7ZA-03

Proof of Human-work 実現に向けた CAPTCHA の検討

寺田 崇倫 † 西垣 正勝 † 大木 哲史 †

† 静岡大学情報学部

1 はじめに

ブロックチェーン技術における Proof of Work（以下，PoW）は，ある参加者が取引データの追加権利を賭けて複雑な計算問題を解き，解答を他の参加者が検証する仕組みで，ブロックチェーンにおける情報の不可逆性と真正性を担保する．一方で，情報社会においては，人の行為そのものを証明するべき場面が多く存在する．そのような領域に対してもブロックチェーン技術の適用範囲を広げるために，近年では Proof of HumanWork（以下，PoH）というコンセプトが提案され [1]，その実現に期待が集まっている．本研究では，PoH を用いた合意形成手法の実現要件を整理するとともに，その一要件である機械学習による解読に耐性を持つ，CAPTCHA を提案することを目的とする．

CAPTCHA とは人間と機械を判別するチューリングテストであり，PoH においては，PoW による計算タスクの代替手段として利用される．本稿では，識別器を誤認識させる入力を意図的に生成する Adversarial Examples（以下 A.E.）を CAPTCHA に応用し，人間による CAPTCHA 解読性を保ちつつ，機械による自動解読への高度な耐性（以下，機械解読耐性）および A.E. 除去耐性を有する，周波数領域限定型 A.E. について検討した結果を報告する．

2 関連研究

2.1 Proof of Human-work

PoH[1] は，PoW における計算問題を，機械では解読困難だが，人であれば解読が容易な問題に置き換える手法である．文献 [1] では，計算問題を置き換える手法として CAPTCHA を用いており，次のような要件が定義されている．

1. 機械による解読が困難であること
2. 人間が行う場合でも，一定程度の労力を要するものであること
3. 生成が容易でかつ，生成した計算機であっても容易には解けないこと
4. CAPTCHA の解答は，第三者により解答の検証が可能であり，かつその検証者は，答え自体を知ることなく， 解答が正解であるかどうかを検証できること

2.2 Adversarial Examples

A.E. とは、入力に対して、人の目には知覚できない程度の微小なノイズを加えることによって、機械学習識別器を誤認識させる手法として提案された．文献 [2] で用いられている Fast Gradient Sign Method（FGSM）は A.E. の作成手法の 1 つである．FGSM で付与するノイズ η は，θ をモデルのパラメータ，入力 x に対応するラベルを y，固定値 ϵ とすると次式で表される．

$$\boldsymbol{\eta} = \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right) \tag{1}$$

本稿では，文献 [2] に従い，複数ステップによりネットワークに対して任意の信頼度を満たす A.E. を作成する手法 [3] を用いる．

2.3 DeepCAPTCHA

DeepCAPTCHA[2] は，機械解読耐性を有する，A.E. を応用した CAPTCHA を提案している．ここで，A.E. 作成時に加えるノイズは微小であるため，メディアンフィルタ等のノイズ除去フィルタによって容易に取り除くことができるという問題がある．DeepCAPTCHA では，A.E 作成手法として FGSM を用い，メディアンフィルタによる A.E. 除去耐性の判定と，A.E. の再適用を反復的に行うことで，A.E. 除去耐性のある CAPTCHA の生成を可能としている．一方で，A.E. を複数回適用するという手法の特性上，CAPTCHA の画質が劣化することや，ノイズ除去耐性の付与に必ずしも成功しない

An investigative study on CAPTCHA toward realizing Proof of Human-work

†Takamichi Terada, Masakatsu Nishigaki, Tetsushi Ohki

†Shizuoka University

という問題がある．

3　周波数領域限定型 A.E. の提案

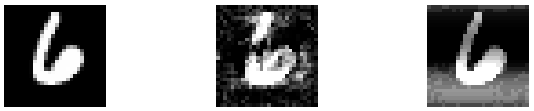
既存手法 [2] では，式 (1) で示すように，入力 $\boldsymbol{x}$ を任意のラベル y に対する損失を最大化させる方向へとノイズを加算する．この場合，加算されるノイズは一様な高周波ノイズに近くなるため，複数回のノイズ加算による画質劣化や，メディアンフィルタ等の高周波ノイズ除去フィルタによる A.E. 除去耐性の低下を引き起こす．そこで，本稿では，低周波成分のみを有するノイズを用いて A.E. を作成する周波数領域限定型 A.E. を提案する．周波数領域限定型 A.E. は，DeepCAPTCHA のアルゴリズムを元に，以下の手順で作成する．

1. 入力画像 $\boldsymbol{x}$ から FGSM を用いて A.E. を生成する
2. 生成した A.E. のノイズ除去耐性を，メディアンフィルタ（カーネルサイズ：k）を用いて評価する
3. ノイズ除去耐性が存在しない場合，A.E. を入力として，式 (1) に基づいてノイズ η を作成する
4. η をフーリエ変換し，フーリエ画像 $\mathcal{F}(\eta)$ を作成
5. フーリエ画像 $\mathcal{F}(\eta)$ の高周波をカット量 c でカット
6. フーリエ画像 $\mathcal{F}(\eta)$ を逆フーリエ変換し，$\tilde{\eta}$ とする
7. 逆フーリエ変換後のノイズ $\tilde{\eta}$ を加算して新しい A.E. にする
8. ノイズ除去耐性を再評価し，ノイズ除去耐性が存在しなければ，同様にノイズ加算を行う

4　実験

提案手法の基礎的な検討として，本手法により作成した CAPTCHA の画質を既存手法 [2] と比較することで，本手法の有用性を検証する．ネットワークとして画像サイズ 28×28 を入力とし，(300, 100, 10) 個のノードを持つ 3 つの全結合層から構成される多層パーセプトロンを用い，損失関数として交差エントロピーを用いた．MNIST データセットに含まれる 60,000 枚の画像のうち，テストデータを 5,000 枚とし，残りの画像を学習用画像として用いた．また，本実験では高周波カット量 $c = 13$，カーネルサイズ $k = 5$ として実験を行い，生成された画像の画質を比較するとともに，分類ラベルで誤分類が起きているかを確認することで，A.E. として機能しているかを検証する．

図 1(a)〜(c) に，元画像，既存手法 [2] および提案手法によって生成した CAPTCHA 画像の代表的な例をそれぞれ示す．併せて，識別器による分類結果を記載する．既存手法により生成された CAPTCHA は広範囲にノイズがかかり，数字の輪郭が霞んでいることから，人による判定の困難性が上がっていることがわかる．一方，提案手法で作成された CAPTCHA は，ノイズが均質に付加されることにより，数字の輪郭が明確に確認できている．また，識別器による分類は，既存手法と提案手法の両者ともに誤分類が生じていることが確認できる．以上のことから，提案手法は，画質を向上させるとともに A.E. としての機能も保つことが確認できた．



(a) 元画像（判定結果: 6）　(b) 既存手法（判定結果: 2）　(c) 提案手法（判定結果 2）

図 1　元画像と DeepCAPTCHA，提案手法との比較

5　議論とまとめ

本実験で，周波数領域限定型ノイズを加えることにより，画質の劣化を抑えた A.E. を作成可能となった．画質劣化を抑えた機械解読耐性を持つ CAPTCHA は，真の意味で人の作業証明に用いることが可能となる．MNIST 以外の一般画像での評価を行うことで，本提案手法の信頼性を検証するとともに，PoH のその他の要件を満たすアルゴリズムへと拡張を行うことが今後の課題である．

参考文献

[1] Jeremiah Blocki, Hong-Sheng Zhou, ”Designing Proof of Human-work Puzzles for Cryptocurrency and Beyond,” Theory of Cryptography Conference, 2016.

[2] Margarita Osadchy, Julio Hernandez-Castro, Stuart Gibson, Orr Dunkelman, Daniel Perez-Cabo, “No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples with Applications to CAPTCHA,” IEEE Transactions on Information Forensics and Security, 2017.

[3] Alexey Kurakin, Ian Goodfellow, Samy Bengio, ”Adversarial Machine Learning at Scale”, arXiv:1611.01236, 2016.