標的型敵対的サンプルを用いたCAPTCHAシステム

A CAPTCHA System Based on Targeted Adversarial Examples

福井恵悟・暗号分科会・情報セキュリティ大学院大学

研究背景

Webサイトをボットから保護するCAPTCHAは、機械学習の進歩により破られやすくなっている。 敵対的サンプルを用いた敵対的CAPTCHAは、この問題に対処する有望な手法である。 しかし、既存の敵対的CAPTCHAには、ユーザビリティの低下、ブラックボックス攻撃への研究不足、攻撃モデルの特定が困難といった課題がある。

提案手法

本研究では、敵対的サンプルの標的型攻撃を応用したCAPTCHAシステムを提案する。 複数の敵対的画像を用意し、各画像に異なるモデルを誤認識させることで、攻撃モデルの特定を可能にする。 また、人間には正解が明らかな画像を選択肢に加えることで、ユーザビリティを確保する

サンプル

今後の方針

今後は、提案手法に基づいたCAPTCHAシステムを開発し、その有効性を検証する予定。 具体的には、以下のような評価を行う。

- セキュリティ評価：ボットによる攻撃に対する耐性を評価
- ユーザビリティ評価：人間にとっての使いやすさを評価