Some issues to be addressed.
but easy to fix.

Seminar for Master Students in Data Science

Master Thesis Proposal Draft

Rethinking Adversarial Examples

Yahya Jabary, 11912007

September 23, 2024

Advisor:
Prof. Dr. Schahram Dustdar

Co-Supervisor Alireza Furutanpey:
Dipl. Ing. Alireza Furutanpey

Based on: `https://arxiv.org/abs/2409.05558`

Problem Statement

Modern CAPTCHA systems heavily rely on visual tasks that are intended to be challenging for automated systems but straightforward for human users. However, advancements in image recognition models have significantly diminished the effectiveness of these CAPTCHAs. Specifically, state-of-the-art models can be deceived by introducing visible geometric masks, leading to a substantial drop in their accuracy while maintaining the image's readability for humans. This undermines the fundamental purpose of CAPTCHAs in differentiating between human users and automated bots.

≠ scientific problem statement

Goals and Expected Outcome

The primary objective of this research is to enhance the robustness of CAPTCHA systems against advanced image recognition models by employing geometric masks. The proposed solution involves overlaying CAPTCHA images with various geometric patterns at different opacity levels to disrupt machine recognition without impairing human comprehension.

Characteristics of the solution include:

- **Geometric Mask Types:** Circle, Diamond, Square, and Knit.
- **Opacity Levels:** Ranging from 20% to 50% to balance between machine disruption and human readability
- **Performance Metrics:** Drop in Accuracy @1 (Acc@1) and Accuracy @5 (Acc@5) of vision models, coupled with perceptual quality assessments.

Success criteria include achieving a reduction of over 50 percentage points in Acc@1 for all targeted models while maintaining a perceptual quality score above 0.4 to ensure human users can still interpret the CAPTCHA effectively.

OK ✓

Research Questions

The research will address the following questions:

- **Effectiveness of Geometric Masks:** How do different geometric mask types and opacity levels affect the accuracy of state-of-the-art vision models in solving CAPTCHAs?
- **Human vs. Machine Perception:** Can geometric masks be designed to maximize the accuracy drop in machine models while preserving image semantics for human users? reformulate
- **Model Robustness:** Which vision model architectures are more resilient to geometric mask-induced adversarial attacks, and what characteristics contribute to their robustness?
- **Scalability and Generalizability:** How well do the proposed masking techniques generalize across various datasets and CAPTCHA implementations?

Research Methods

How do you guarantee scientific validity?

The research will involve applying different geometric masks to standardized image datasets and evaluating the performance of multiple vision models against these altered CAPTCHAs.

Models selected include Convolutional Networks (ConvNeXt_XXLarge, ResNet50x64), Transformer-Based Models (Vision Transformers, EVA-02, DFN5B-CLIP-ViT-H), and Robust Models (RoBERTa-B and RoBERTa-L). Datasets used include ImageNet-Enriched, ImageNette, and subsets created for varied experimental depth (SubSet200, SubSet500, and ResizedAll).

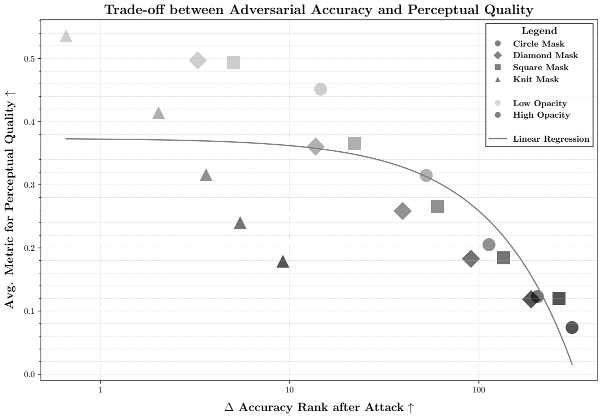
Four types of geometric masks—Circle, Diamond, Square, and Knit—will be applied at varying opacity levels (20%, 30%, 40%, 50%) to assess their impact on model accuracy and image perceptual quality.

The evaluation will focus on model accuracy metrics (Acc@1, Acc@5) and perceptual quality scores (Cosine Similarity, PSNR, SSIM, LPIPS) using a weighted sum to quantify the trade-off between machine disruption based on mask fidelity and human interpretability.

Evaluation

The research will evaluate the effectiveness and therefore the performance in generalizability of geometric masks based on the following criteria:

- **Accuracy Reduction:** Quantifying the drop in Acc@1 and Acc@5 across different models when masks are applied.
- **Perceptual Quality Maintenance:** Ensuring that the masked images retain high perceptual quality scores to remain user-friendly.



For a comparative analysis, the research will assess the resilience of different vision models to geometric masks and identify the most robust architectures. The study will also determine the optimal mask types and opacity levels that offer the best trade-off between machine disruption and human readability.

A successful outcome will demonstrate that geometric masks can consistently lower model accuracy by over 50 percentage points while keeping perceptual quality above acceptable thresholds for human users.

State of the Art

Current CAPTCHA systems, such as Google's reCAPTCHA, predominantly utilize visual tasks that are increasingly vulnerable to sophisticated image recognition models. Recent studies have explored adversarial machine learning to create more resilient CAPTCHAs, but many approaches suffer from being model-specific or compromising human usability.

Advancements in Vision Transformers and robust model architectures like RoBERTa have set new benchmarks in image classification tasks. However, their susceptibility to visible geometric perturbations highlights a gap in developing truly adversarial-resistant CAPTCHA systems. This research builds upon existing adversarial example strategies, emphasizing the need for masks that disrupt machine perception without degrading human interpretability.

No references

Relevance to the Curriculum

This research is highly relevant to curricula focused on computer vision, machine learning, and cybersecurity. It bridges theoretical concepts of adversarial machine learning with practical applications in web security through CAPTCHA systems. The study enhances understanding of model vulnerabilities and robustness, providing valuable insights for courses on:

- **Machine Learning:** Exploring adversarial attacks and defenses.
- **Computer Vision:** Applying geometric transformations and evaluating model performance.
- **Cybersecurity:** Developing secure user authentication mechanisms.

By integrating multidisciplinary approaches, the research aligns with the knowledge and skills imparted in related courses, preparing us for real-world challenges in AI safety and security.

Conclusion

This research aims to fortify CAPTCHA systems against evolving image recognition models by leveraging geometric masks to create robust, human-friendly adversarial examples. By systematically evaluating the impact of different masks and opacity levels on various state-of-the-art models, the study seeks to establish effective strategies for maintaining CAPTCHA security in an AI-advancing landscape. The findings will contribute to the development of more resilient CAPTCHA mechanisms, ensuring continued efficacy in distinguishing human users from automated bots.