Final Report: CS 4501
Avery Witherow (adw5nej)

THE RISKS AND FLAWS OF A PERCEPTUAL HASHING ALGORITHM

THE INITIAL IDEA

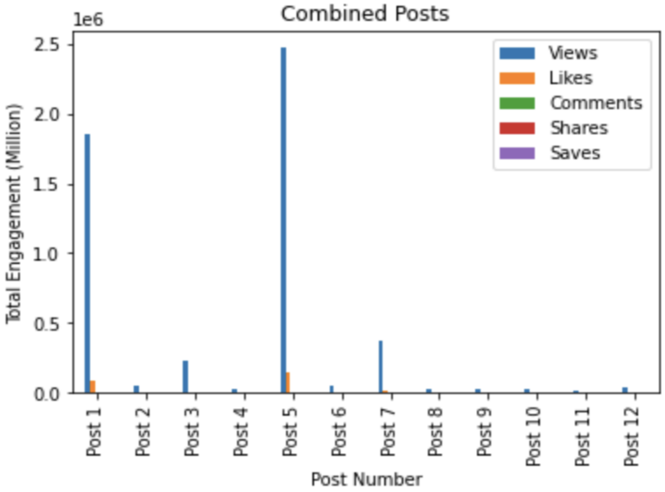
Beginning this project, neural models sparked my interest in current side work being conducted on social media. Specifically, I target the underlying structure behind platforms to achieve accelerated rates of follower growth and audience reach across Instagram and TikTok. Blindly, I had been using a technique of mirroring an image or video across social media platforms that resulted in absurd amounts of views for reposted content. Questioning the taxed reach of reposted content in the past, experiencing this stark change with a simple modification, it generated more curiosity in exploring this rapidly growing context with what causes this. Additionally, the announcement of Apple's CSAM detection, utilizing NeuralHash, a perceptual hashing algorithm, can directly be integrated into this project context. The intended use of NeuralHash presented possible rationale for the result and presented multiple issues of possible adversarial attacks and flaws to explore.

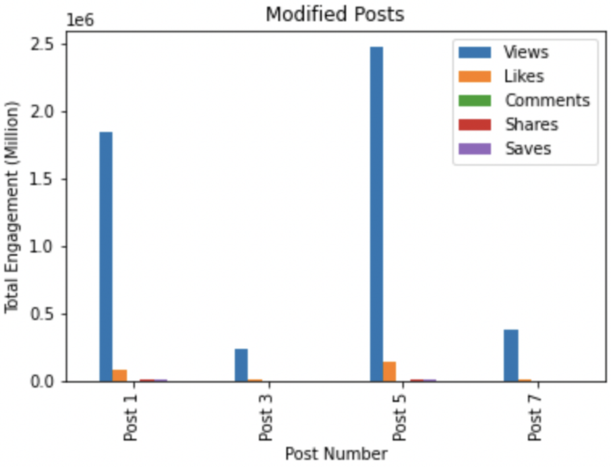
HOW DOES THIS CONTEXTUALLY CONNECT?

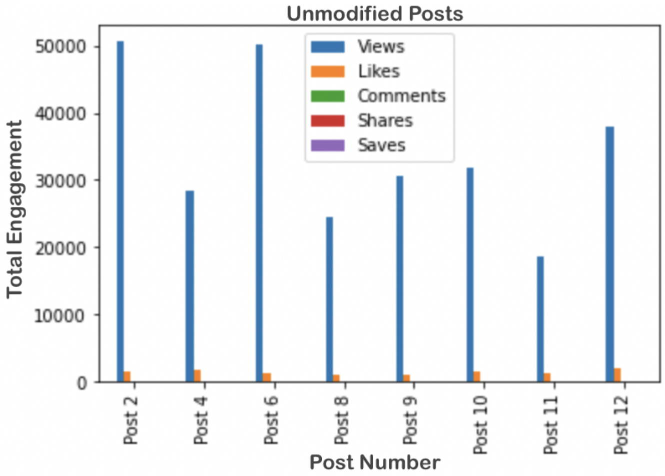
The announcement of Apple's CSAM detection, utilizing a neural model fitted with known child pornography to target sexual predators, provided fuel to understanding to why the mirror modification bypassed a taxed audience reach on social media. NeuralHash uniquely assigns a hash to each image or video and is highlighted for decreased sensitivity to slight modifications such as cropping, pixelation, or resizing. If a sexual predator's device had an image or video fitted into the neural model, Apple would be notified of the matching hash upon an iCloud scan, landing the sexual predator in prison. Similarly, on social media, if an individual posts an image that has already been uploaded or fitted into the model, it will be taxed in audience reach due to the matching hash. The platforms encourage creativity and originality, furthering rationale for this content being taxed, even through cropping or resizing. The real danger, and the biggest flaw noticed throughout this project, is seen through the results of the mirror modification. The bits of the resulting hash greatly differ additionally offering little pattern resemblance as other modifications while maintaining the integrity of the image or video. As demonstrated on social media platform Instagram, this modification resulted in video performance differing starkly compared to the unmodified videos posted on the test account. Through utilizing this evidence and directly integrating it into a resembling detection system, it can be stated that sexual predators could abuse and bypass Apple's CSAM detection through the mirror modification.

TESTING ON INSTAGRAM, CHALLENGES, RESULTS

To test the mirror modification, Instagram served as the primary social media platform of choice. I purchased an account with 10,000 followers to conduct the tests on, posting videos additionally onto Instagram Reels (TikTok competitor). I initially felt Instagram Reels would best combine the features of Instagram and TikTok to centralize the tests and provide similar results, rather than individually testing both platforms. I created a Python script to automate the process of downloading the videos for posting. I had to utilize the Selenium library to automate this download process to avoid the watermark of videos retrieved from TikTok. Learning an unfamiliar library of web automation took time, but was overall useful in maintaining video integrity. I additionally extended the script to download videos from other sources, such as Instagram, allowing the user to download videos without watermarks from other social media platforms. This permitted more accurate testing, resembling the video without the watermark as if it were the original. Below show graphs, retrieved from the script, showcasing the views as a metric for each individual video. The y-axis displays the total engagement of the video, the x-axis displays the post number of a video. Upon each run of the script, I manually noted the mirrored or unmirrored random selection, indicated by 0 or 1, into a text file. The data was further separated into two separate graphs based on if the modification was made.



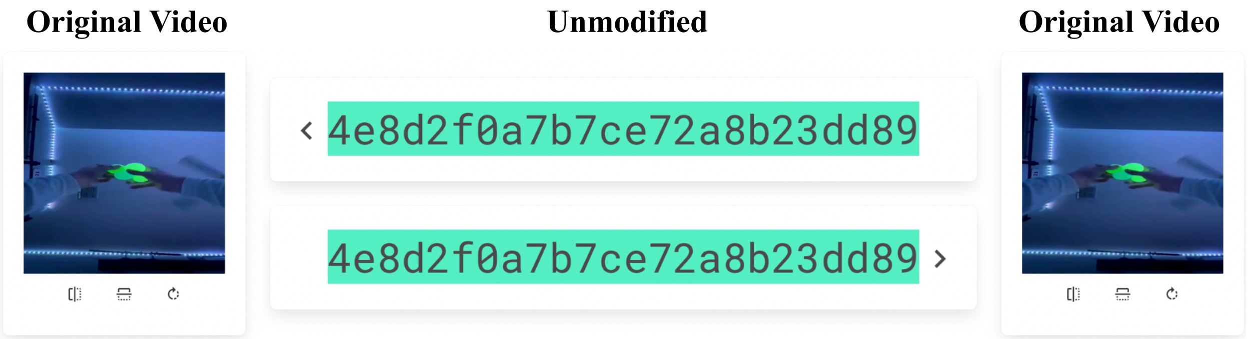




In the graph combining both mirrored and unmirrored videos, there is no particular trend in the data. The data points can be seen starkly differing, proving the effectiveness of the modification. It can be indicated through the large spikes in audience reach of the videos receiving the mirrored modification- even without separation of the data. Instagram continues to push the video to a larger audience reach due to the falsely interpreted originality rather than capping the audience reach of the repost. It can be indicated by posts not undergoing modification that the engagement tends to be fairly constant demonstrating the capped audience reach. This can be supported by each of the posts with large spikes of audience reach matching on the mirrored individual graph in Post 1, Post 3, Post 5, and Post 7. Similarly, on the posts without large spikes of audience reach, the post numbers can be seen matching the unmirrored individual graph in Post 2, Post 4, Post 6, Post 8, Post 10, and Post 12.

MIRROR MODIFICATION ON NEURAL HASH

At the original due date of the project, the Intel-based Macbook had the framework associated with Apple's CSAM detection to make it fully functional to integrate the videos into with the Python script. However, upon upgrading to the M1 Macbook, the framework was removed from the system for a future update due to backlash in potential for adversarial attacks and the privacy violation posed. Due to this unexpected change and unsupported code, I found another solution hosted on Github by user 'greentfrapp', utilizing the same code and framework in Javascript. Ultimately, the site achieved the desired result, though it had to be manually conducted. To compare the videos, the Python script selects the first frame of both outputted videos and additionally downloads the pertaining labeled frames. As mentioned, a goal of this was maintaining video and image integrity, leaving the mirror modification as one of the only modifications to be compared against. The resulting hash can even provide more rationale behind the audience reach on each post. Below, the first frame of a video, retrieved and posted on December 21, 2021, is utilized to highlight the similarities and differences in the hash generated by NeuralHash, based on the modification of the original video.



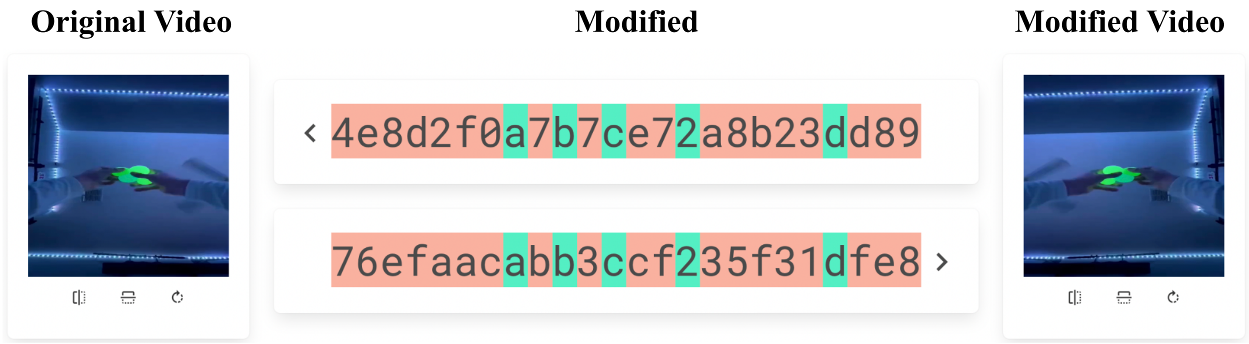
Original Video

Unmodified

Original Video

‹ 4e8d2f0a7b7ce72a8b23dd89

4e8d2f0a7b7ce72a8b23dd89 ›



‹ 4e8d2f0a7b7ce72a8b23dd89

76efaacabb3ccf235f31dfe8 ›

The hash in the first example indicates the similarity between the first frame of the same videos, displaying “4e8d2f0a7b7ce72a8b23dd89” for both first frames and matching all 24/24 bits. The hash in the second example indicates the difference between the first frame of the different videos, displaying “4e8d2f0a7b7ce72a8b23dd89” for the original video and “76efaacabb3ccf235f31dfe8” for the modified video and only matching 5/24 bits. The hash of the modified video having a 79.2% difference and 20.8% similarity in bits is a very weak correlation, allowing the modification to be effectively utilized in any system that takes on a perceptual hashing algorithm.

WHAT CAN BE DONE? WHAT IS A FIX?

The threat of the mirror modification on perceptual hashing algorithms presents a flaw displayed on social media platform, Instagram, along with Apple’s CSAM detection system. Both lack the coverage in a model that carries out a large component of their system. For Instagram, determining the originality of a post and further determining the audience reach of the post, for Apple, protecting against the distribution of child pornography across Apple devices. The largest surface argument behind Apple’s system involved the instance of a hash collision resulting in a false positive. The probability of this occurring is very low, but is still of relevance as a counterargument. On Instagram may not have much of an implication other than a taxed audience reach, but in Apple’s system for detecting the distribution of child pornography, this false positive could land an innocent individual in prison. Still remaining an area of concern, the bigger concern is sexual predators utilizing the flaw, due to lacking coverage, in Apple’s system to bypass and continue the distribution of child pornography. As a new and uncovered area of concern for models utilizing perceptual hashing algorithms, it is difficult to provide an optimal solution to fix the issue. However, a solution could involve fitting the model with a counter-mirrored image. When fitting the model with known child pornography, a mirror of the image or video would also be included. NeuralHash, designed to be insensitive to smaller changes, would have full coverage for image modifications that preserve image or video integrity. As a result, this would seal the existing lack of coverage and provide the desired full functionality behind the system. Through increasing the accuracy of the model, it does resurface the once smaller issue of false positives occurring- further increasing by a factor of two. As a temporary collision resolution, matching hashes should be reviewed by the human eye to ensure accuracy. As artificial intelligence continues to grow and expand, there could potentially be a more effective solution.

CLOSING THOUGHTS AND CONCLUSION

This project helped to give an explanation of why the mirror modification entailed past results across Instagram and Tiktok. It additionally allowed exploration of the flaws behind Apple's CSAM detection system and how the two examples contextually weave due to their use of perceptual hashing algorithms. A solution to combat the use of the mirror modification could include fitting a counter-mirrored image into the model, sealing the lack of coverage and increasing detection accuracy, but also resurfacing the issue of hash collisions resulting in false positives. In a new and growing context, I hope to gather more information and ideas to provide an accurate and cost–effective solution of collision resolution to future systems as means to achieve their desired goal.