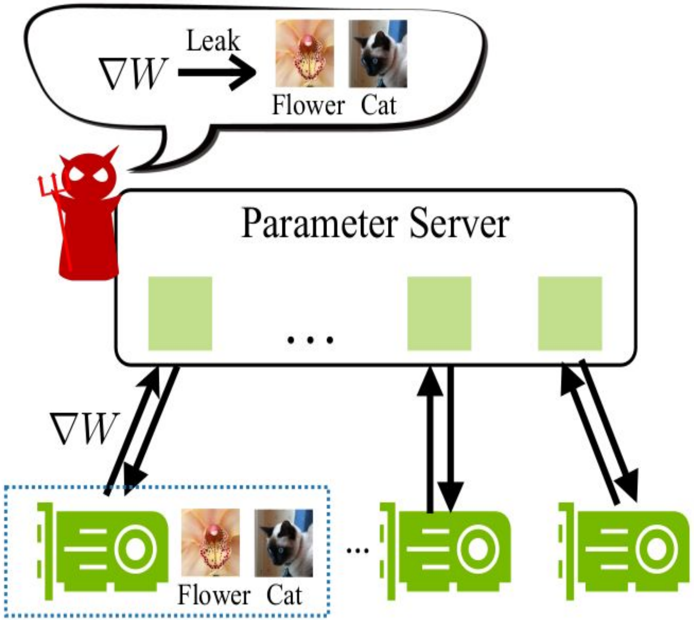
Deep Leakage from Gradients

By: Wafaa Mohammed

The Idea

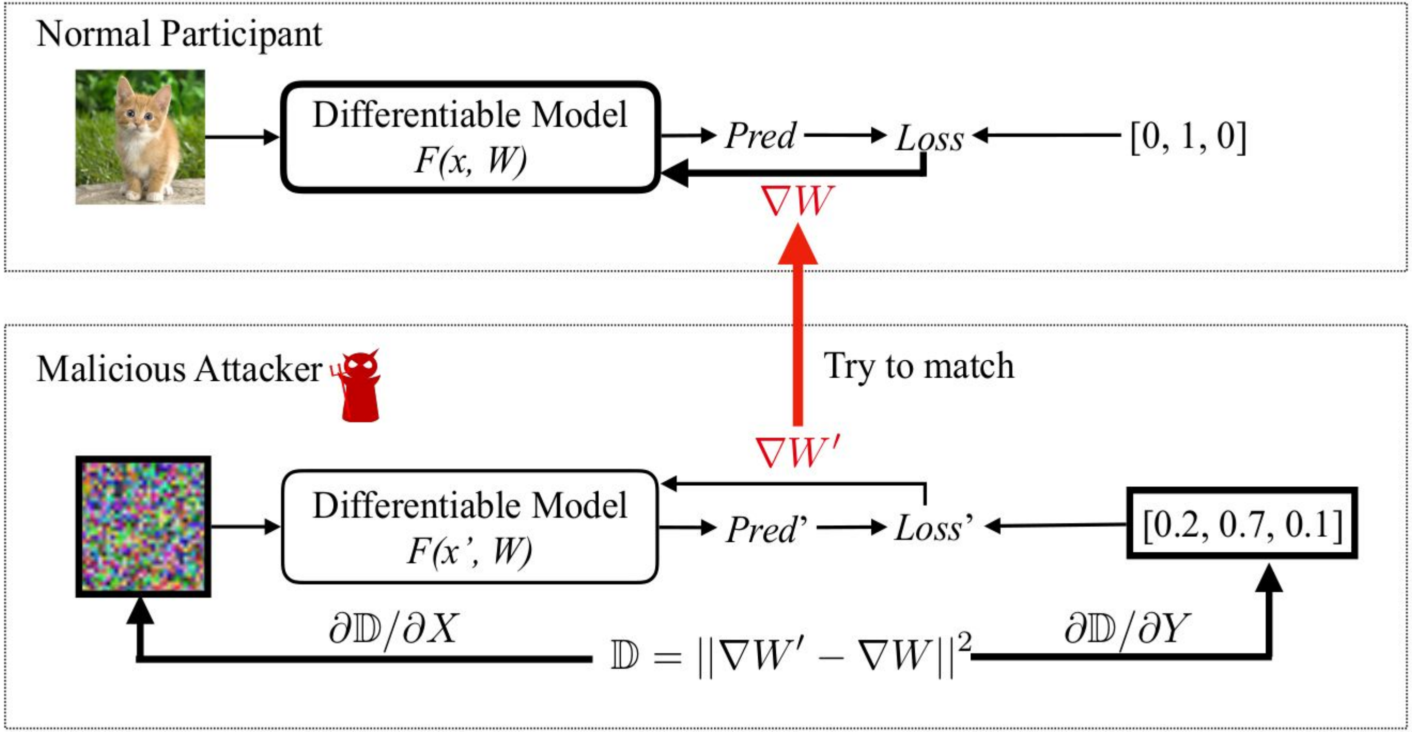
Main Ideas

- For a long time, people believed that gradients are safe to share (the training data will not be leaked).
- The paper shows that it is possible to obtain the private training data from the publicly shared gradients.
- Without changes on training setting, the most effective defense method is gradient pruning.



(a) Distributed training with a centralized server (b) Distributed training without a centralized server

The DLG Algorithm

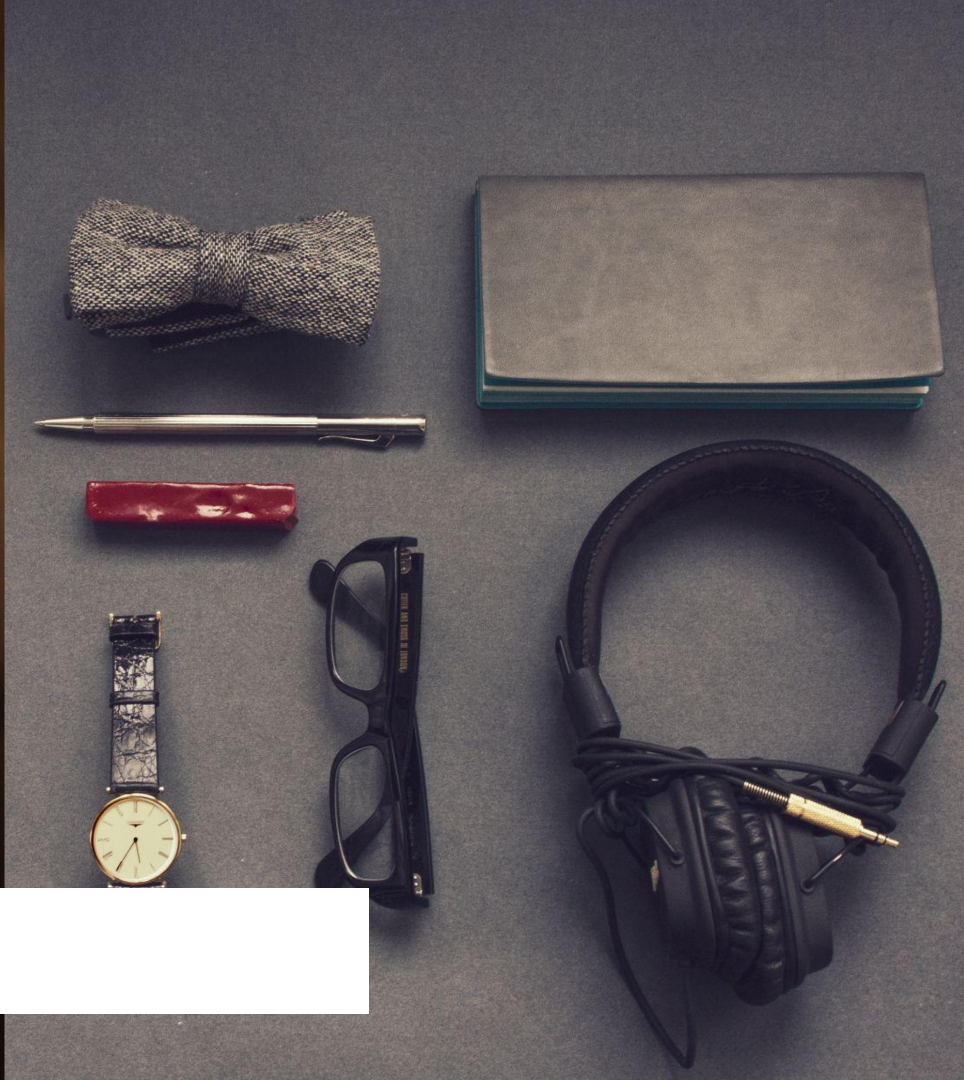

Normal Participant
Differentiable Model
F(x, W)
Pred
Loss
[0, 1, 0]
∇W
Try to match
Malicious Attacker
$\nabla W'$
Differentiable Model
F(x', W)
Pred'
Loss'
[0.2, 0.7, 0.1]
$\partial \mathbb{D} / \partial X$
$\mathbb{D} = ||\nabla W' - \nabla W||^2$
$\partial \mathbb{D} / \partial Y$

Algorithm 1 Deep Leakage from Gradients.

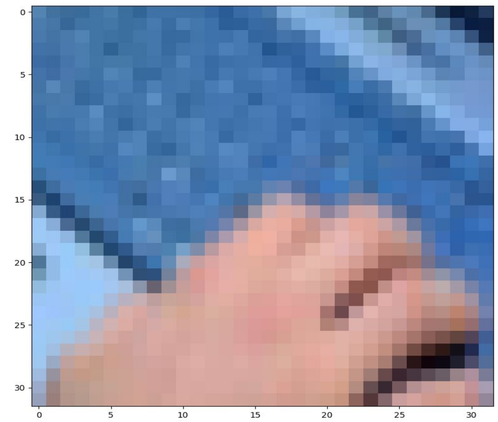
Input: $F(\mathbf{x}; W)$: Differentiable machine learning model; W: parameter weights; ∇W: gradients calculated by training data

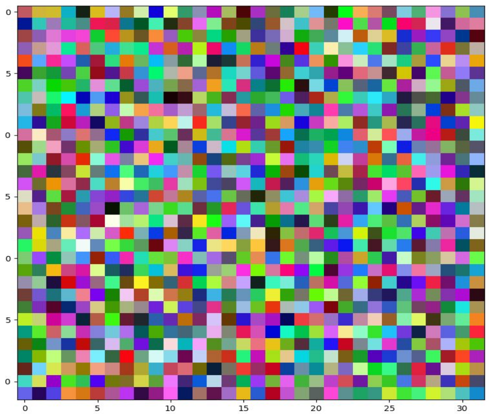
Output: private training data $\mathbf{x}, \mathbf{y}$

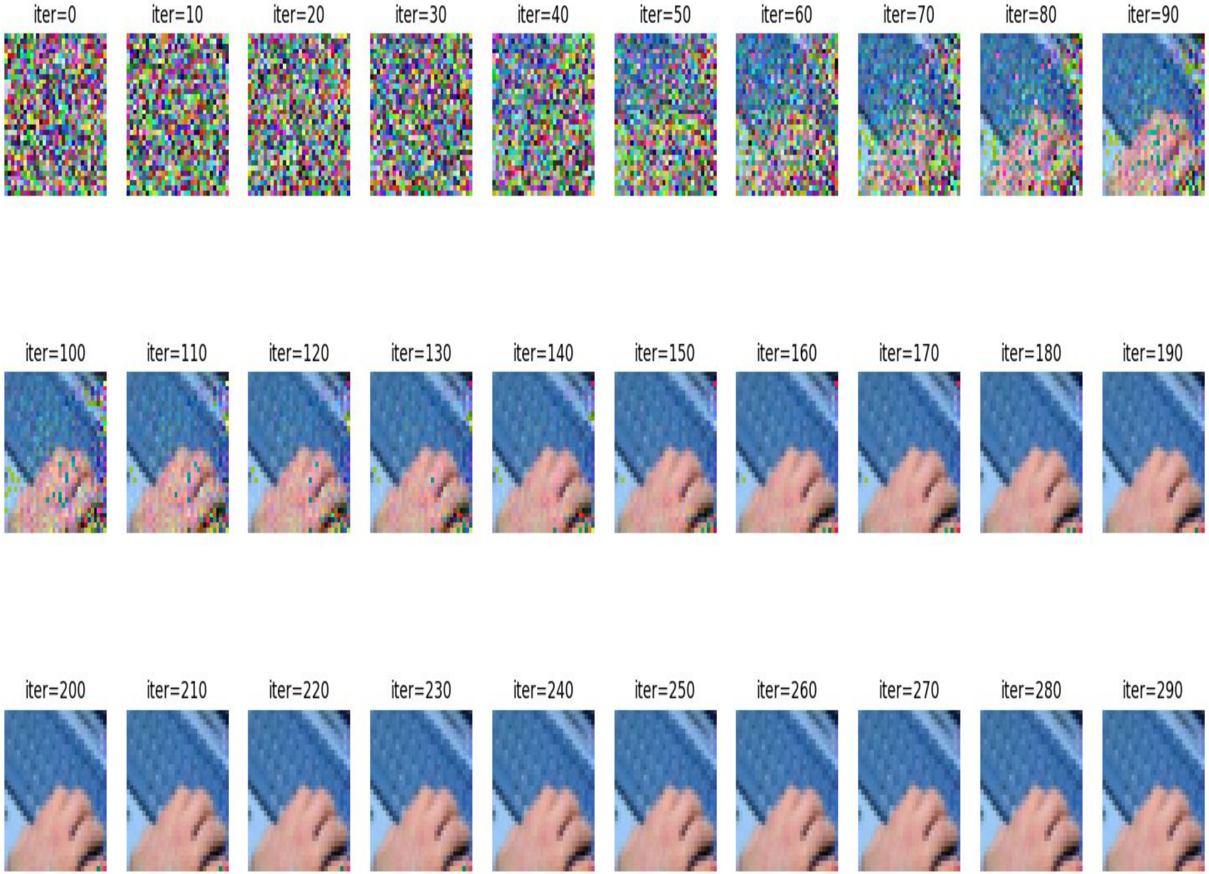
procedure DLG(F, W, ∇W)
    $\mathbf{x}'_1 \leftarrow \mathcal{N}(0,1)$, $\mathbf{y}'_1 \leftarrow \mathcal{N}(0,1)$ ▷ Initialize dummy inputs and labels.
    for $i \leftarrow 1$ to n **do**
        $\nabla W'_i \leftarrow \partial \ell(F(\mathbf{x}'_i, W_t), \mathbf{y}'_i)/\partial W_t$ ▷ Compute dummy 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

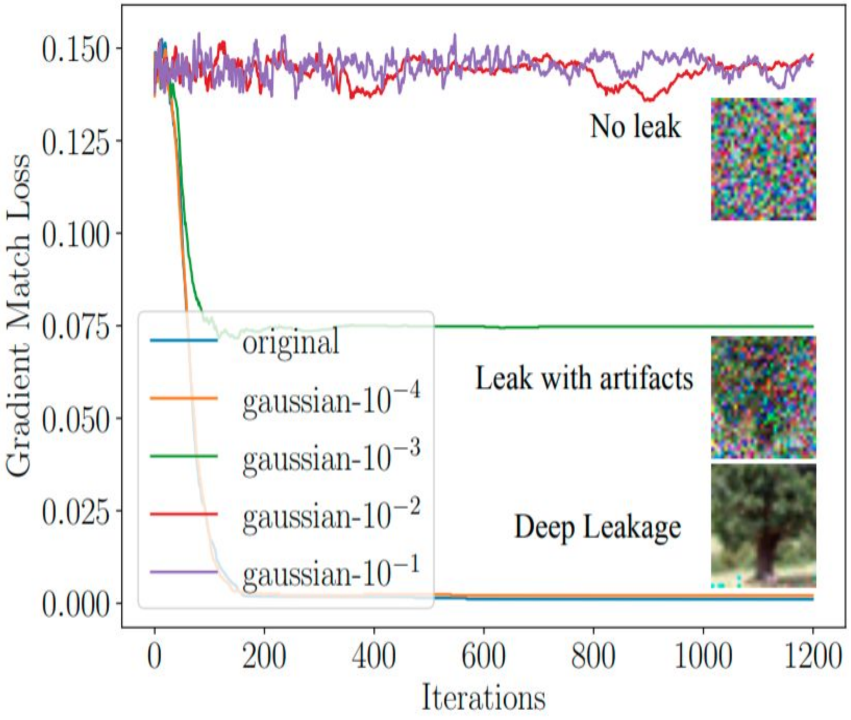


Experiments and results

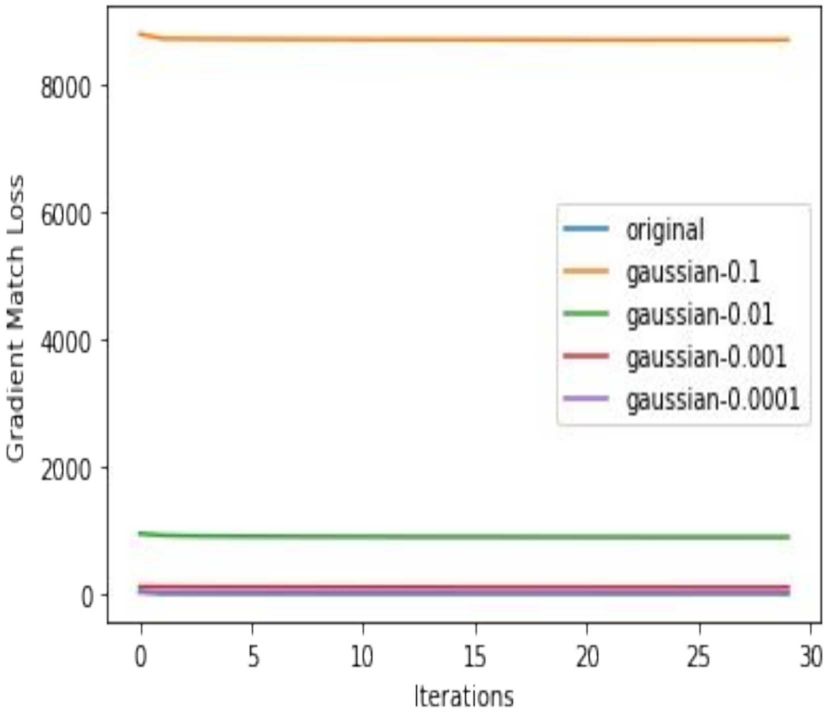

0 5 10 15 20 25 30
0 5 10 15 20 25 30


0 5 10 15 20 25 30


iter=0
iter=10
iter=20
iter=30
iter=40
iter=50
iter=60
iter=70
iter=80
iter=90
iter=100
iter=110
iter=120
iter=130
iter=140
iter=150
iter=160
iter=170
iter=180
iter=190
iter=200
iter=210
iter=220
iter=230
iter=240
iter=250
iter=260
iter=270
iter=280
iter=290



Gradient Match Loss
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
200
400
600
800
1000
1200
Iterations
original
gaussian-10^{-4}
gaussian-10^{-3}
gaussian-10^{-2}
gaussian-10^{-1}
No leak
Leak with artifacts
Deep Leakage



Gradient Match Loss
8000
6000
4000
2000
0
0
5
10
15
20
25
30
Iterations
original
gaussian-0.1
gaussian-0.01
gaussian-0.001
gaussian-0.0001

Conclusions

- Deep Leakage from Gradients (DLG): an algorithm that can obtain the local training data from public shared gradients.
- DLG does not rely on any generative model or extra prior about the data.
- Defense strategies include noisy gradients, gradient perturbation (half precision), and gradient compression (pruning).