Blogpost: Deep perceptual hashing is not robust to adversarial detection avoidance attacks

Avril Wong, Shubham Jain, Ana-Maria Cretu, Yves-Alexandre de Montjoye

Transcript of the blogpost at: https://cpg.doc.ic.ac.uk/blog/deephash-not-robust-to-detection-avoidance/

Abstract – *Perceptual-hashing based client-side scanning is promoted by governments[1] and technology companies[2] as a privacy-preserving solution to detect illegal content on end-to-end encrypted platforms. The client-side scanning solutions, if deployed, would scan the media (such as images and videos) on the user device before they are encrypted and sent on a messaging platform (e.g. WhatsApp or Signal) or uploaded to a cloud service (e.g. iCloud). In a paper presented this week at 31st USENIX Security Symposium[3] in Boston, USA we show that existing perceptual hashing algorithms are not robust to adversarial evasion attacks. More specifically, an attacker with only black-box access to the system could modify an image such that the modified image evades detection by the client-side scanning system while remaining visually similar to the original illegal image. We here extend the attack to the state-of-the-art deep perceptual hashing algorithms for image copy detection and show that even they are vulnerable our detection avoidance attack.*

August 11, 2022

Perceptual-hashing based client-side scanning is promoted by governments[1] and technology companies[2] as a privacy-preserving solution to detect illegal content on end-to-end encrypted platforms. The client-side scanning solutions, if deployed, would scan the media (such as images and videos) on the user device before they are encrypted and sent on a messaging platform (e.g. WhatsApp or Signal) or uploaded to a cloud service (e.g. iCloud). In a paper presented this week at 31st USENIX Security Symposium[3] in Boston, USA we show that existing perceptual hashing algorithms are not robust to adversarial evasion attacks. More specifically, an attacker with only black-box access to the system could modify an image such that the modified image evades detection by the client-side scanning system while remaining visually similar to the original illegal image.

We here extend these results to deep perceptual hashing algorithms, perceptual hashing algorithms learned by machine learning models. NeuralHash[2] which was proposed by Apple as part of their client-side scanning solution is a deep perceptual hashing algorithm. More specifically we demonstrate the applicability of the attack against the winning model[4] of Facebook's Image Similarity Challenge 2021[5] by Shuhei Yokoo. Based on EfficientNetv2[6], the Yokoo model outputs a 256-sized embedding for each image. In line with the strategy used by NeuralHash, we turn Yokoo's embeddings into a 256-bit binary hash using a

locality-sensitive hashing[7] algorithm and use the normalized Hamming distance for detection. We call this the Yokoo hash.

Detecting illegal content using client-side scanning requires finding the illegal content and minimizing the number of false positives. This is done using a threshold, any image at a distance $\leq T$ from an illegal image would be flagged as illegal. We here picked three thresholds such that the lowest threshold would result in 1% of false positives against a database of size 1 million, while the second and third one would result in 10% and 20% false positives[1].

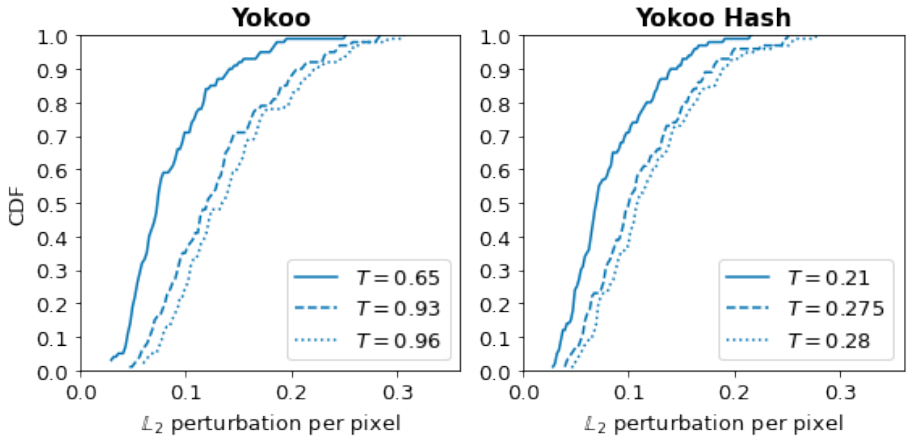

Figure 1: Fig. 1: Cumulative distribution function (CDF) for the L_2 perturbations for different algorithms and thresholds and for 100 successfully attacked images.

We performed our attack on 100 random images from the ISC 2021 reference dataset and different thresholds, **succeeding in finding a perturbation 100%** of the time. Figure 1 shows the results of our attack against Yokoo (directly using embeddings for matching) and Yokoo Hash model. To quantify the visual difference between the original and modified image obtained using our attack, we report the L_2 distance between the two, normalized over the number of pixels (to enable comparison across images of different sizes).

Our results show that the attack[2] works very well, as the hash of an illegal image can be pushed away from the original with very small changes to the image. We can see that for the lowest threshold (yielding an FPR of 1% and typical of

[1]The false positive rate was calculated by using ISC 2021 training dataset as a set of query images, and ISC 2021 reference dataset as the database.

[2]To speed up the attack, here we use slightly different attack parameters compared to our attack on shallow perceptual hashing algorithms (i.e. pHash, aHash, dHash and PDQ). More precisely, we set the momentum parameter (μ) for gradient calculation as 0.1, step size for updating perturbation (η) as 0.04, initial maximum allowed perturbation (ϵ_0) as 0.2, and step size for incrementing the perturbation budget (η_ϵ) as 0.01. Details of these parameters and the attack algorithm can be found in our USENIX paper[3].

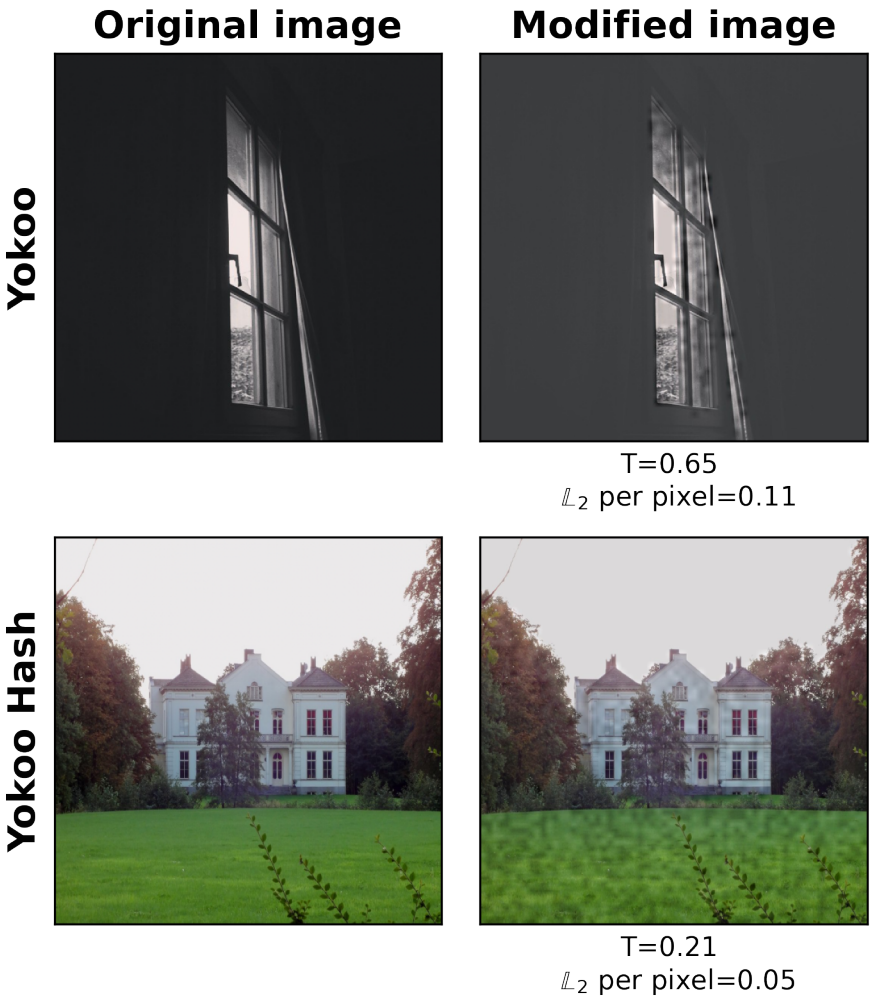

Figure 2: Fig. 2: Examples of original and modified images for different hashing functions. All modified images evade detection for the threshold T while maintaining a small L_2 perturbation per pixel and high visual similarity to the original image.

what is likely to be used in practice), **more than 71%** of the modified images differ by less than 0.1 from the original for Yokoo (73% for Yokoo hash). As the threshold increases, a larger perturbation would have to be added, e.g., for the largest threshold, 43% of images would differ by an L_2 perturbation per pixel of 0.1 for Yokoo Hash (correspondingly 31% for Yokoo). Figure 2 shows that the modified images would evade detection while remaining visually similar to the original image. In practice, the lowest threshold we attack would result in almost **45M images being falsely flagged** (and possibly decrypted) daily. While the number of images flagged would go up to 450M (1 in every 10 images sent) and 900M (1 in every 5 images sent) for the largest two thresholds respectively[3].

Taken together, our results show that the state-of-the-art deep hashing algorithm for edited copy detection is vulnerable to evasion attacks.

Bibliography

[1] Hearns A., 2022, July 21. UK cybersecurity chiefs back plan to scan phones for child abuse images, The Guardian.

[2] Apple Inc., 2021, August. CSAM Detection, Technical Summary, Apple Inc.

[3] Jain, S., Cretu A., and de Montjoye, Y.A., 2021. Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning, 31st USENIX Security Symposium.

[4] Yokoo, S., 2021, December 8. Contrastive Learning with Large Memory Bank and Negative Embedding Subtraction for Accurate Copy Detection, *arXiv preprint arXiv:2112.04323.*.

[5] Douze, M., Tolias, G., Pizzi, E., Papakipos, Z., Chanussot, L., Radenovic, F., Jenicek, T., Maximov, M., Leal-Taixé, L., Elezi, I. and Chum, O., 2021. The 2021 image similarity dataset and challenge, *arXiv preprint arXiv:2106.09672.*

[6] Tan, M. and Le, Q., 2021, July. Efficientnetv2: Smaller models and faster training. In International Conference on Machine Learning (pp. 10096-10106). PMLR.

[7] Dasgupta, A., Kumar, R. and Sarlós, T., 2011, August. Fast locality-sensitive hashing. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1073-1081).

[8] WhatsApp blog, 2017. Connecting one billion users every day, Accessed on June 6, 2021.

[3]WhatsApp reported[8] 4.5 billion images were shared on its platform daily, and we report numbers assuming the prevalence rate of illegal content to be 0.0001.