Deep Leakage from Gradients

- Deep Leakage from Gradients
- 1. Introduction
- 2. Related Work
  - 2.1 Distributed Training
  - 2.2 “Shallow” Leakage from Gradients
- 3. Method
- 4. Experiments
  - 4.1 Deep Leakage on Image Classification
  - 4.2 Deep Leakage on Masked Language Model
  - 4.3 Deep Leakage for Batched Data
- 5. Defense Strategies
  - 5.1 Noisy Gradients
  - 5.2 Gradient Compression and Sparsification
  - 5.3 Large Batch, High Resolution and Cryptology
    - batch size
    - cryptology
- 6. Conclusions

提出了几种可能的策略来阻止 deep leakage，最有效的是 gradient pruning.

1. Introduction

在分布式机器学习系统上，computation 是在每个 worker 上面并行执行并通过交换梯度来进行同步的。

case: can we completely steal the training data from gradients?

Deep Leakage from Gradients (DLG): 在共享梯度的时候会泄露训练数据的隐私

obtain both the training inputs and the labels in just a few iterations.

在获取了 dummy gradients 后，不是直接去优化 model weights, 而是优化dummy inputs and labels 。去最小化在 dummy gradients 和 real gradients 之间的 distance（使得 dummy data 更接近于 original ones.）

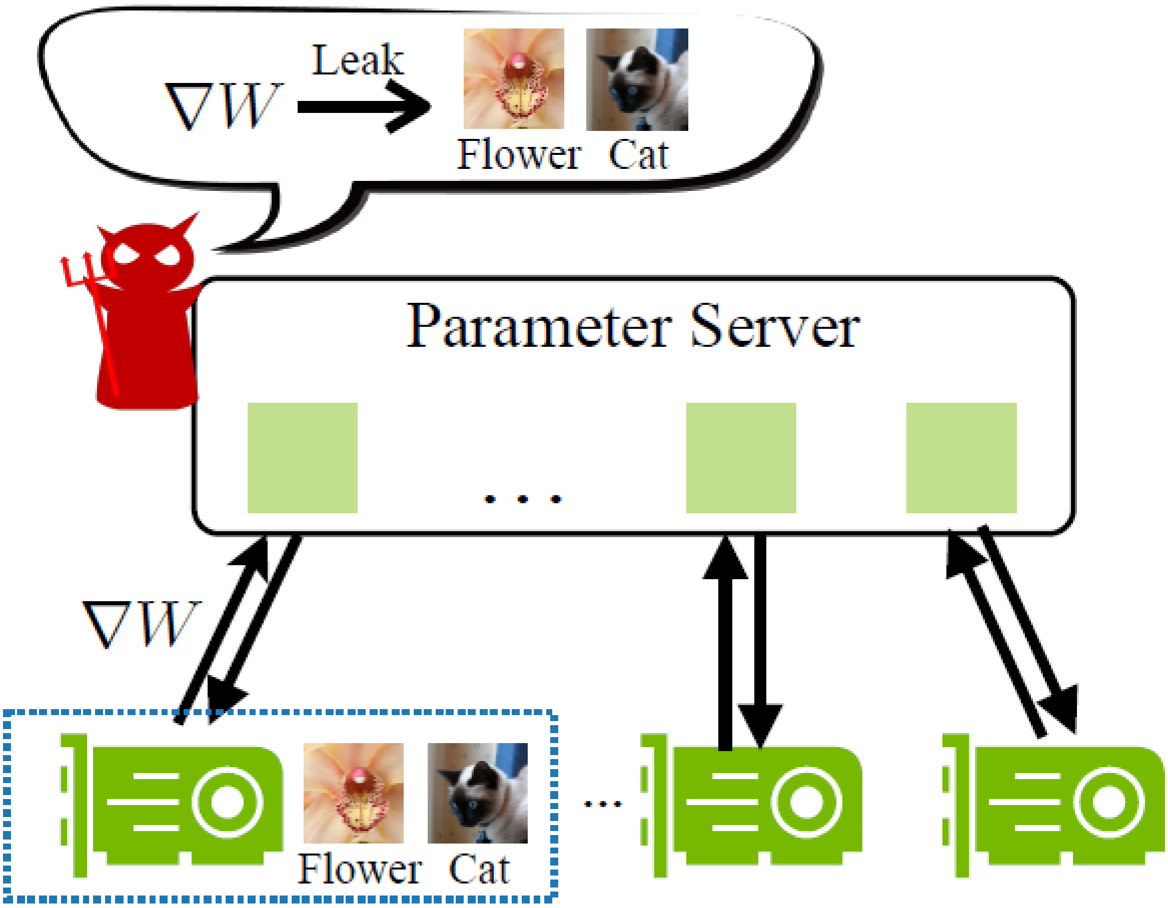
优化过程完成后，隐私数据(inputs and labels) 都会被 恢复出来

evaluate the effectiveness:

1. vision (image classification)
2. language tasks (masked language model)

Centralized distributed training:

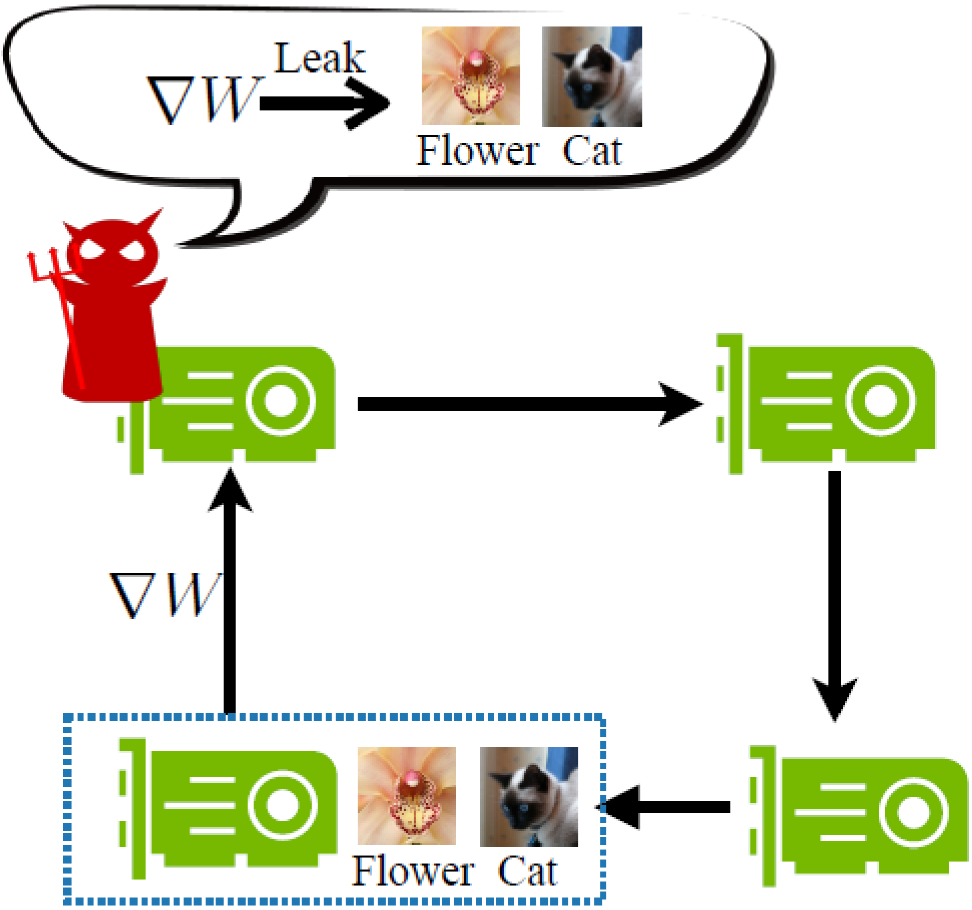
the parameter server, 不存储数据，能够从参与者中 steal local training data.



(a) Distributed training with a centralized server

Decentralized distributed training:

任何的参与者都可以 steal its neighbors' private training data.



(b) Distributed training without a centralized server

Defense strategies:

- gradient perturbation

  Gaussian and Laplacian noise 10^{-2}
- low precision

  half precision fails
- gradient compression

  successfully defends the attack with the pruned gradient is more than 20%

Contributions:

- DLG 首次提出可以从公开共享的 gradients 获取 private training data.
- DLG 只需要 gradients 就可以恢复 pixel-wise accurate images and token-wise matching texts. 传统的攻击方法需要额外的信息，而且只能恢复部分信息
- 分析了攻击的困难性，并提出了几种对抗此攻击的防御措施

2. Related Work

2.1 Distributed Training

many studies worked on distributed training to speedup，广泛使用的是 synchronous SGD.

分布式训练分为两类：

1. 有参数服务器 centralized

   gradients 首先被聚合，并分发给每个 node

2. 参数服务器 decentralized

   gradients 在 neighboring nodes 之间交换

（相同点：都要先在node上进行本地更新，并把 gradients 发送给别的nodes）

collaborative learning,数据不离开本地，只有梯度被共享

2.2 “Shallow” Leakage from Gradients

对于某些层， the gradients 泄露了某种程度的信息。

- the embedding layer

  language tasks, 揭露了什么 words 已经在其他参与者的训练集中被使用

  但是过于“shallow”，因为 words 是 unordered 并且由于歧义性很难推断出原始的 sentence

- the fully connected layers

  infer output feature values. 但是不能扩展到 卷积层，因为 features 的规模远大于 weights 的规模

Recently:

- learning-based methods.

  infer properties of the batch, 可以识别是否一个 data record 或者 data record properties 被包含在其他 participants' batch.

3. Method

标准的 synchronous distributed training:
在 step t，每个 node i 从自己的 local dataset 取样一个 minibatch $(x_{t,i}, y_{t,i})$, 并计算 gradient.

$$\nabla W_{t,i} = \frac{\partial \ell(F(\mathbf{x}_{t,i}, W_t), \mathbf{y}_{t,i})}{\partial W_t} \tag{1}$$

接着 gradients 在 N 个 servers 中取平均值聚合，并用于更新 weights:

$$\overline{\nabla W_t} = \frac{1}{N}\sum_{j}^{N} \nabla W_{t,j}; \quad W_{t+1} = W_t - \eta \overline{\nabla W_t} \tag{2}$$

已知参与者 k 的 $\nabla W_{t,k}$ ，目的是获取参与者 k 的 training data $(x_{t,k}, y_{t,k})$

steps:

- randomly initialize a dummy input x' and label input y'
- get "dummy" gradients

$$\nabla W' = \frac{\partial \ell(F(\mathbf{x}', W), \mathbf{y}')}{\partial W} \tag{3}$$

- 优化 dummy gradients 接近 original 的同时，就是 make the dummy data close to the real training data.
- minimize the following objective

$$\mathbf{x}'^{*}, \mathbf{y}'^{*} = \underset{\mathbf{x}', \mathbf{y}'}{\arg\min} ||\nabla W' - \nabla W||^2 = \underset{\mathbf{x}', \mathbf{y}'}{\arg\min} ||\frac{\partial \ell(F(\mathbf{x}', W), \mathbf{y}')}{\partial W} - \nabla W||^2 \tag{4}$$

Algorithm 1 Deep Leakage from Gradients.

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

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

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

4. Experiments

> optimize for 1200 iterations for image task
>
> optimize for 100 iterations for text tasks

DLG attack can happen any time

randomly initialized weights.

4.1 Deep Leakage on Image Classification

modern CNN architectures ResNet-56

pictures from MNIST, CIFAR-100, SVHN, LFW

models changes:

(1) replacing activation ReLU to Sigmoid

(2) removing strides.

(因为模型需要二阶可导)

泄露过程如下图所示，从 random Gaussian noise开始，尝试 match the gradients

> 黑白图像（MNIST）更容易被识别
>
> 复杂 的人脸图像需要更多的 iterations 去恢复
>
> optimization 完成后，recover 的结果和真实结果很接近

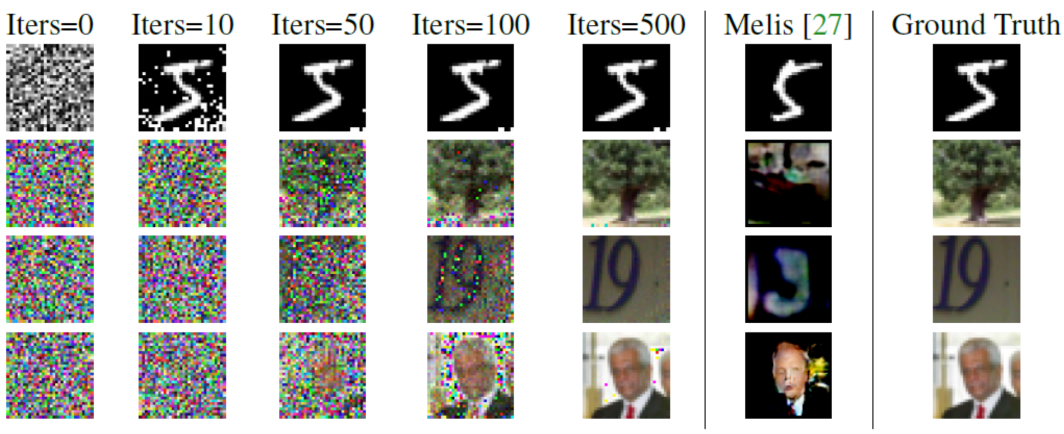


Figure 3: The visualization showing the deep leakage on images from MNIST [22], CIFAR-100 [21], SVHN [28] and LFW [14] respectively. Our algorithm fully recovers the four images while previous work only succeeds on simple images with clean backgrounds.

minimizing the distance between gradients 同样 reduces the gap between data.

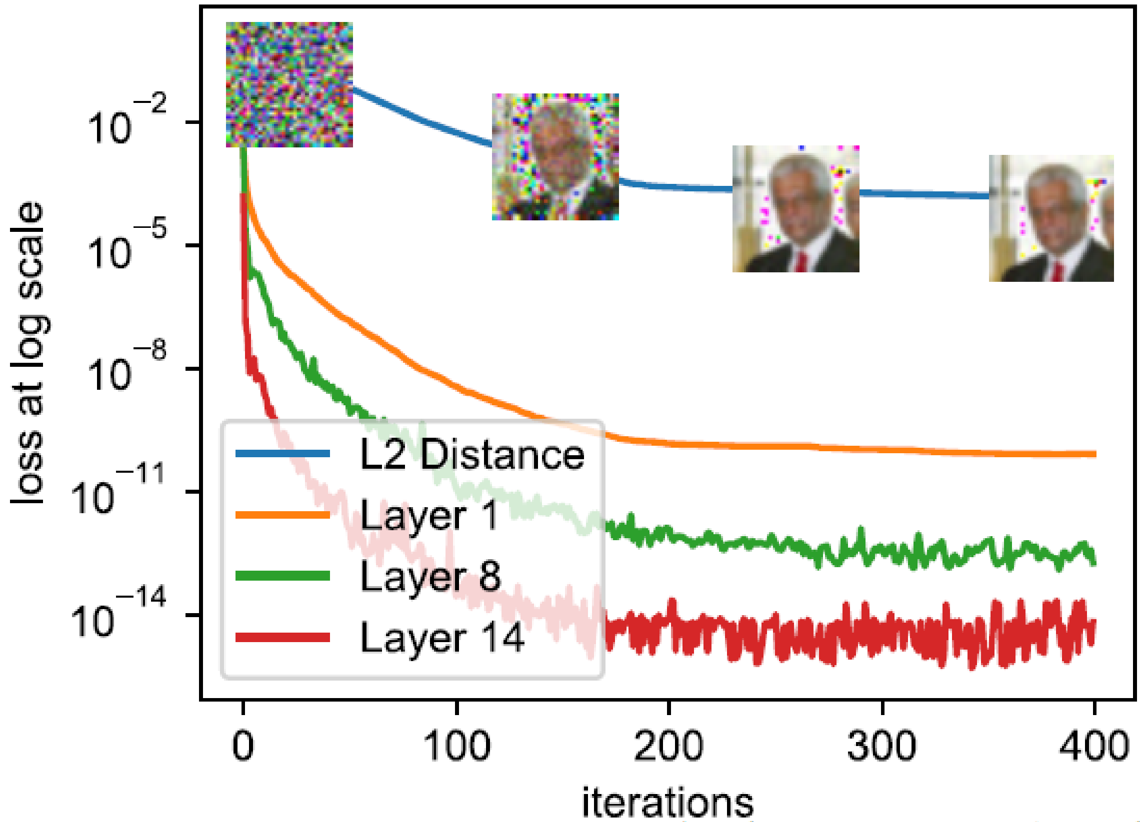


Figure 4: Layer-i means MSE between real and dummy gradients of i^{th} layer. When the gradients' distance gets smaller, the MSE between leaked image and the original image also gets smaller.

之前的 method 使用 GAN models, class label 已经给出，并且只在 MNIST 上有效

在 SVHN 上的结果已经不是原来的 training image.

在LFW上面结果更差

在CIFAR上面甚至 collapse

测试leaking并且测量所有 dataset images 的 MSE，效果如下：

> 图像被归一化到[0,1]的范围
>
> 算法效果很好ours < 0:03 v.s. previous > 0:2

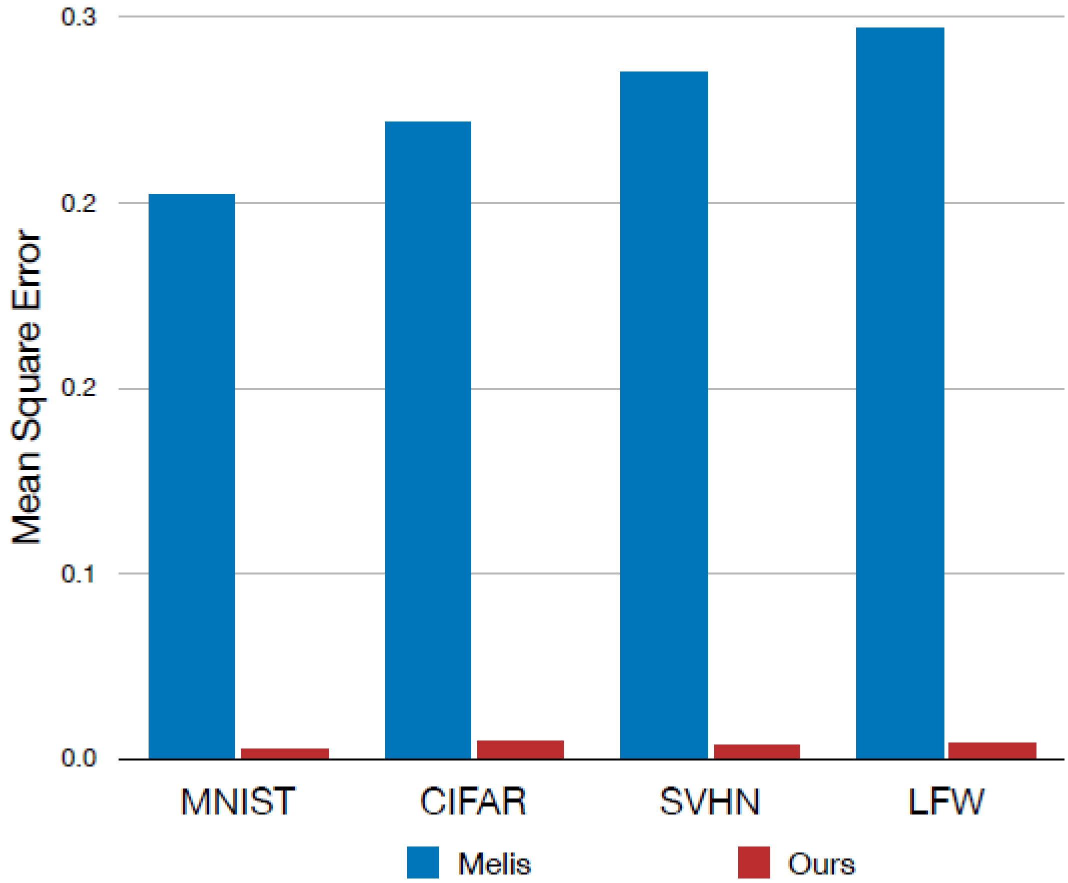


Figure 5: Compassion of the MSE of images leaked by different algorithms and the ground truth. Our method consistently outperforms previous approach by a large margin.

4.2 Deep Leakage on Masked Language Model

Masked Language Model (MLM) task

15% words 被用 [MASK] token 代替

MLM model 去预测 masked words 的原始 value 值

不同于 vision tasks，输入值是连续的 RGB values.

language model 需要去处理 **discrete words** into embeddings.

将DLG 应用于 embedding space，并且 minimize the gradients distance between dummy embeddings and real ones.

下表展示了在*NeurIPS*会议上的三条语句的泄露问题：

- randomly initialized embedding (语句在iteration 0 is meaningless)

- the gradients gradually match the original ones.
- 之后的 iterations中. leaked sentence 逐渐接近于 the original one.
- DLG 完成后，即使有 ambiguity，主体语句也能够完全 leaked.

	Example 1	Example 2	Example 3
Initial Sentence	tilting fill given **less word **itude fine **nton overheard living vegas **vac **vation *f forte **dis cerambycidae ellison **don yards marne **kali	toni **enting asbestos cutler 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 lifelong 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 identified major relief gin dive displaced **lice doll us apps _ **face space
Iters = 20	registration , volunteer applications , 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 importance for machine learning .	one **ry toppled hold major ritual ' dive annual conference days 1924 apps novelist dude space
Iters = 30	registration , volunteer applications , 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 importance for machine learning .	we invite submissions for the thirty - third annual conference on neural information 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 1: The progress of deep leakage on language tasks.

4.3 Deep Leakage for Batched Data

算法之前应用于的是在一个 batch 中，只有一对 input and label

但是对于 batch size > 1 的情况，算法进行很慢，难以收敛

Reason：

batch data 有 *N!* 种不同的排列组合，这使得 optimizer 很难去选择 gradient directions.

改进：

不更新整个 batch，而是更新一个 sample.

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.

$$
\begin{aligned}
\mathbf{x}'^{\,i \bmod N}_{t+1} &\leftarrow \mathbf{x}'^{\,i \bmod N}_{t} - \nabla_{\mathbf{x}'^{\,i \bmod N}_{t+1}} \mathbb{D} \\
\mathbf{y}'^{\,i \bmod N}_{t+1} &\leftarrow \mathbf{y}'^{\,i \bmod N}_{t} - \nabla_{\mathbf{y}'^{\,i \bmod N}_{t+1}} \mathbb{D}
\end{aligned} \tag{5}
$$

如下表，展示对于不同的 batch size，达到收敛时需要的 iterations 的数量的关系

batch size 越大，*DLG* 算法需要越多的 *iterations* 去执行攻击

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

Table 2: The iterations required for restore batched data on CIFAR [21] dataset.

即使 order 并不相同，而且加入了更多的人为噪声，但是DLG 仍然可以生成非常接近原始图像的 images.

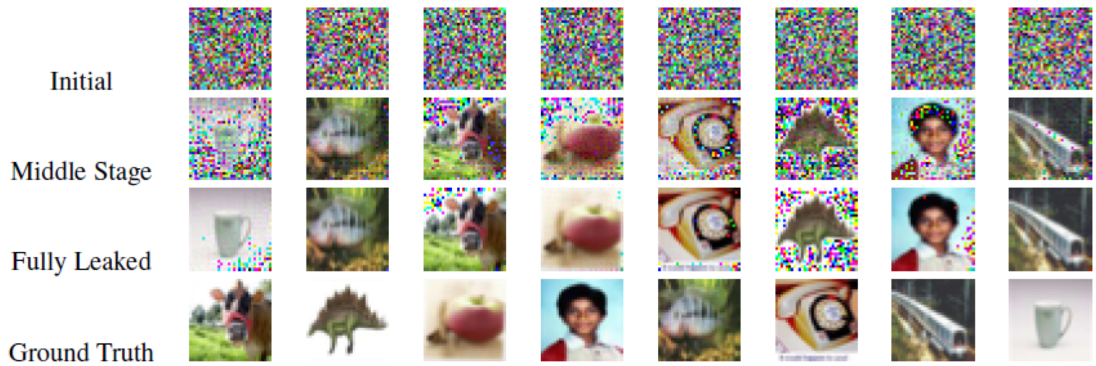


Figure 6: Results of deep leakage of batched data. Though the order may not be the same and there are more artifact pixels, DLG still produces images very close to the original ones.

5. Defense Strategies

5.1 Noisy Gradients

(1) **add noise on gradients before sharing**

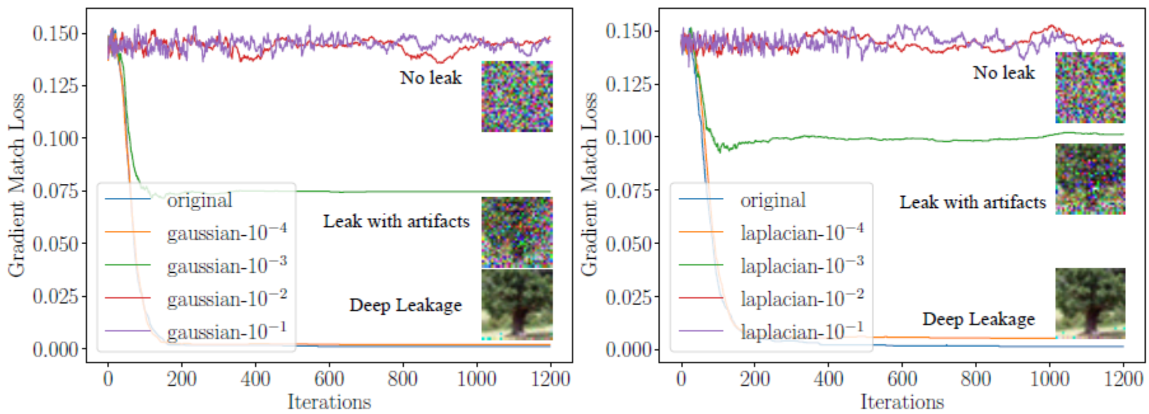
应用了Gaussian and Laplacian noise，分布的方差大小范围是 10^{-1}到10^{-4}

如图所示，防御的效果取决于 distribution variance 的大小，而不是与添加的 noise 的类别有关

方差在10^{-4} 和10^{-3}时候，效果不是很明显，攻击仍然可以执行

大于10^{-2}时候，noise开始影响准确性

然而，noise方差大于 10^{-2} 会显著降低 ACC



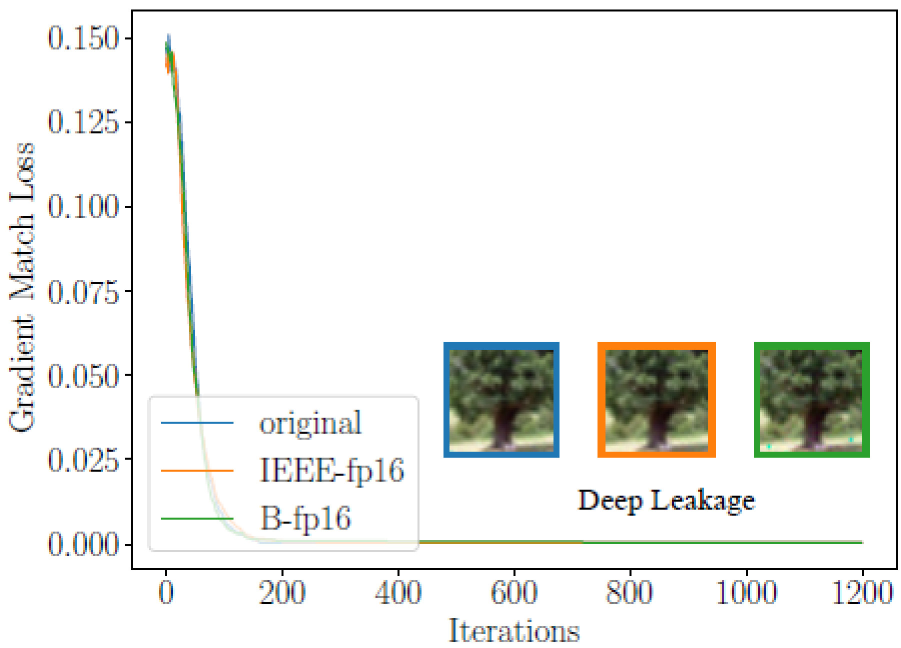
(a) Defend with different magnitude Gaussian noise. (b) Defend with different magnitude Laplacian noise.

(2) **half-precision**，起初用于节约 CPU memory 以及减少 communication bandwidth.

测试：

- EEE float16 (Single-precision floating-point format)
- bfloat16 (Brain Floating Point [35], a truncated version of 32 bit float).

如图所示，half-precision 都不能够保护 training data.



(c) Defend with fp16 convertion.

(3) **low-bit representation Int-8**, 可以成功保护隐私泄露，但是性能下降非常严重，

	Original	G-10^{-4}	G-10^{-3}	G-10^{-2}	G-10^{-1}	FP-16
Accuracy	76.3%	75.6%	73.3%	45.3%	≤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%	≤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.

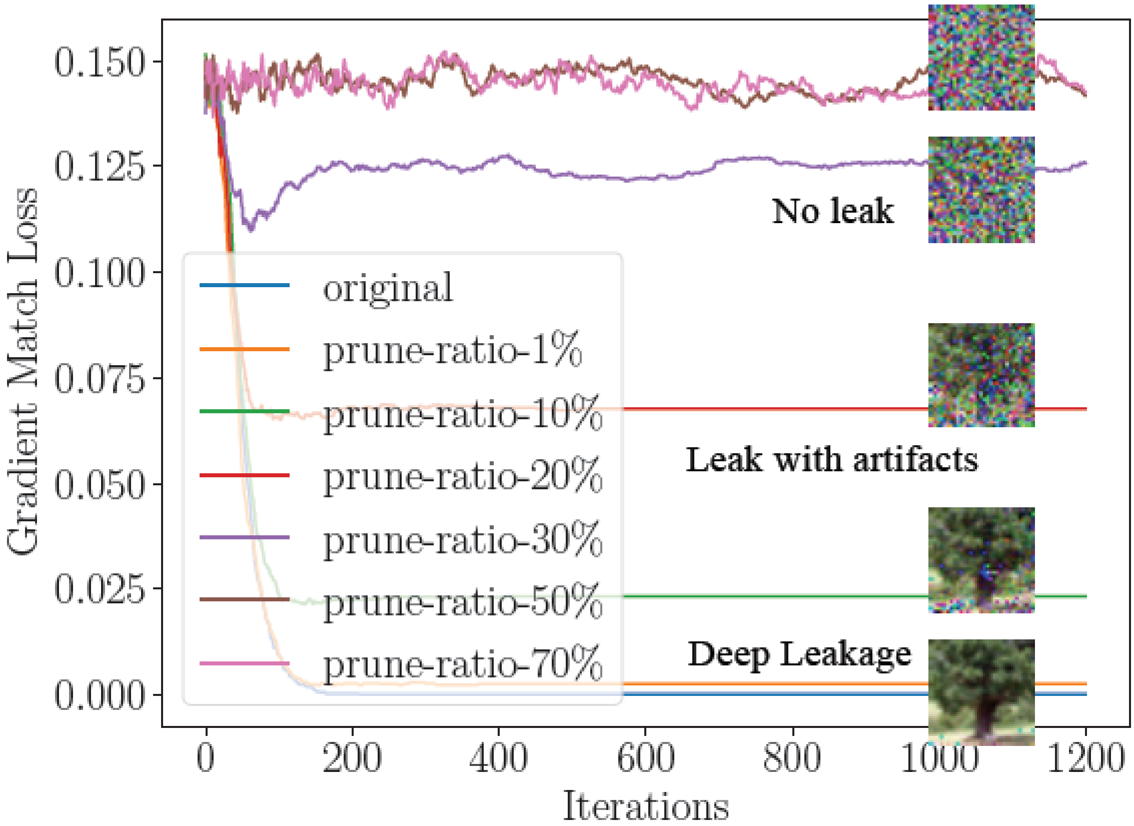
5.2 Gradient Compression and Sparsification

DLG 很难去 match gradients, 当 optimization targets 是 pruned

sparsity

> 1% to 10%, almost no effects against DLG.
>
> 20%. obvious artifact pixels on the recover images

最大容忍度是 20%, 当 purning ratio 更大时，恢复的 images 不能识别，此时 gradient compression 成功保护了隐私泄露。



(d) Defend with gradient pruning.

研究表明，the gradients 在有 error compensation techniques 情况下，最多可以被压缩 300 倍。这种情况下，sparsity 接近 99%，已经超过了 DLG 可以容忍的最大程度 (20%), 完全可以保护隐私。

5.3 Large Batch, High Resolution and Cryptology

batch size

增大 batch size 可以使得 leakage 更加困难，因为在 optimization 过程中，有更多的 variables to solve

由此， upscaling the input image 同样是一个好的 defense.

cryptology

- secure aggregation protocol

limitations: requires gradients to be integers.

- encrypt the gradients before sending

limitations: homomorphic encryption 仅能抵抗 parameter server.

6. Conclusions

- 介绍了 Deep Leakage from Gradients (DLG)，可以从 public shared gradients 获取 local training data.
- DLG 不依赖于任意的生成模型或者额外的 prior data
- 在 vision 和 language tasks 都证明了这种 deep leakage 的风险，并只有通过降低准确性的防御策略来抵抗这种攻击