Deep Leakage from Gradients

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

Visualization, code and more examples on dlg.mit.edu!

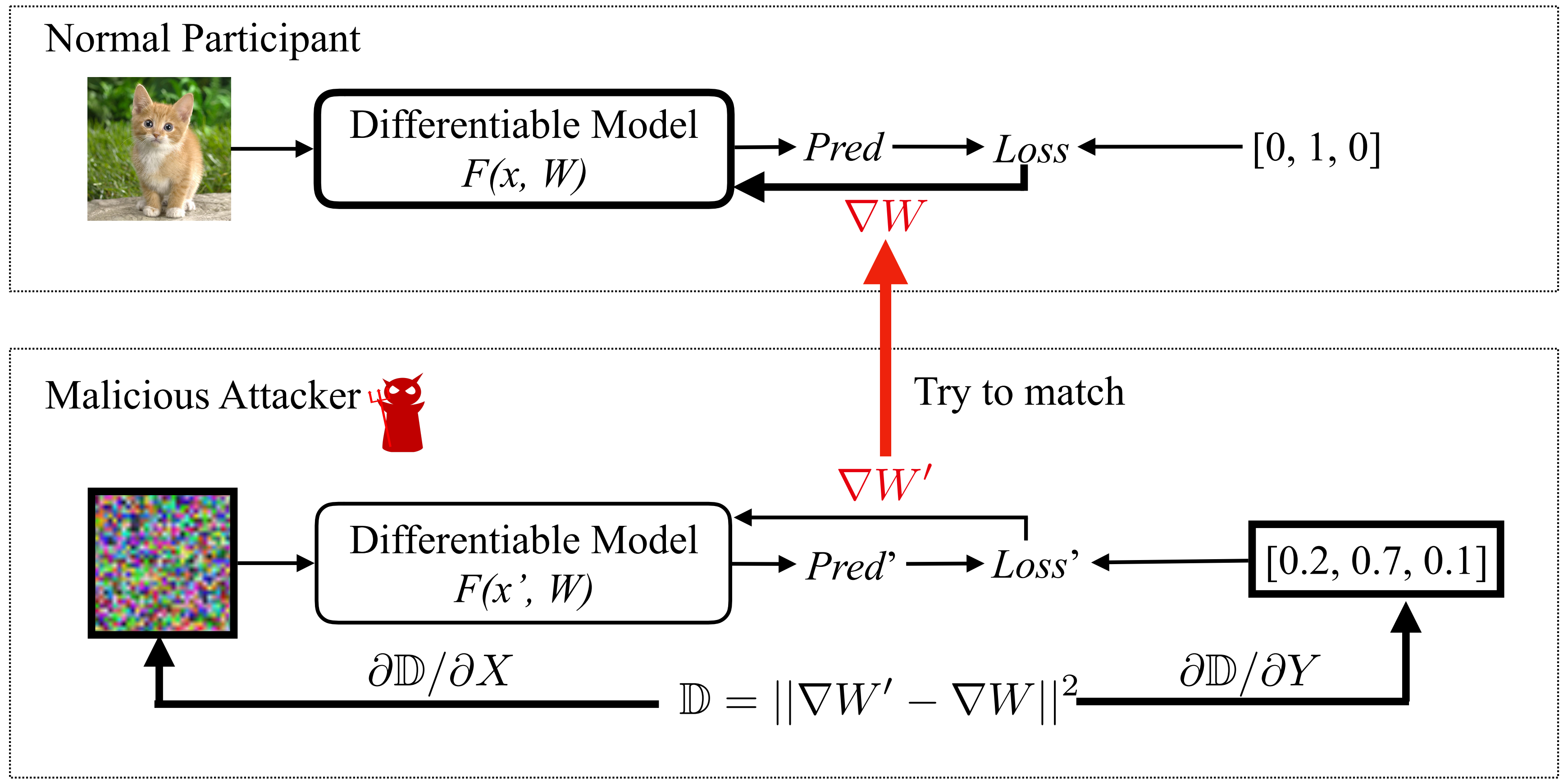
Rethink the Safety of Gradients Exchange

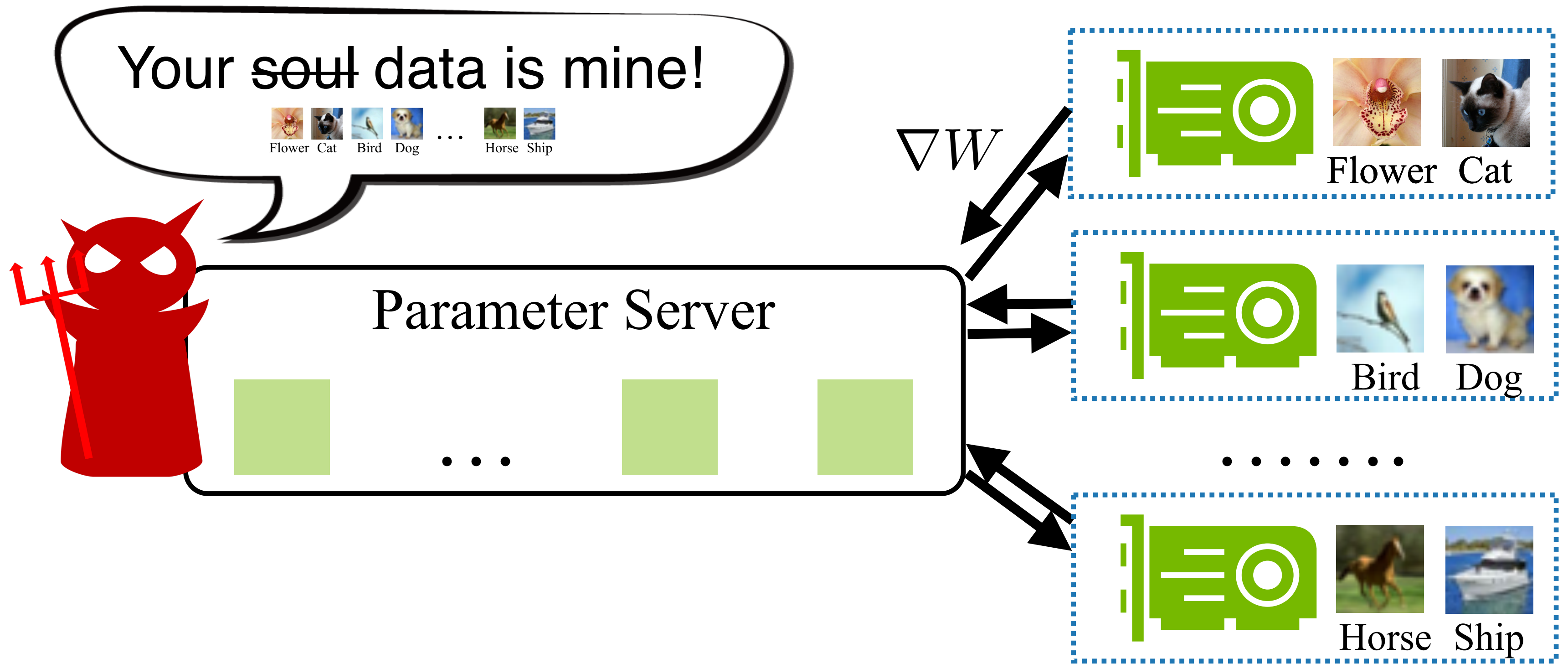
People believe it is safe to share gradients since numerical gradients seem have nothing related with semantic training set.

Based on the believe, collaborative & federated learning are developed where ***private training data** never leaves local device* and only ***public gradients** are communicating between*.

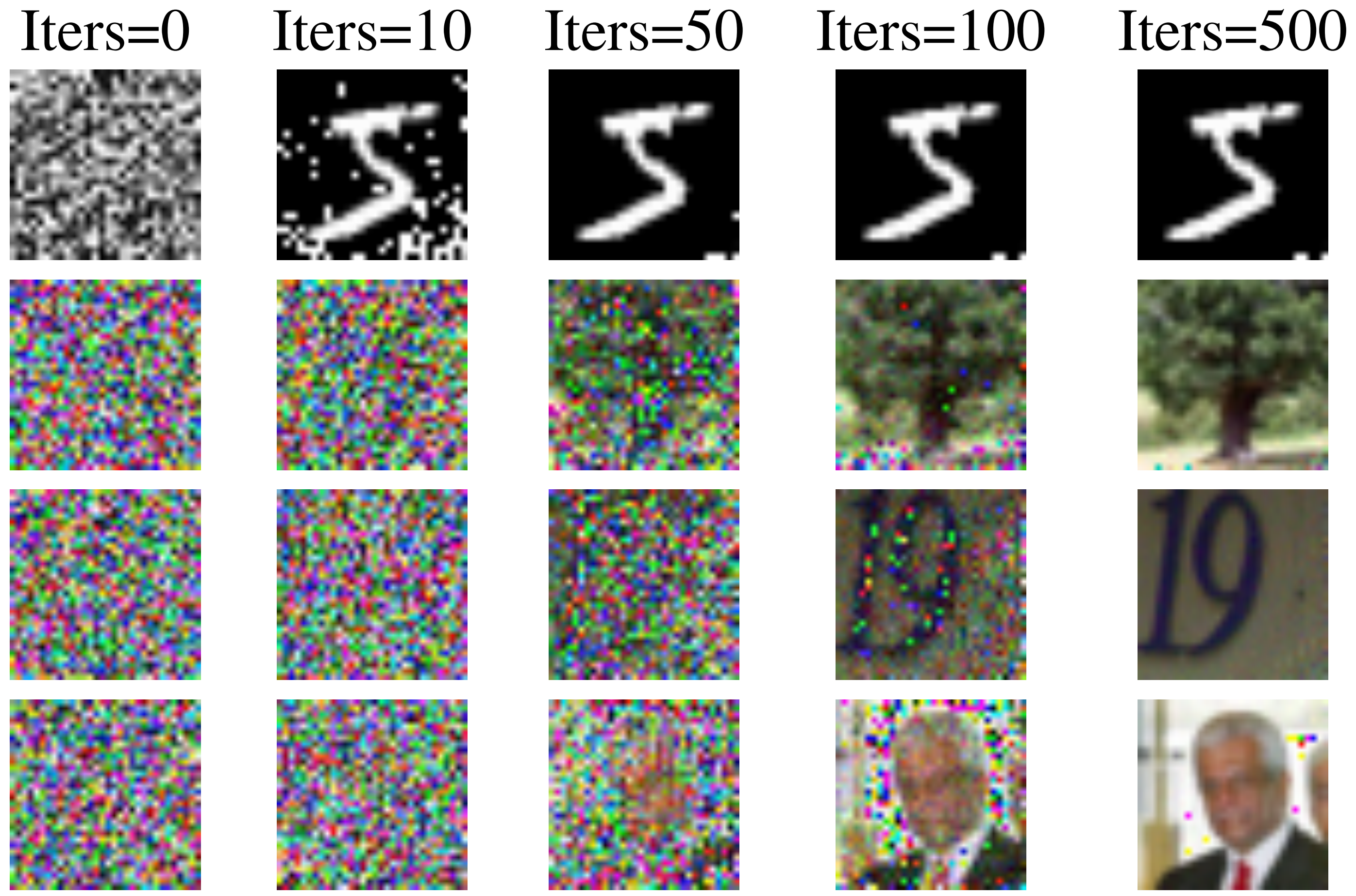
But does the scheme really protect training data? Conventionally wisdom suggests yes. But we show that actually **it is NOT safe**.

Deep ***L***eakage from ***G***radients (DLG)

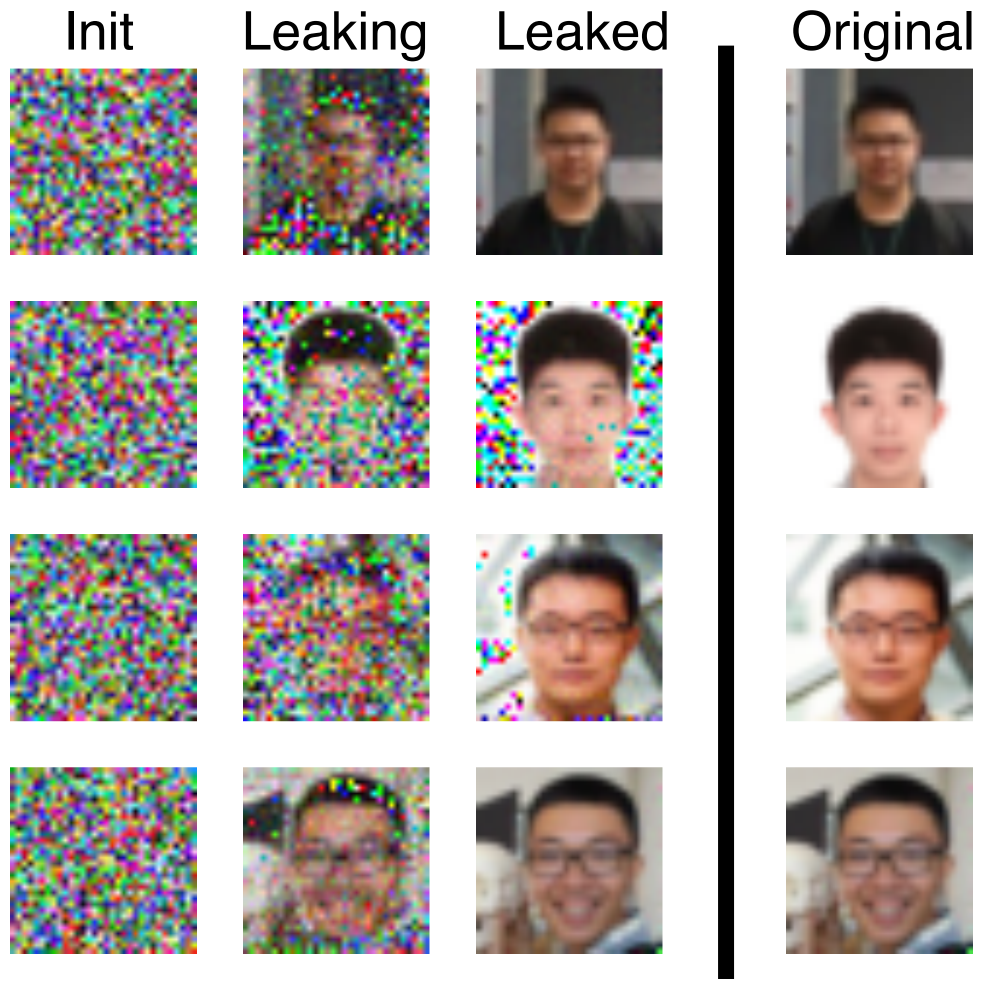




DLG for vision tasks



DLG for batched data



Note: the order may not be same as original's.

DLG for NLP task

[iter=0]

tilting fill given **less word **itude fine **nton over- heard living vegas **vac **vation *f forte **dis ce- rambycidae ellison **don yards marne **kali

[iter=10]

tilting fill given **less full solicitor other ligue shrill living vegas rider treatment carry played sculptures life- long ellison net yards marne **kali

[Iter=20]

registration , volunteer applications , at student travel application open the ; week of played ; child care will be glare

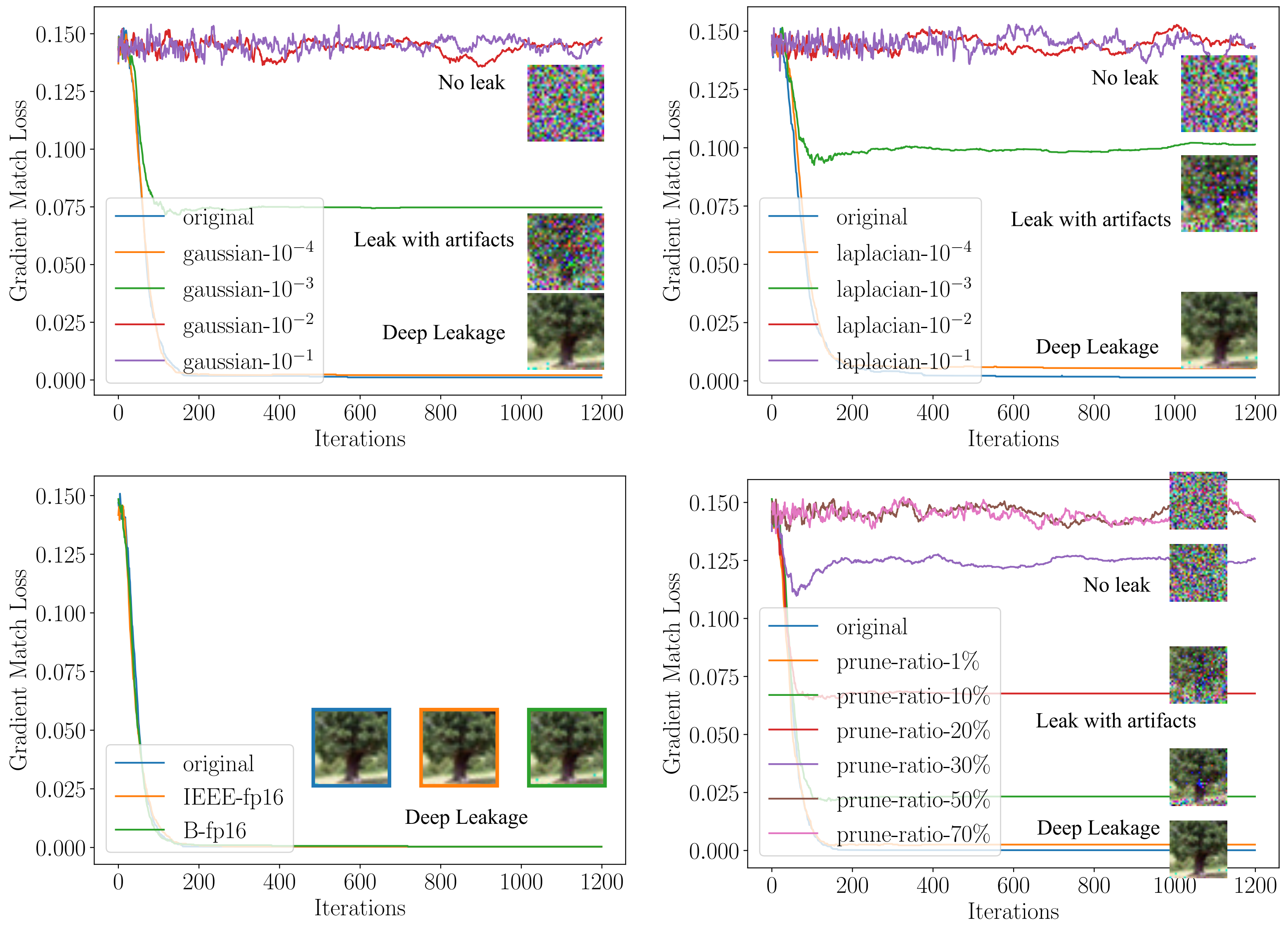
[Iter=30]

registration , volunteer applications , and student travel application open the first week of september . child care will be available .

[Ground Truth]

Registration, volunteer applications, and student travel application open the first week of September. Child care will be available.

DLG is hard to defend


Gradient Match Loss
Iterations
No leak
Leak with artifacts
Deep Leakage
original
gaussian-10^{-4}
gaussian-10^{-3}
gaussian-10^{-2}
gaussian-10^{-1}
laplacian-10^{-4}
laplacian-10^{-3}
laplacian-10^{-2}
laplacian-10^{-1}
IEEE-fp16
B-fp16
prune-ratio-1%
prune-ratio-10%
prune-ratio-20%
prune-ratio-30%
prune-ratio-50%
prune-ratio-70%