Efficient Differentially Private Fine-Tuning of Diffusion Models

*Jing Liu ♦, Andrew Lowy *, Toshiaki Koike-Akino ♦, Kieran Parsons ♦, and Ye Wang ♦*

♦ Mitsubishi Electric Research Laboratories   * University of Wisconsin-Madison

Contact: {jiliu, yewang}@merl.com

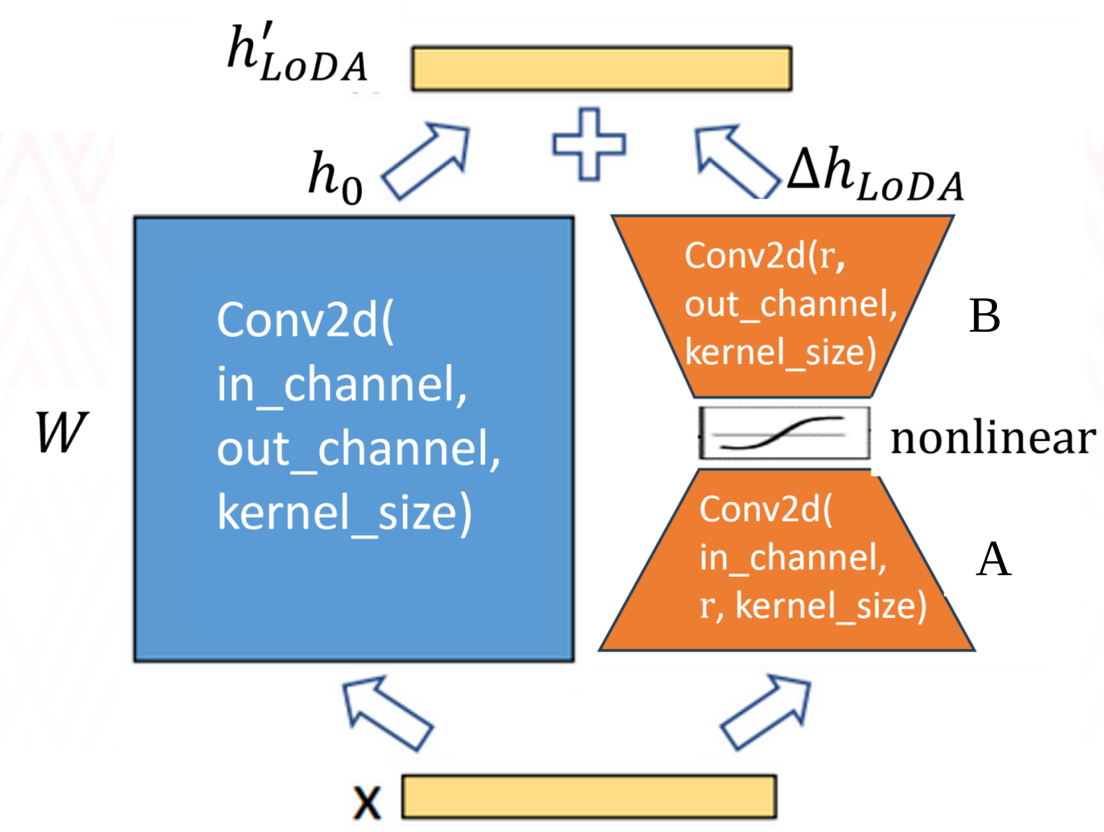
MITSUBISHI ELECTRIC
Changes for the Better

NextGenAISafety @ ICML 2024

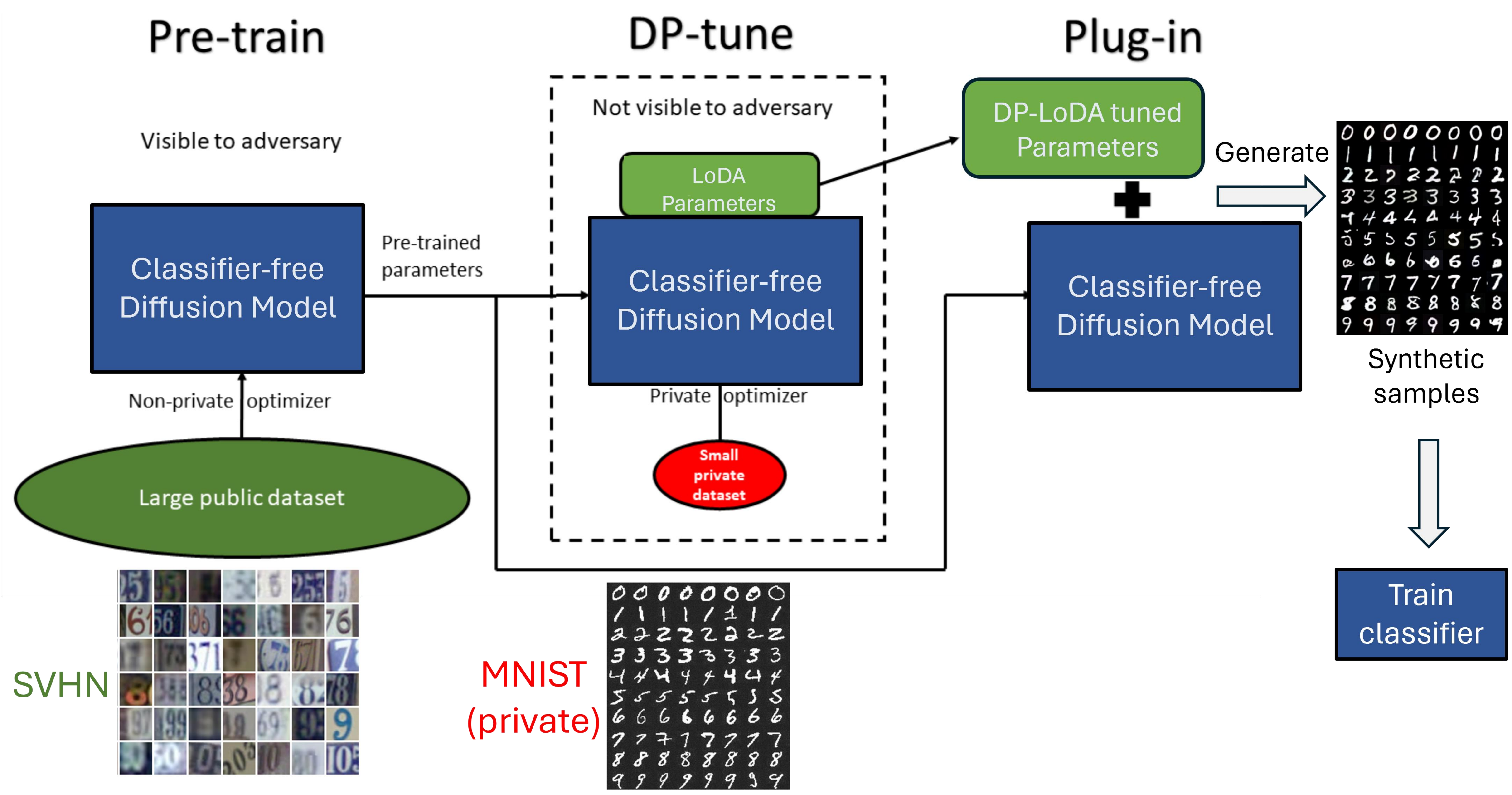
Motivation

- **Differential Privacy usually comes with a significant utility cost**
- **Diffusion Models (DM) enable high-quality synthetic generation**
- **Can we leverage synthetic samples to protect privacy?**
- **[Ghalebikesabi *et al.*, 23] shows that fully fine-tuned DM (with DP-SGD) on private data can generate useful synthetic images**
- **However, full fine-tuning DM with DP-SGD is resource-demanding in terms of memory and computation**
- **Parameter-Efficient Fine-Tuning (PEFT) is popular in LLMs, can we leverage PEFT for finetuning DM with DP-SGD?**

Low-Dimensional Adaptation (LoDA) for convolution layer



DP-LoDA Framework



Empirical Results

Table 1. MNIST test accuracy of CNN classifier for each DP training method (with access to full MNIST train set).

Method	$(\epsilon = 10, \delta = 10^{-5})$
DP-LDM	94.3
DP-LoDA	95.0
DP-Diffusion	95.9
DP-SGD	79.3
No DP	99.4

Table 2. CIFAR-10 test accuracy of ResNet9 for each DP training method (with access to full CIFAR-10 training set). $\delta = 10^{-5}$

Method	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$
DP-LDM	51.3 ± 0.1	59.1 ± 0.2	65.3 ± 0.3
DP-LoDA	60.2 ± 0.2	62.2 ± 0.4	63.5 ± 1.8
DP-Diffusion	66.3 ± 0.4	69.6 ± 0.2	69.7 ± 1.4
DP-SGD	36.5 ± 0.9	47.4 ± 0.9	48.3 ± 0.2
DP-MEPF (ϕ_1, ϕ_2)	28.9	47.9	48.9
DP-MEPF (ϕ_1)	29.4	48.5	51.0
DP-MERF	13.8	13.4	13.2
No DP		90.7	

Table 3. CIFAR-10 test accuracy of ResNet9 for each DP training method (with access to 1% CIFAR-10 training set). $\delta = 10^{-5}$

Method	$\epsilon = 1$	$\epsilon = 10$
DP-LoDA	54.2	53.6
DP-Diffusion	54.6	55.9
DP-SGD	11.5	21.2
No DP	52.5	

*Dimension r is set to 4 for DP-LoDA in all experiments.

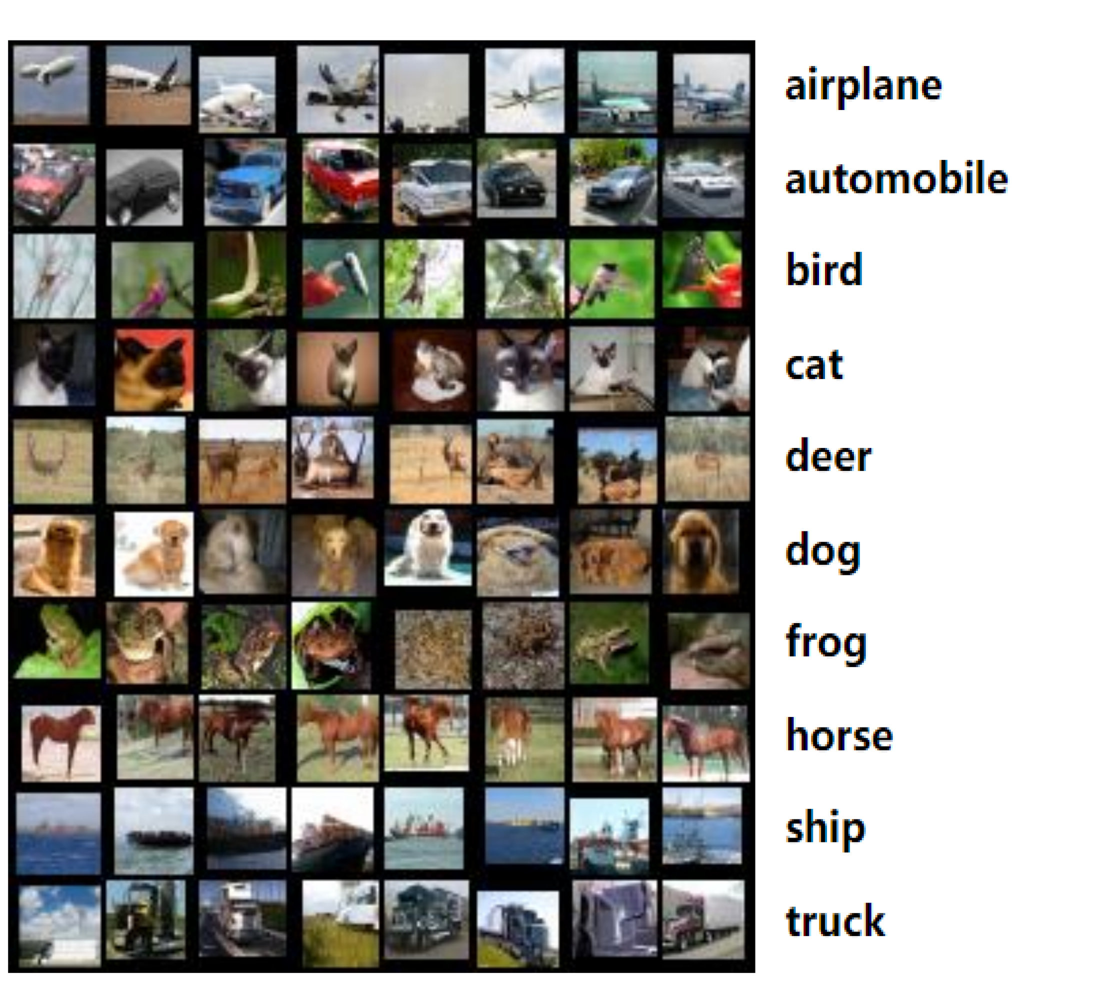


Figure. Generated images by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on 1% CIFAR-10 training set.

References

LoDA: Liu, J., Koike-Akino, T., Wang, P., Brand, M., Wang, Y., Parsons, K., "LoDA: Low-Dimensional Adaptation of Large Language Models", NeurIPS'23 workshop, December 2023.

DP-Diffusion: Ghalebikesabi, S., Berrada, L., Gowal, S., Ktena, I., Stanforth, R., Hayes, J., De, S., Smith, S. L., Wiles, O., and Balle, B. Differentially private diffusion models generate useful synthetic images. arXiv preprint arXiv:2302.13861, 2023a..

DP-LDM: Lyu, S., Vinaroz, M., Liu, M. F., and Park, M. Differentially private latent diffusion models. arXiv preprint arXiv:2305.15759, 2023.

DP-MERF: Harder, F., Adamczewski, K., and Park, M. DP-MERF: Differentially private mean embeddings with random features for practical privacy-preserving data generation. Proceedings of Machine Learning Research, 130:1819–1827, 2021a.

DP-MEPF: Harder, F., Jalali, M., Sutherland, D. J., and Park, M. Pretrained perceptual features improve differentially private image generation. Transactions on Machine Learning Research, 2023.