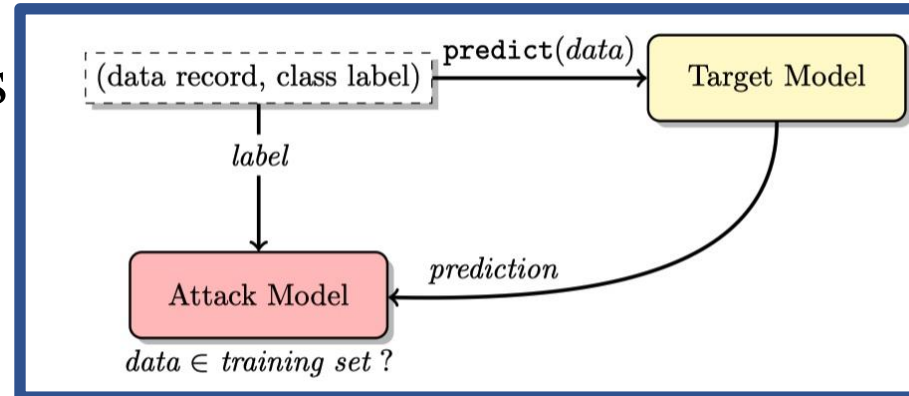
Related works and motivation

Privacy leakage

◆ Model Extraction Attacks



◆ Membership Inference Attacks



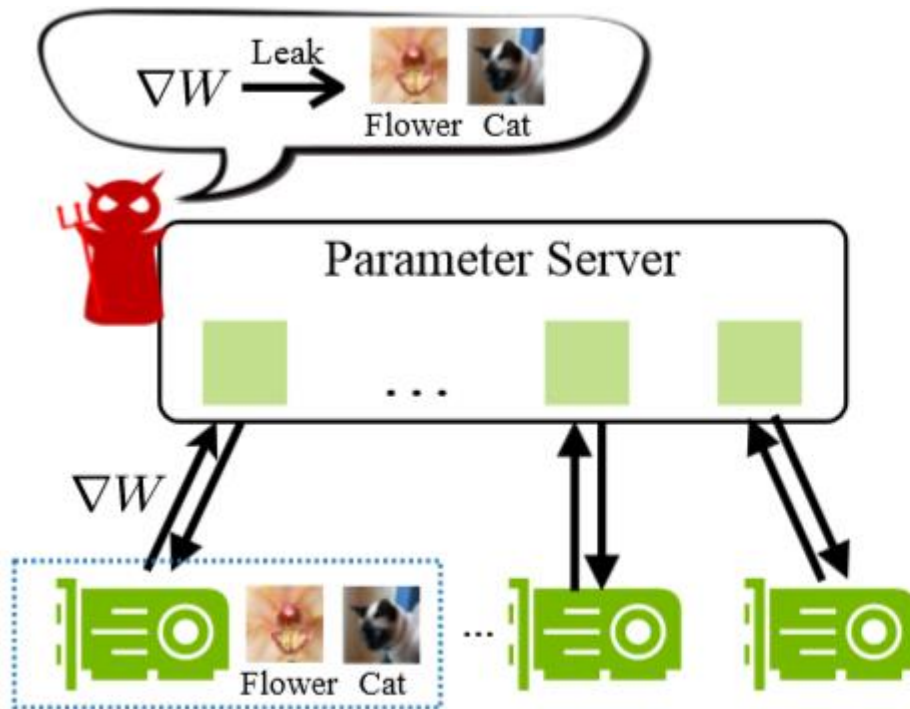
◆ Model Inversion Attacks

Related works and motivation

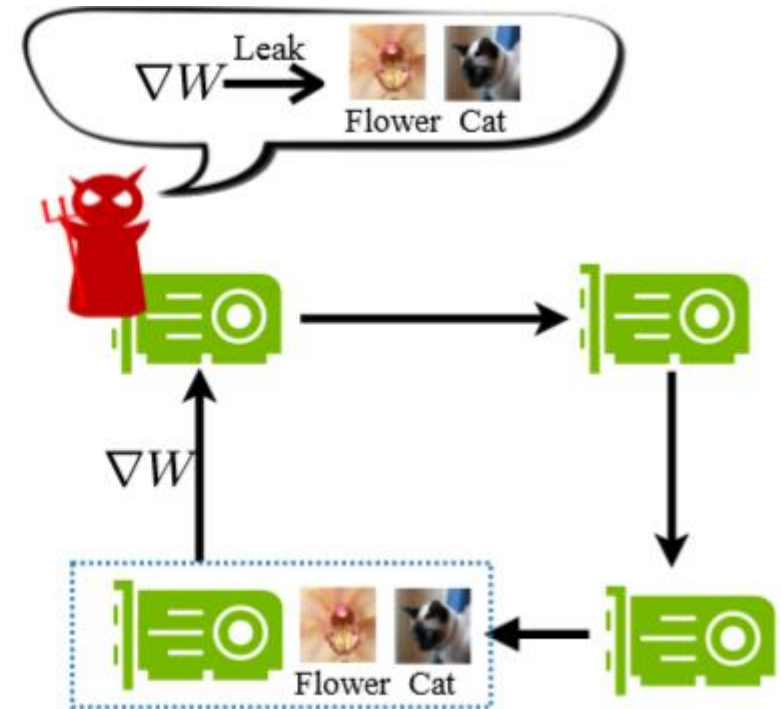
- Shallow:
 - Partial information
- Deep:
 - Get the original training dataset from gradients

Deep leakage scenarios

- With or without a centralized server



(a) Distributed training with a centralized server

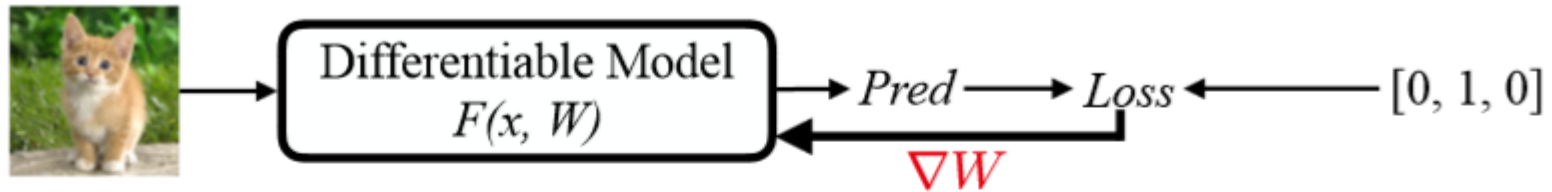



(b) Distributed training without a centralized server

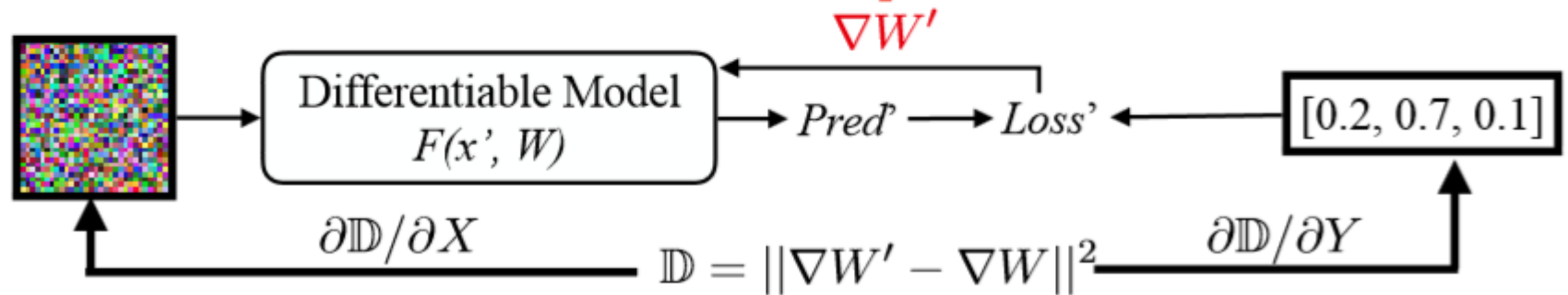
Methods

- Standard synchronous distributed training

Normal Participant



Malicious Attacker 



Methods

F is twice differentiable



$$\mathbf{x}'^*, \mathbf{y}'^* = \arg \min_{\mathbf{x}', \mathbf{y}'} \|\nabla W' - \nabla W\|^2 = \arg \min_{\mathbf{x}', \mathbf{y}'} \left\| \frac{\partial \ell(F(\mathbf{x}', W), \mathbf{y}')}{\partial W} - \nabla W \right\|^2$$

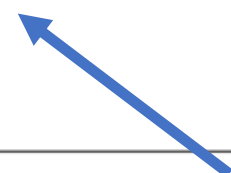
Algorithm 1 Deep Leakage from Gradients.

Input: $F(\mathbf{x}; W)$: Differentiable machine learning model; W : parameter weights; ∇W : gradients calculated by training data

Output: private training data \mathbf{x}, \mathbf{y}

```

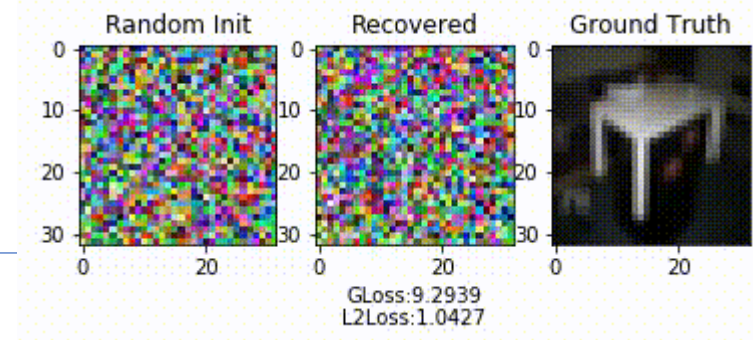
1: procedure DLG( $F, W, \nabla W$ )
2:    $\mathbf{x}'_1 \leftarrow \mathcal{N}(0, 1), \mathbf{y}'_1 \leftarrow \mathcal{N}(0, 1)$                                 ▷ Initialize dummy inputs and labels.
3:   for  $i \leftarrow 1$  to  $n$  do
4:      $\nabla W'_i \leftarrow \partial \ell(F(\mathbf{x}'_i, W_t), \mathbf{y}'_i) / \partial W_t$                                 ▷ Compute dummy gradients.
5:      $\mathbb{D}_i \leftarrow \|\nabla W'_i - \nabla W\|^2$ 
6:      $\mathbf{x}'_{i+1} \leftarrow \mathbf{x}'_i - \eta \nabla_{\mathbf{x}'_i} \mathbb{D}_i, \mathbf{y}'_{i+1} \leftarrow \mathbf{y}'_i - \eta \nabla_{\mathbf{y}'_i} \mathbb{D}_i$     ▷ Update data to match gradients.
7:   end for
8:   return  $\mathbf{x}'_{n+1}, \mathbf{y}'_{n+1}$ 
9: end procedure
  
```



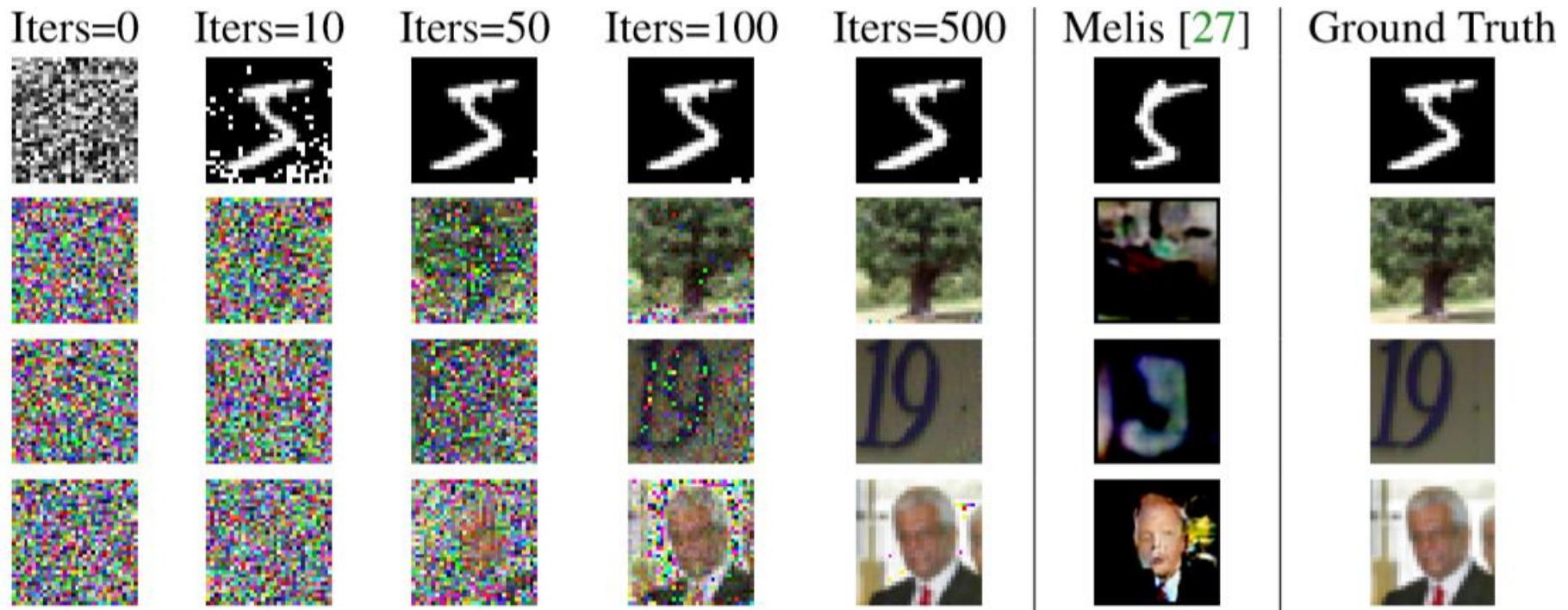
$$\begin{aligned} \mathbf{x}'_{t+1}^{i \bmod N} &\leftarrow \mathbf{x}'_t^{i \bmod N} - \nabla_{\mathbf{x}'_{t+1}^{i \bmod N}} \mathbb{D} \\ \mathbf{y}'_{t+1}^{i \bmod N} &\leftarrow \mathbf{y}'_t^{i \bmod N} - \nabla_{\mathbf{y}'_{t+1}^{i \bmod N}} \mathbb{D} \end{aligned}$$

Batched Data:

Experiments



The deep leakage on images from MNIST, CIFAR-100, SVHN and LFW, respectively



Experiments

- Deep leakage from batched data

	BS=1	BS=2	BS=4	BS=8
ResNet-20	270	602	1173	2711

Table 1: The iterations required for restore batched data on CIFAR [21] dataset.

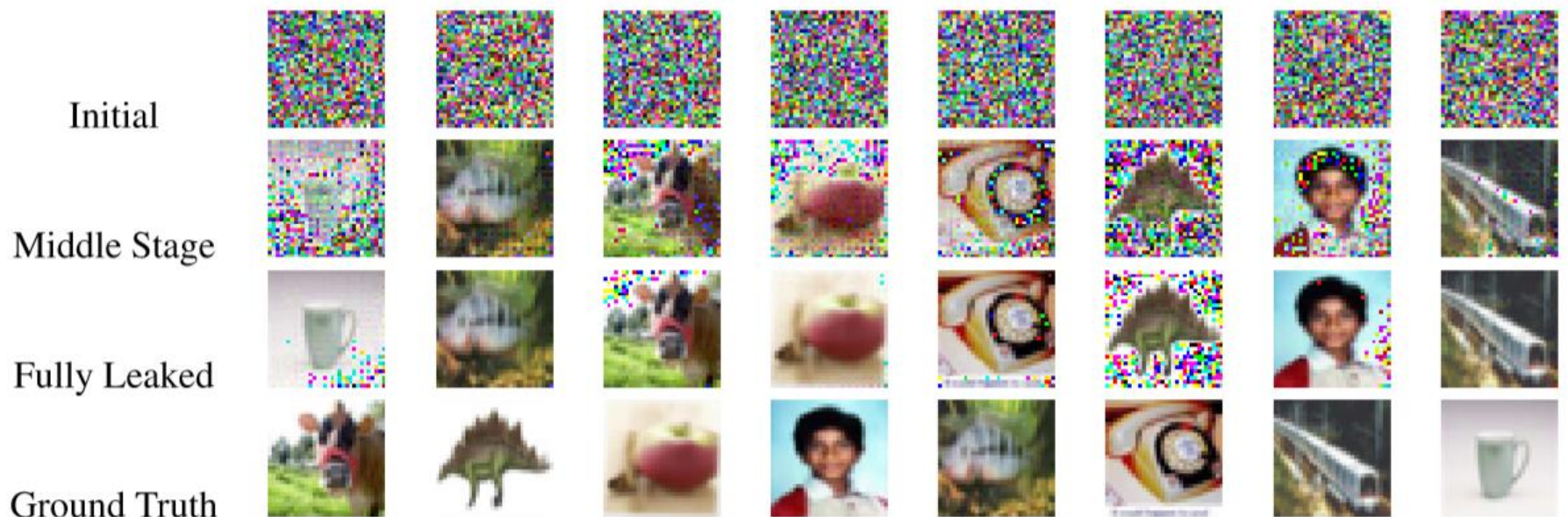


Figure 4: Results of deep leakage of batched data. Though the order may not be the same and there are more artifact pixels, DLG still produces images very close to the original ones.

Experiments

Deep Leakage on Masked Language Model

	Example 1	Example 2	Example 3
Initial Sentence	tilting fill given **less word **itude fine **nton over- heard living vegas **vac **vation *f forte **dis ce- rambycidae ellison **don yards marne **kali	toni **enting asbestos cut- ler km nail **oof **dation **ori righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled **wled major relief dive displaced **lice [CLS] us apps _ **face **bet
Iters = 10	tilting fill given **less full solicitor other ligue shrill living vegas rider treatment carry played sculptures life- long ellison net yards marne **kali	toni **enting asbestos cutter km nail undefeated **dation hole righteous **xie lucan **hot **ery at **tle ordered pa **eit smashing proto	[MASK] **ry toppled iden- tified major relief gin dive displaced **lice doll us apps _ **face space
Iters = 20	registration , volunteer ap- plications , at student travel application open the ; week of played ; child care will be glare .	we welcome proposals for tutor **ials on either core machine denver softly or topics of emerging impor- tance for machine learning .	one **ry toppled hold major ritual ' dive annual confer- ence days 1924 apps novel- ist dude space
Iters = 30	registration , volunteer ap- plications , and student travel application open the first week of september . child care will be available .	we welcome proposals for tutor **ials on either core machine learning topics or topics of emerging impor- tance for machine learning .	we invite submissions for the thirty - third annual con- ference on neural informa- tion processing systems .
Original Text	Registration, volunteer applications, and student travel application open the first week of September. Child care will be available.	We welcome proposals for tutorials on either core machine learning topics or topics of emerging importance for machine learning.	We invite submissions for the Thirty-Third Annual Conference on Neural Information Processing Systems.

Table 2: The progress of deep leakage on language tasks.

Defense strategies

- Noisy Gradients (**effective when the accuracy decreased**)
 - Gaussian noise
 - Laplacian noise
- Quantization
 - Float 16 (**noneffective**)
 - Int 8 (**effective when the accuracy decreased**)
- Gradient Compression and Sparsification (**most recommended**)
 - Gradients with small magnitudes are pruned to zero
- Large Batch, High Resolution and Cryptology (**discussed**)
 - DLG currently only works for batch size up to 8 and image resolution up to 64×64 .
 - Among all defenses, cryptology is the most secured one, but have their limitations and not general enough

Defense strategies

Experiments on noises and half precision

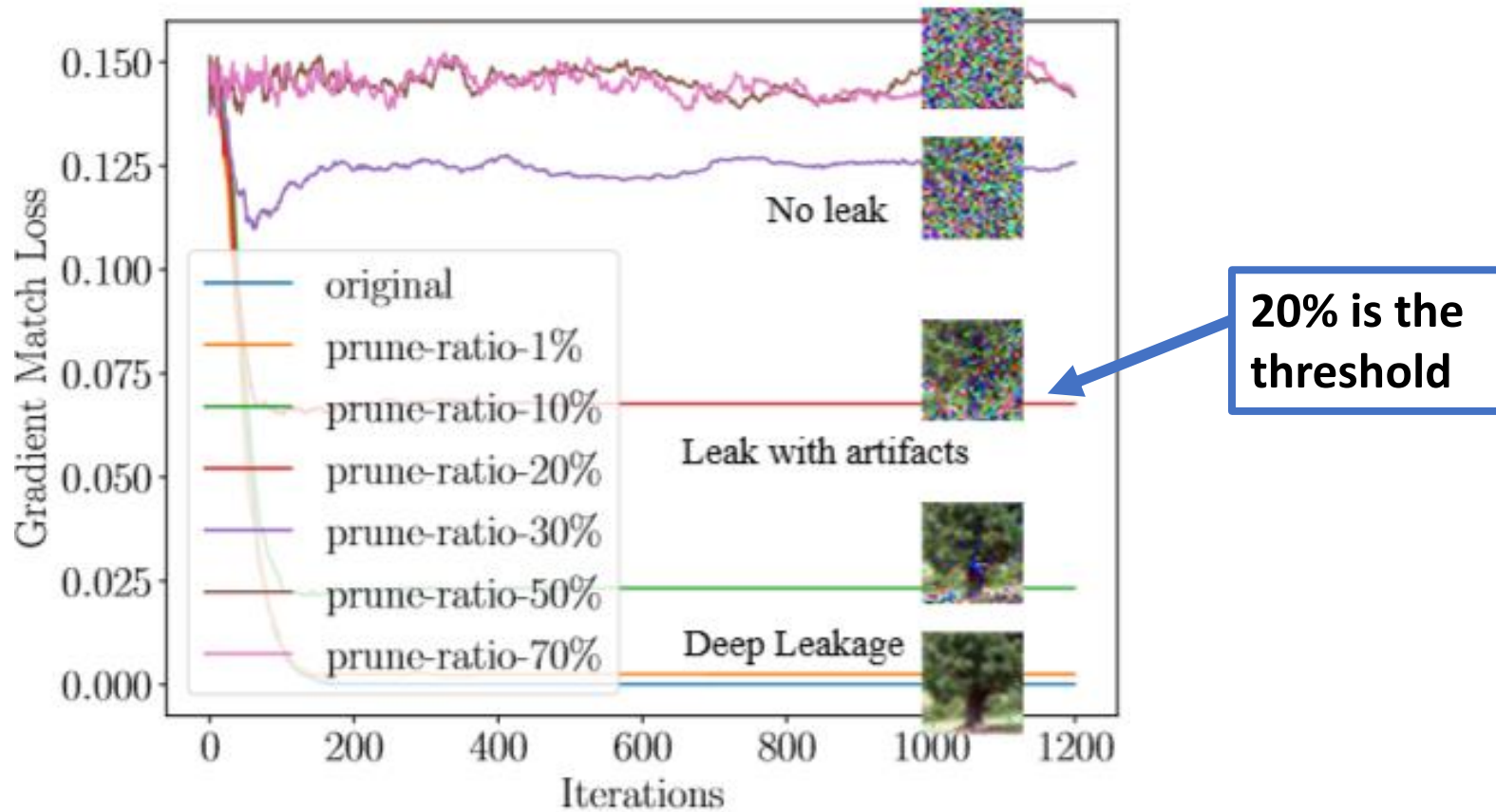
	Original	G - 10^{-4}	G - 10^{-3}	G - 10^{-2}	G - 10^{-1}	FP -16
Accuracy	76.3%	75.6%	73.3%	45.3%	$\leq 1\%$	76.1%
Defendability	—	✗	✗	✓	✓	✗
		L - 10^{-4}	L - 10^{-3}	L - 10^{-2}	L - 10^{-1}	Int -8
Accuracy	—	75.6%	73.4%	46.2%	$\leq 1\%$	53.7%
Defendability	—	✗	✗	✓	✓	✓

Table 3: The trade-off between accuracy and defendability. **G**: Gaussian noise, **L**: Laplacian noise, **FP**: Floating number, **Int**: Integer quantization. ✓ means it successfully defends against DLG while ✗ means fails to defend (whether the results are visually recognizable). The accuracy is evaluated on CIFAR-100.

Only when the variance is larger than 10^{-2} and the noise is starting affect the accuracy, DLG will fail

Defense strategies

■ Experiments on pruning



(d) Defend with gradient pruning.

Conclusion and Discussions

- An algorithm that can completely obtain the local training data from public shared gradients.
- It's effective on both computer vision and natural language processing tasks, with single data and batched data.
- Such deep leakage can be only prevented when defense strategies starts to degrade the accuracy.
- The most effective defense method is gradient pruning.
- My comments:
 - It may not work when the dimension of Parameters is small than inputs. E.g., $w_1(x_1+x_2)^2+w_2x_3=y$;
 - It may not work on FL over the air.

Key references

- Zhao B, Mopuri K R, Bilen H. iDLG: Improved Deep Leakage from Gradients [J]. arXiv preprint arXiv:2001.02610, 2020.
 - Their proposed iDLG can extract ground-truth labels from the gradients and empirically demonstrate the advantages over DLG.

Algorithm 1 Improved Deep Leakage from Gradients (iDLG)

Require:

$F(\mathbf{x}; \mathbf{W})$: Differentiable learning model, \mathbf{W} : Model parameters, $\nabla \mathbf{W}$: Gradients produced by private training datum (\mathbf{x}, c) , N : maximum number of iterations. η : learning rate.

Ensure:

(\mathbf{x}', c') : Dummy datum and label.

- 1: $c' \leftarrow i$ s.t. $\nabla \mathbf{W}_L^i \cdot \nabla \mathbf{W}_L^j \leq 0, \forall j \neq i$ ➤ Extract the ground-truth label.
- 2: $\mathbf{x}' \leftarrow \mathcal{N}(0, 1)$ ➤ Initialize the dummy datum.
- 3: **for** $i \leftarrow 1$ to N **do**
- 4: $\nabla \mathbf{W}' \leftarrow \partial l(F(\mathbf{x}'; \mathbf{W}), c') / \partial \mathbf{W}$ ➤ Calculate the dummy gradients.
- 5: $L_G = \|\nabla \mathbf{W}' - \nabla \mathbf{W}\|_F^2$ ➤ Calculate the loss (difference between gradients).
- 6: $\mathbf{x}' \leftarrow \mathbf{x}' - \eta \nabla_{\mathbf{x}'} L_G$ ➤ Update the dummy datum.
- 7: **end for**