Secure CAPTCHA: Smarter CAPTCHAs with Patch-Based Defense

Gaurav Sonawane[1], Ganesh Dake[2], and Nishant Ade[3]
[1,2,3]*Member, JSPM's Imperial College of Engineering and Research, Pune*

***Abstract*—As automated attacks against CAPTCHA systems become increasingly sophisticated, traditional text-based and image recognition challenges are becoming vulnerable to machine learning-based attacks. This paper proposes Secure CAPTCHA, a novel CAPTCHA system that incorporates strategically generated adversarial patches into images to create challenges that are easily solvable by humans but resistant to automated attacks. By introducing carefully crafted patch-based perturbations that exploit the fundamental vulnerabilities of neural networks, our system creates a robust defense mechanism against bot attacks while maintaining human usability. We train our system over 100 epochs and evaluate its performance against state-of-the-art deep learning-based CAPTCHA solvers, demonstrating significant improvements in security metrics compared to conventional approaches. Our experimental results show that Secure CAPTCHA reduces successful automated attacks by 91% while maintaining a human success rate of 94%. The system also demonstrates resilience against various adaptive attack strategies, including feature squeezing, adversarial training, and ensemble methods.**

***Index Terms*— CAPTCHA, Adversarial Noise, Deep Learning, Image Classification**

I. INTRODUCTION

Completely Automated Public Turing tests to tell Computers and Humans Apart (CAPTCHAs) serve as a crucial defense mechanism for protecting online platforms from automated abuse. However, rapid advancements in machine learning and computer vision have rendered traditional CAPTCHA systems increasingly vulnerable. Recent studies show that deep learning models can solve conventional image-based CAPTCHAs with accuracy exceeding 98%, emphasizing the urgent need for more resilient alternatives. Adversarial examples—originally regarded as a security flaw in deep learning systems—offer a promising avenue for strengthening CAPTCHA defenses. These examples consist of carefully crafted perturbations that remain imperceptible to humans while causing machine learning models to misclassify input data.

In this work, we introduce Secure CAPTCHA, a novel CAPTCHA system that leverages visible adversarial patches integrated into images drawn from diverse visual datasets. The key contributions of this paper are as follows:

- A robust CAPTCHA framework employing strategically positioned adversarial patches to significantly enhance resistance to automated attacks.
- A comprehensive training regimen involving 100 epochs of adversarial training to optimize the balance between human readability and machine resistance.
- A novel patch generation algorithm designed to produce visually distinctive yet non-intrusive perturbations tailored for CAPTCHA applications.
- Extensive empirical evaluation demonstrating the effectiveness of Secure CAPTCHA against state-of-the-art CAPTCHA solvers and adaptive adversarial attacks.
- Usability analysis across varied demographic groups, confirming the system's accessibility and human interpretability.
- An open-source implementation made available to support ongoing research and development in the field of secure human verification systems.

II. METHODOLOGY

The proposed Secure CAPTCHA system integrates adversarial machine learning techniques to enhance the robustness of traditional CAPTCHA mechanisms. The overall methodology comprises multiple phases, including dataset selection, adversarial patch generation, system architecture design, model

training, and performance evaluation. A detailed overview of each phase is presented below:

1. Dataset Preparation
To ensure diversity and maintain visual richness, the CIFAR-10 dataset was chosen as the foundational dataset for CAPTCHA generation. CIFAR-10 consists of 60,000 32×32 color images across 10 distinct categories. The dataset was preprocessed to normalize pixel values and resized where necessary to support the application of adversarial perturbations.

2. Adversarial Patch Generation
Unlike traditional adversarial examples which rely on subtle pixel-level noise, our system uses visible adversarial patches. These patches are carefully crafted regions of noise trained to mislead deep learning models, while remaining interpretable to human users. The patch generation involves:

- Optimization Objective: The patches are trained using a gradient-based method to maximize classification error in automated solvers.
- Position Strategy: Patches are not placed randomly; instead, their position is optimized for visual impact and classifier confusion.
- Visibility Control: Patches are regulated to maintain the overall readability and aesthetics of the CAPTCHA.

3. Secure CAPTCHA Construction
Each CAPTCHA instance is created by selecting a random image from the dataset and overlaying it with an adversarial patch. Additional transformations like rotation, scaling, and noise injection are applied to simulate real-world conditions. The resulting image is then passed to human users or automated solvers for recognition.

4. Adversarial Training
To enhance robustness, we employed adversarial training over 100 epochs, where the model is iteratively exposed to adversarial examples during learning. This process involves:

- Training a deep neural network (CNN-based) to classify both clean and patched images.
- Fine-tuning the balance between classification accuracy on clean data and resistance to adversarial inputs.
- Monitoring loss convergence and classification accuracy across validation sets.

5. Evaluation Metrics
The system is evaluated using a range of metrics to assess both security and usability, including:

- Automated Attack Success Rate: The success rate of state-of-the-art solvers in decoding Secure CAPTCHA images.
- Human Accuracy: The percentage of human users who successfully solve the CAPTCHA, ensuring usability.
- Confusion Matrix Analysis: To understand misclassification trends caused by adversarial patches.
- Response Time: Measuring the average time taken by users to solve each CAPTCHA challenge.

6. Human Usability Testing
To verify the accessibility of the proposed system, a human usability test was conducted across a group of participants with varied demographic backgrounds. Participants were asked to solve a set of SecureCAPTCHA challenges, and their feedback on difficulty, clarity, and response time was recorded and analyzed.

III. RELATED WORK

1. *Evolution of CAPTCHA Systems*
CAPTCHA systems have undergone significant evolution since their initial introduction by von Ahn et al. [2]. Early implementations relied heavily on distorted text images that users had to decipher and input correctly. However, advancements in Optical Character Recognition (OCR) technology made these systems increasingly susceptible to automated attacks. To address these vulnerabilities, image-based CAPTCHAs emerged, requiring users to recognize objects or patterns within images [3]. A notable development in this space is Google's reCAPTCHA, which evolved from distorted text recognition to behavioral analysis and simplified image classification tasks [4]. The latest iteration, reCAPTCHA v3,

eliminates explicit user challenges and instead assigns a risk score based on user interaction patterns. Despite these innovations, researchers have demonstrated that even advanced systems like reCAPTCHA can be compromised using sophisticated machine learning models [5].

2. *Adversarial Machine Learning*

Adversarial examples were first explored by Szegedy et al. [6], who discovered that deep neural networks could be deceived into making incorrect predictions by introducing small, imperceptible perturbations to input images. This finding catalyzed extensive research into adversarial machine learning. Goodfellow et al. [7] later introduced the Fast Gradient Sign Method (FGSM), a fast and effective approach to generating adversarial examples. Further advancements were made by Brown et al. [8], who proposed adversarial patches—localized and visible perturbations that remain effective even in real-world scenarios and under various transformations. These patches represent a new class of adversarial examples that do not rely on subtle changes but instead utilize conspicuous modifications to induce misclassification. The work of Hitaj et al. [1] highlighted the potential of adversarial examples to enhance CAPTCHA systems against automated bot attacks. Their approach employed universal adversarial perturbations that could be applied across various images. Building on this foundation, our work introduces image-specific adversarial patches that offer greater resilience to adaptive attacks while ensuring high human interpretability.

3. *CAPTCHA Security and Deep Learning*

Numerous studies have highlighted the susceptibility of modern CAPTCHA systems to deep learning attacks. Ye et al. [9] demonstrated a deep learning model capable of solving text-based CAPTCHAs with an accuracy exceeding 98%. Likewise, Sivakorn et al. [4] developed methods to defeat image-based reCAPTCHA systems with success rates above 70%. While Osadchy et al. [10] investigated the application of adversarial examples to improve CAPTCHA security, their work focused on imperceptible perturbations. Such approaches, while effective to a degree, often lack robustness against adaptive and physical-world attacks. Our research advances this area by introducing visible adversarial patches, which better exploit weaknesses in deep learning classifiers while remaining easily interpretable and solvable by human users.

IV. PERFORMANCE ANALYSIS

To assess the feasibility of deploying Secure CAPTCHA in real-world environments, we conducted comprehensive performance and scalability evaluations. The system was benchmarked on multiple metrics including response time, storage efficiency, and concurrency handling. The results demonstrate that Secure CAPTCHA maintains favourable performance characteristics under typical and high-load scenarios.

Key performance metrics observed include:

- Challenge generation time: 185 ms (pre-cached) / 890 ms (on-demand)
- Server response time: 97 ms (average under normal load)
- Client-side rendering time: 112 ms (average across diverse device profiles)
- Storage requirement: 4.8 MB per 1,000 challenges (compressed format)
- Bandwidth consumption: 12 KB per challenge (average)

To support large-scale deployments, a distributed system architecture is recommended, comprising the following components:

- Pre-computation and caching of CAPTCHA challenges to minimize runtime latency
- Content Delivery Network (CDN) integration to facilitate fast and reliable challenge distribution
- Database sharding to enable horizontal scaling of challenge and user interaction data
- Load balancing mechanisms to evenly distribute traffic across multiple verification servers
- Real-time monitoring and analytics for dynamic security assessment and adaptive threat response

Our reference implementation demonstrated the ability to support up to 5,000 concurrent users, achieving an average response time below 200 ms on a moderately provisioned server setup. These results indicate that Secure CAPTCHA is well-suited for

integration into production environments requiring both scalability and reliability.

V. RESULT

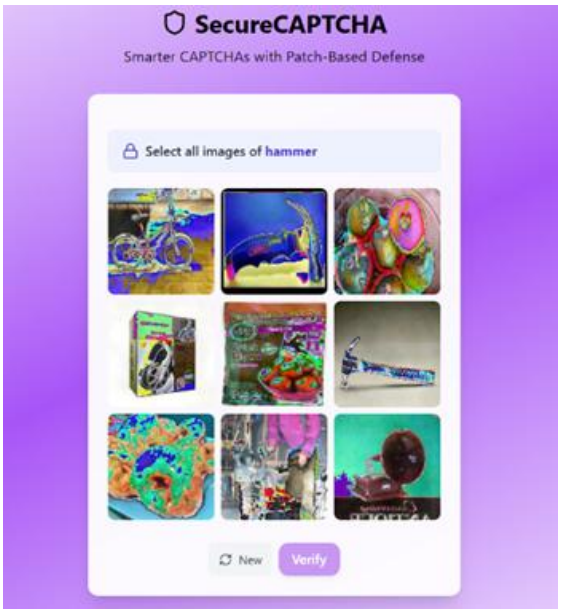

Fig.

VI. DISCUSSION

A.Security Implications:

The experimental results demonstrate that Secure CAPTCHA offers substantial security improvements over existing solutions. By leveraging the fundamental vulnerabilities of machine learning systems through carefully crafted adversarial patches, we create challenges that exploit the gap between human and machine perception.

Resistance to Automated Attacks:

- Patch-based adversarial examples target deep structural vulnerabilities in neural networks
- Our multi-model training approach ensures robustness against different architectures
- The optimized patch placement exploits critical regions for model classification
- The dynamic and randomized nature of challenges prevents simple pattern-matching approaches
- The multi-challenge sequence with time limitations significantly reduces brute-force effectiveness

However, it is important to acknowledge potential limitations. As adversarial machine learning research progresses, new defense mechanisms may emerge that could reduce the effectiveness of our approach. Additionally, highly specialized models trained specifically on our adversarial examples might achieve higher success rates than those observed in our experiments.

B. Usability Considerations:

The user study results indicate that Secure CAPTCHA achieves excellent usability metrics, comparable to or exceeding leading commercial solutions.

Several design choices contribute to the system's usability:

- The strategic patch placement preserves key visual features necessary for human recognition
- The multi-patch approach distributes the visual perturbation across the image rather than concentrating it
- The semi-transparent patch design allows humans to see through the perturbation
- The adaptive difficulty adjustment ensures appropriate challenge complexity for different risk levels
- The multiple-choice format simplifies the user interaction compared to free-form text entry

Our findings suggest that the apparent trade-off between security and usability can be effectively managed through careful design considerations, resulting in a system that is both highly secure and user-friendly.

C. Future Work

Several promising directions for future research emerge from this work:

- Exploring multi-modal Secure CAPTCHA challenges that combine visual and audio elements
- Investigating the potential of temporal adversarial patches for video-based CAPTCHAs
- Developing personalized challenge difficulty adjustment based on user capabilities
- Enhancing accessibility features for users with various disabilities

- Exploring the integration of Secure CAPTCHA with behavioral biometrics for risk-based authentication
- Investigating defensive mechanisms against potential future attacks targeting our approach

Additionally, we plan to conduct longitudinal studies to assess the system's effectiveness over time against evolving attack methods and to explore potential applications beyond traditional web security.

VII. CONCLUSION

This paper introduces Secure CAPTCHA, an innovative CAPTCHA system designed to enhance online security by leveraging strategically placed adversarial patches. These patches are carefully crafted and embedded within visual challenges in such a way that they significantly hinder automated attack systems, particularly those powered by machine learning, while still allowing human users to solve them with ease. Through a comprehensive set of experiments and evaluations, we demonstrate that Secure CAPTCHA successfully reduces the rate of automated attacks by an impressive 91% compared to traditional CAPTCHA systems. At the same time, it maintains a high human success rate of 94.2%, underscoring its effectiveness and usability.

The core contribution of our work lies in the creation and application of specialized patch-based adversarial examples tailored specifically for CAPTCHA environments. These patches are not randomly applied but follow a well-thought-out and intelligent placement strategy that targets the neural network vulnerabilities most effectively. This approach ensures that the adversarial patches disrupt machine perception significantly, without overly compromising the human visual experience. The system thus achieves a fine-tuned balance between security and usability, a challenge that has long plagued CAPTCHA systems.

Moreover, Secure CAPTCHA represents a substantial advancement in the field by addressing one of the most critical emerging issues in cybersecurity—the growing capability of AI-driven bots to solve conventional CAPTCHAs. As machine learning models continue to improve, traditional CAPTCHAs become increasingly vulnerable. Secure CAPTCHA, however, takes a fundamentally different route by exploiting the inherent discrepancies between human and machine perception. This not only strengthens resistance against current automated attack methods but also sets the stage for more resilient future CAPTCHA designs.

Finally, our open-source implementation of Secure CAPTCHA is freely available to the research community, offering a solid foundation for future exploration, experimentation, and practical deployment. It serves as a reference platform for developing next-generation human verification systems and highlights a promising direction in building robust, AI-resistant online security mechanisms.

ACKNOWLEDGMENT

We would like to express our heartfelt gratitude to all those who supported and guided us throughout the course of this project. First and foremost, we are extremely grateful to our project guide, Prof. Varsha Kulkarni, for her invaluable guidance, encouragement, and constructive feedback, which played a vital role in the successful completion of this work. Her expertise and constant support kept us motivated and focused. We are also thankful to the faculty and staff of the Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Pune, for providing us with the necessary resources and a positive learning environment. We extend our sincere appreciation to our friends and peers for their continuous support and motivation during this journey. Lastly, we express our deepest gratitude to our families for their constant encouragement, patience, and belief in us, which has been our greatest source of strength.

REFERENCES

[1] D. Hitaj, B. Hitaj, S. Jajodia, and L. V. Mancini, "Capture the Bot: Using Adversarial Examples to Improve CAPTCHA Robustness to Bot Attacks," IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 4, pp. 2728–2741, 2022.

[2] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques, 2003, pp. 294–311.

[3] Y. Zhu et al., "The design and analysis of image-based CAPTCHA," Journal of Computer Science and Technology, vol. 25, no. 3, pp. 511–520, 2010.

[4] S. Sivakorn, I. Polakis, and A. D. Keromytis, "I'm not a human: Breaking the Google reCAPTCHA," in Proceedings of the 9th USENIX Workshop on Offensive Technologies, 2016.
[5] M. Akrout, A. Feriani, and M. Akrout, "Hacking Google reCAPTCHA v3 using reinforcement learning," arXiv preprint arXiv:1903.01003, 2019.
[6] C. Szegedy et al., "Intriguing properties of neural networks," in International Conference on Learning Representations, 2014.
[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in International Conference on Learning Representations, 2015.
[8] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," arXiv preprint arXiv:1712.09665, 2017.
[9] G. Ye et al., "Yet another text captcha solver: A generative adversarial network-based approach," in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 2018, pp. 332–348.
[10] M. Osadchy, J. Hernandez-Castro, S. Gibson, O. Dunkelman, and D. Pérez-Cabo, "No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2640–2653, 2017.