Erasing Concepts from Diffusion Models

Rohit Gandikota and Joanna Materzynska and Jaden Fiotto-Kaufman and David Bau

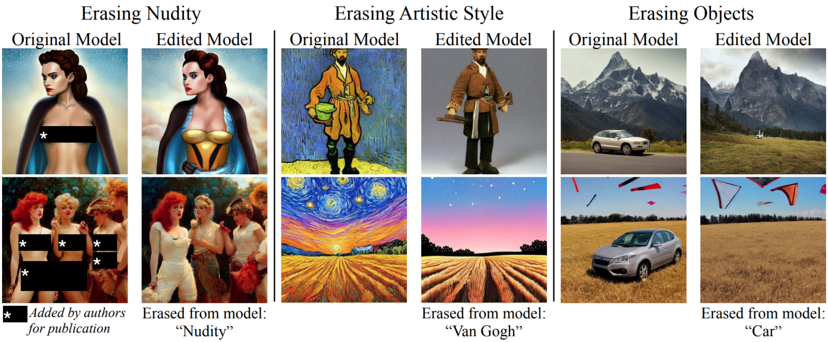
ICCV 2023

Department of Statistics, Seoul National University
Presented by Sangmoon Han

2025-09-07

Table of Contents

Overview



- Given only a short text description of an undesired visual concept and no additional data, Erased Stable Diffusion(ESD) fine-tunes model weights to erase the targeted concept.
- ESD can avoid NSFW("Not Safe For Work") content, stop imitation of a specific artist's style, or even erase a whole object class from model output.

noise-based Diffusion models for Text-to-Image

- Let x_0 denote image data, c denote text data, $T \in \mathbb{N}$
- Let $\epsilon \sim \mathcal{N}(0, I)$, $t \in \{0, \ldots, T\}$.
- Refer to diffusion models as noise(ϵ)-based diffusion models(NBDM) that share the following patterns :
  1. $f(x_0, t, \epsilon)$, called the forward operator, is specified so that $x_t = f(x_0, t, \epsilon)$ and, at $t = T$, x_T is approximately pure noise, e.g., DDPM :

     $$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

     , where $\bar{\alpha}_t$ are hyperparameters.
  2. Training is performed by minimizing $\mathcal{L}$ with function $\epsilon_\theta(x_t, t, c)$ with learnable parameters θ :

     $$\mathcal{L} \propto \mathbb{E}_{x_0, c, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right],$$

- For simplicity, let $\mathcal{L} = \mathbb{E}_{x_0, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2\right]$.

noise-based Diffusion models for Text-to-Image

3. $g(x_t, t, \epsilon_\theta(x_t, t, c))$, called the reverse operator, is specified so that $x_{t-1} = g(x_t, t, \epsilon_\theta(x_t, t, c))$ and, at $t = T$, $x_T \sim \mathcal{N}(0, I)$, e.g., DDPM :

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t, c)\right) + \sigma_t z, \quad z \sim \mathcal{N}(0, I).$$

   , where $\bar{\alpha}_t, \alpha_t, \sigma_t$ are hyperparameters.
4. For sampling, initialize $x_T \sim \mathcal{N}(0, I)$ and, iteratively denoise using ϵ_θ.

$$x_{t-1} = g(x_t, t, \epsilon_\theta(x_t, t, c)), \quad \text{For} \quad t = T, \cdots, 1$$

- In short, an NBDM is diffusion model whose $\epsilon_\theta(x_t, t, c)$ estimates the injected noise ϵ and is used both for training and inference.

Classifier-free guidance

- Classifier-free guidance(CFG) is a technique employed to regulate image generation, as described in [1, Ho et al].
- Compared to a NBDM, CFG introduces two changes with $0 < p_{\text{uncond}} < 1$, a guidance scale $s \geq 0$:
  1. Randomly drop the condition with p_{uncond}, training is performed by minimizing $\mathcal{L}'$:

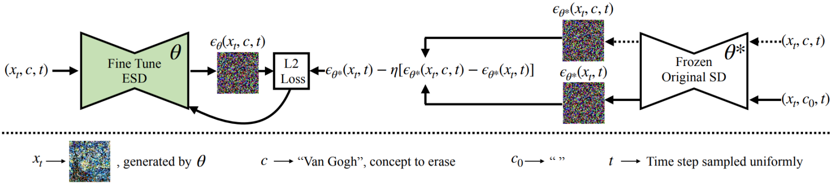
$$c' = \begin{cases} \varnothing & \text{with prob. } p_{\text{uncond}}, \\ c & \text{otherwise,} \end{cases} \qquad \mathcal{L}' = \mathbb{E}_{x_0, c', t}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, c')\right\|_2^2\right].$$

  2. Form a guided prediction with a guidance scale $s \geq 0$ and use it in the usual update:

$$\begin{aligned} \hat{\epsilon}_\theta(x_t, t, c) &= \epsilon_\theta(x_t, t, \varnothing) + s\big(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \varnothing)\big) \\ &= \begin{cases} \epsilon_\theta(x_t, t, \varnothing), & s = 0 \\ \epsilon_\theta(x_t, t, c), & s = 1 \\ \epsilon_\theta(x_t, t, \varnothing) + s\big(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \varnothing)\big), & s > 1 \end{cases} \end{aligned}$$

  , then reverse operator is $g(x_t, t, \hat{\epsilon}_\theta(x_t, t, c))$

Method of Erasing Concepts from Diffusion Models (ESD)



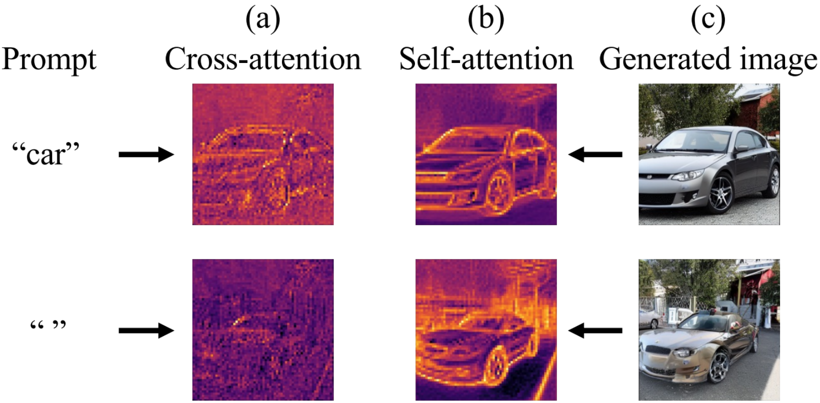
- Let ϵ_{θ^*} denotes pretrained NBDM, ϵ_θ denotes the NBDM to be fine-tuned.
- Let $\mathcal{C}_{\text{erase}}$ denotes a set of the concepts to erase, $c_0 = \varnothing$ the null prompt, and $\eta > 0$ the erase strength.
- For time step t and noised input x_t,

$$\tilde{\epsilon}(x_t, t, c) = \epsilon_{\theta^*}(x_t, t, c_0) - \eta(\epsilon_{\theta^*}(x_t, t, c) - \epsilon_{\theta^*}(x_t, t, c_0)).$$

- Training is performed by minimizing $\mathcal{L}_{\text{erase}}(\theta)$:

$$\mathcal{L}_{\text{erase}}(\theta) = \mathbb{E}_{x_0, t, c \in \mathcal{C}_{\text{erase}}}\left[\left\|\epsilon_\theta(x_t, t, c) - \tilde{\epsilon}(x_t, t, c)\right\|_2^2\right].$$

Importance of Parameter Choice

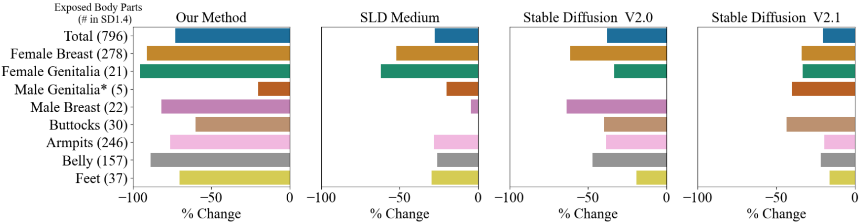


- Cross-attention parameters, illustrated in (a), directly depending on the text of the prompt.
- Other parameters, illustrated in (b), tend to contribute to a visual concept even if the concept is not mentioned in the prompt.

Importance of Parameter Choice

- Tuning the the cross-attention parameters only (ESD-x) erases the distinctive style of Van Gogh specifically when his name is mentioned in the prompt, keeping the interference with other artistic styles to a minimum.

Experiments



- When removing NSFW content it is important that the visual concept of "nudity" is removed globally, especially in cases when nudity is not mentioned in the prompt.
- I2P is a collection of 4,703 diverse text prompts that can generate harmful or inappropriate (NSFW) images, but do not explicitly mention NSFW terms.
- Using [2, Nudenet] detector, The figure shows the percentage change in the nudity-classified samples compared to the original model.

References I

Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: 2207.12598 [cs.LG]. URL: https://arxiv.org/abs/2207.12598.

Bedapudi Praneeth, brett koonce, and Alireza Ayinmehr. *bedapudi6788/NudeNet: place for checkpoint files*. Version v0. Dec. 2019. DOI: 10.5281/zenodo.3584720. URL: https://doi.org/10.5281/zenodo.3584720.