Using AI to detect AI

Creating a System to Detect AI Generated Images

Scott Morel

December 2023

Table of Contents

Introduction……………………………………………………………………………………3
Motivation………………………………………………………………………………………3
Description……………………………………………………………………………………..3
Conclusion………………………………………………………………………………………4
Algorithm………………………………………………………………………………………..5
Further Information…………………………………………………………………………..6

Introduction

With the introduction of new generative AI technologies, there is a substantial need for security and safety related to these technologies. One of the biggest dangers posed by AI is the ability to manipulate audio, video, and images, usually to aid in vishing attacks or to push a specific (and usually false) narrative. This report has been written to pose a solution for the use of AI generated images being used to defraud or mislead.

Motivation

The motivation for this report is to propose a machine learning application to detect and report AI generated or manipulated images that are being used for nefarious or dishonest purposes. The big question is what we need to accomplish this. To answer this, we must ask a few more questions:

What do we need to look for to detect AI generated images?
How do we determine a threshold for flagging and reporting?
Who should we report these images to?

Description

Now to answer the above questions. As far as what to look for when detecting AI generated images, I have noticed that 3 dead giveaways are looking at the eyes, mouth, and hands. There are many AI images out there where people have glassy, warped looking eyes as well as hands with 4 or 6 fingers and mouths that don't look right. These images become even more detectable if there are any background images (posters, billboards, etc.) as well as any text in the background. Most of the time these items appear smudged and just look wrong.

For determining a threshold, I suggest using a points-based scoring system. This will help flag images correctly, as well as reinforce what the application already learned. There will also need to be some form of human interaction to verify the findings at first. This is to prevent the application from trying to gain as many points as possible, thus eliminating any reason to detect the images correctly.

To report these images, we will need to program in a way to determine which website the photo is on and which email address to contact. This should be easy. We would just need to scrape the URL/domain registration data from the page, as well as the correct email address.

Conclusion

To create the application to detect the images, we would need to create data sets. The data sets would include tens to hundreds of thousands of AI generated images, as well as complete copies of all the letters in all the alphabets throughout the world. We would then need to train it to look for any imperfections (smudges, smearing, unrealistic looking lighting/reflections) or patterns that would help it determine its authenticity. It would also need to be trained to detect any characters or numbers that do not look real. Also, we would need to collect data about all the big AI image generators (StableDiffusion, Midjourney, Firefly etc.)

The scoring system would be based on the criteria mentioned previously. Once the image has been assigned all its points, the application would then determine what to do with it. This can be accomplished by setting a threshold and if the image has points that are equal to or above that threshold, the get flagged and reported to whichever platform the image was found on. Again, this could be done by scraping the URL/domain/IP information of the site, as well as the correct email address for reporting. Below I have included an algorithm with some pseudocode to illustrate the basic concept.

A I Image Detector

```
let r;
let f.                          r is real
let s;

    //Detection

      //Check for any signs that would imply image is fake
        //check fore ground
        //Check background
       //check any images or text
let r=0
let f=1
let s=sum(r+f)
        for(item=r; r≤0; r=r)
        for(item=f; f>0; f++)
    if (S≥{threshold1}){
                flag&report}
      else if({threshold2}≤ S ≤ {threshold1}){
                flag for review}
      else if(0≤ S≤{threshold2}){
                image is real}
//if image is flagged for reporting, scrape site info for IP/Domain
// and correct email address
```

Further Information

Midjourney: https://www.midjourney.com/home?callbackUrl=%2Fexplore
Stable Diffusion: https://stability.ai/stable-diffusion
Adobe Firefly: https://www.adobe.com/products/firefly.html