# Differential Privacy vs Detecting Copyright Infringement: A Case Study with Normalizing Flows

Saba Amiri [1] Eric Nalisnick [1] Adam Belloum [1] Sander Klous [1] Leon Gommans [2]

## 1. Introduction

Generative Models (GMs) are becoming increasingly popular for synthesizing data for a diverse range of modalities. One common aspect of training GMs is that they need large amounts of training data. The era of big data has made such huge troves of data available to us. However, utilizing such a diverse range of datasets to train GMs has its own technical and ethical challenges, one of the most prominent ones being privacy. Data used to train GMs could potentially contain personal and sensitive information, e.g. social security number in the training set of a large language model. Moving beyond identifying features, more nuanced issues could arise such as a model learning the specific style of an artist to the point of infringing on their intellectual rights or it could memorize copyrighted material such as books and reiterate them (Carlini et al., 2023; Zhang et al., 2022; Hu & Pang, 2023). Thus, we need to make sure the GMs are learning as much as possible about the population to generate high quality results while learning as little as possible about discriminating and/or protected *features* of individual *members* of the training data - both definition of a *feature* and a *member* subjective to the specifics of the problem.

Differential Privacy (DP) (Dwork et al., 2014) has emerged as the de facto standard for privacy preservation in machine learning. The adoption of DP entails constraining the influence that the inclusion or exclusion of individuals, or groups thereof, within the training dataset may have on the model's output. This is typically achieved through introduction of some sort of stochasticity during the training process, with noise addition being the most prevalent mechanism. The selection of an appropriate noise distribution and precise calibration of DP parameters are crucial for attaining desired outcomes while ensuring stringent privacy guarantees.

In this work, we present early results for a new method for making Normalizing Flows, a powerful family of GMs, differentially private without adding noise. Based on the definition of pure $\varepsilon$-DP, we show that using the ability of flow-based models for exact density evaluation we can add differential privacy to flows by limiting their expressivity instead of adding noise to them.

We show, through this methodology, that enforcing privacy can lead to the obfuscation of private material (denoted as such with a watermark). This implies the benefits of privacy preserving methods for removing discriminant features. But it could have negative side effects: preserving privacy can potentially hide the fact that a model was trained on protected material. We show that when features identifying copyrighted materials are non-discriminative and prevalent in the dataset, the DP model is still able to capture them.

## 2. Results

We train a volume preserving non linear independent components estimation (NICE) flow (Dinh et al., 2014) on MNIST dataset. We use 50000 samples for training and 5000 samples for validation. We train all models for 20 iterations with a batch size of 256 and the total privacy budget spent for DP flows is $\epsilon = 0.32$.

### 2.1. Utility

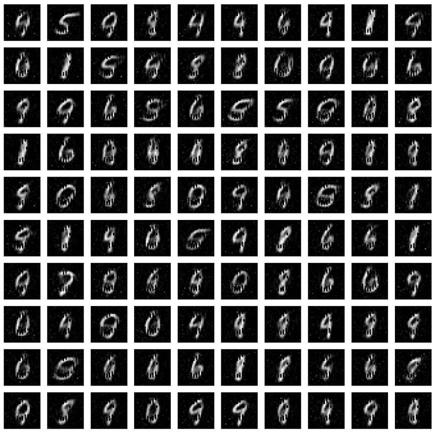

*Figure 1.* Samples generated by the DP model with $\epsilon = 0.32$

[*] Equal contribution [1] Informatics Institute, Universiteit van Amsterdam, Amsterdam, The Netherlands [2] KLM - Air France. Correspondence to: Saba Amiri <s.amiri@uva.nl>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

We train our model with and without DP and visualize samples generated by the model. Figure 1 shows samples generated with the DP flow. We can see that training the flow using out method in high privacy regime with a low privacy budget still yields good results.

### 2.2. Privacy for Watermarked Samples

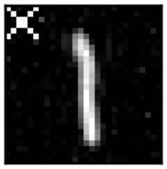

*Figure 2.* Sample digit with watermark X on the top left corner

To test the capabilities of our method not to learn discriminative features while capturing common patterns such as watermarks, we randomly add a distinct cross-shaped watermark (Figure 2) to 25% and 100% of the samples in the training data. The goal is for the model to still generate coherent digits with the watermark in the case of the 100% dataset while not learning and generating watermarks for the 25% dataset as that could be counted as a discriminant feature.

Figure 3 shows the average pixel values for DP and non-DP flows on the watermark sign for 0%, 25% and 100% datasets. We observe that the difference in average values of 0% and 100% datasets is negligible, while for the 25% dataset there is a noticeable difference between them, meaning the non-DP flow is generating watermarked samples while the pixel values for the DP flow remain almost the same as the DP flow trained on 0% dataset. This can be interpreted as the DP model not removing watermarks while removing personal and identifying features from individual samples.

## 3. Conclusion

In this work we present early results of our research on a novel method to add differential privacy to flow-based GMs. We demonstrate a simple case in which the model doesn't learn member-level discriminant features while learning both the target density and the prevalent feature - in this case a watermark. Therefore, we are lead to conclude that privacy preserving methods, especially DP, while not the ultimate solution and admittedly highly dependent on the type of model, definitions of privacy level and prevalent features and the training pipeline, are a step in the right direction for preserving privacy while letting the model learn population-level identifying features such as watermarks which would help regulate the use of copyrighted material to train generative models.

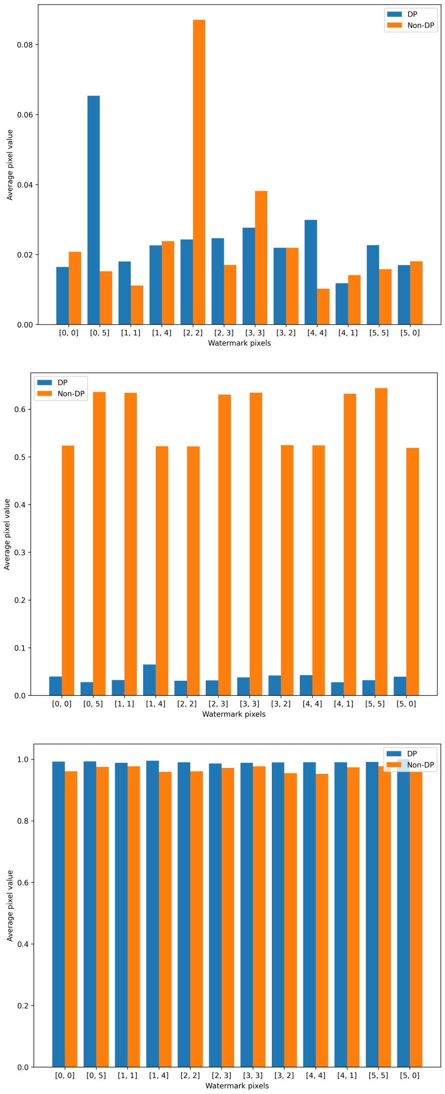

*Figure 3.* Average pixel values for Top: 0% dataset, Middle: 25% dataset and Bottom: 100% dataset

## References

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Hu, H. and Pang, J. Membership inference of diffusion models. *arXiv preprint arXiv:2301.09956*, 2023.

Zhang, Z., Yan, C., and Malin, B. A. Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125:103977, 2022.