ScuDo
Scuola di Dottorato – Doctoral School
WHAT YOU ARE, TAKES YOU FAR

XXXVI Cycle

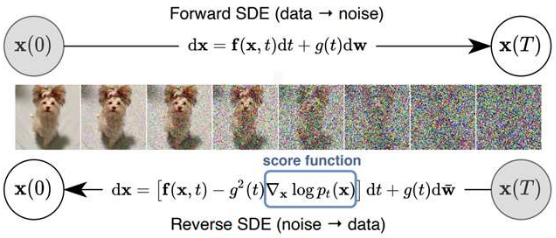
Privacy-preserving image generation with diffusion models

Nikhil Jha

Supervisors: Prof. Marco Mellia, Martino Trevisan, Ph.D.

Research context and motivation

- Diffusion models represent a trending alternative to adversarial frameworks in the **generation of synthetic data**, e.g., images. Diffusion models work over a Markov chain. The **forward diffusion process** gradually corrupts a training point x_0 with noise, until after T steps it becomes equivalent to a Gaussian distribution. Data is generated following the Markov chain in the opposite direction, the **backward diffusion process**: going from T to 0, we move from noise to generated, synthetic data. Song et al. (2021) state that the two processes can be regulated by two **stochastic differential equations** (SDEs), related one to the other by a **score** (i.e., the gradient of the log probability density with respect to data). Learning the score with a neural network, one can then sample following the Markov chain to generate new, synthetic data.



- However, generative models (as any other data-driven model) face the risk of leaking the privacy of the users whose data have been employed as training data. Two types of attacks are possible: in **membership attacks**, one with access to the model could be able to trace back the presence of some user's data in the training set, while in **reconstruction attacks** the model output could be exploited to retrieve an intelligible version of a target training point.

- **Differential Privacy** (DP) is the *de-facto* standard approach to tackle privacy-related risks in Machine Learning. While DP covers a wide set of algorithms, its core concept stands in the injection of noise during the training process: privacy leaks from the model outputs and the model parameters are thus mathematically prevented. Abadi et al., 2016, conceived the seminal framework to bring privacy-preserving techniques into deep learning, based on the **DP-Stochastic Gradient Descent** (DP-SGD), proving that injecting a controllable amount of noise in the backpropagation phase while training a neural network offers rigorous privacy guarantees.

Addressed research questions/problems

- Existing privacy-preserving image generation models **do not meet the standard** in output resolution and quality that non-private models have set in recent years. Most of the studies fail to obtain satisfying result when facing more complex datasets than MNIST or F-MNIST.

- The ambition of this research project is to move from one of the best-performing generative diffusion model, such as the **score-based generative modeling via SDE** proposed by Song et al., 2021, and to use it as the basic framework for a privacy-preserving learning. This way, we aim towards two separate goals: to improve the landscape of privacy-preserving image generation models as a whole, and to offer a privacy-preserving competitive alternative to score-based generative modeling via SDE.

- The input to the score-learning neural network of Song et al., 2021, are an image and its noise-corrupted version. A side aspect of out research project focuses on understanding whether this noise-based procedure already in place could be linked to differentially-privacy-related privacy guarantees.

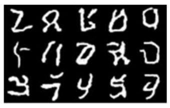
Submitted and published works

- Jha, N. *et al.*, "*z-anonymity: Zero-Delay Anonymization for Data Streams*", IEEE International Conference on Big Data, Atlanta, GA, USA, 2020, pp. 3996 – 4005.
- Jha, N. *et al.*, "*A PIMS Development Kit for New Personal Data Platforms*", IEEE Internet Computing, vol. 26, issue. 3, 2022, pp. 79-84.
- Jha, N., *et al.*, "The Internet with Privacy Policies: Measuring The Web Upon Consent", ACM Transactions on the Web, vol. 16, issue. 3, article no. 15, pp. 1-24.
- Jha, N., *et al.*, "*Practical Anonymization for Data Streams: z-anonymity and relation with k-anonymity*", under review at Performance Evaluation, 2021.

Novel contributions

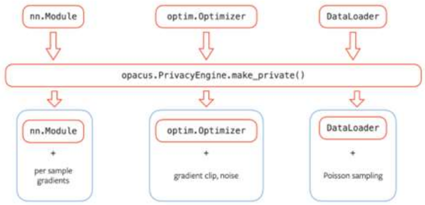
- To the best of our knowledge, our effort is the first one to combine DP-SGD and diffusion models. To do this, we inject noise to the gradient learnt during the backpropagation process. The amount of privacy guaranteed by this mechanism is proportional to the amount of noise injected: the more noise added, the more the privacy obtained.

- Although presenting multiple advantages with respect to adversarial networks, diffusion models are high-demanding in terms of time and computing resources. Before moving to larger and more complex datasets, we started to look for preliminary results on MNIST. This is a sample of the results, with the privacy budget set at $\epsilon = 10$:



Adopted methodologies

- We relied on the Python library Opacus to perform noise injection in a convenient way, both from a computational and a programming point of view.



- Opacus comes as a convenient wrapper to your standard PyTorch training procedure, by transparently implementing key features to perform privacy-preserving analytics in deep learning.

Future work

- The preliminary results on MNIST show that there is no silver bullet when it comes to provide a differentially-private solution to generative models. Previous research suggests indeed that privacy learning should be treated independently and its techniques fine tuned according to this specific use case. In future work, we aim at moving from vanilla-wrapping diffusion models with privacy-preserving mechanisms to improve the results of the model by a more in-depth understanding of better practices for private learning.

- During the last months of research, we came across the work of Jia-Wei et al., 2022, which propose a novel approach in training energy-based models, called DP-GEN. In DP-GEN, the stochastic element is added not when learning the gradient of the network, but in the choice of the training points themselves. While in standard diffusion models the training point is coupled to a noisy version of it to learn how to reverse the corruption, DP-GEN proposes a stochastic coupling between images, providing privacy. We are currently in touch with the authors of the paper to speed up the transition of DP-GEN to score-based generative modeling with SDE.

List of attended classes

- 03QTIIU – Mimetic Learning (26/01/2021, 20h)
- 01PJMRV – Etica informatica (03/05/2021, 20h)
- 01UNWRV – Intercultural & interpersonal management (03/06/2021,8h)
- 01UJBRV – Adversarial training of neural networks(03/06/2021, 15h)
- 01UNXRV – Thinking out of the box (15/07/2021,1h)
- 01TSBRV – Scienza dei dati applicata alle reti complesse (23/07/2021, 20h)
- 01RISRV – Public speaking(09/09/2021, 5h)
- 01SCVIU – Data analytics for science and society (30/09/2021,15h)
- 01SYBRV – Research integrity (09/03/2022, 5h)
- BigSec – Big Data Security (09/12/2021, 21h) @EURECOM
- MobMod – Mobility Modeling (16/12/2021, 21h) @EURECOM
- WebSem – Semantic Web and Information Extraction technologies (14/04/2022, 21h) @EURECOM

POLITECNICO DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering