Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash




Lukas Struppek* [1], Dominik Hintersdorf* [1], Daniel Neider [2 3], Kristian Kersting [1 4]

* Equal Contribution
[1] Technical University of Darmstadt, Germany
[2] Carl von Ossietzky University of Oldenburg, Germany
[3] Max Planck Institute for Software Systems, Kaiserslautern, Germany
[4] Hessian Center for AI (hessian.AI), Germany

Hot Topic In the News

TechCrunch
Apple confirms it will begin scanning iCloud Photos for child abuse images
5 Aug 2021

CPO Magazine
Apple's New Plan To Scan iCloud Photos Raises Concerns About Mass Surveillance
17 Aug 2021

Input Mag
Sneaky Apple scrubbed all mention of widely hated CSAM scanning from its site
15 Dec 2021

WIRED
EU Plan to Scan Private Messages for Child Abuse Images Puts Encryption at Risk
1 month ago

At a Glance

Apple recently revealed its **deep perceptual hashing system NeuralHash to detect child sexual abuse material (CSAM) on user devices** before files are uploaded to its iCloud service

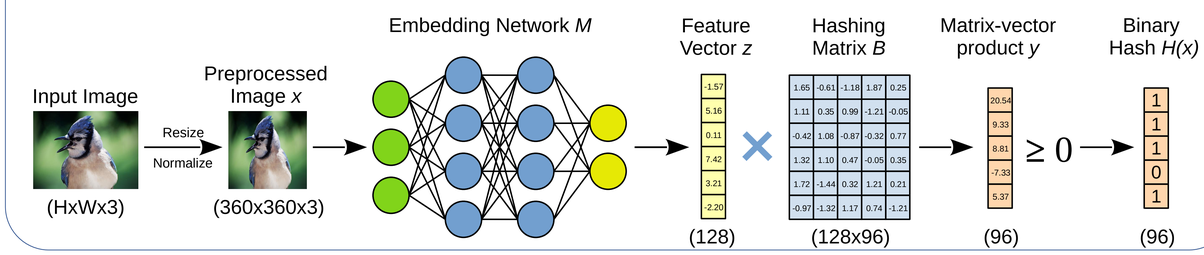
Investigation of the **technical vulnerabilities of client-side scanning based on deep perceptual hashing** from a machine learning perspective.

- Investigation of Apple's NeuralHash as a use case for client-side scanning systems.
- Innocent users could be framed with hash collision attacks.
- A simple image editor is sufficient to avoid detection by the system

➥ **NeuralHash and related deep perceptual client-side scanning systems are not robust and do not provide a safe method for crime detection on user devices.**

Deep Perceptual Hashing – A Safe Way for Crime Detection?

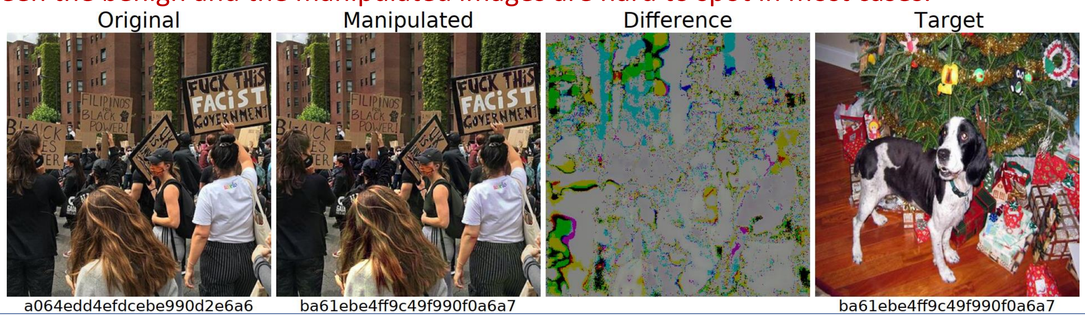
Neural Network-based hashing algorithms compute similar hash values (fingerprints) for visually similar inputs. Hashes are compared to a database with hashes from known illegal material.



1.) Collision Attacks – Framing Innocent Users

Goal: The attacker manipulates benign images so that the hash assigned to them matches a hash from the CSAM database to force a false-positive detection by the system.

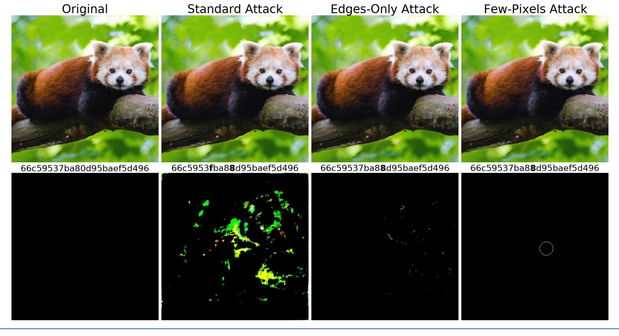
Results: Hash collisions could be forced in >90% of all evaluated cases. Visual differences between the benign and the manipulated images are hard to spot in most cases.

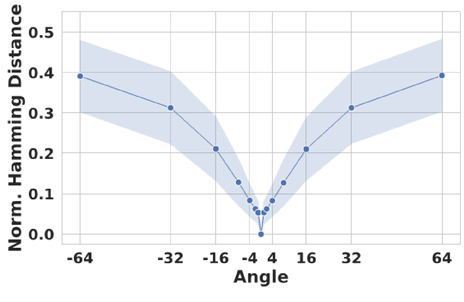


2.) Evasion Attacks – Outsmarting the System

Goal: The adversary manipulates images with malicious content, such as CSAM, to avoid detection. Manipulations are either fine-tuned perturbations or standard image transformations.

Results: NeuralHash is not robust against simple fine-tuned perturbations – changing a few pixels is sufficient to avoid detection. Moreover, using a simple image editor to slightly rotate, crop or mirror an image already leads to strong hash changes.





Lessons and Implications

Current systems are not robust!

- **Simple image manipulations are sufficient to avoid detection.**
- **No deep technical knowledge required.**

System misuse for malicious purposes!

- **Framing or monitoring of innocent users with hash collision attacks is possible.**
- **Governments or organizations with control over the system might extend the database with additional, non-criminal content for surveillance.**

➥ **NeuralHash and related deep perceptual client-side scanning systems should not be deployed on user devices in their current form!**

- ! **Systems are easy to manipulate, pose a risk of misuse, and lack robustness.**
- ! **No safe method for legal violation detection.**

Contact

Lukas Struppek
Technical University of Darmstadt
lukas.struppek@cs.tu-darmstadt.de
@LukasStruppek

Dominik Hintersdorf
Technical University of Darmstadt
dominik.hintersdorf@cs.tu-darmstadt.de
@D0miH

Code

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