How Real Is Real?
A Human Evaluation Framework for Unrestricted Adversarial Examples

Dren Fazlija[1], Arkadij Orlov[2], Johanna Schrader[1], Monty-Maximilian Zühlke[1], Michael Rohs[2], Daniel Kudenko[1]

[1]L3S Research Center, Appelstr. 4, 30167 Hannover, Germany
[2]Leibniz University Hannover, Welfengarten 1, 30167 Hannover, Germany
{dren.fazlija, schrader, zuehlke, kudenko}@l3s.de,
arkadij.orlov@stud.uni-hannover.de, michael.rohs@hci.uni-hannover.de

Abstract

With an ever-increasing reliance on machine learning (ML) models in the real world, *adversarial examples* threaten the safety of AI-based systems such as autonomous vehicles. In the image domain, they represent maliciously perturbed data points that look benign to humans (i.e., the image modification is not noticeable) but greatly mislead state-of-the-art ML models. Previously, researchers ensured the imperceptibility of their altered data points by *restricting* perturbations via ℓ_p norms. However, recent publications claim that creating natural-looking adversarial examples without such restrictions is also possible. With much more freedom to instill malicious information into data, these *unrestricted* adversarial examples can potentially overcome traditional defense strategies as they are not constrained by the limitations or patterns these defenses typically recognize and mitigate. This allows attackers to operate outside of expected threat models. However, surveying existing image-based methods, we noticed a need for more human evaluations of the proposed image modifications. Based on existing human-assessment frameworks for image generation quality, we propose SCOOTER – an evaluation framework for unrestricted image-based attacks. It provides researchers with guidelines for conducting statistically significant human experiments, standardized questions, and a ready-to-use implementation. We propose a framework that allows researchers to analyze how imperceptible their unrestricted attacks truly are.

Introduction

Since 2013, it has been well-established that malicious entities can mislead sophisticated computer vision models by adding imperceptible noise to images (Szegedy et al. 2013). The resulting *adversarial examples* (AEs) look benign to humans. However, due to the restricted nature of the attacks, their effectiveness can be limited significantly through image pre-processing (Dziugaite, Ghahramani, and Roy 2016) and certified robust defense strategies (Li, Xie, and Li 2020). Hence, *unrestricted* AEs have garnered increasingly more interest in the last few years. Such attacks alter images significantly through modifications that humans easily overlook, which is achieved by taking semantic information into account. For example, (Shamsabadi, Sanchez-Matilla, and Cavallaro 2020) perform significant color changes only to non-sensitive areas (e.g., furniture) that look natural to the human eye across a wide range of colors. Works on unrestricted attacks should rigorously assess the imperceptibility of the resulting images, as this characteristic can no longer be assumed. However, few publications employ human evaluation experiments to support their claim – none of which offer statistically significant insights. Hence, providing the research community with a statistically significant human evaluation protocol based on well-established study design recommendations is crucial.

Contributions. To facilitate research into unrestricted adversarial examples, we propose SCOOTER (**S**ystemizing **C**onfusion **O**ver **O**bservations **T**o **E**valuate **R**ealness) – a human evaluation framework for examining the quality of unrestricted adversarial images (i.e., the *imperceptibility* of modifications). Drawing inspiration from existing tools (Otani et al. 2023) and following study design recommendations (Aguinis, Villamor, and Ramani 2021), SCOOTER enables researchers to make statistically significant claims about the imperceptibility of image-based attacks. The SCOOTER framework encompasses (i) a ready-to-use web application with a modular design allowing researchers to integrate their images easily; (ii) a carefully crafted study protocol that guides researchers in every step of performing online studies; (iii) an online leaderboard enabling the comparisons of state-of-the-art attacks across different target models; (iv) an image database containing all generated AEs for further analyses.

State-Of-The-Art Assessment Protocols

The most similar work to ours is the evaluation protocol of (Otani et al. 2023) for analyzing the quality of Text-To-Image generators. The authors provide researchers with well-designed domain-specific questions and user interfaces, recommendations for several design choices (e.g., requirements that participants need to fulfill), and templates for reporting human evaluation results. While our goals align with the authors', their protocol does not sufficiently guide inexperienced researchers in difficult aspects of experiment design. Most notable is the lack of methods to guarantee high-quality evaluation data. For example, their protocol does not cover standard measures like attention and instruction manipulation checks. The researchers also publicly state their eligibility requirements, which is not recommended as

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

it increases the self-misrepresentation of participants (Aguinis, Villamor, and Ramani 2021; Bauer et al. 2020). In contrast to their work, we aim to support inexperienced researchers by rigorously defining every detail of the study design. Another adjacent publication (Zhou et al. 2019) provides a basic framework for collecting human image quality assessments. While the resulting protocols $\text{HYPE}_{\text{time}}$ and HYPE_{∞} are widely used in subjective image quality assessment tasks, they display similar weaknesses to (Otani et al. 2023). We build on the insights of both publications while considering these flaws.

Framework Design

Online Study Design. To evaluate the imperceptibility of unrestricted attacks, we propose to conduct a 13-minute online study on Prolific[1]. We use Prolific because it can prescreen participants without publicly sharing eligibility requirements while providing higher-quality data than other services (Douglas, Ewell, and Brauer 2023). We use the built-in prescreeners to filter out workers who are colorblind and those who self-report to be not fluent in English. We further ensure workers' capabilities by performing a short colorblindness and comprehension check. In line with Prolific compensation guidelines and related publications, we offer an average compensation of £7.60 per hour. To support our protocol's critical design decisions (e.g., the slider-based input as in Figure 3), we developed a Flask-based[2] web application to perform these studies. Research groups can use the web app to replicate our study design for their experiments.

Colorblindness Check. The most prominent attack vectors for unrestricted AEs are the colors of an image. As such, colorblind annotators will likely overestimate the imperceptibility of most attacks. Thus, participants must correctly classify five different Ishihara-like images (Ishihara et al. 1918) before accessing the study's central portion (see Figure 1). These images emulate so-called Ishihara plates, which are widely used diagnostic tools for vision deficiencies. Four images show a digit, while one image always displays no digit. Failing this check will end the study, and the user will be compensated for 1 minute of work.

Comprehension Check. After passing the colorblindness check, we provide users with a brief explanation outlining common modification strategies. Examples include image filters and the change of colors in a particular area of the image. After reading the explanation, the user must pass a comprehension check. We display six image pairs, each containing a random unmodified and one random modified image (see Figure 2). To move on to the main portion of the study, participants must correctly classify the modified image of at least five pairs. Failing this check will end the study, and the user will be paid for 2.5 minutes of work.

Main Study. Here, we aim to analyze the imperceptibility of an attack strategy for one specific victim model. We evaluate generated AEs by asking users to rate the degree of modification for 106 images, 50 of which are unmodified ImageNet (Russakovsky et al. 2015) samples, that the victim classifies correctly with high confidence. Another 50 images represent random AEs generated by the assessed attack. The remaining six images represent attention checks to ensure the attentiveness of participants. Instead of a forced binary choice between ”modified” and ”unmodified”, we want users to rate how confident they feel about the degree of modification via a slider input (see Figure 3).

The input ranges from -100, *I am 100% certain that this image is unmodified*, to +100, *I am 100% certain that this image is modified*. Collecting continuous ratings from many participants will lead to better-informed comparisons between unrestricted attacks, as such ratings can capture finer nuances of attack perceptibility (Chyung et al. 2018). For instance, attack A, which produces successful attacks with imperceptible modifications, could be rated the same as attack B, whose successful modifications ”barely convince” the average user.

Empirical Sample Size Estimation. A key factor for conducting statistically significant studies is the choice of an appropriate sample size. However, we cannot perform an apriori sample size estimation due to this domain's lack of an established effect size. Hence, we must perform studies to empirically determine a suitable number of participants per (attack, model) pair. For this purpose, we plan to collect large amounts of data for three attacks on one single victim model. Given 3,000 modified and 3,000 unmodified images per (attack, model) pair, which we distribute across 60 unique datasets, we want to collect at least ten samples per image. Hence, we must invite at least 600 people per (attack, model) pair. To create a recommended buffer of 15% for low-quality annotations (Aguinis, Villamor, and Ramani 2021), we will invite another 90 annotators per attack-model pair. Using the adversarially trained ResNet-50 (He et al. 2015) model of (Salman et al. 2020), we will collect data from 690 annotators for three attacks, which vary in complexity. Based on these results, we will then be able to determine the sufficient (and ideally much smaller) sample size needed for analyzing unrestricted attacks.

Conclusions

Unrestricted image-based attacks will play a significant role in the near future, especially considering the rapid progress made in AI-based image generation (Croitoru et al. 2023). We provide an accessible toolbox that supports high-quality research into this developing field while increasing awareness for unrestricted AE research.

Acknowledgments

This work is supported by the Center for Digital Innovations of Lower Saxony (ZDIN) and has received funding from the Lower Saxony Ministry of Science and Culture under grant numbers ZN3492 (”Zukunftslabor Gesellschaft & Arbeit”) and 51171389 (”CAIMed - Niedersächsisches Zentrum für KI & Kausale Methoden in der Medizin”) within the Lower Saxony “zukunft.niedersachsen” funding program of the Volkswagen Foundation.

[1] https://www.prolific.com/

[2] https://flask.palletsprojects.com/

References

Aguinis, H.; Villamor, I.; and Ramani, R. S. 2021. MTurk Research: Review and Recommendations. *Journal of Management*, 47(4): 823–837.

Bauer, B.; Larsen, K. L.; Caulfield, N.; Elder, D.; Jordan, S.; and Capron, D. 2020. Review of best practice recommendations for ensuring high quality data with amazon's mechanical turk.

Chyung, S. Y.; Swanson, I.; Roberts, K.; and Hankinson, A. 2018. Evidence-based survey design: The use of continuous rating scales in surveys. *Performance Improvement*, 57(5): 38–48.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Douglas, B. D.; Ewell, P. J.; and Brauer, M. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *Plos one*, 18(3): e0279720.

Dziugaite, G. K.; Ghahramani, Z.; and Roy, D. M. 2016. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385.

Ishihara, S.; et al. 1918. Tests for color blindness. *American Journal of Ophthalmology*, 1(5): 376.

Li, L.; Xie, T.; and Li, B. 2020. Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*.

Otani, M.; Togashi, R.; Sawai, Y.; Ishigami, R.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; and Satoh, S. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14277–14286.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

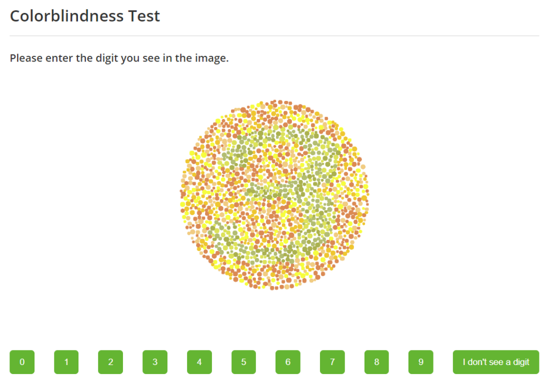
Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; and Madry, A. 2020. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33: 3533–3545.

Shamsabadi, A. S.; Sanchez-Matilla, R.; and Cavallaro, A. 2020. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1151–1160.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Zhou, S.; Gordon, M.; Krishna, R.; Narcomey, A.; Fei-Fei, L. F.; and Bernstein, M. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.

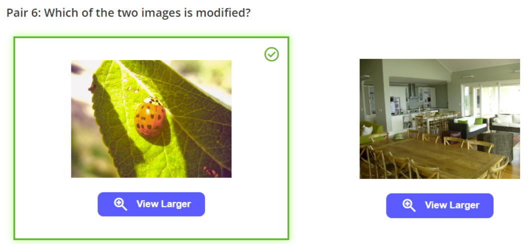
Appendix



Colorblindness Test

Please enter the digit you see in the image.

Figure 1: Prototype of the Colorblindness Check Interface



Pair 6: Which of the two images is modified?

Figure 2: Prototype of the Comprehension Check Interface

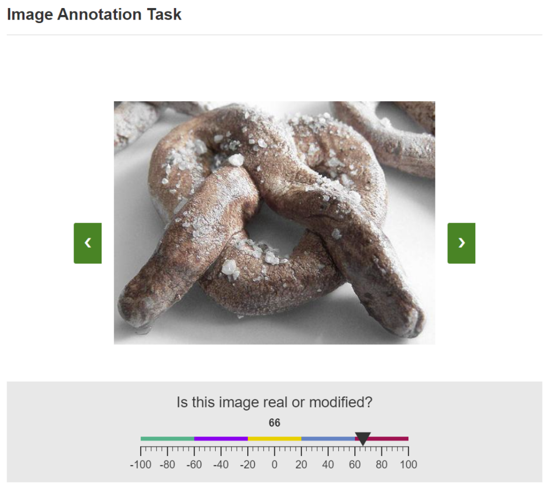


Image Annotation Task

Is this image real or modified?

66

Figure 3: Prototype of the Main Study Interface