Michele Mora

CIS129

Brittany Griwzow

05/06/2025

CIS129 Final Project Rough Draft

Note: this document is a compilation/draft that will be used to create a slideshow format presentation

PART I

Concept: Protect digital art from diffusion

Pertinent questions:

What is AI art?

.Art generated through artificial intelligence machine learning.

What is the problem with AI art?

.AI art is created with information learned from preexisting art and images of art. This process is called diffusion and has advanced to the point where generative AI can mimic the styles of artists. Many ethical and legal debates are currently being conducted as to whether or not AI art is ethical and if an artist's "style" can be legally protected (as quantifiable stylistic elements are difficult to conclude). This project will have nothing to do with this debate.

The fact remains that artists do not want their works or styles to be replicable. Consequences and protections are a deterrent, not a preventor. A real demand for a preventative measure and protection against artistic works being used in diffusion for machine learning exists. This project explores a possible solution to this need.

Who does this effect?

.Everyone. Art touches everyone's' lives. The nature of art in the world we live in impacts our own ideas and identities. Art echoes culture and speaks more into existence.

.Small businesses- Many small businesses construct their own logos, imagery, and designs to sell their products and stand out in the market. Saturation of branding and products weakens one of the most important tools they have.

.Small businesses- Many small businesses construct their own logos, imagery, and designs to sell their products and stand out in the market. Saturation of branding and products weakens one of the most important tools they have

.Digital artists – Demand for their work lowers as more generators can copy their styles. A lot of digital artists make money from teaching as much from if not more than they do from their actual pieces. People may develop less incentive to learn how to draw like them if a machine can produce similar work to theirs. Digital artists also have to share more of their work for free to keep momentum going in their careers. This momentum becomes less effective as digital art becomes increasingly easier to reproduce with generative AI.

Why I care:

Digital art has become important to the world. It is used in animation, video game and movie concepts, commercial use, and much more. It is also used by smaller indie artists for both personal enjoyment and profit. I have spent my life enjoying these projects. They mean a lot to me, and the artists that make them have put more happiness in my life than they will ever know. I am seeing less and less enthusiasm towards these projects. I think all the listed ways that AI art impacts hold equal precedence in the conversation. I am prioritizing digital art to hone the project and focus on what I care about on a personal level.

Key terms:

Diffusion- the method of teaching AI image generators with pre-existing images

LAION Dataset - “a collection of 5.6 billion images scraped, without permission, from the internet. Almost every digital artist has images in LAION, given that DeviantArt and ArtStation were lifted wholesale, along with Getty Images and Pinterest.” - Mattei, Escalante-De

Stability AI- Company that creates open-source generative AI Models, best known for a text to image product called Stable Diffusion

Style mimicry- the ability to replicate an artist’s style through stable diffusion-enabling the recreation of an image in a particular artist’s style.

Steganography- Steganography is the practice of hiding a secret message inside of (or even on top of) something that is not secret.

Midjourney- Prompt based generator of artificial images.

Pre-existing solutions: Watermarks are not enough anymore

Blurb: In the past the main concern was an image being copied and directly used for profit or personal use by someone that did not possess the rights to the image. Now that diffusion can directly

."Opting out": Sites like DeviantArt and platforms like Meta have options to manually select "opting out" of user uploads being used for AI machine learning. OpenAI even has an option on the site to opt out of uploads being used to to train the DALL-E 3 open AI generator.

Pros:

.some level of protection against diffusion

Cons:

.Site/platform specific

-opting out from one platform is not a universal solution

-images copies to other platforms are still at risk

-opting out from images used for one machine will not guarantee no use by another

.tedious

-too much effort to "opt out" on every possible platform a user wants to upload to

-"opting out" usually requires deep navigation into the platform to find and use

-scrapers can bypass opt-outs

Glaze: "***Glaze*** is a system designed to protect human artists by disrupting style mimicry. At a high level, Glaze works by understanding the AI models that are training on human art, and using machine learning algorithms, computing a set of minimal changes to artworks, such that it appears unchanged to human eyes, but appears to AI models like a dramatically different art style. For example, human eyes might find a *glazed* charcoal portrait with a realism style to be unchanged, but an AI model might see the glazed version as a modern abstract style, a la Jackson Pollock. So when someone then prompts the model to generate art mimicking the charcoal artist, they will get something quite different from what they expected." -official site

Pros:

.robust to change

-resistant to redownloading, compression, smoothing effects

-not reliant on steganography

.not a watermark

-operates as a dimension within the piece

-each use of glaze produces a unique dimension

-attack would have to know the specific dimensions

.accessibility

-webglaze can be used on any phone, tablet, computer, or device with internet

Cons:

.Effects on art

-glaze can be somewhat detectable on art styles with flat colors and simple backgrounds

.not futureproof

-earlier versions of glaze have lost robustness against some image purifiers

-the current version could also be overcome by new advancements

.only works for current works

-Glaze can prevent style mimicry from pieces it is used on

-Glaze cannot undo any mimicry already trained into base models

(like SDL or SD3)

Nightshade**: "*Nightshade*** works similarly as Glaze, but instead of a defense against style mimicry, it is designed as an offense tool to distort feature representations inside generative AI image models. Like Glaze, Nightshade is computed as a multi-objective optimization that minimizes visible changes to the original image. While human eyes see a shaded image that is largely unchanged from the original, the AI model sees a dramatically different composition in the image." -official site

Pros:

.offense

-use by artists as a group attack mimicry ability by misleading scraping tools

.robust

.robust to change

-resistant to redownloading, compression, smoothing effects

-not reliant on steganography

.not a watermark

-operates as a dimension within the piece

-each use of glaze produces a unique dimension

-attack would have to know the specific dimensions

Cons:

.Effects on art

-detectable on pieces with simple backgrounds and flat colors

-lower levels are possible but less robust

.Not Future Proof

-no guarantee that Nightshade will work in the long term against attack

Kudurru: “Kudurru monitors popular AI datasets for scraping behavior, and coordinates amongst the network to quickly identify scrapers. When a scraper is identified, its identity is broadcast to all protected Kudurru sites. All Kudurru sites then collectively block the scraper from downloading content from their respective host. When the scraper is finished, Kudurru informs the network and traffic is allowed to proceed as normal.” -official site

Pros: WIP

.detects scrapers

.notifies multiple platforms

.

Cons: WIP

.Still in Beta testing

.membership required

.only so many sites are protected

PART II

Proposition:

Make protecting digital pieces and styles easier with software that can perform both tasks of defense against diffusion and attack to model learning.

In addition to

A combination of these pseudocodes:

Code B

Pseudocode for Add Adversarial Noise to an Image (FGSM Attack):

1. Load a pre-trained image classification model (e.g., ResNet)
2. Load an image from disk
3. Preprocess the image: a. Resize to model's input size b. Normalize pixel values c. Convert to tensor or array format
4. Pass the preprocessed image through the model to get its predicted label
5. Calculate the loss between the model's output and the predicted label
6. Compute the gradient of the loss with respect to the input image
7. Generate adversarial noise: a. Take the sign of the gradient b. Multiply by a small value (epsilon) to control noise strength
8. Add the adversarial noise to the original image to create a perturbed image
9. Clip the pixel values of the perturbed image to stay within valid bounds (e.g., [0, 1])
10. Output or display: a. Original image b. Perturbed (adversarial) image c. Compare the model's predictions before and after

And

Code A

Pseudocode for Adversarial Image Perturbation Targeting CLIP (used in diffusion):

1. Load a pre-trained feature encoder used in diffusion models (e.g., CLIP image encoder)
2. Load the image you want to protect
3. Preprocess the image to match the model's input requirements
4. Define a misleading or irrelevant target concept as text (e.g., "a photo of trash")
5. Convert the image into a tensor and enable gradient computation
6. Initialize an optimizer to update the image tensor
7. For a number of steps:
   a. Encode the image using the image encoder
   b. Encode the misleading text using the text encoder
   c. Compute similarity between image and text embeddings

d. Calculate loss to *maximize* similarity (i.e., make image appear like the wrong concept)
e. Backpropagate the loss to compute gradients
f. Update the image tensor using the optimizer

8. After optimization, clip pixel values to valid range
9. Save or display the perturbed (protected) image

By making a program that can run an image through both processes, diffusion could potentially be better prevented. Also, we will have to wrap this code like Tutankhamun.

Advanced mode:

Allows user to manually adjust the amount of noise added to image

Allows batch processing

WIP: defense against bypassing/scraping ???

User Interaction:

Goals:

- Comprehensive
- Straightforward
- Little to no navigation necessary
- Accessible
- Wham bam thank you sam experience

The software would ideally be used as both a site and an application. The site would suit users that want to use the software once or rarely. The application would be useful for users that want to regularly use the software. Allowing users to try the software on the site may lead to more downloads and also lowers the barrier to usage from those that may be unable or unwilling to download.

This is ideal, because a key factor in an element of the software's effectiveness is the number of people using it. (The more pollution to model training, the better.)

A clearly visible field for image downloads would ideally be front and center on the home page. Uploading should resemble familiar mainstream file storers like google drive. A drag and drop file option and any other means to make uploading a file as easily as possible should be used.

A straightforward result should be produced. The processed image should present itself on the same page and in tandem with the original image. Having the second image beside the original with clear labelling will help prevent any confusion for which is which and if the process actually happened for the user.

```
//Site home
//open/minimalistic design
//two fields, large boxes
//tab above fields labeled “advanced mode”
//Within left field
  //prompt user to upload with multiple options
      //drag and drop option
      //upload from device
      // upload from drive
      //upload from dropbox
  //Obfuscate image
     //run through code A
    //create New Image object
   //run through code B
 //display original image in left field
//display new image in right field
```

For advanced mode:

```
//Selectable tab
```

//Effect minimizer

//Blurb explaining Effect Minimizer (layers a somewhat translucent image with half the obfuscation of the new image to preserve some of the original traits of the piece over the new image to lessen the changes

//Select opacity of new image layer

//Processes images like home pseudocode

//Copy New image and make Layer Image

//half the desired level of obfuscation

//lower opacity to desired level

//layer over New Image

//Display in Left field (under new image)

//Display original image in right field (under original image)

//Batch processor

//Prompts users to upload file of images/ multiple images

//Processes images like home pseudocode

//returns a new file/ images with

//Display left field (under new image)

//Display in right field (under original image)

Open Questions:

How effective is scraping?

What does a typical bypass of cloaking and adversarial noise and image perturbation look like? The image can have minefields embedded in it, but how do we prevent any crossing?

Sources

https://research.ebsco.com/c/k4o4aa/viewer/html/zi463e3u55

https://www.artworkflowhq.com/resources/stability-ai-guide

https://glaze.cs.uchicago.edu/

https://nightshade.cs.uchicago.edu/whatis.html#:~:text=The%20answer%20is%20that%20Glaze%20is%20a%20defensive,consent%20%28thus%20protecting%20all%20artists%20against%20these%20models%29.

https://www.comptia.org/blog/what-is-steganography

https://kudurru.ai

https://arxiv.org/pdf/2302.10149

https://automatetheboringstuff.com/2e/chapter19/