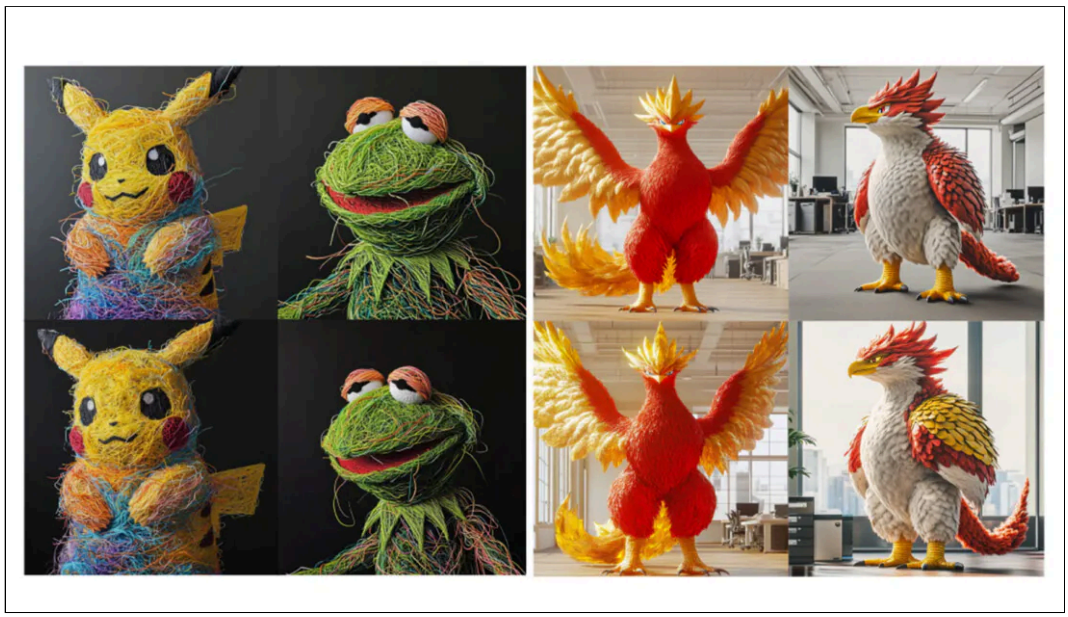
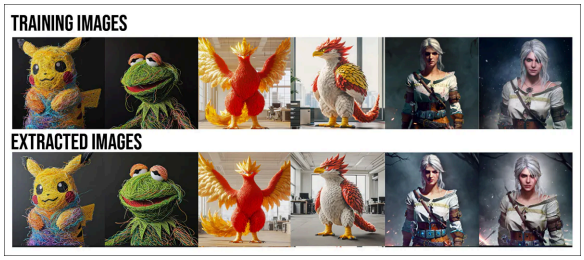
Extracting Training Data From Fine-Tuned Stable Diffusion Models

Originally published at https://www.unite.ai/extracting-training-data-from-fine-tuned-stable-diffusion-models/



New research from the US presents a method to extract significant portions of training data from fine-tuned models.

This could potentially provide legal evidence in cases where an artist's style has been copied, or where copyrighted images have been used to train generative models of public figures, IP-protected characters, or other content.



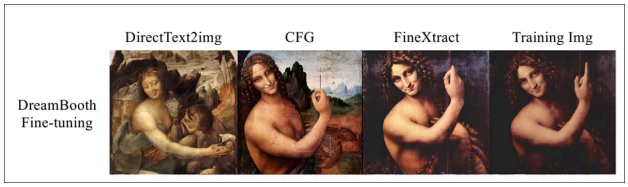
From the new paper: original training images are seen in the row above, and the extracted images are depicted in the row below. Source: https://arxiv.org/pdf/2410.03039

Such models are widely and freely available on the internet, primarily through the enormous user-contributed archives of civit.ai, and, to a lesser extent, on the Hugging Face repository platform.

The new model developed by the researchers is called *FineXtract*, and the authors contend that it achieves state-of-the-art results in this task.

The paper observes:

> *'[Our framework] effectively addresses the challenge of extracting fine-tuning data from publicly available DM fine-tuned checkpoints. By leveraging the transition from pretrained DM distributions to fine-tuning data distributions, FineXtract accurately guides the generation process toward high-probability regions of the fine-tuned data distribution, enabling successful data extraction.'*

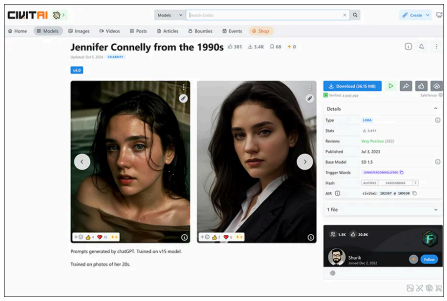


Far right, the original image used in training. Second from right, the image extracted via FineXtract. The other columns represent alternative, prior methods. Please refer to the source paper for better resolution.

Why It Matters

The *original* trained models for text-to-image generative systems as Stable Diffusion and Flux can be downloaded and fine-tuned by end-users, using techniques such as the 2022 DreamBooth implementation.

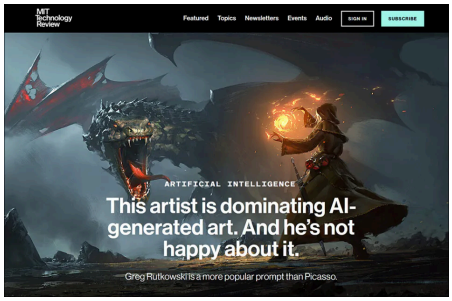
Easier still, the user can create a much smaller LoRA model that is almost as effective as a fully fine-tuned model.



An example of a trained LORA, offered for free download at the hugely popular civitai domain. Such a model can be created in anything from minutes to a few hours, by enthusiasts using locally-installed open source software – and online, through some of the more permissive API-driven training systems. Source: civitai.com

Since 2022 it has been trivial to create identity-specific fine-tuned checkpoints and LoRAs, by providing only a small (average 5-50) number of captioned images, and training the checkpoint (or LoRA) locally, on an open source framework such as Kohya ss, or using online services.

This facile method of deepfaking has attained notoriety in the media over the last few years. Many artists have also had their work ingested into generative models that replicate their style. The controversy around these issues has gathered momentum over the last 18 months.



The ease with which users can create AI systems that replicate the work of real artists has caused furor and diverse campaigns over the last two years. Source: https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/

It is difficult to prove which images were used in a fine-tuned checkpoint or in a LoRA, since the process of generalization 'abstracts' the identity from the small training datasets, and is not likely to ever reproduce examples from the training data (except in the case of overfitting, where one can consider the training to have failed).

This is where FineXtract comes into the picture. By comparing the state of the 'template' diffusion model that the user downloaded to the model that they subsequently created through fine-tuning or through LoRA, the researchers have been able to create highly accurate reconstructions of training data.

Though FineXtract has only been able to recreate 20% of the data from a fine-tune*, this is more than would usually be needed to provide evidence that the user had utilized copyrighted or otherwise protected or banned material in the production of a generative model. In most of the provided examples, the extracted image is extremely close to the known source material.

While captions are needed to extract the source images, this is not a significant barrier for two reasons: a) the uploader generally wants to facilitate the use of the model among a community and will usually provide apposite prompt examples; and b) it is not that difficult, the researchers found, to extract the pivotal terms blindly, from the fine-tuned model:

Training Caption	Extracted Caption ($\leq$ 3 words)	Extracted Caption ($\leq$7 words)
art style of Post Impressionism vincent	attributed **impressionism vincent**	plein **impressionism vincent** demonstrating! fantastic :))
art style of Fauvism henri matisse	**henri** paintings **matisse**	donneinarte hemingway **henri matisse** throughout fineart paintings
art style of High Renaissance leonardo da vinci	**leonardo** confident paintings	**leonardo** onda elengrembrandt pre picasso artwork
art style of Impressionism claude monet	suggestions **impressionism monet**	cassini gustave **monet impressionist monet** etosuggesti

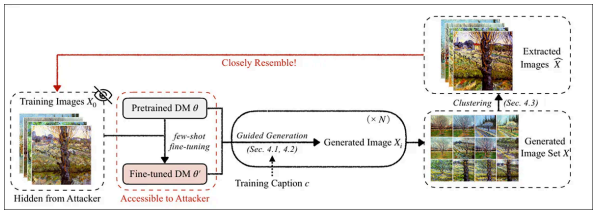
Essential keywords can usually be extracted blindly from the fine-tuned model using an L2-PGD attack over 1000 iterations, from a random prompt.

Users frequently avoid making their training datasets available alongside the 'black box'-style trained model. For the research, the authors collaborated with machine learning enthusiasts who did actually provide datasets.

The new paper is titled *Revealing the Unseen: Guiding Personalized Diffusion Models to Expose Training Data*, and comes from three researchers across Carnegie Mellon and Purdue universities.

Method

The 'attacker' (in this case, the FineXtract system) compares estimated data distributions across the original and fine-tuned model, in a process the authors dub 'model guidance'.



Through 'model guidance', developed by the researchers of the new paper, the fine-tuning characteristics can be mapped, allowing for extraction of the training data.

The authors explain:

> *'During the fine-tuning process, the [diffusion models] progressively shift their learned distribution from the pretrained DMs' [distribution] toward the fine-tuned data [distribution].*
>
> *'Thus, we parametrically approximate [the] learned distribution of the fine-tuned [diffusion models].'*

In this way, the sum of difference between the core and fine-tuned models provides the guidance process.

The authors further comment:

> *'With model guidance, we can effectively simulate a "pseudo-"[denoiser], which can be used to steer the sampling process toward the high-probability region within fine-tuned data distribution.'*

The guidance relies in part on a time-varying noising process similar to the 2023 outing *Erasing Concepts from Diffusion Models*.

The denoising prediction obtained also provide a likely Classifier-Free Guidance (CFG) scale. This is important, as CFG significantly affects picture quality and fidelity to the user's text prompt.

To improve accuracy of extracted images, FineXtract draws on the acclaimed 2023 collaboration *Extracting Training Data from Diffusion Models*. The method utilized is to compute the similarity of each pair of generated images, based on a threshold defined by the Self-Supervised Descriptor (SSCD) score.

In this way, the clustering algorithm helps FineXtract to identify the subset of extracted images that accord with the training data.

In this case, the researchers collaborated with users who had made the data available. One could reasonably say that, ***absent*** such data, it would be impossible to prove that any particular generated image was actually used in training in the original. However, it is now relatively trivial to match uploaded images either against live images on the web, or images that are also in known and published datasets, based solely on image content.

Data and Tests

To test FineXtract, the authors conducted experiments on few-shot fine-tuned models across the two most common fine-tuning scenarios, within the scope of the project: *artistic styles*, and *object-driven* generation (the latter effectively encompassing face-based subjects).

They randomly selected 20 artists (each with 10 images) from the WikiArt dataset, and 30 subjects (each with 5-6 images) from the DreamBooth dataset, to address these respective scenarios.

DreamBooth and LoRA were the targeted fine-tuning methods, and Stable Diffusion V1/.4 was used for the tests.

If the clustering algorithm returned no results after thirty seconds, the threshold was amended until images were returned.

The two metrics used for the generated images were Average Similarity (AS) under SSCD, and Average Extraction Success Rate (A-ESR) – a measure broadly in line with prior works, where a score of 0.7 represents the minimum to denote a completely successful extraction of training data.

Since previous approaches have used either direct text-to-image generation or CFG, the researchers compared FineXtract with these two methods.

Style-Driven Generation: WikiArt Dataset

Metrics and Settings	DreamBooth			LoRA		
	AS↑	A-ESR$_{0.7}$↑	A-ESR$_{0.6}$↑	AS↑	A-ESR$_{0.7}$↑	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			LoRA		
	AS↑	A-ESR$_{0.7}$↑	A-ESR$_{0.6}$↑	AS↑	A-ESR$_{0.7}$↑	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**

Results for comparisons of FineXtract against the two most popular prior methods.

The authors comment:

> *'The [results] demonstrate a significant advantage of FineXtract over previous methods, with an improvement of approximately 0.02 to 0.05 in AS and a doubling of the A-ESR in most cases.'*

To test the method's ability to generalize to novel data, the researchers conducted a further test, using Stable Diffusion (V1.4), Stable Diffusion XL, and AltDiffusion.

Metrics and Settings	SD (V1.4)		SDXL (V1.0)		AltDiffusion	
	AS↑	A-ESR$_{0.6}$↑	AS↑	A-ESR$_{0.6}$↑	AS↑	A-ESR$_{0.6}$↑
Direct Text2img+Clustering	0.341	0.03	0.335	0.05	0.282	0.00
CFG+Clustering	0.434	0.23	0.360	0.10	0.364	0.03
FineXtract	**0.501**	**0.35**	**0.467**	**0.25**	**0.388**	**0.05**

FineXtract applied across a range of diffusion models. For the WikiArt component, the test focused on four classes in WikiArt.

As seen in the results shown above, FineXtract was able to achieve an improvement over prior methods also in this broader test.

(a) Comparison on WikiArt dataset. (b) Comparison on DreamBooth dataset.

A qualitative comparison of extracted results from FineXtract and prior approaches. Please refer to the source paper for better resolution.

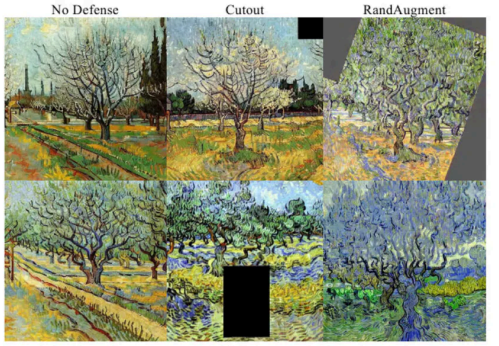
The authors observe that when an increased number of images is used in the dataset for a fine-tuned model, the clustering algorithm needs to be run for a longer period of time in order to remain effective.

They additionally observe that a variety of methods have been developed in recent years designed to impede this kind of extraction, under the aegis of privacy protection. They therefore tested FineXtract against data augmented by the Cutout and RandAugment methods.

Defense Methods	AS↑	A-ESR$_{0.6}$↑
No Defense	0.501	0.35
Cutout	0.397	0.08
RandAugment	0.267	0.03

FineXtract's performance against images protected; by Cutout and RandAugment.

While the authors concede that the two protection systems perform quite well in obfuscating the training data sources, they note that this comes at the cost of a decline in output quality so severe as to render the protection pointless:



Images produced under Stable Diffusion V1.4, fine-tuned with defensive measures – which drastically lower image quality. Please refer to the source paper for better resolution.

The paper concludes:

> *'Our experiments demonstrate the method's robustness across various datasets and real-world checkpoints, highlighting the potential risks of data leakage and providing strong evidence for copyright infringements.'*

Conclusion

2024 has proved the year that corporations' interest in 'clean' training data ramped up significantly, in the face of ongoing media coverage of AI's propensity to replace humans, and the prospect of legally protecting the generative models that they themselves are so keen to exploit.

It is easy to claim that your training data is clean, but it's getting easier too for similar technologies to prove that it isn't – as Runway ML, Stability.ai and MidJourney (amongst others) have found out in recent days.

Projects such as FineXtract are arguably portents of the absolute end of the 'wild west' era of AI, where even the apparently occult nature of a trained latent space could be held to account.

** For the sake of convenience, we will now assume 'fine-tune and LoRA', where necessary.*

First published Monday, October 7, 2024

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More