Leveraging Model Guidance to Extract Training Data from Personalized Diffusion Models

Xiaoyu Wu, Jiaru Zhang, Zhiwei Steven Wu

PURDUE UNIVERSITY · Carnegie Mellon University

TL;DR: We extract ~20% training data from real-world fine-tuned diffusion model checkpoints!

1. Motivation

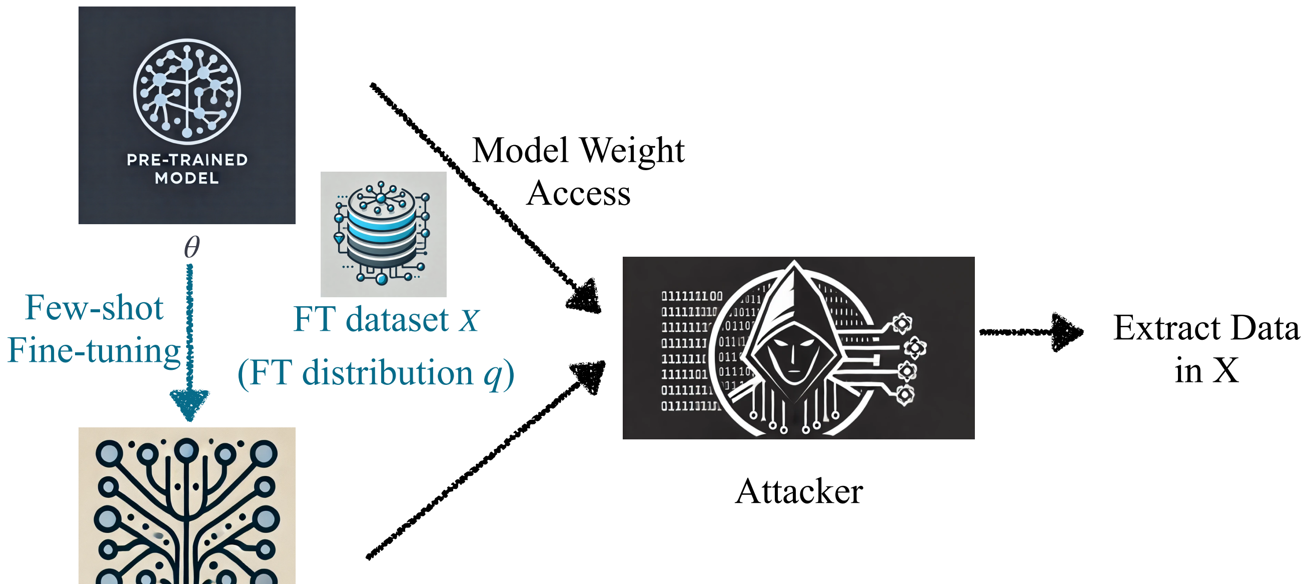
- **Few-shot Fine-tuning:**
  - ➢ Quickly adapt a pretrained DM to given **subjects or objects**
  - ➢ Low computational costs
  - ➢ Fostering larger platforms

CIVITAI · Hugging Face

- **Released fine-tuned checkpoints' risks:**
  - ➢ Copyright Risks: Unauthorized use of artists' work
  - ➢ Privacy Risks: Sensitive data such as human faces included during fine-tuning

Is it possible to extract fine-tuning data from these fine-tuned Diffusion Model checkpoints released online?

2. Threat Models



* Access to both pretrained and fine-tuned models is practical in real-world cases for checkpoints on websites like Civitai and HuggingFace

3. Methodology

- **Model Guidance**

Parametric approximation:

$$P_{\theta'}(x) \propto P_\theta(x)^{1-\lambda} \cdot q(x)^\lambda$$

Guidance towards q using *scores* of θ and θ'

$$\underbrace{\nabla_x \log q(x)}_{\text{Denoising function}} = \nabla_x \log P_{\theta'}(x) + \frac{1-\lambda}{\lambda}\underbrace{(\nabla_x \log P_{\theta'}(x) - \nabla_x \log P_\theta(x))}_{\text{Difference between two models}}$$

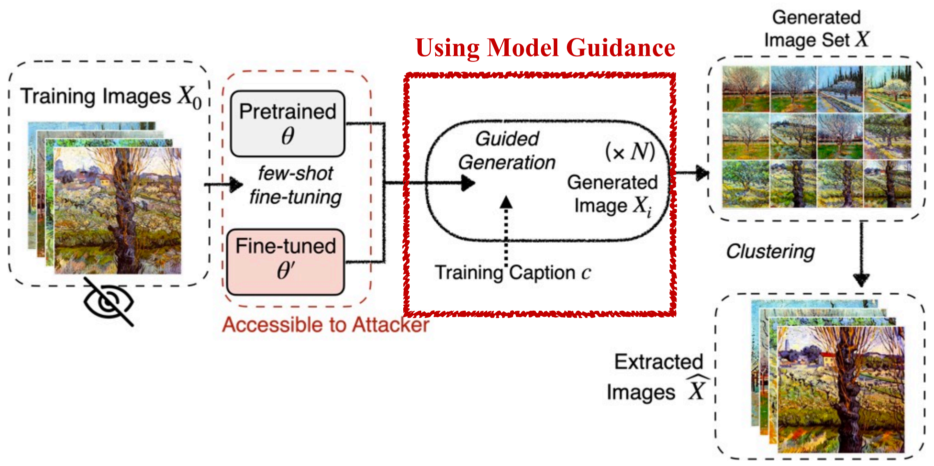
Using equivalence between score $\nabla_x \log p(x)$ and denoiser $\epsilon_q(x_t, t)$

$$\epsilon_q(x_t,t) = \epsilon_{\theta'}(x_t,t) + (w-1)(\epsilon_{\theta'}(x_t,t) - \epsilon_\theta(x_t,t)) \quad \text{where } w = \frac{1}{\lambda}$$

Extending to caption c available:

$$\epsilon_q(x_t,t,c) \approx \epsilon_{\theta'}(x_t,t,c) + \underbrace{(w'-1)(\epsilon_{\theta'}(x_t,t,c) - \epsilon_\theta(x_t,t))}_{\text{Guidance: unconditional pretrained DM -> conditional fine-tuned DM}} + \underbrace{k\epsilon_\theta(x_t,t)}_{\text{Correction Term}}$$

- **Our Framework FineXtract**



Step 1: Generate N images with model guidance.

Step 2: Find best candidates within N images using clustering.

Step 3 (Evaluation) : Compare selected images with training images.

4. Experiments

- **Real-world Results:**

Extracted Images

Training Images

- **Quantitative Results:**

Style-Driven Generation: WikiArt Dataset

Metrics and Settings	DreamBooth AS↑	DreamBooth A-ESR$_{0.7}$↑	DreamBooth A-ESR$_{0.6}$↑	LoRA AS↑	LoRA A-ESR$_{0.7}$↑	LoRA A-ESR$_{0.6}$↑
Direct Text2img+Clustering	0.317	0.00	0.01	0.299	0.00	0.00
CFG+Clustering	0.396	0.03	0.11	0.357	0.00	0.01
FineXtract	**0.449**	**0.06**	**0.22**	**0.376**	**0.01**	**0.05**

Object-Driven Generation: DreamBooth Dataset

Metrics and Settings	DreamBooth AS↑	DreamBooth A-ESR$_{0.7}$↑	DreamBooth A-ESR$_{0.6}$↑	LoRA AS↑	LoRA A-ESR$_{0.7}$↑	LoRA A-ESR$_{0.6}$↑
Direct Text2img+Clustering	0.418	0.03	0.11	0.347	0.00	0.02
CFG+Clustering	0.528	0.15	0.36	0.379	0.01	0.05
FineXtract	**0.557**	**0.25**	**0.45**	**0.466**	**0.04**	**0.18**

5. Summary

- ➢ We show that it is possible to extract fine-tuned data largely used for few-shot fine-tuning.
- ➢ We parametrically approximate the fine-tuning process and apply **model guidance** to effectively extract data.
- ➢ We show our extraction is successful in real-world scenarios.

Project homepage:
https://github.com/Nicholas0228/FineXtract