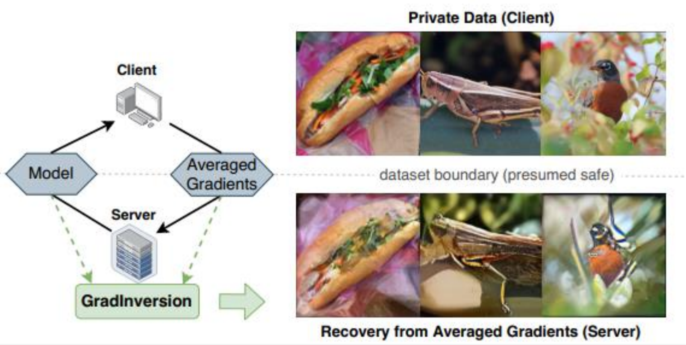
中国科学院自动化研究所
INSTITUTE OF AUTOMATION CHINESE ACADEMY OF SCIENCES

APRIL: Finding the Achilles' Heel on Privacy for Vision Transformers

Jiahao Lu, Xi Sheryl Zhang, Tianli Zhao, Xiangyu He, Jian Cheng

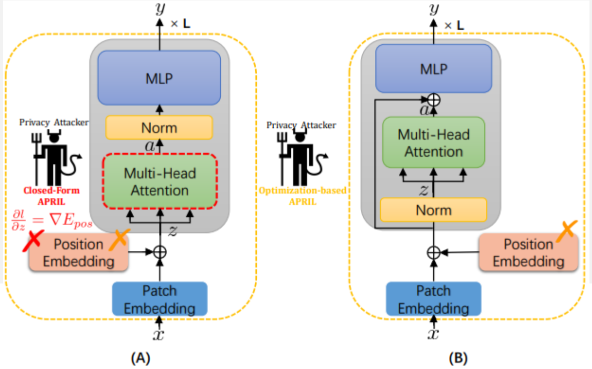
Gradient Inversion Attacks

What:
In a Federated Learning scenario, gradients shared by collaborative trainers can expose data privacy.
Where:
Past works conduct gradient inversion (gradient leakage attacks) mainly on CNNs and FC-NNs.


Private Data (Client)
Client
Model
Averaged Gradients
Server
dataset boundary (presumed safe)
GradInversion
Recovery from Averaged Gradients (Server)

Hongxu Yin et al. "See through Gradients: Image Batch Recovery via GradInversion" (CVPR 2021)

In this paper, we analyse the gradient leakage risk of self-attention based mechanism in both theoretical and practical manners. We propose **APRIL** – **A**ttention **PRI**vacy **L**eakage which poses strong threat to self-attention based models such as ViT.



Closed-Form APRIL attack for attention modules

For a self-attention module expressed as following:

$$Qz = q; Kz = k; Vz = v; \quad \frac{softmax(q\cdot k^T)}{\sqrt{d_k}}\cdot v = h \quad Wh = a;$$

If the derivative of loss *w.r.t.* input *z* is known, we can recover input by solving following linear system:

$$\frac{\partial l}{\partial z} z^T = Q^T\frac{\partial l}{\partial Q} + K^T\frac{\partial l}{\partial K} + V^T\frac{\partial l}{\partial V}$$

Where the RHS of the equation is shared in FL and exposed to the attacker. For a Vision Transformer who stack attention block on position embedding in a VGG-style, sensitive information (derivative of input *z*) is exposed by the gradient of learnable position embedding:

$$\boxed{\frac{\partial l}{\partial z}} = \frac{\partial l}{\partial E_{pos}}$$

$$\boxed{\frac{\partial l}{\partial z}} z^T = Q^T\frac{\partial l}{\partial q}z^T + K^T\frac{\partial l}{\partial k}z^T + V^T\frac{\partial l}{\partial v}z^T$$

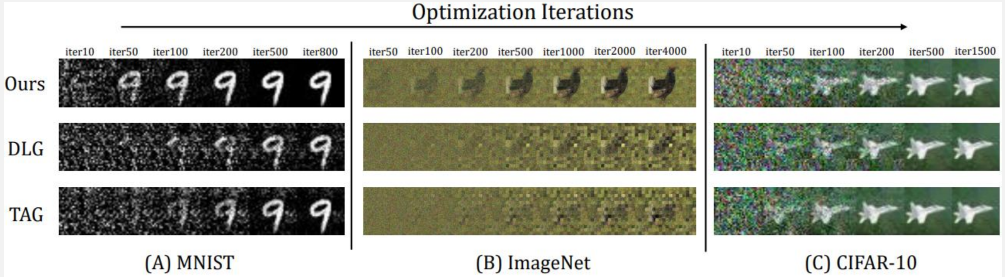
$$= \boxed{Q^T\frac{\partial l}{\partial Q} + K^T\frac{\partial l}{\partial K} + V^T\frac{\partial l}{\partial V}}$$

Optimization-based APRIL for real world ViTs

Matching the updating direction of position embedding with an direction caused by dummy data can do benetis on the recovery.

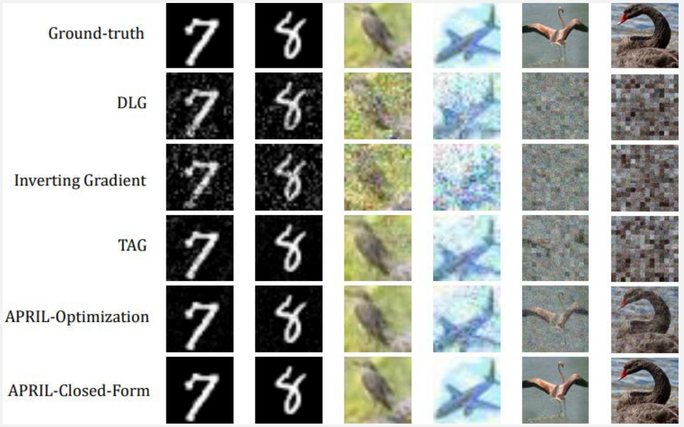
$$\mathcal{L} = \mathcal{L}_G + \alpha\mathcal{L}_A$$

$$= \|\nabla w' - \nabla w\|_F^2 - \alpha\cdot\frac{<\nabla E_{pos}, \nabla E'_{pos}>}{\|\nabla E_{pos}\|\cdot\|\nabla E'_{pos}\|}.$$



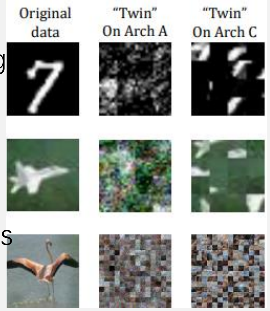
Our optimization-based APRIL attack shows faster convergence speed and does not easily fall into bad local minima, providing a better reconstruction result.

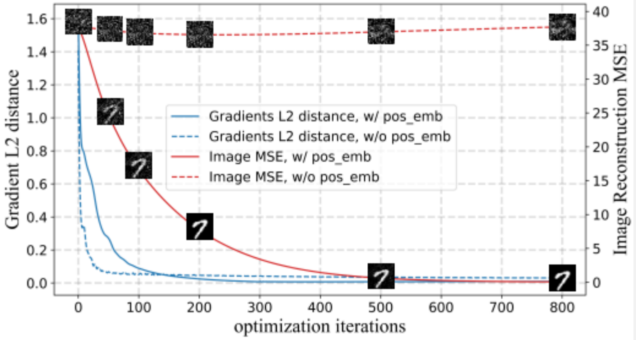
Results Comparison



APRIL-Inspired Defense Strategy

Closing the sharing of gradients from learnable position embedding will result in "twin" data, which is not visually similar to original data.
In other words, replacing learnable position embedding with fixed ones can protect privacy.





(A) Gradient l2 loss and image MSE on Architecture A