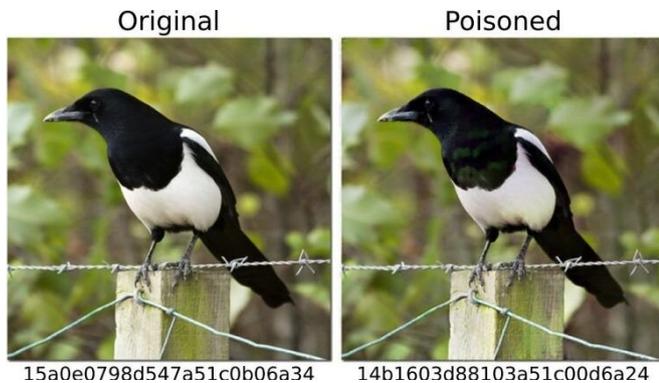


Learning to break deep perceptual hashing

September 12 2022, by Ute Kehse



To put the system to the test, the researchers defined images of dogs as "dangerous" and changed other photo motifs - such as that of the magpie shown on the left - so that (as on the right) its fingerprint corresponded to that of a dog image. To the human eye, both photos look identical. Credit: <https://arxiv.org/abs/2111.06628>

Can artificial intelligence (AI) methods reliably detect child pornography images on end devices?

A study in which Oldenburg computer scientist Daniel Neider was involved raises doubts about whether this is currently possible. The research was published in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Here, Neider discusses his findings:

Mr. Neider, do you have a virus scanner on your computer?

I think every Windows computer comes with an antivirus program—so yes.

Apple installed its NeuralHash program, which automatically scans image files for child pornography, on end devices last year. Does it work in a similar way to an antivirus program?

NeuralHash does something similar, though the

way it works is different: the software scans end devices for a specific type of content—not, as with antivirus programs, for malware, but for illegal images. This is known as client-side scanning, which refers to the scanning of files on the user's device.

How exactly does NeuralHash work?

The program is based on [artificial intelligence](#) methods and uses so-called [neural networks](#). Put simply, it's a computer program that is trained to recognize certain patterns in images. The program assigns a kind of code to each image, basically a sequence of numbers and letters. These codes are called hashes. You can imagine them as fingerprints that are generated for each image. The trick is that images that look similar are assigned the same hash—so, for example, all images featuring black cats could be assigned the hash 3x580ac97e. Apple has a large database of such hashes, which can be assigned to child pornography images. And whenever a user tries to upload an image with a hash that is in the database, the image is marked without the user noticing. Such images cannot be forwarded.

Does that mean Apple knows what images are on my mobile phone?

No, Apple doesn't look at the images. It only has the database with the codes. The idea is that the company works together with child protection organizations. And on the basis of material that comes from [law enforcement agencies](#), these organizations use a program to generate hashes for the database.

You analyzed NeuralHash in a research project with colleagues from the Technical University of Darmstadt. How did the project come about?

Neural networks don't always work the way we think they do. The technology is very promising, but it isn't always one hundred percent accurate. It's often difficult to find out why it delivers a certain

result, because the procedure has not been explicitly programmed. In principle, this technology has simply learned to recognize certain patterns in the data. However, this can also be exploited to trick the program—and it works with alarming frequency. So we asked ourselves: how does this affect a system that is intended to be used to assess illegal content? What happens if you slightly modify images, for example?

Why did you focus on NeuralHash?

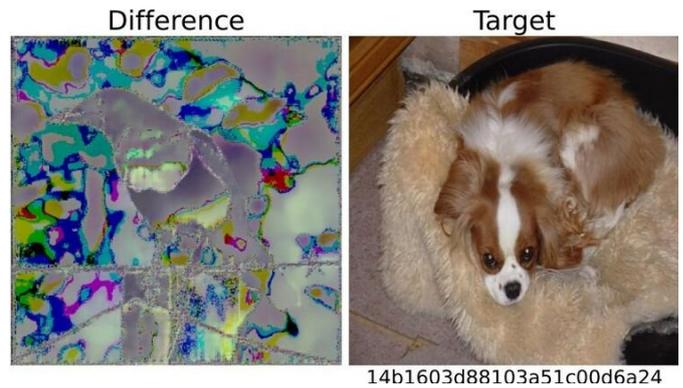
In 2021, Apple delivered a prototype of NeuralHash together with an operating system update to end devices—basically all devices that can send photos to the iCloud cloud storage service, like iPhones or Macs. The prototype was not yet activated, so the program didn't start checking images on the Apple devices. But this move made the technology available to us; we were able to extract the program and thus gain access to the neural network. We wanted to take a look at how a big company would go about such a task. Apple later refrained from officially rolling out NeuralHash due to massive criticism of the mass surveillance and invasion of privacy it entailed.

What exactly did you test?

We tested how the system could be abused. To avoid having to work with child pornography material, we defined images of dogs as "dangerous". Then we calculated their "digital fingerprints". In the first scenario, we took images of other things, for example of a cat, and tried to modify them slightly so that the result was a "fingerprint" of a dog.

Did it work?

Yes, and it turned out to be relatively easy. You need access to the neural network—which we had because the program was installed on the devices—and you need some knowledge of how machine learning works. But then it's quite easy to alter the cat images so that they generate any other hash. To the [human eye](#), the manipulated photos look almost like the original, you can't really tell the difference.



On the right, the dog image defined as "dangerous" on a test basis. Because of the visible differences between the two magpie images on the left, the system sounded the alarm for the photo that had been altered in this way. Anyone familiar with machine learning could thus "relatively easily" play tricks with the system, Neider says. Credit: <https://arxiv.org/abs/2111.06628>

And that's when things get problematic...

Right. Because I could send you a manipulated photo, and the moment you try to upload it to your cloud or send it to someone else via a messaging app, the system is triggered without you noticing. You don't even know why the upload or forwarding function is blocked. But the real problem is that Apple also notices that you've tried to send a suspicious image. And if this happens too often, Apple decrypts the material and, if deemed necessary, reports it to the local law enforcement authorities. This means that material could be planted on someone to incriminate them.

What else did you test?

We also posed the opposite question: can I bypass the system? Can I manipulate an image with a fingerprint in the database in such a way that it generates a different fingerprint? In one scenario, we again assumed that the user has access to the AI via their device, is familiar with the neural network and has some knowledge of machine learning.

And can the images be manipulated to make

them look unsuspecting?

Yes, it works very well. But what we also discovered is that even if you don't have access to the system and make very simple changes to a photo that anyone can make with their [mobile phone](#), it's possible to trick the program. For instance, simply by rotating an image by 90 degrees you can substantially alter the "fingerprint". This, of course, is not good, because you can undo this change just by rotating the image 90 degrees in the other direction. The entire information contained in the image is retained. This shows that it's relatively easy to trick the system.

What conclusions do you draw from the study?

In my view, we don't know enough about neural networks at the moment to be able to use them safely. These programs are not robust enough for such sensitive tasks—as we saw in this case study. Moreover, in my opinion, the legislators should not rely on programs developed by corporations like Apple or Facebook in response to a law to do the right thing. For example, there is the danger that these companies will block more content than necessary, as a pre-emptive measure, so to speak, to avoid getting into trouble and having to pay high fines. Something similar is already happening in reaction to the Network Enforcement Law (also known as the Facebook Act).

So should we not use technology to automatically prevent the uploading of indexed images?

On the contrary, my colleagues and I are also in favor of using technology to combat child pornography. But we think it's important that there is a public discourse about what image recognition using neural networks can do, what it can't do, and what we are prepared to accept as collateral damage. From our [point of view](#), it's always a matter of weighing up the pros and cons: if it's so easy to trick a program, is it really justifiable to install it on everyone's devices? After all, there is a risk of false alarms. At the same time, anyone who wants to can bypass the system relatively easily. So doesn't it actually do more harm than good? Of course, it's not up to us computer scientists to make

the decisions here. Our contribution is to point out the problems with the technology so that a meaningful discussion can take place on that basis.

More information: Lukas Struppek et al, Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash, *2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). [DOI: 10.1145/3531146.3533073](https://doi.org/10.1145/3531146.3533073)

GitHub: github.com/ml-research/Learning-to-Break-Deep-Perceptual-Hashing

Provided by Carl von Ossietzky Universität Oldenburg

APA citation: Learning to break deep perceptual hashing (2022, September 12) retrieved 8 December 2022 from <https://techxplore.com/news/2022-09-deep-perceptual-hashing.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.