Positive-Unlabeled Diffusion Models for Preventing Sensitive Data Generation

Hiroshi Takahashi, Tomoharu Iwata, Atsutoshi Kumagai,

Yuuki Yamanaka, Tomoya Yamashita

NTT Corporation

In Short

We prevent diffusion models from generating sensitive data by using positive-unlabeled (PU) learning.

Task: Generate a generic middle-aged man (without reproducing Brad Pitt)

Unlabeled data

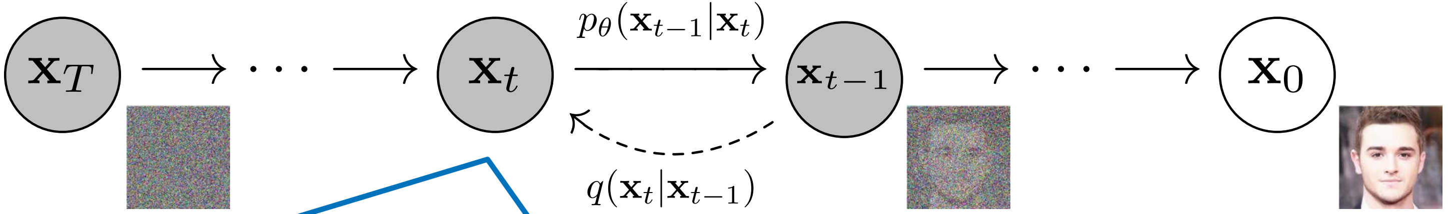
Sensitive data

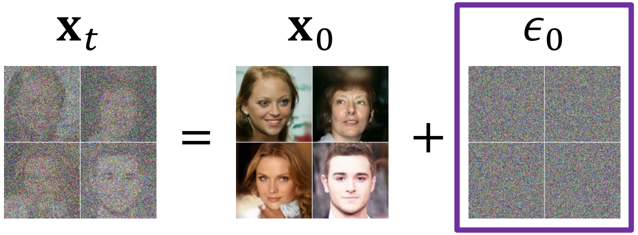
Existing method

Proposed

Diffusion Model

- Diffusion process gradually adds noise to data until it becomes Gaussian noise, and denoising process removes the noise using the estimator $\epsilon_\theta(\mathbf{x}_t, t)$, thereby generating data from noise [1].





The noise estimator $\epsilon_\theta(\mathbf{x}_t, t)$ is trained to minimize the squared error between the predicted noise and the true noise ϵ_0:

$$\ell_\theta(\mathbf{x}_0) \equiv \left\| \epsilon_0 - \epsilon_\theta\left(\mathbf{x}_t, t\right) \right\|^2$$

Problem and Goal

- If the training dataset contains sensitive data, diffusion models may reproduce them.
  - For example, LAION-5B had over 1,000 child abuse images [2].
  - Since the dataset is large, removing all sensitive data is impractical.
  - On the other hand, preparing a small set of such data is feasible.

Our goal

Preventing sensitive data generation using unlabeled training data along with a small amount of labeled sensitive data.

Existing Method: Supervised Approach

- Existing method [3] tries to prevent sensitive data generation by training the estimator to accurately estimate noise for normal data, while failing to estimate noise for sensitive data.
- Since collecting only normal data is difficult, unlabeled data is often treated as normal, although it may contain sensitive data.

Minimize the loss for normal data

$$\min_{\theta} \mathbb{E}_{\underline{p_{\mathcal{N}}}}[\ell_{\theta}(\mathbf{x}_0)]$$

Normal data distribution

and

Maximize the loss for sensitive data

$$\max_{\theta} \mathbb{E}_{\underline{p_{\mathcal{S}}}}[\ell_{\theta}(\mathbf{x}_0)]$$

Sensitive data distribution

intractable in practice

Propose Method: PU Learning Approach

- Based on PU learning, we approximate training on normal data by using unlabeled data and sensitive data as follows:

$$\min_{\theta} \underbrace{\mathbb{E}_{p_{\mathcal{N}}}}_{\text{Normal data distribution}} [\ell_\theta(\mathbf{x}_0)] = \min_{\theta} \frac{1}{1-\beta} \{ \underbrace{\mathbb{E}_{p_{\mathcal{U}}}}_{\text{Unlabeled data distribution}} [\ell_\theta(\mathbf{x}_0)] - \beta \underbrace{\mathbb{E}_{p_{\mathcal{S}}}}_{\text{Sensitive data distribution}} [\ell_\theta(\mathbf{x}_0)] \}$$

$$p_{\mathcal{U}}(\mathbf{x}) = \beta p_{\mathcal{S}}(\mathbf{x}) + (1-\beta) p_{\mathcal{N}}(\mathbf{x}) \Leftrightarrow p_{\mathcal{N}}(\mathbf{x}) = \frac{1}{1-\beta} \{ p_{\mathcal{U}}(\mathbf{x}) - \beta p_{\mathcal{S}}(\mathbf{x}) \}$$

β: proportion of sensitive data (estimable hyperparameter)

Experiments with Stable Diffusion 1.4

Task: Generate a generic middle-aged man (without reproducing Brad Pitt)
Data: 80 unlabeled (20% sensitive) & 20 labeled sensitive

Unsupervised [1]	Supervised [3]	Proposed
		
Non-sensitive rate: 0.80	Non-sensitive rate: 0.84	Non-sensitive rate: **0.99**

Reference

1. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
2. David Thiel, "Investigation Finds AI Image Generation Models Trained on Child Abuse", Stanford Internet Observatory, Dec. 2023.
3. Heng, Alvin, and Harold Soh. "Selective amnesia: A continual learning approach to forgetting in deep generative models." Advances in Neural Information Processing Systems 36 (2024).