標的型敵対的サンプルを用いたCAPTCHAシステム

CAPTCHA System Using Targeted Adversarial Examples

福井 恵悟・ネットワーク分科会・情報セキュリティ大学院大学

Abstract: Existing image-based CAPTCHAs often impose excessive distortion and ambiguous boundary judgments on users to counter the evolution of AI, significantly degrading usability. In this study, we propose a new CAPTCHA system that applies “Targeted Adversarial Examples”—designed to induce specific incorrect responses in bots—as a defensive measure. By exploiting the cognitive vulnerabilities of AI, this method actively eliminates bots (i.e., setting a "trap") while presenting humans with clear, undistorted images that allow for authentication through intuitive classification alone. Experimental results with human subjects demonstrated that the proposed system has a statistically significant lower workload compared to reCAPTCHA v2, achieving high practical utility with an average response time of 8.40 seconds and a success rate of 90.7%.

1. 背景・目的

- **既存手法の限界：** CNNや物体検出(YOLO)による既存CAPTCHAの自動突破。
- **ユーザビリティの低下：** セキュリティ強化による過度なノイズが認知負荷を増大。
- **境界判定の課題：** 厳密な領域選択タスクによる正解の不透明性とストレス。
- **研究目的：** ボットを意図した誤答（罠）へ誘導し、直感的で高ユーザビリティな認証を実現。

貢献： 標的型敵対的サンプル攻撃の特性を利用したCAPTCHAシステムにより、ユーザビリティとセキュリティの両立を実現。

2. 提案手法

システム設計と認証ロジック

- **生成フェーズ：** 最新手法RFCoAを採用。堅牢な特徴を標的クラスで上書き。
- **提示フェーズ：** 3×3グリッド。直感的なカテゴリ分類タスク（4種）に統一。
- **判定フェーズ：** 誘導された標的クラス（罠）を選択した挙動を検知し排除。
- **カテゴリ定義：** 動物、乗り物、楽器、食べ物の自明な4分類を選定。

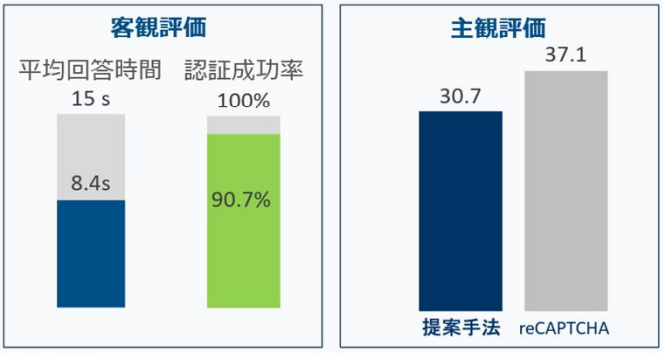
提案CAPTCHAの例（解説用）

実用的な堅牢性基準

- **前処理耐性評価：** 5×5メディアンフィルタ後も標的誘導を維持する画像を選別。
- **カテゴリ内堅牢性：** 予測が標的クラスから逸脱しても標的クラスの属するカテゴリ内なら防御成功。

3. 実験と評価

- **実験設定：** IISEC関係者の被験者31名。実環境でのreCAPTCHA v2との比較実験。
- **客観評価：** 平均回答時間（実用的なしきい値 15s）& 認証成功率（独自の目標 90%）。
- **主観評価：** NASA-TLX（簡便法）による6尺度の平均からユーザの負荷を評価。



実験結果

- 客観評価において、平均解答時間・認証成功率ともに設定した目標値を達成。
- 主観評価では、reCAPTCHA v2と比較して統計的有意差($p=0.049<0.05$)が認められ、作業負荷の低減を確認。

4. 考察・結論

- **ユーザビリティ：** 鮮明な画像と自明なタスクにより、知的要求・フラストレーションが有意に改善。
- **セキュリティ：** ボットを意図的な誤答（罠）へ誘導することで、機械学習モデルを効率的に検知・排除可能。

今後の展望

- **画像生成：** 生成AI（Diffusion Model等）を活用し、視覚的な自然さと多様な前処理への堅牢性を高度に両立する手法を確立する。
- **評価拡充：** 多様な属性の被験者を対象とした、より大規模で厳密な比較実験の実施。

参考文献

• CGCL-codes. RFCoA: Source code for Breaking Barriers in Physical-World Adversar-ial Examples. https://github.com/CGCL-codes/RFCoA, 2025. GitHub repository;source code accompanying the AAAI 2025 paper “Breaking Barriers in Physical-World Adversarial Examples: Improving Robustness and Transferability via RobustFeature”.
• Margarita Osadchy, Julio Hernandez-Castro, Stuart Gibson, Orr Dunkelman, andDaniel P´erez-Cabo. No bot expects the deepcaptcha! introducing immutable ad-versarial examples, with applications to captcha generation. IEEE Transactions onInformation Forensics and Security, Vol. 12, No. 11, pp. 2640-2653, 2017.
• Takamichi Terada, Vo Ngoc Khoi Nguyen, Masakatsu Nishigaki, and Tetsushi Ohki.Improving robustness and visibility of adversarial captcha using low-frequency per-turbation. In International Conference on Advanced Information Networking andApplications, pp. 586–597. Springer, 2022.