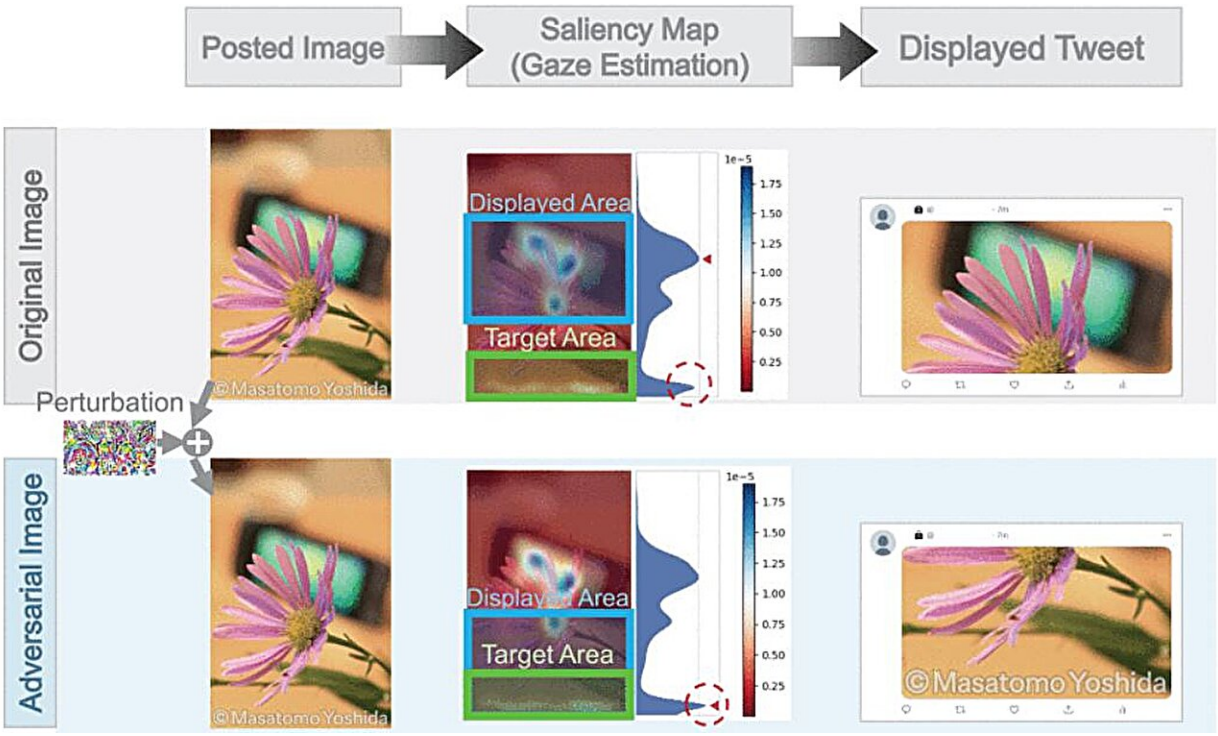
Enhancing automatic image cropping models with advanced adversarial techniques

August 1 2024



Many commercial image cropping models utilize saliency maps (also known as gaze estimation) to identify the most critical areas within an image. In this study, researchers developed innovative techniques to introduce imperceptible noisy perturbations into images, thus influencing the output of cropping models. This approach aims to prevent essential parts of images, such as copyright information or watermarks, from being inadvertently cropped, thus promoting fairness in AI models. Credit: Masatomo Yoshida / Doshisha University

Image cropping is an essential task in many contexts, right from social media and e-commerce to advanced computer vision applications. Cropping helps maintain image quality by avoiding unnecessary resizing, which can degrade the image and consume computational resources. It is also useful when an image needs to conform to a predetermined aspect ratio, such as in thumbnails.

Over the past decade, engineers around the world have developed various machine learning (ML) models to automatically crop images. These models aim to crop an input image in a way that preserves its most relevant parts.

However, these models can make mistakes and exhibit biases that, in the worst cases, can put users at legal risk. For example, in 2020, a lawsuit was filed against X (formerly Twitter) because its automatic cropping function hid the copyright information in a retweeted image.

Therefore, it is crucial to understand the reason image cropping ML models fail so as to train and use them accordingly and avoid such problems.

Against this background, a research team from Doshisha University, Japan, set out to develop new techniques to generate adversarial examples for the task of image cropping.

As explained in their paper, published in *IEEE Access* on June 17, 2024, their methods can introduce imperceptible noisy perturbations into an image to trick models into cropping regions that align with user intentions, even if the original model would have missed it.

Doctoral student Masatomo Yoshida, the first author and lead researcher of the study, said, "To the best of our knowledge, there is very little research on adversarial attacks on image cropping models, as most

previous research has focused on image classification and detection. These models need to be refined to ensure they respect user intentions and eliminate biases as much as possible while cropping images."

Masatomo Yoshida and Haruto Namura from the Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan and Masahiro Okuda from the Faculty of Science and Engineering at Doshisha University, were also involved in the study.

The researchers developed and implemented two distinct approaches for generating adversarial examples—a white-box approach and a black-box approach.

The white-box method, requiring access to the internal workings of the target model, involves iteratively calculating perturbations to input images based on the model's gradients.

By employing a gaze prediction model to identify salient points within an image, this approach manipulates gaze saliency maps to achieve effective adversarial examples. It significantly reduces perturbation sizes, achieving a minimum size 62.5% smaller than baseline methods across an experimental image dataset.

The black-box approach utilizes Bayesian optimization to effectively narrow the search space and target specific image regions. Similar to the white-box strategy, this approach involves iterative procedures based on gaze saliency maps.

Instead of using internal gradients, it employs a tree-structured Parzen estimator to select and optimize pixel coordinates that influence gaze saliency, ultimately producing desired adversarial images. Notably, black-box techniques are more broadly applicable in real-world scenarios and hold greater relevance in cybersecurity contexts.

Both approaches show promise based on experimental outcomes. As graduate student Haruto Namura, a participant in the study, explains, "Our findings indicate that our methods not only surpass existing techniques but also show potential as effective solutions for real-world applications, such as those on platforms like Twitter."

Overall, this study represents a significant advancement toward more reliable AI systems, crucial for meeting public expectations and earning their trust. Enhancing the efficiency of generating adversarial examples for image cropping will propel research in ML and inspire solutions to its pressing challenges.

Professor Masahiro Okuda, advisor to Namura and Yoshida, concludes, "By identifying vulnerabilities in increasingly deployed AI models, our research contributes to the development of fairer AI systems and addresses the growing need for AI governance."

More information: Masatomo Yoshida et al, Adversarial Examples for Image Cropping: Gradient-Based and Bayesian-Optimized Approaches for Effective Adversarial Attack, *IEEE Access* (2024). DOI: 10.1109/ACCESS.2024.3415356

Provided by Doshisha University