A Forensic Data Method for a New Generation of Deepfakes

Originally published at https://www.unite.ai/a-forensic-data-method-for-a-new-generation-of-deepfakes/

Although the deepfaking of private individuals has become a growing public concern and is increasingly being outlawed in various regions, actually proving that a user-created model – such as one enabling revenge porn – was specifically trained on a particular person's images remains extremely challenging.

To put the problem in context: a key element of a deepfake attack is falsely claiming that an image or video depicts a specific person. Simply stating that someone in a video is identity #A, rather than just a lookalike, is enough to create harm, and no AI is necessary in this scenario.

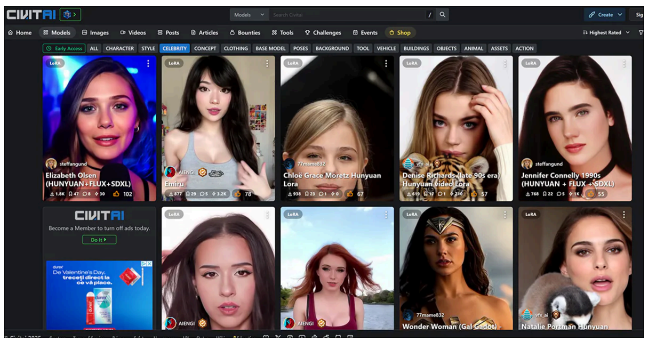
However, if an attacker generates AI images or videos using models trained on real person's data, social media and search engine face recognition systems will automatically link the faked content to the victim –without requiring names in posts or metadata. The AI-generated visuals alone ensure the association.

The more distinct the person's appearance, the more inevitable this becomes, until the fabricated content appears in image searches and ultimately reaches the victim.

Face to Face

The most common means of disseminating identity-focused models is currently through Low-Rank Adaptation (LoRA), wherein the user trains a small number of images for a few hours against the weights of a far larger foundation model such as Stable Diffusion (for static images, mostly) or Hunyuan Video, for video deepfakes.

The most common *targets* of LoRAs, including the new breed of video-based LoRAs, are female celebrities, whose fame exposes them to this kind of treatment with less public criticism than in the case of 'unknown' victims, due to the assumption that such derivative works are covered under 'fair use' (at least in the USA and Europe).



Female celebrities dominate the LoRA and Dreambooth listings at the civit.ai portal. The most popular such LoRA currently has more than 66,000 downloads, which is considerable, given that this use of AI remains seen as a 'fringe' activity.

There is no such public forum for the non-celebrity victims of deepfaking, who only surface in the media when prosecution cases arise, or the victims speak out in popular outlets.

However, in both scenarios, the models used to fake the target identities have 'distilled' their training data so completely into the latent space of the model that it is difficult to identify the source images that were used.

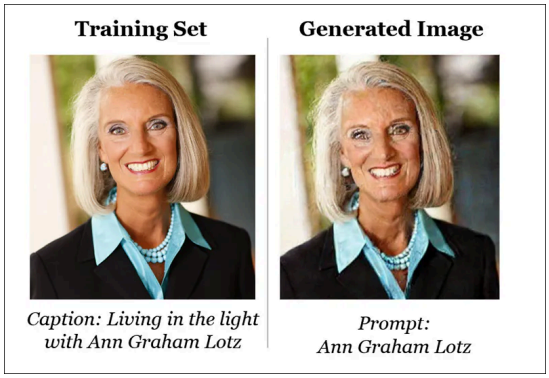
If it *were* possible to do so within an acceptable margin of error, this would enable the prosecution of those who share LoRAs, since it not only proves the intent to deepfake a particular identity (i.e., that of a specfic 'unknown' person, even if the malefactor never names them during the defamation process), but also exposes the uploader to copyright infringement charges, where applicable.

The latter would be useful in jurisdictions where legal regulation of deepfaking technologies is lacking or lagging behind.

Over-Exposed

The objective of training a foundation model, such as the multi-gigabyte base model that a user might download from Hugging Face, is that the model should become well-generalized, and ductile. This involves training on an adequate number of diverse images, and with appropriate settings, and ending training before the model 'overfits' to the data.

An overfitted model has seen the data so many (excessive) times during the training process that it will tend to reproduce images that are very similar, thereby exposing the source of training data.



The identity 'Ann Graham Lotz' can be almost perfectly reproduced in the Stable Diffusion V1.5 model. The reconstruction is nearly identical to the training data (on the left in the image above). Source: https://arxiv.org/pdf/2301.13188

However, overfitted models are generally discarded by their creators rather than distributed, since they are in any case unfit for purpose. Therefore this is an unlikely forensic 'windfall'. In any case, the principle applies more to the expensive and high-volume training of foundation models, where multiple versions of the same image that have crept into a huge source dataset may make certain training images easy to invoke (see image and example above).

Things are a little different in the case of LoRA and Dreambooth models (though Dreambooth has fallen out of fashion due to its large file sizes). Here, the user selects a very limited number of diverse images of a subject, and uses these to train a LoRA.

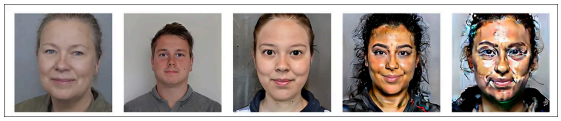
On the left, output from a Hunyuan Video LoRA. On the right, the data that made the resemblance possible (images used with permission of the person depicted).

Frequently the LoRA will have a trained-in trigger-word, such as *[nameofcelebrity]*. However, very often the specifically-trained subject will appear in generated output *even without such prompts*, because even a well-balanced (i.e., not overfitted) LoRA is somewhat 'fixated' on the material it was trained on, and will tend to include it in any output.

This predisposition, combined with the limited image numbers that are optimal for a LoRA dataset, expose the model to forensic analysis, as we shall see.

Unmasking the Data

These matters are addressed in a new paper from Denmark, which offers a methodology to identify source images (or groups of source images) in a black-box Membership Inference Attack (MIA). The technique at least in part involves the use of custom-trained models that are designed to help expose source data by generating their own 'deepfakes':



Examples of 'fake' images generated by the new approach, at ever-increasing levels of Classifier-Free Guidance (CFG), up to the point of destruction. Source: https://arxiv.org/pdf/2502.11619

Though the work, titled *Membership Inference Attacks for Face Images Against Fine-Tuned Latent Diffusion Models*, is a most interesting contribution to the literature around this particular topic, it is also an inaccessible and tersely-written paper that needs considerable decoding. Therefore we'll cover at least the basic principles behind the project here, and a selection of the results obtained.

In effect, if someone fine-tunes an AI model on your face, the authors' method can help prove it by looking for telltale signs of memorization in the model's generated images.

In the first instance, a target AI model is fine-tuned on a dataset of face images, making it more likely to reproduce details from those images in its outputs. Subsequently, a classifier attack mode is trained using AI-generated images from the target model as 'positives' (suspected members of the training set) and other images from a different dataset as 'negatives' (non-members).

By learning the subtle differences between these groups, the attack model can predict whether a given image was part of the original fine-tuning dataset.

The attack is most effective in cases where the AI model has been fine-tuned extensively, meaning that the more a model is specialized, the easier it is to detect if certain images were used. This generally applies to LoRAs designed to recreate celebrities or private individuals.

The authors also found that adding visible watermarks to training images makes detection easier still – though hidden watermarks don't help as much.

Impressively, the approach is tested in a black-box setting, meaning it works without access to the model's internal details, only its outputs.

The method arrived at is computationally intense, as the authors concede; however, the value of this work is in indicating the direction for additional research, and to prove that data can be realistically extracted to an acceptable tolerance; therefore, given its seminal nature, it need not run on a smartphone at this stage.

Method/Data

Several datasets from the Technical University of Denmark (DTU, the host institution for the paper's three researchers) were used in the study, for fine-tuning the target model and for training and testing the attack mode.

Datasets used were derived from DTU Orbit:

DseenDTU The base image set.

DDTU Images scraped from DTU Orbit.

DseenDTU A partition of DDTU used to fine-tune the target model.

DunseenDTU A partition of DDTU that was not used to fine-tune any image generation model and was instead used to test or train the attack model.

wmDseenDTU A partition of DDTU with visible watermarks used to fine-tune the target model.

hwmDseenDTU A partition of DDTU with hidden watermarks used to fine-tune the target model.

DgenDTU Images generated by a Latent Diffusion Model (LDM) which has been fine-tuned on the DseenDTU image set.

The datasets used to fine-tune the target model consist of image-text pairs captioned by the BLIP captioning model (perhaps not by coincidence one of the most popular uncensored models in the casual AI community).

BLIP was set to prepend the phrase *'a dtu headshot of a'* to each description.

Additionally, several datasets from Aalborg University (AAU) were employed in the tests, all derived from the AU VBN corpus:

DAAU Images scraped from AAU vbn.

DseenAAU A partition of DAAU used to fine-tune the target model.

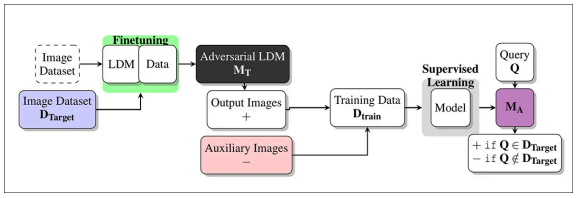
DunseenAAU A partition of DAAU that is not used to fine-tune any image generation model, but rather is used to test or train the attack model.

DgenAAU Images generated by an LDM fine-tuned on the DseenAAU image set.

Equivalent to the earlier sets, the phrase *'a aau headshot of a'* was used. This ensured that all labels in the DTU dataset followed the format *'a dtu headshot of a (…)'*, reinforcing the dataset's core characteristics during fine-tuning.

Tests

Multiple experiments were conducted to evaluate how well the membership inference attacks performed against the target model. Each test aimed to determine whether it was possible to carry out a successful attack within the schema shown below, where the target model is fine-tuned on an image dataset that was obtained without authorization.



Schema for the approach.

With the fine-tuned model queried to generate output images, these images are then used as positive examples for training the attack model, while additional unrelated images are included as negative examples.

The attack model is trained using supervised learning and is then tested on new images to determine whether they were originally part of the dataset used to fine-tune the target model. To evaluate the accuracy of the attack, 15% of the test data is set aside for validation.

Because the target model is fine-tuned on a known dataset, the actual membership status of each image is already established when creating the training data for the attack model. This controlled setup allows for a clear assessment of how effectively the attack model can distinguish between images that were part of the fine-tuning dataset and those that were not.

For these tests, Stable Diffusion V1.5 was used. Though this rather old model crops up a lot in research due to the need for consistent testing, and the extensive corpus of prior work that uses it, this is an appropriate use case; V1.5 remained popular for LoRA creation in the Stable Diffusion hobbyist community for a long time, despite multiple subsequent version releases, and even in spite of the advent of Flux – because the model is completely uncensored.

The researchers' attack model was based on Resnet-18, with the model's pretrained weights retained. ResNet-18's 1000-neuron last layer was substituted with a fully-connected layer with two neurons. Training loss was categorical cross-entropy, and the Adam optimizer was used.

For each test, the attack model was trained five times using different random seeds to compute 95% confidence intervals for the key metrics. Zero-shot classification with the CLIP model was used as the baseline.

(Please note that the original primary results table in the paper is terse and unusually difficult to understand. Therefore I have reformulated it below in a more user-friendly fashion. Please click on the image to see it in better resolution)

Test No.	Comparison Type	Training Data (Positives)	Training Data (Negatives)	Test Data (Positives - "Members")	Test Data (Negatives - "Non-Members")	Attack Accuracy (AUC - ResNet-18)	Baseline Accuracy (CLIP AUC)
1	AI vs. AI (Generated DTU vs. AAU)	AI-generated DTU images	AI-generated AAU images	AI-generated DTU images	AI-generated AAU images	**1.00** (Perfect detection)	0.94
2	AI vs. Real (DTU vs. Non-Generated AAU)	AI-generated DTU images	Real AAU images	Real DTU images	Real LFW images (random faces)	0.71	**0.99** (Baseline performs better)
3	Fine-Tuned Model (DTU vs. AAU)	AI-generated DTU images	AI-generated AAU images	Real DTU images	Real AAU images	**0.86** (Strong detection)	0.63
4	Fine-Tuning Duration (DTU vs. AAU, trained for 50, 100, 400 epochs)	AI-generated DTU images	AI-generated AAU images	Real DTU images	Real AAU images	**0.86, 0.82, 0.86** (More training helps, but even 50 epochs works)	0.63, 0.62, 0.63
5	Effect of Changing Prompts	AI-generated DTU images (with different prompts)	AI-generated AAU images	Real DTU images	Real AAU images	**0.82** (Slightly weaker but still effective)	0.56
6	Effect of Guidance Scale (0, 4, 8, 12, 16)	AI-generated DTU images	AI-generated AAU images	Real DTU images	Real AAU images	**Varies: Best around 8 (0.89), worst at 16 (0.58)**	-
7	Fine-Tuned Model (DTU vs. LFW)	AI-generated DTU images	AI-generated AAU images	Real DTU images	Real LFW images (random faces)	**0.89** (Strong detection)	**0.99** (Baseline performs much better)
8	Same Dataset (DTU "Seen" vs. "Unseen")	AI-generated DTU images	AI-generated AAU images	Real DTU images (used in training)	Real DTU images (never used in training)	**0.53** (Fails - barely better than guessing)	0.52
9	Non-Fine-Tuned Model vs. AAU	AI-generated DTU images (from an unmodified model)	AI-generated AAU images	Real DTU images	Real AAU images	**0.66** (Weaker but still detects some leakage)	0.55
10	Non-Fine-Tuned Model vs. Itself	AI-generated DTU images (from an unmodified model)	AI-generated DTU images (from another unmodified model)	Real DTU images	Real AAU images	**0.54** (Almost random guessing)	0.55
11	Effect of Visible Watermarks	AI-generated DTU images (with visible watermark)	AI-generated AAU images	Real DTU images (with watermark)	Real AAU images	**1.00** (Perfect detection)	0.95
12	Effect of Hidden Watermarks	AI-generated DTU images (with hidden watermark)	AI-generated AAU images	Real DTU images (with hidden watermark)	Real AAU images	**0.83** (No significant improvement)	0.57
13	Mixed Dataset (DTU + LFW vs. AAU)	AI-generated DTU + real LFW images	AI-generated AAU images	Real DTU images	Real AAU images	**0.83** (Still effective even when mixing datasets)	0.64

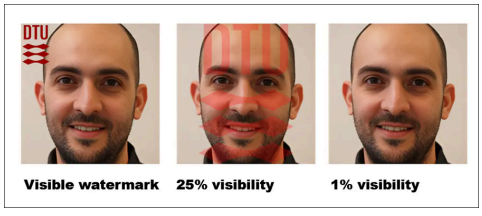
Summary of results from all tests. Click on the image to see higher resolution

The researchers' attack method proved most effective when targeting fine-tuned models, particularly those trained on a specific set of images, such as an individual's face. However, while the attack can determine whether a dataset was used, it struggles to identify individual images within that dataset.

In practical terms, the latter is not necessarily a hindrance to using an approach such as this forensically; while there is relatively little value in establishing that a famous dataset such as ImageNet was used in a model, an attacker on a private individual (not a celebrity) will tend to have far less choice of source data, and need to fully exploit available data groups such as social media albums and other online collections. These effectively create a 'hash' which can be uncovered by the methods outlined.

The paper notes that another way to improve accuracy is to use AI-generated images as 'non-members', rather than relying solely on real images. This prevents artificially high success rates that could otherwise mislead the results.

An additional factor that significantly influences detection, the authors note, is watermarking. When training images contain visible watermarks, the attack becomes highly effective, while hidden watermarks offer little to no advantage.



The right-most figure shows the actual 'hidden' watermark used in the tests.

Finally, the level of guidance in text-to-image generation also plays a role, with the ideal balance found at a guidance scale of around 8. Even when no direct prompt is used, a fine-tuned model still tends to produce outputs that resemble its training data, reinforcing the effectiveness of the attack.

Conclusion

It is a shame that this interesting paper has been written in such an inaccessible manner, as it should be of some interest to privacy advocates and casual AI researchers alike.

Though membership inference attacks may turn out to be an interesting and fruitful forensic tool, it is more important, perhaps, for this research strand to develop applicable broad principles, to prevent it ending up in the same game of whack-a-mole that has occurred for deepfake detection in general, when the release of a newer model adversely affects detection and similar forensic systems.

Since there is some evidence of a higher-level guiding principle cleaned in this new research, we can hope to see more work in this direction.

First published Friday, February 21, 2025

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