Supervisors:

Lydia Y. Chen
Jeroen Galjaard
Chaoyi Zhu

Ellipse: Robust and imperceptible watermarking for tabular diffusion models

Author:
Toma Volentir
T.Volentir@student.tudelft.nl
TU Delft

1. Introduction, motivation, background

Motivation:

Tables are widely used in science and research for compact, logical data storage. Diffusion models have improved the quality of generated tabular data but raise concerns about misuse and copyright. Therefore, controlling and monitoring data from diffusion models is essential for harm mitigation and protection.

Research question:

How to adapt Tree-Ring[1] watermarking for synthetic tables generated by diffusion models, having negligible impact on synthetic data quality, is imperceptible by humans, but robustly detectable by model owner?

Background:

Diffusion model: A diffusion model consists of an *encoder*, a *diffusion pipeline*, and a *decoder*. The *encoder* converts clean data into vectors called latents. The *decoder* takes latents and converts them back to data. The *diffusion pipeline* simulates two Markov chains: a forward and a backward process. The forward process corrupts latents with Gaussian noise until they become pure noise. The backward process uses a neural network to remove the noise and reconstruct clean data.

Tree-Ring: Tree-Ring embeds a pattern into the Fourier space of latents before diffusion. The Fourier space maps data values to frequencies, making the watermark invisible. A circular mask is applied on the latents, and the key is overlaid. Tree-Ring offers 3 patterns:

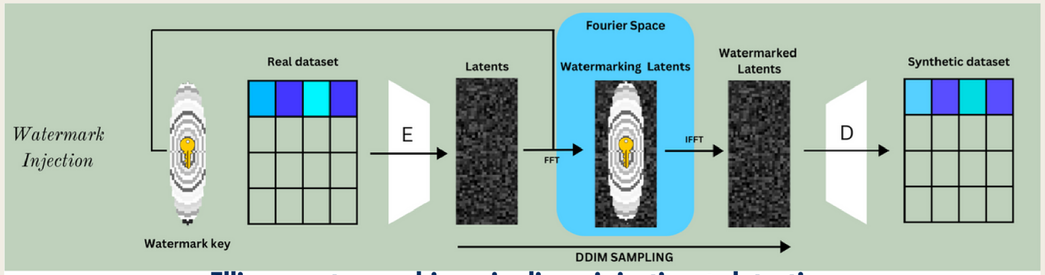
- rand - random values from Gaussian distribution
- zeros - array of zeros
- ring - multiple concentrical ellipses with random values from Gaussian distribution

Problem:

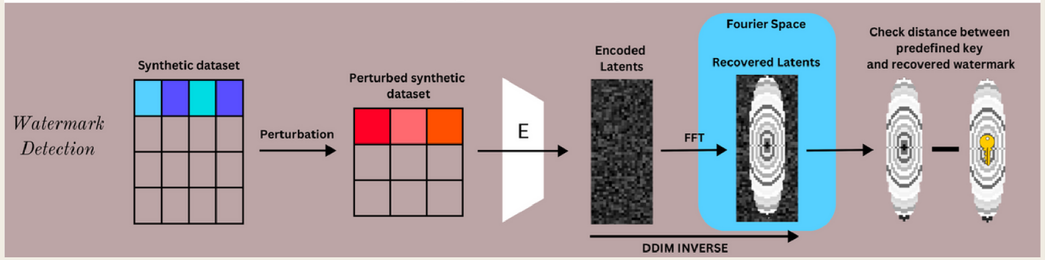
Tree-Ring is designed for square-shaped image data, but tabular data is rectangular shaped => original Tree-Ring is not suitable

Our proposal:

Ellipse - a generalisation of Tree-Ring for rectangular shaped data, by using oval shaped watermark instead of circle



Ellipse watermarking pipeline - injection + detection



2. Methodology

- Implement invertible sampling process for latents - **Denoising Diffusion Implicit Models[2] (DDIM)**
- Adapt the shape of Tree-Ring watermark from circle to oval
- Generate a watermark patch using one of the patterns (*rand, zeros* or *ring*)
- In Fourier space, apply watermark patch over latents, then return to latent space
- Sample new data using DDIM
- To detect, we invert data using DDIM inverse, check if the L1 distance between the recovered latents and the generated watermark patch is below an empirically chosen threshold

3. Evaluation

Experimental Setup:

- **4 datasets:** Abalone, Adult, Default, Diabetes
- **5 metrics for quality and detection:**
  - data resemblance - **quality**
  - machine learning efficiency (MLE) - **quality**
  - data discriminability - **quality**
  - area under the ROC curve (AUC) - **detection**
  - true positive rate (TPR) when the false positive rate (FPR) is 1% - **detection**
- **4 attacks:**
  - **numerical skew**: add Gaussian random noise to numerical values
  - **categorical skew**: replace categorical values with values from same column
  - **row deletion**: delete a part of total rows
  - **column deletion**: delete a part of total columns
- **720** trials for data quality, **2.400** trials for detection, **19.200** trials for attacks detection

	Circle watermark	Ellipse - oval watermark
Impact on quality	12.46%	**3.50%**
Detection efficiency - Clean	89.90%	**90.33%**
Detection efficiency - Attacked	84.13%	**90%**

4. Conclusions and Future Work

- **Ellipse offers a negligible drop in data quality while still being detectable with avg. AUC of 0.9 out of 1.**
- **Ellipse's oval watermark performs better than a circular watermark on all metrics, having the best relative performance on data quality**
- **Ellipse performs better on binary classification task datasets, with a small number of columns (<15)**

- **Limitations**
  - Only verifiable by the model owner, but this offers another layer of robustness
  - Requires an invertible diffusion process, which might bring lower data quality
  - Only tested on Tabsyn[3], a state-of-the-art score-based diffusion model
- **Future work**
  - Implementation of exact diffusion inversion[4] for better detection efficiency
  - Testing on other tabular diffusion models
  - Creating a dedicated watermarking pattern for tabular data

References

[1] Jonas Geiping Yuxin Wen, John Kirchenbauer and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[3] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srini- vasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space, 2023.

[4] Guoqiang Zhang, John P. Lewis, and W. Bastiaan Kleijn. Exact diffusion inversion via bi-directional integration approximation, 2023.