Supplementary Material: Decouple-and-Sample: Protecting sensitive information in task agnostic data release

Abhishek Singh, Ethan Garza, Ayush Chopra,
Praneeth Vepakomma, Vivek Sharma, and Ramesh Raskar

Massachusetts Institute of Technology
abhi24@mit.edu

1 Reproducibility

All of our experiments are implemented using PyTorch [6] and conducted using NVIDIA 1080 Ti GPUs. We use Adam optimizer [3] for training all of the neural networks. Throughout all the experiments, we use a ResNet-50 [1] architecture for prediction related tasks and transpose convolution-based architectures for generative tasks. Unless noted otherwise, we use $\epsilon = 1.0$ for all the experiments. The value of ϵ is divided as 0.3 for the mean and 0.7 for the variance parameter of DP-sampling mechanism. For all of our evaluations we choose $k = 8$ and $m = 32$. Our training pipeline requires training base model for all techniques for a given set of hyper-parameters. This base model is used to obtain sanitized datasets. We use sanitized datasets to train a separate utility model and adversary model. Each utility and adversary model is trained independently for every trade-off parameter corresponding to every baseline. This results in a privacy-utility trade-off curve for each technique. We use $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 100, \alpha_4 = 1, \beta = 5$ for all sanitizer experiments except the ones where these parameters were changed to obtain a trade-off. We keep $\alpha_3 = 100$ since the distance correlation is typically a small quantity and requires scaling in order to influence the overall loss term.

1.1 Extended Results

We analyze the visual results for different techniques for a given sample (Fig 1) and different samples for our technique (Fig 2). We note that some of the techniques such as TIPRDC [4], while applicable for the sanitization task, do not release their output in the image domain therefore qualitative results can not be compared with them. Results demonstrate that *sanitizer* provides improved sample quality in comparison to other techniques. The key reason for improved sample quality is that sanitizer randomizes the sensitive attribute instead of removing it. We note that the sample quality can be further improved by extending our work for hierarchical VAEs that are known for high quality image synthesis.

For multiple sensitive attributes, we evaluated ours and strongest baseline TIPRDC on CelebA. We use (*gender*, *smiling*) as sensitive and (*mouth-open*, *high-cheekbones*) as utility. *Sanitizer* gets AuC of 0.452 and *TIPRDC* gets 0.434. A slight reduction in performance of both techniques due to multiple sensitive attributes, however *sanitizer* performs better.

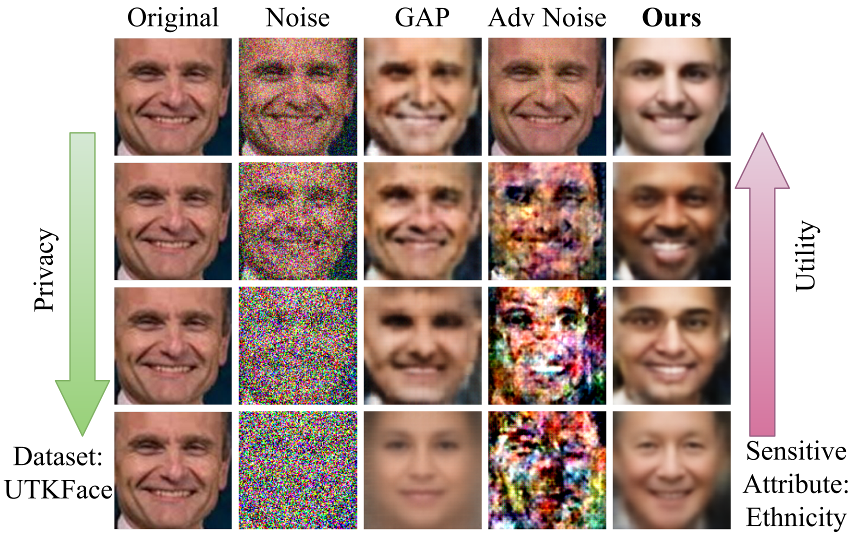

Fig. 1: Comparison of visual quality of different techniques.

1.2 Datasets

UTKFace [7] consists of 20,000 face images. We use the cropped and aligned version of the dataset and generate a random split of 90% – 10%, training and testing data. The dataset has "ethnicity", "gender", and "age" as categorical labels. For our experiments, we keep the sensitive attribute as ethnicity which has 5 unique labels and due to class imbalance, the best possible performance without access to the image is 44%. We use "gender" as the utility attribute for the evaluation.

CelebA [5] is a large scale dataset of 202,599 celebrity face images 10,177 unique identities, each with 40 binary attribute annotations. For our experiments, we define gender as the sensitive attribute. We use "mouth open", "smiling" and "high cheekbones" as the utility attribute for evaluation evaluation.

FairFace [2] dataset consists of 108,501 images, with three different attributes "ethnicity", "gender", and "age". The dataset was curated to overcome the class imbalance in ethnicity present in the majority of the existing Faces datasets. We use "ethnicity" as a sensitive attribute. We use "gender" as the attribute for the utility evaluation.

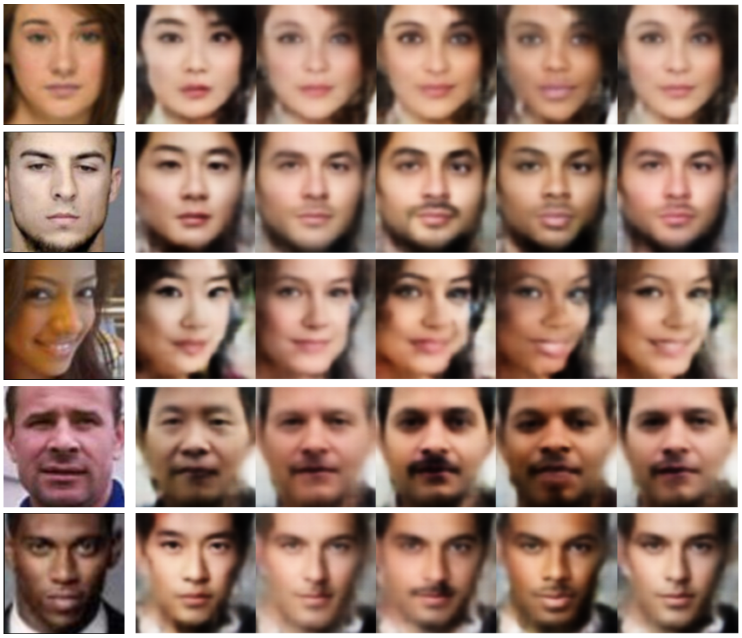

Fig. 2: Comparison across different samples: We sample results for different image samples multiple times to qualitatively evaluate the performance of sanitizer. The first column represents the original image. We use “ethnicity” as the sensitive attribute in this setup and arrange the columns accordingly. Note that a majority of the sampled images preserve attributes independent of the chosen sensitive attribute.

Dataset and benchmark release: To encourage further work in sanitization techniques, we create a benchmark dataset of 1-million sanitized images by applying baseline and our technique on the existing datasets. This will enable rigorous evaluation of different mechanisms and their privacy-utility trade-off. The benchmark will serve as a continuously improving evaluation pipeline for researchers to study both attack or defense techniques. Current implementation of the work, including preliminary dataset can be found at https://github.com/splitlearning/sanitizer. We plan to release the complete benchmark and datasets after receiving feedback from reviewers.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
2. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age. arXiv:1908.04913 (2019)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
4. Li, A., Duan, Y., Yang, H., Chen, Y., Yang, J.: Tiprdc: Task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations. In: ACM SIGKDD (2020)
5. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August (2018)
6. Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
7. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5810–5818 (2017)