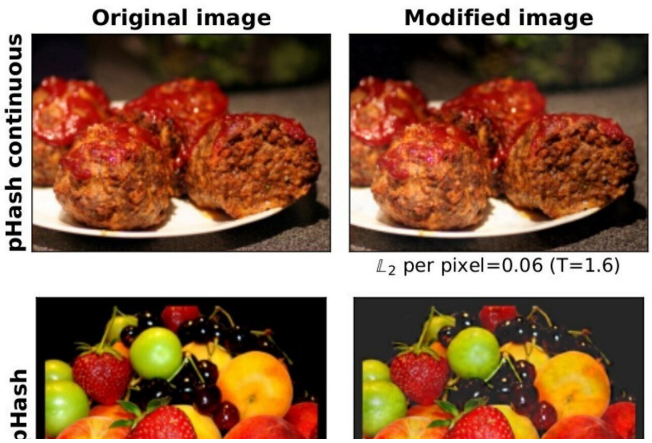
Proposed illegal image detectors on devices are 'easily fooled'

9 November 2021, by Caroline Brogan



These images have been hashed, so that they look different to detection algorithms but nearly identical to us. Credit: Imperial College London

Proposed algorithms that detect illegal images on devices can be easily fooled with imperceptible changes to images, Imperial research has found.

Companies and governments have proposed using built-in scanners on devices like phones, tablets and laptops to detect illegal images, such as child sexual abuse material (CSAM). However the new findings from Imperial College London raise questions about how well these scanners might work in practice.

Researchers who tested the robustness of five similar algorithms found that altering an 'illegal' image's unique 'signature' on a device meant it would fly under the algorithm's radar 99.9 percent of the time.

The scientists behind the peer-reviewed study say their testing demonstrates that in its current form, so-called perceptual hashing based client-side scanning (PH-CSS) algorithms will not be a 'magic bullet' for detecting illegal content like CSAM on personal devices. It also raises serious questions about how effective, and therefore proportional, current plans to tackle illegal material through on-device scanning really are.

The findings are published as part of the USENIX Security Conference in Boston, U.S..

Senior author Dr. Yves-Alexandre de Montjoye, of Imperial's Department of Computing and Data Science Institute, said: "By simply applying a specifically designed filter mostly imperceptible to the human eye, we misled the algorithm into thinking that two near-identical images were different. Importantly, our algorithm is able to generate a large number of diverse filters, making the development of countermeasures difficult.

"Our findings raise serious questions about the robustness of such invasive approaches."

Apple recently proposed, and then postponed due to privacy concerns, plans to introduce PH-CSS on all its personal devices. There are also reports that certain governments are considering using PH-CSS as a law enforcement technique by passing end-to-end encryption.

Under the radar

PH-CSS algorithms can be built into devices to scan for illegal material. The algorithms sift through a device's images and compare their signatures with those of known illegal material. Upon finding an image that matches a known illegal image, the device would quietly report this to the company behind the algorithm and, ultimately, law enforcement authorities.

To test the robustness of the algorithms, the researchers used a new class of tests called avoidance detection attacks to see whether applying their filter to simulated 'illegal' images would let them slip under the radar of PH-CSS and

avoid detection. Their image-specific filters are designed to ensure the image avoids detection even when the attacker does not know how the algorithm works.

They tagged several everyday images as 'illegal' and fed them through the algorithms, which were similar to Apple's proposed systems, and measured whether or not they flagged an image as illegal. They then applied a visually imperceptible filter to the images' signatures and fed them through again.

After applying a filter, the image looked different to the algorithm 99.9 percent of the time, despite them looking nearly identical to the human eye.

The researchers say this highlights just how easily people with illegal material could fool the surveillance. For this reason, the team have decided to not make their filter-generation software public.

Co-lead author Ana-Maria Cretu, Ph.D. candidate at the Department of Computing, said: "Two images that look alike to us can look completely different to a computer. Our job as scientists is to test whether privacy-preserving algorithms really do what their champions claim they do.

"Our findings suggest that, in its current form, PH-CCS won't be the magic bullet some hope for."

Co-lead author Shubham Jain, also a Ph.D. candidate from the Department of Computing added: "This realization, combined with the privacy concerns attached to such invasive surveillance mechanisms, suggest that even the best PH-CSS proposals today are not ready for deployment."

"Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning" by Shubham Jain, Ana-Maria Cretu, and Yves-Alexandre de Montjoye. Published 9 November 2021 as part of USENIX Security Conference in Boston, U.S..

More information: Shubham Jain, Ana-Maria Cretu, Yves-Alexandre de Montjoye, Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. arXiv:2106.09820v1 [cs.CR], arxiv.org/abs/2106.09820

Provided by Imperial College London