Efficient Differentially Private Fine-Tuning of Diffusion Models

Jing Liu [1] **Andrew Lowy** [2] **Toshiaki Koike-Akino** [1] **Kieran Parsons** [1] **Ye Wang** [1]

Abstract

The recent developments of Diffusion Models (DMs) enable generation of astonishingly high-quality synthetic samples. Recent work showed that the synthetic samples generated by the diffusion model, which is pre-trained on public data and fully fine-tuned with differential privacy on private data, can train a downstream classifier, while achieving a good privacy-utility trade-off. However, fully fine-tuning such large diffusion models with DP-SGD can be very resource-demanding in terms of memory usage and computation. In this work, we investigate Parameter-Efficient Fine-Tuning (PEFT) of diffusion models using Low-Dimensional Adaptation (LoDA) with Differential Privacy. We evaluate the proposed method with the MNIST and CIFAR-10 datasets and demonstrate that such efficient fine-tuning can also generate useful synthetic samples for training downstream classifiers, with guaranteed privacy protection of fine-tuning data. Our source code will be made available on GitHub.

1. Introduction

Differentially Private (DP) training, via DP-SGD (Abadi et al., 2016), of a task-specific classifier is commonly used for protecting the privacy of sensitive training data. However, it often comes with a significant cost in terms of model utility. The recent developments of Diffusion Models (DMs) enable generation of astonishingly high-quality synthetic samples, which can be used to train downstream classifiers. However, the synthetic samples generated by diffusion models do not inherently preserve training data privacy. In fact, Carlini et al. (2023) demonstrates that diffusion models memorize individual images from their training data and emit them at generation time. It is vital to train the diffusion model with privacy guarantees. The recent work of Ghalebikesabi et al. (2023b) showed that the synthetic samples generated by the diffusion model, which is fully fine-tuned with differential privacy, can train a downstream classifier that achieves very good utility. However, fully fine-tuning such large diffusion model with DP-SGD can be very resource-demanding.

Parameter-Efficient Fine-Tuning (PEFT) updates only a small set of model parameters, which may be a subset of the existing model parameters or a set of newly added parameters, can greatly reduce the computation and memory costs, and has become popular in fine-tuning Large Language Models (LLMs). The most widely-used PEFT approach is Low-Rank Adaptation (LoRA), which constrains the updates of weights to be low-rank. In LLMs, the various modules are mostly built from linear layers, where it is straightforward to apply LoRA to the weight matrix. However, in Diffusion Models, there are many convolutional layers instead of linear layer, and the appropriate generalization of "Low-Rank Adaptation" becomes somewhat unclear when applied to convolutional layers. For instance, in the official implementation of LoRA[1], the weight parameters of the 2D convolutional layer with dimension $(\text{out_channels}, \text{in_channels}, \text{kernel_size}, \text{kernel_size})$ is reshaped/viewed as a 2D matrix with shape $(\text{out_channels} \times \text{kernel_size}, \text{in_channels} \times \text{kernel_size})$, and the low-rank adapter is applied to this giant matrix. The recent work (Liu et al., 2023) proposed Low-Dimensional Adaptation (LoDA), which is a generalization of LoRA from a linear low-rank mapping to a nonlinear low-dimensional mapping, which is more suitable for convolutional layers than LoRA. We will discuss the details of LoDA applied to convolutional layers in Section 2.1.

This motivates us to investigate whether using LoDA to fine-tune Diffusion Models with Differential Privacy (a.k.a. DP-LoDA) can also generate synthetic samples for training a downstream classifier with good utility. We notice an interesting concurrent work (Lyu et al., 2023) that uses a Latent Diffusion Model to reduce the number of fine-tuning parameters, and they further propose to fine-tune the

[1]Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA. {jiliu, koike, parsons, yewang}@merl.com
[2]Wisconsin Institute for Discovery, University of Wisconsin-Madison. This work was performed while A.L. was an intern at MERL.
Correspondence to: Jing Liu <jiliu@merl.com>.
Proceedings of the 41st International Conference on Machine Learning, Next Generation of AI Safety workshop, Vienna, Austria.
Copyright 2024 by the author(s).

[1]https://github.com/microsoft/LoRA

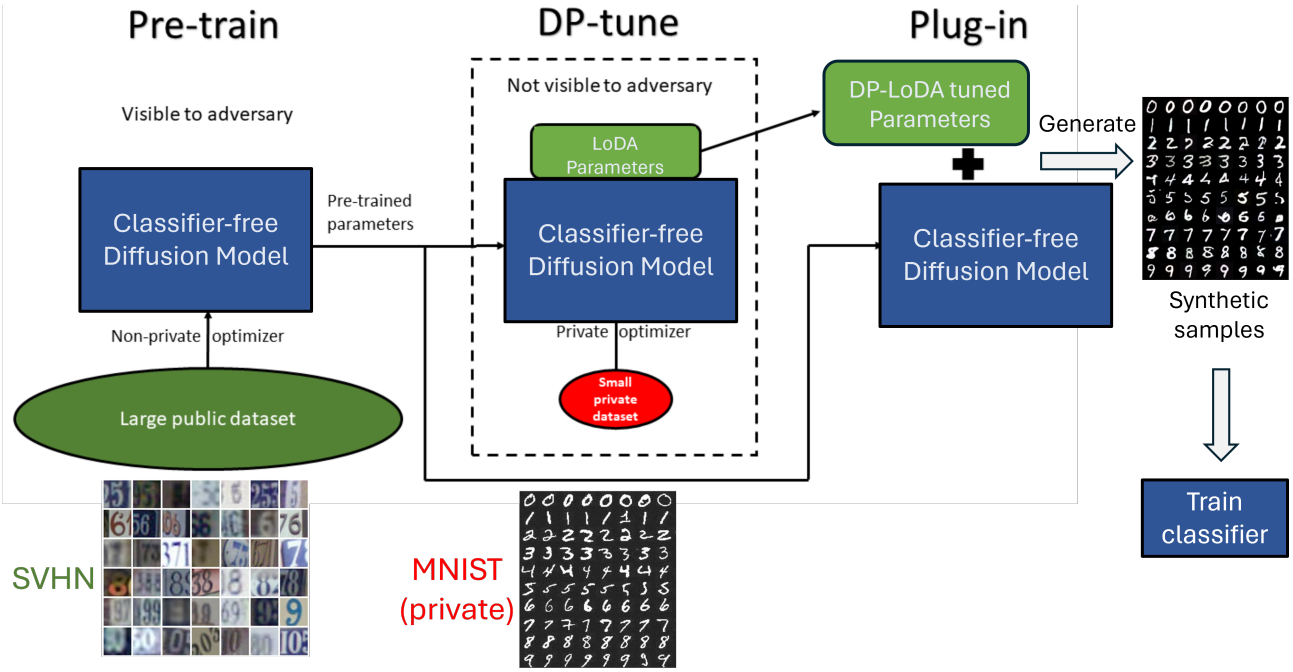

Figure 1. An illustration of the proposed DP-LoDA framework for training a downstream classifier with differential privacy. Figure is modified from Yu et al. (2022).

attention module only[2]. It is an interesting future direction to investigate DP-LoDA combined with Latent Diffusion Models. Note that the high-level idea of combining DP with PEFT was proposed in Yu et al. (2022) for LLMs, but with the purpose of only fine-tuning the LLM. No synthetic samples were generated for training a downstream classifier in that work.

It is also worth mentioning that there is a large body of work using conventional generative models, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) or Variational AutoEncoders (VAEs) (Kingma & Welling, 2019) to generate differentially private synthetic samples (Xie et al., 2018; Tantipongpipat et al., 2020; Jordon et al., 2019; Chen et al., 2020; Harder et al., 2021a; Papernot et al., 2021; Cao et al., 2021; Liew et al., 2022; Harder et al., 2023; 2021b; Acs et al., 2018; Jiang et al., 2022; Pfitzner & Arnrich, 2022). We compare with two recent approaches (DP-MEPF and DP-MERF) in our experiments.

In the following sections, we first introduce the overall framework of DP-LoDA, and the details of LoDA for convolutional layers. We then conduct experiments on benchmark datasets and compare with state-of-the-art methods. Finally, we conclude and discuss future directions as well as potential broader impact.

[2]In the latest version of Lyu et al. (2023), we notice the authors tried LoRA fine-tuning, but the FID score of generated synthetic samples is worse than fully fine-tuning the attention module, and the downstream classification accuracy is not reported for LoRA, and remains unclear.

2. Efficiently fine-tuning of diffusion models via DP-LoDA

The overall framework of the proposed **DP-LoDA** method is illustrated in Fig. 1. First, the classifier-free diffusion model (Ho & Salimans, 2021) is pre-trained on some public dataset. Next, LoDA adapters are attached to the Diffusion Model and are differential privately fine-tuned (via DP-SGD) on a smaller private dataset. The original parameters of the pre-trained diffusion models are frozen during this process. Then, the DP-LoDA fine-tuned, classifier-free diffusion model is used to generate synthetic samples with class conditioning. Finally, the downstream classifier is trained with the generated synthetic samples.

2.1. Illustration of diffusion model and LoDA adapter

The classifier-free diffusion model (Ho & Salimans, 2021) that we used for image generation is adapted from a public codebase[3]. Its U-Net structure is illustrated in Fig. 2, which contains 21 attention modules and 22 ResNet mod-

[3]https://github.com/coderpiaobozhe/classifier-free-diffusion-guidance-Pytorch

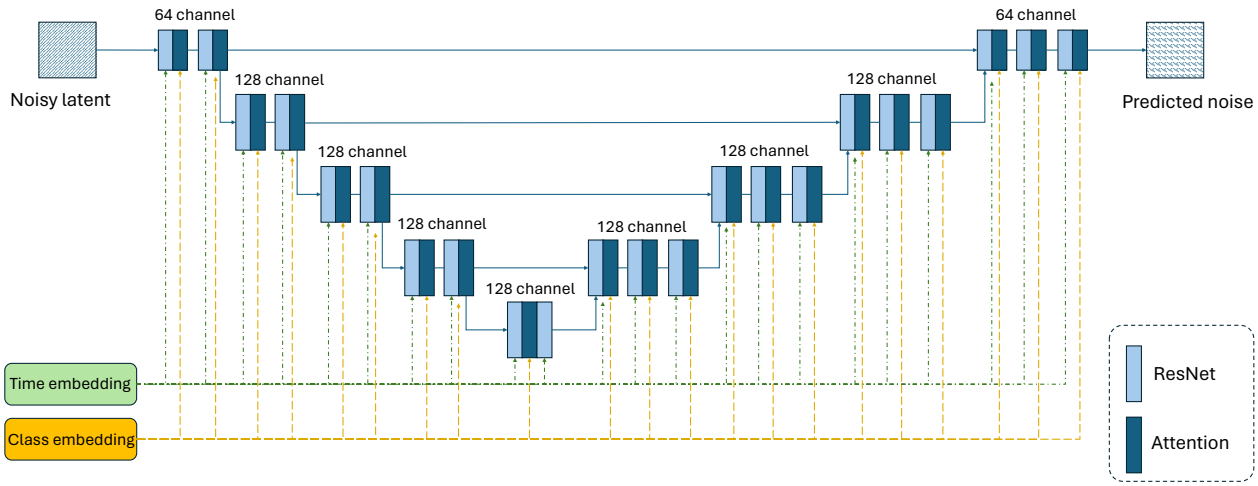

Figure 2. Classifier-free diffusion model structure.

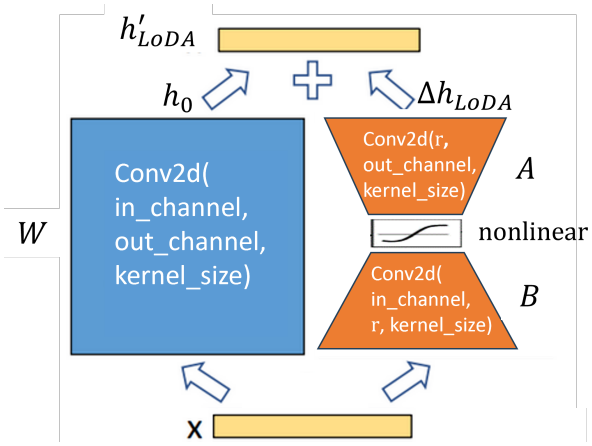

Figure 3. Low-Dimensional Adaptation (LoDA) for convolutional layer.

ules, where most of the layers (including the query/value/key projection layers) are convolutional layers. The total number of parameters in the U-Net base model is 10.42M. Following the idea of Low-Dimensional Adaptation (LoDA), we attach a low-dimensional convolutional adapter in parallel with the original convolutional layer. As illustrated in Fig. 3, it consists of 1) a convolutional layer A which has output channel size r that is much smaller than the input channel size; 2) one or more nonlinear function(s); 3) a convolutional layer B which maps the r channel latent back to the original output channel size. During the LoDA fine-tuning, the original convolutional layer W is frozen and only the LoDA adapter is fine-tuned. As the value of r is much smaller than the original input/output channel size, the number of tunable parameters in A and B is much smaller than the original convolutional layer W.

3. Experiments

We first compare the proposed method with the widely used standard DP method, which simply trains the classifier using DP-SGD. We also compare with several state-of-the-art generative model based approaches, such as DP-Diffusion (Ghalebikesabi et al., 2023b) that fully fine-tunes the Diffusion Model, DP-LDM (Lyu et al., 2023) that fine-tunes the attention module of the Latent Diffusion Model, DP-MEPF (Harder et al., 2023), and DP-MERF (Harder et al., 2021a). As a reference, we also report the accuracy of the classifier when it is trained on private data without any privacy protection, denoted as "No DP".

We consider MNIST (LeCun & Cortes, 2010) and CIFAR-10 (Krizhevsky & Hinton, 2009) as the private datasets, and leverage SVHN (Netzer et al., 2011) and ImageNet32 (Deng et al., 2009) as the corresponding public datasets. We also consider the setting that each class of CIFAR-10 has only 1% of the original training samples, to simulate the situation of a small private dataset in practice, which may often be the case for private medical or biosignal datasets, for example.

For our proposed DP-LoDA, the dimension r is set to 4 in all of the experiments. We use LeakyReLU with negative slope of 0.1 as the nonlinear function between A and B in LoDA. Our implementations of the LoDA adapter for the convolutional layer is modified from a public codebase[4], and the DP accounting tool we used is from Bu et al. (2023). We fine-tune for 200 epochs on the CIFAR-10 training set, and 100 epochs on the MNIST training set. We report the downstream classification accuracies on the the true CIFAR-10/MNIST testing set, under various privacy levels.

[4] https://github.com/cloneofsimo/lora

3.1. Experiments on CIFAR-10

In this first set of experiments, we use the training set of CIFAR-10 (Krizhevsky & Hinton, 2009) as private dataset, and leverage ImageNet32 (Deng et al., 2009) as the corresponding public dataset. However, ImageNet32 has 1000 classes, and the 10 classes of CIFAR-10 are not exactly covered by the ImageNet32 classes. Motivated by Huang et al. (2021), which identifies the ImageNet classes that are close to CIFAR-10 classes, we identified 10 classes of ImageNet32 that are similar to CIFAR-10 classes, listed in Table 1.

Table 1. The 10 ImageNet classes similar to the CIFAR-10 classes.

CIFAR-10 Class	ImageNet32 Class
Airplane	Airliner
Car	Wagon
Bird	Humming Bird
Cat	Siamese Cat
Deer	Hartebeest
Dog	Golden Retriever
Frog	Tailed Frog
Horse	Sorrel Horse
Ship	Container Ship
Truck	Trailer Truck

However, for those corresponding 10 ImageNet32 classes, there are only 13k training samples in total, and we empirically found that they are not sufficient to train the Classifier-free Diffusion Model. Since ImageNet32 has more than 1281k training samples overall, we propose the following pre-training procedure to leverage the entire ImageNet32 dataset:

Step 1: Pre-train the diffusion model without class conditioning using all ImageNet32 samples;

Step 2: Continue pre-training using the corresponding 10 ImageNet32 classes only, with class conditioning.

Fig. 4 in the Appendix shows the images generated by the Diffusion Model following Step 1, after pre-training for 300 epochs on all ImageNet32 training samples without class conditioning. Fig. 5 in the Appendix shows the images generated by the Diffusion Model following Step 2, after further pre-training of 1200 epochs with class conditioning on the corresponding 10 ImageNet32 classes. We can see that Step 2 is very effective at training class conditioning.

3.1.1. CIFAR-10 AS THE PRIVATE DATA

The full training set of CIFAR-10 has 50K samples. DP-LDM generates 50K synthetic samples for training the classifier, while DP-Diffusion generates 1 million synthetic samples. In our proposed method, we follow the conventions of DP-Diffusion by generating 1 million synthetic samples. Table 2 shows the test accuracies (on the true CIFAR-10 testing set) of ResNet9 by each DP training method (with access to the full CIFAR-10 training set). Various levels of (ϵ, δ)-DP are reported, with δ fixed at 10^{-5}.

Table 2. Test accuracies (on the true CIFAR-10 testing set) of ResNet9 by each DP training method (with access to the full CIFAR-10 training set), under different levels of privacy cost (ϵ, δ), with δ fixed to be 10^{-5}. The values of DP-LDM, DP-MEPF and DP-MERF are from Lyu et al. (2023).

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	

First of all, we can see that DP-LDM, DP-Diffusion, and the proposed DP-LoDA significantly outperform DP-SGD, DP-MERF and DP-MEPF. When ϵ is relatively small, the proposed DP-LoDA outperforms DP-LDM. There is still some gap between parameter-efficiently fine-tuned Diffusion Model and fully fine-tuned Diffusion Model. Note that the classifier accuracy can be further improved by using the Wide ResNet (WRN-40-4) architecture instead of ResNet9, as demonstrated in Lyu et al. (2023). Further, Ghalebikesabi et al. (2023a) showed that the accuracy can be further improved by using the augmentation multiplicity technique (De et al., 2022). Finally, as a reference, Fig. 6 in the Appendix shows the images generated by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on the full CIFAR-10 training set.

3.1.2. 1% OF CIFAR-10 AS THE PRIVATE DATA

In some cases of private datasets, such as medical or biosignal data, the number of training samples may be quite limited. To simulate this scenario, we conduct an experiment with only 1% of the CIFAR-10 training set as the private dataset.

Table 3 shows the test accuracies (on the true CIFAR-10 testing set) of ResNet9 using each DP training method (with access to 1% of the CIFAR-10 training set), under the common DP settings $\epsilon = 1, 10$ and $\delta = 10^{-5}$. First of all, we can see that DP-SGD works poorly under this setting. The gap between the accuracies of DP-Diffusion and DP-LoDA becomes very small, which aligns with the findings in the LLM literature that PEFT approaches perform similarly as full fine-tuning when the fine-tuning dataset is small. More interestingly, *even at a very small privacy cost of*

Table 3. Test accuracies (on the true CIFAR-10 testing set) of ResNet9 by each DP training method (with access to 1% CIFAR-10 training set), under different levels of privacy cost (ϵ, δ), with δ fixed to be 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	

Table 4. Test accuracies (on true MNIST testing set) of a CNN classifier by each DP training method (with access to the full MNIST training set). The value of DP-LDM is from Lyu et al. (2023).

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

$(\epsilon = 1, \delta = 10^{-5})$, *both DP-Diffusion and DP-LoDA outperform the accuracy of the classifier trained without DP* ("No DP"). Finally, as a reference, Fig. 7 in the Appendix shows the generated images by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ using 1% of the CIFAR-10 training set. Even with such a very limited private dataset, the quality of the generated synthetic samples are still good.

3.2. Experiments on MNIST

In this section, we consider the 60K sample training set of MNIST (LeCun & Cortes, 2010) as the private dataset, and we use SVHN (Netzer et al., 2011) as the public dataset. Note that SVHN has exactly the same classes as MNIST, and therefore we directly pre-train the diffusion model with class conditioning. Fig. 8 in the Appendix shows the images generated by the pre-trained Diffusion Model. Fig. 9 in the Appendix shows the images generated by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on MNIST training set. It is clear that DP-LoDA fine-tuning can successfully transfer from SVHN to MNIST.

For the tuned Diffusion Models, we generate 60K synthetic samples for both DP-Diffusion and DP-LoDA methods. We fix $(\epsilon = 10, \delta = 10^{-5})$ and report the test accuracy of different methods in Table 4. DP-Diffusion, DP-LDM, and the proposed DP-LoDA methods significantly outperform DP-SGD, and achieve very good utility. DP-LoDA performs slightly better than DP-LDM and close to DP-Diffusion.

Since SVHN dataset has exactly the same classes as MNIST, as a reference, we also tried directly training the classifier on the SVHN dataset and testing on the MNIST testing set, which yields a classification accuracy of only 60.22%.

4. Conclusion and Future Work

We have proposed and evaluated DP-LoDA for efficiently fine tuning a diffusion model with differential privacy, which is then used to generate synthetic samples for training a downstream classifier. We have evaluated the proposed method on the MNIST and CIFAR-10 datasets, and demonstrated that such efficient fine-tuning can generate useful synthetic samples for training downstream classifiers, while protecting the privacy of the fine-tuning data. Empirical studies on the CIFAR-10 dataset further show that such parameter-efficient method gains more advantage when the private fine-tuning data is limited. An interesting future exploration is to test DP-LoDA on Latent Diffusion Models, which can further improve parameter-efficiency.

Impact Statement

This paper presents work whose goal is to propose and investigate whether the efficient fine-tuning of vision foundation model under Differential Privacy is effective for data privacy protection and useful for training a downstream classifier. There are many positive social impacts in terms of privacy protection and computational efficiency. We do not feel there are any major negative impacts.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Acs, G., Melis, L., Castelluccia, C., and Cristofaro, E. D. Differentially private mixture of generative neural networks, 2018.

Bu, Z., Wang, Y.-X., Zha, S., and Karypis, G. Differentially private optimization on large model at small cost. In *International Conference on Machine Learning*, pp. 3192–3218. PMLR, 2023.

Cao, T., Bie, A., Vahdat, A., Fidler, S., and Kreis, K. Don't generate me: Training differentially private generative models with sinkhorn divergence. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wal-

lace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chen, D., Orekondy, T., and Fritz, M. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

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:2301.13188*, 2023a.

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, 2023b.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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.

Harder, F., Adamczewski, K., and Park, M. DP-MERF: Differentially private mean embeddings with random features for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pp. 1819–1827. PMLR, 2021b.

Harder, F., Jalali, M., Sutherland, D. J., and Park, M. Pre-trained perceptual features improve differentially private image generation. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=R6W7zkMz0P.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Huang, H., Ma, X., Erfani, S. M., Bailey, J., and Wang, Y. Unlearnable examples: Making personal data unexploitable. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=iAmZUo0DxC0.

Jiang, D., Zhang, G., Karami, M., Chen, X., Shao, Y., and Yu, Y. DP^2-VAE: Differentially private pre-trained variational autoencoders. *arXiv preprint arXiv:2208.03409*, 2022.

Jordon, J., Yoon, J., and van der Schaar, M. Pate-gan: Generating synthetic data with differential privacy guarantees. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010.

Liew, S. P., Takahashi, T., and Ueno, M. PEARL: data synthesis via private embeddings and adversarial reconstruction learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=M6M8BEmd6dq.

Liu, J., Koike-Akino, T., Wang, P., Brand, M., Wang, Y., and Parsons, K. LoDA: Low-dimensional adaptation of large language models. *NeurIPS'23 Workshop on on Efficient Natural Language and Speech Processing*, 2023.

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

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, Spain, 2011.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, U. Differentially private hyperparameter tuning for deep learning. *arXiv preprint arXiv:2102.06387*, 2021.

Pfitzner, B. and Arnrich, B. DPD-FVAE: Synthetic data generation using federated variational autoencoders

with differentially-private decoder. *arXiv preprint arXiv:2211.11591*, 2022.

Tantipongpipat, U., Liu, Z. S., and Cui, Y. Differentially private conditional gans. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 15–24. IEEE, 2020.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H., Kamath, G., Kulkarni, J. J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. In *ICLR 2022*, April 2022.

A. Appendix

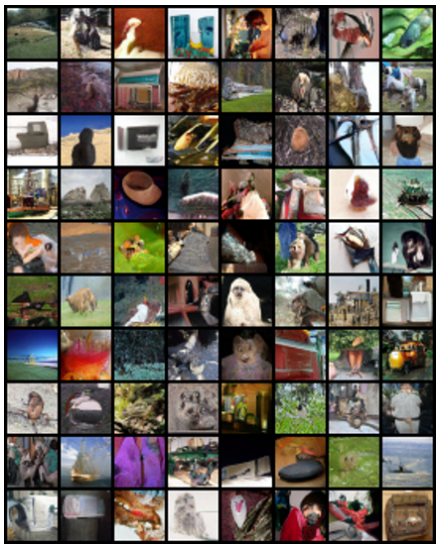

Figure 4. Generated images by Diffusion Model from Step 1 of pre-training (epoch 300).

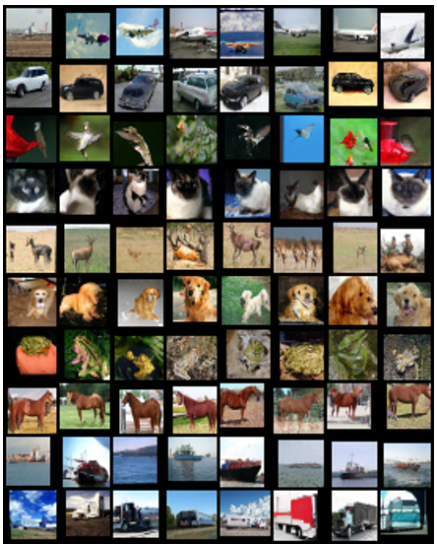

Figure 5. Generated images by Diffusion Model from Step 2 of pre-training (epoch 1200).

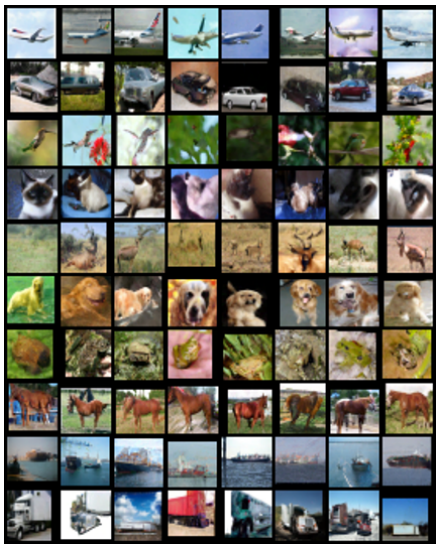

Figure 6. Generated images by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on full CIFAR-10 training set (epoch 200).

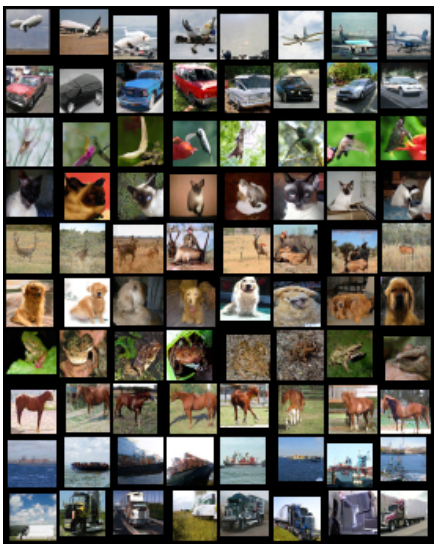

Figure 7. Generated images by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on 1% CIFAR-10 training set (epoch 200).

Figure 8. Generated images by Diffusion Model after pre-trained 720 epochs on SVHN dataset.

Figure 9. Generated images by Diffusion Model after DP-LoDA fine-tuning with $(\epsilon = 10, \delta = 10^{-5})$ on MNIST training set (epoch 100).