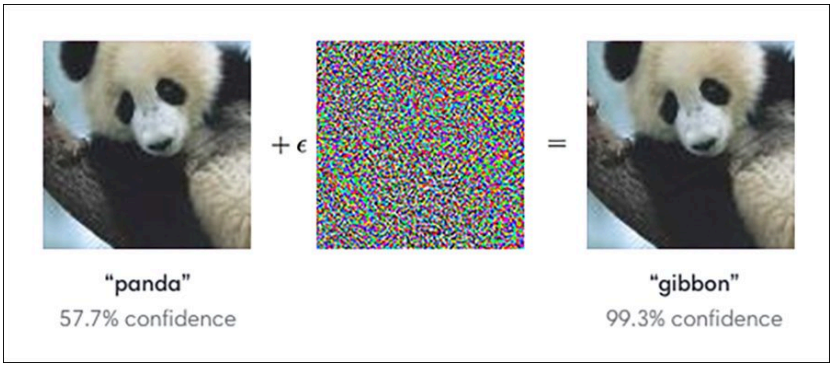
Preventing Stable Diffusion ‘Copyright Infringement’ by Poisoning The Source Data

Originally published at https://arquivo.pt/wayback/20240203185200oe_/https://blog.metaphysic.ai/preventing-stable-diffusion-copyright-infringement-by-poisoning-the-source-data/

February 13, 2023

A new research collaboration between China, the UK and the US has come up with a preventative measure for artists concerned about their style and works being made reproducible (or at least imitable) in the new breed of latent diffusion-based text-to-image AI systems, such as Stable Diffusion and DALL-E 2.

The new system, titled AdvDM, is an algorithm that injects adversarial data into images on which generative AI systems might be trained.



Adversarial data introduces highly crafted and disruptive noise into images. The perturbations target specific characteristics of machine learning systems, so that with very few noticeable changes to the source image, the usage of the image in those systems is adversely affected. Source: https://openai.com/blog/adversarial-example-research/

Effectively, tiny perturbations are added to the source data images, which are designed to undermine the capability of a machine learning system to extract features from the image, and to thereafter be able to reproduce or imitate the image or the style contained in the original data.

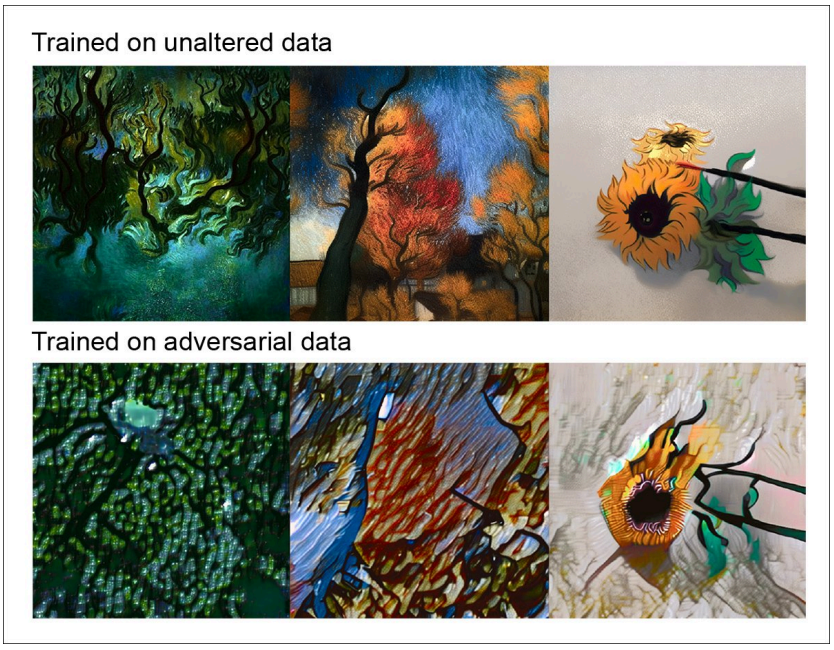


Image generations from Stable Diffusion, with models trained on uncompromised data (above) and perturbed data (below). Source: http://arxiv.org/pdf/2302.04578

In the upper row of the image above, from the examples given in the new paper, we see Stable Diffusion making a pretty good job of imitating the style of Vincent Van Gogh. The source data for the model used was unaltered.

In the lower part of the image, where the source data was perturbed via AdvDM prior to training, we see that the style of Van Gogh is scarcely represented, if at all.

The new paper states:

'Extensive experiments show that the estimated adversarial examples can effectively hinder DMs from extracting their features. Our method can be a powerful tool for human artists to protect their copyright against infringers with DM-based AI-for-Art applications.'

Approach

AdvDM, like prior approaches, exploits the gradient of the optimization goal for the target system – the mapping of the architecture's trained transformative powers, which is expecting non-'optimized' novel data, rather than data that actually knows something about the way the system operates.

Previous works using this method, cited in the new paper, include *Explaining and Harnessing Adversarial Examples* (2014), MIT's 2018 *Towards Deep Learning Models Resistant to Adversarial Attacks*, and *Towards Evaluating the Robustness of Neural Networks* (2016).

However, this method, as it stands, will not work with a latent diffusion system, since it would only be able to target the gradient for an *expected function* (rather than known functions in the gradient), because latent diffusion is more iterative and granular and less explicit than GANs and other types of generative network.

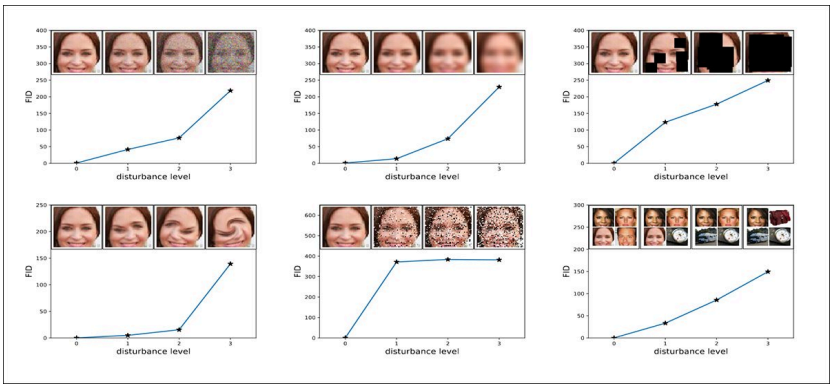
Therefore AdvDM obtains its 'attack vector' by Monte Carlo estimation, building up a usable profile with each step of gradient ascent.

The paper notes that unlike attacks on classification models, which have dominated headlines in adversarial research, diffusion models don't use images as direct input, but rather extract features from them, in order to develop abstracted concepts that can be used to generate similar (but not identical) images.

The authors comment:

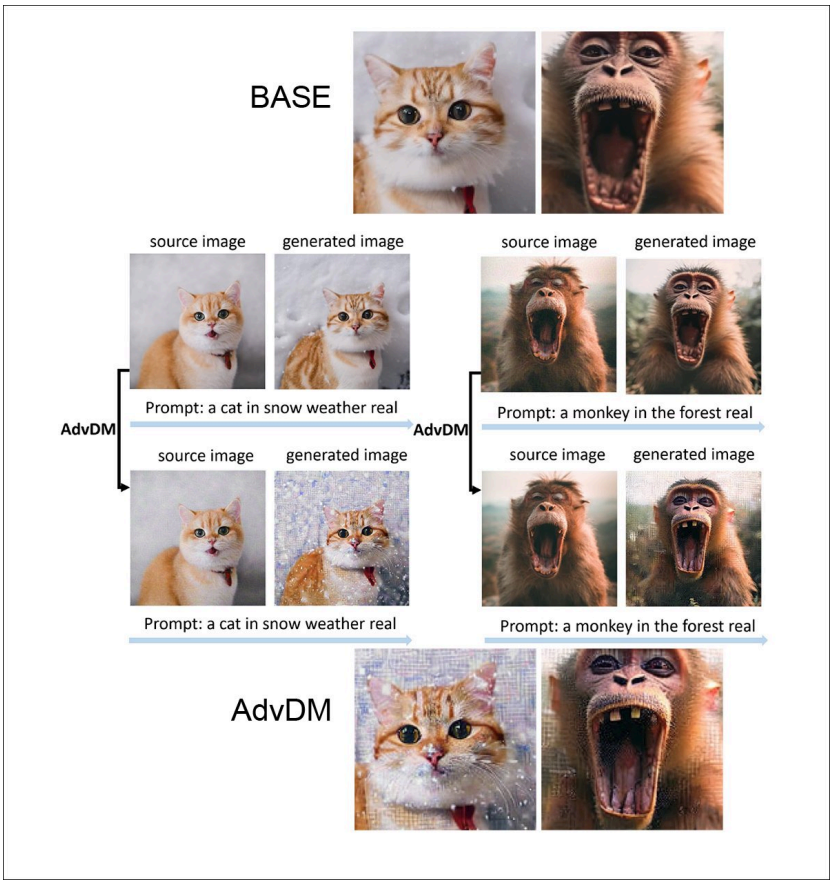
'We mainly focus our evaluation scenario on [conditional] inference, where copyright violations have taken place. For unconditional inference, the model samples a noise and generates images. This process has no input images and does not raise copyright concerns, thus [is] not included in our evaluation.'

The latent diffusion model evaluates its success in generating images by the distance of the proposed generated image from the features in the source material, and this method is targeted under AdvDM, which uses Fréchet Inception Distance (FID) and Precision (prec.) to establish approaches by which the native conditioning of the system can be subtly subverted.



Measuring image prediction accuracy with Fréchet Inception Distance (FID). Source: https://arxiv.org/pdf/1706.08500.pdf

The result is that AdvDM can attack latent diffusion models on three fronts: via *text-to-image* generation, where the user provides a text-prompt and the system finds and utilizes related latent variables to formulate a new image; through *style transfer*, which is more concerned with changing the appearance and styling of the generated image; and, interestingly, through *image-to-image* (Img2Img) functionality, where the user provides a source photo and requires the system to reinterpret the photo according to a text prompt:



Img2Img functionality in Stable Diffusion degraded by AdvDM. See paper for clearer reference images.

As can be seen in the above image, AdvDM is able to interfere adequately with the Img2Img process in order to make the use of these 'adulterated' source images unappealing to the end user.

Tests

The system was tested on Stable Diffusion. The architecture was trained on eight NVIDIA RTX A4000 GPUs across all experiments, on a batch size of 4 – settings and conditions which, the authors state, are broadly in line with prior works aimed at non-LD systems.

For the text-to-image component, the researchers selected 1000 random images from three LSUN datasets related to cats, sheep and airplanes. The aim was to subvert a common and popular method of developing user-specific models – textual inversion. Under this approach, as few as five images can be used to create a generalized model that can generate images related to subjects not present in the original, base Stable Diffusion model.

The researchers generated 10,000 images for each dataset, using unaffected and perturbed approaches.

Dataset	LSUN-Cat			LSUN-Sheep			LSUN-Airplane		
Metric	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*	FID ↑	*prec.* ↓	*recall.*
No attack	34.94	0.5643	0.1531	32.81	0.6378	0.1228	39.22	0.5016	0.2765
AdvDM	127.04	0.1708	0.061	203.5	0.0058	0.378	169.67	0.0263	0.3235

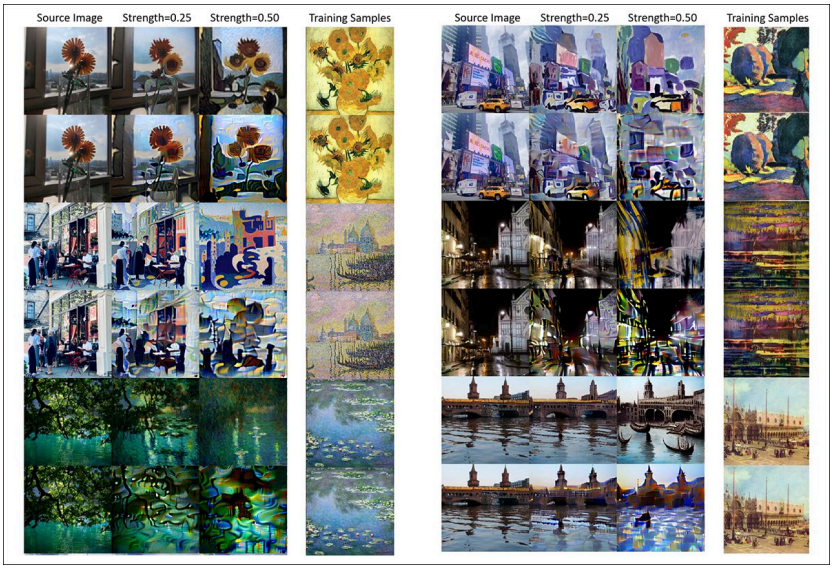
Subverting textual inversion: results from the text-to-image round.

Of the results, the authors comment:

'Our adversarial examples significantly increase FID and decrease Precision of the conditionally-generated images.'

Next, the researchers tested the ability of AdvDM to interfere with the 'theft' of artists' styles, by examining the extent to which style transfer can be distorted. To test this, the authors selected 20 paintings from 10 artists, from the WikiArt dataset, and trained a disruptive element for each artist.

Select results from this round are shown below (please see the paper for more accurate resolution and clarity).



Images in each group are derived from the same source image. Images extracted from the adversarially-perturbed examples are in the bottom.

Regarding these results, the authors state:

'[The] results demonstrate that the style of the conditionally-generated images is significantly different from the input images when conditioning on S_* training on adversarial examples. This suggests that AdvDM can be effectively used for copyright protection against illegal style transfer.'

For the Img2Img tests, the researchers applied AdvDM on open-source photos from Pexels, generating both 'clean' and perturbed examples using Stable Diffusion's image-to-image methodology (for results, see above image of the cat and the monkey).

Here the researchers assert:

'The generated images based on adversarial examples are unrealistic in comparison with those based on clean images.'

Logistics, Implementation, and Implications

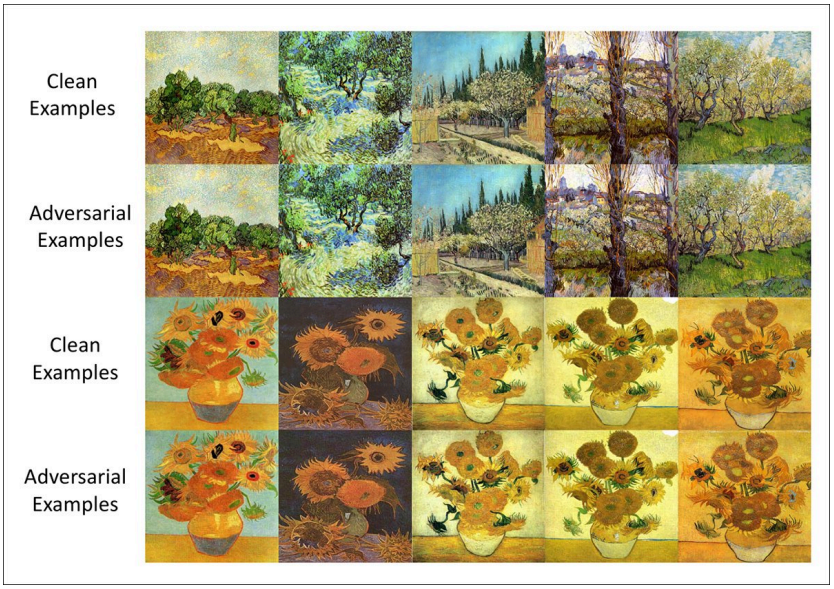
The system follows in the footsteps of many prior works along the same lines in recent years, though it claims to be the first to implement such an approach in a latent diffusion system. Though they are scantly (or not at all) addressed in the new work, we should consider four questions: how badly are the source images affected by the perturbation process; what would actually need to happen, in terms of distribution and infrastructure, if such a method were to be made widely and openly available; does the scope of such a system extend beyond protecting the rights of artists (such as Greg Rutkowski, now arguably the 'poster boy' of the anti-AI art movement) into more general images; and would it also work for video?

Image Quality

Regarding the first issue, this is ground that has been covered by the many prior works that either sought to use perturbed images to prevent image use in AI systems, or to use such images to attack recognition systems. Such projects include *TnT Attacks*, a 2021 initiative to exploit common features in overused datasets; *Optical Adversarial Attack*, which used similar methods to change the meaning of road signs; *FakeTagger*, designed to stop the use of 'stolen' images in deepfake system training; and *Fawkes*, a system designed to allow users to add AI-foiling perturbations to their own social media images before uploading; among many others in a crowded and growing field.

From the Fawkes project, we see examples of original and perturbed or 'cloaked' images. Source: https://arxiv.org/pdf/2002.08327.pdf

Common to all systems that use minor pixel perturbations to cause distress to AI systems trying to use the images as source data, the question *How badly does the perturbation affect the image?* is almost always 'it's barely perceptible', and the authors of AdvDM make the same claim:



Affected and unaffected source images used for the AdvDM experiments. Please see the original paper for a more accurate representation of resolution and quality.

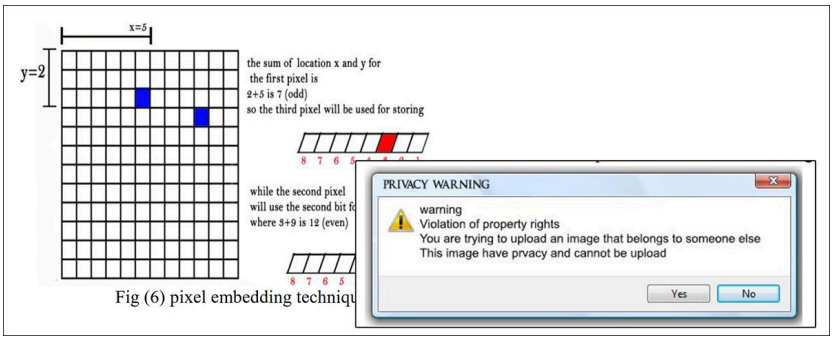
The shared problems of all perturbation-based approaches is that the effect is not *entirely* imperceptible, and, depending on the system, does tend to in some way affect the quality of the output. With the emphasis in recent years on ever-improving image quality, higher resolutions in image/video output and multimedia reproduction, this is a hard pill to swallow, and a traditional obstacle to adoption.

However, faced with the possible choice between ceasing to upload *any* images and uploading slightly-affected perturbed images, it's possible that artists looking to protect their style and individuals hoping to immunize their own media against deepfake exploitation might consider the trade-off to be worth it, if such a system were ever to become widely available.

Feasibility of Deployment

Secondly, and arguably a much greater obstacle, is that watermarking or perturbation systems of this type would effectively need to be deployed in a very intrinsic and integrated way into image-posting systems, with minimal user friction or need for technical prowess on the part of the user, in order to make any major difference.

There are two obvious ways that such a system could gain traction: one is that it becomes adopted by a major provider, and preferably a platform that is a traditional target of web-scraping systems hungry for new AI data, such as Facebook, Flickr, Twitter, Pinterest, or any other truly prominent outlet.

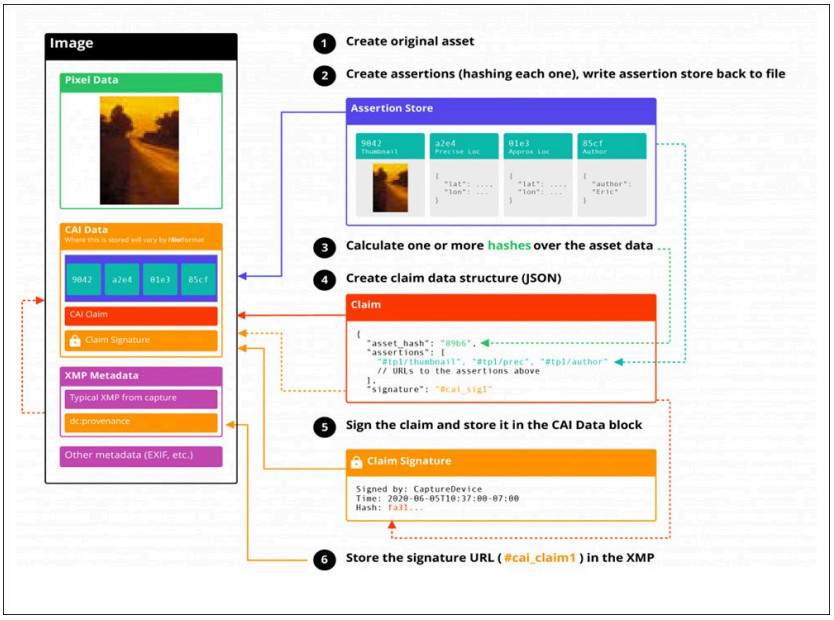


From 2016, one of a number of schemes designed to embed steganographic information into social media image by default, enabling restrictive processes. Source: https://tinyurl.com/av35u63r (Preview before link)

The second likely route to wider adoption would be if one particular smaller platform (such as DeviantArt or ArtStation) were to trial such a system over a long-term period, and be able to prove later that work published with the perturbations features less in new AI models than prior unprotected work.

In either case, the widespread use of this kind of invisible 'watermarking' would then become a *de facto* standard, albeit a 'degraded' one (in respect to prior image quality), and the market would adjust to the *force majeure* of it all.

Changing the infrastructure of digital distribution, whether for professional or casual use, is no small feat. The only major current initiative in this respect is the Content Authenticity Initiative (CAI), which is making some inroads into persuading major image generation technology companies (such as Leica and Nikon) to incorporate new methods of establishing the true provenance of original user content, with a view to combating fake news in general, and image-based deepfakes in particular.



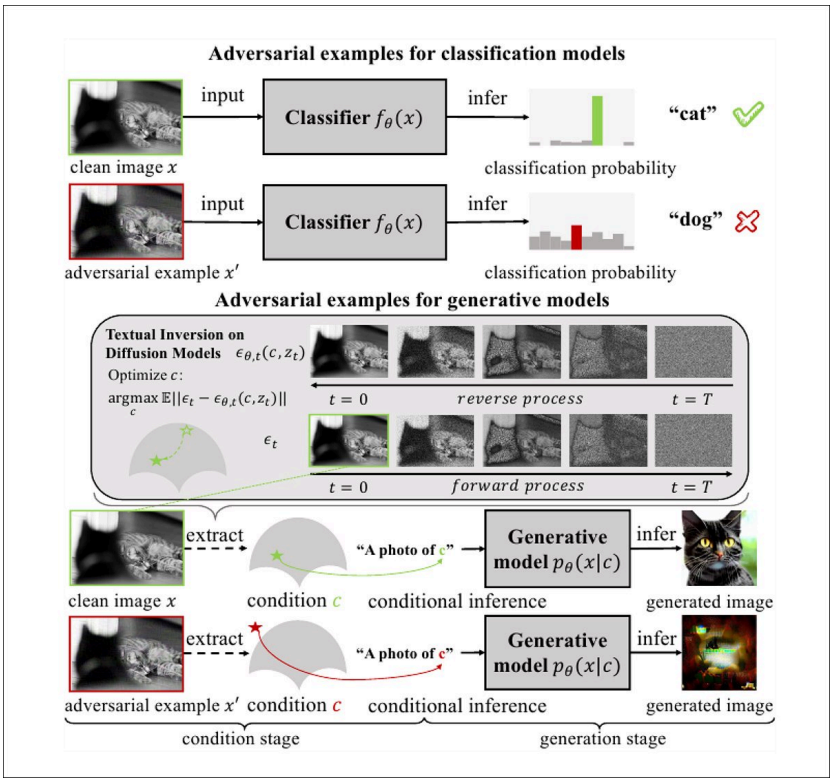
StarlingLab, partners in the Adobe-led Content Authenticity Initiative, are committing to changing the way newly-created images are registered. Source: https://www.starlinglab.org/image-authentication/

Beyond Art

The new work – either sincerely, or because it's hooking onto a 'red hot' topic that attracts funding – concentrates on using perturbation to protect both the specific images and general style of artists. However, broad take-up of such a system would arguably need to offer wider protection and cover a larger part of the market, in order to make the expense and difficulty of changing or amending tried-and-true systems worth the effort.

Therefore the question arises as to whether perturbation-based AI-proofing can work also for general images, and perhaps to protect individuals from being deepfaked, or otherwise having their likeness sucked into the training schedules of new generative image and video systems.

Based on prior works into protections against autoencoder and GAN ingestion of public images, this would seem to be the case, and the new method is directly and avowedly descended from, and beholden to these earlier approaches.



A comparison of 'traditional' methods that use adversarial examples, against the new method proposed by the researchers.

A further question, little explored in that extensive strand of research into perturbation-based AI 'inoculation', is how badly *video* quality would be affected by such perturbations, since, in the case of deepfakes, most of the material used in datasets is gathered from YouTube videos and disc or streaming rips of movies and TV shows.

Most deepfake source material, including many of the images used in DreamBooth training, are pulled from video, which is rendered out into individual frames. Could each frame contain a unique or at least persistent perturbation?

Including video content in a perturbation-based system would mean that wider protection against AI would conceivably have to extend from selfies on Facebook to actual changes in the compression systems used to distribute movies on platforms like TikTok, and television series on streaming and on disc (unless, as some contend, famous people are 'fair game', and the object of protecting *non*-celebrities is considered a higher priority).

This aspect is not addressed in the new paper, since a single 'frame' created by an analog artist like Rutkowski is a high-effort and discrete work, unlike the mass-dumping of terabytes of data from the Blu-ray sources that are used in modern deepfake and general AI training techniques.

A further issue is whether or not such a system would only protect new material, or whether its advantages would be extended to 'historical' data. It's not unknown for major organizations to recompress their entire back-catalogue, usually in pursuit of savings in bandwidth (as Netflix has done in recent years); but it would perhaps be difficult to force smaller organizations, and even some of the larger ones, to devote resources to re-encoding, except if it were to become a legislative requirement.

However, given the extent to which concerns about falling behind in machine learning development is causing governments and courts to prefer lenient policies towards AI-focused web-scraping, the latter case may be unlikely.

High Redundancy..?

One other final consideration in regard to popularizing a scheme of this kind is that it could become defunct quite quickly, for several reasons, despite the potentially enormous expense of implementing it.

Firstly, the features that the perturbations key upon can't necessarily be relied on as the architecture evolves. For instance, Stable Diffusion V2+ brought some severe modifications to the internal workings and rendering quality, with respect to the prior system; there's no guarantee that it, or other latent diffusion systems will have enough consistent and persistent features that can be reliably targeted by a perturbation-based system over the long term.

Secondly, despite the conflict between governmental will to advance their national progress in AI development and adoption, the growing pressure from voters to reign in AI is leading to new laws and proposals for tighter legislation that, potentially, could criminalize or at least penalize some of the practices that systems such as AdvDM wish to target, making its functionality redundant (since users would not be able to disseminate the fruits of these processes any more).

These, and various other imponderable factors, arguably stand against the prospect of introducing AI-resistant image-tampering systems on a wide-scale, or with any level of commitment – though there is nothing to prevent concerned groups from adopting such systems in the short term, while the current state of chaos and uncertainty about where the technology (and the laws related to it) are heading, gradually becomes clearer.

The paper we've been looking at here is titled *Adversarial Example Does Good: Preventing Painting Imitation from Diffusion Models via Adversarial Examples*, and is an equally-contributed work from researchers at Shanghai Jiao Tong University, Queen's University Belfast, and New York University.