A Proposal of Adversarial CAPTCHA with High Perceptibility Using Low Frequency Adversarial Perturbation

SCIS 2020 2020 Symposium on
Cryptography and Information Security
Kochi, Japan, Jan. 28 – 31, 2020
The Institute of Electronics,
Information and Communication Engineers

低周波摂動を利用した視認性の高い Adversarial CAPTCHA の提案
A Proposal of Adversarial CAPTCHA with High Perceptibility Using Low Frequency Adversarial Perturbation

寺田　崇倫 *　　西垣　正勝 *　　大木　哲史 *
Takamichi Terada　　Masakatsu Nishigaki　　Tetsushi Ohki

あらまし CAPTCHA は人間と計算機械とを区別するチューリングテストの一つであるが，機械学習分類器を用いた画像認識によって，容易に解読できるようになった．本稿では，Adversarial Example（以下，A.E.）を用いることによって，機械学習分類器による解読耐性を付与した CAPTCHA を提案する．A.E. は機械学習分類器を誤認識させる攻撃手法であるが，これに対抗した各種防御手法が存在する．防御手法のうち，簡易かつ強力な手法の一つである，フィルタ操作への対策に焦点を当てた既存研究を発展させ，低周波領域のみに摂動を加える A.E. 作成手法を適用することで，摂動除去に対する耐性を保ちつつ，CAPTCHA 画像の画質劣化を低減する手法を提案する．提案手法により作成した Adversarial CAPTCHA を用いて，複数の機械学習分類器を対象として，解読耐性を評価した．また，二重刺激方式を用いた品質評価実験により，既存手法との画質比較を行った．2 つの評価実験により，CAPTCHA として人間が解読できる量の摂動で，機械学習分類器を誤認識させられることが結論づけられた．

キーワード Deep Learning, Adversarial Example, CAPTCHA, 低周波摂動

1　はじめに

CAPTCHA は人間と計算機械とを区別する，完全に自動化されたチューリングテストとして、Luis らによって提案された手法である [1]．Luis らはボットシステムと人間とを区別する，人工知能システムの障壁となる CAPTCHA の提案が，人工知能分野のさらなる発展に寄与するであろうと述べている．特に，人工知能分野の中でも深層学習を用いた機械学習分類器は，画像認識の分野で高い認識精度を示し，画像認識競技会において注目を集めた [2][3]．しかし，機械学習分野の発展により，計算機械が CAPTCHA を解読することが容易になり，CAPTCHA の持つ，人と計算機械とを区別する機能が失われつつあることが指摘されている [4]．

また，近年では機械学習分類器に対して誤分類を引き起こす敵対的なサンプル（Adversarial Examples, 以下 A.E.）に関する研究が活発に行われている．多くの場合，A.E. は機械学習分類器に対する攻撃を目的として利用されるが，この手法を CAPTCHA に対して応用する手法である DeepCAPTCHA は，入力に対して微小な摂動を加えることで，機械学習分類器による解読が困難な CAPTCHA を実現している [4]．

加えて，DeepCAPTCH では，メディアンフィルタ等で A.E. を除去することで，CAPTCHA を解読する攻撃を想定し，メディアンフィルタによる摂動が除去可能な A.E. に対して，再帰的に摂動を加えることによって，摂動除去耐性を向上させている．しかし一方で，複数回に渡って摂動を加えることで，大幅な画質劣化を伴うことが問題となっている．

画質の維持は CAPTCHA が人間に対する解読を阻害しない程度の視認性を担保する上で重要である．しかし，除去耐性を持つ摂動を加えなければ，機械学習分類器に対する解読耐性を入力に付与することができず，CAPTCHA が本来持つべき，人間と計算機械とを区別する能力を取り戻すことができない．

そこで本稿は，DeepCAPTCHA を発展させ，除去耐性と画質維持を両立する，視認性を維持した A.E. による Adversarial CAPTCHA を実現することを目的とする．また本稿では，摂動による画像の変更量を制限するとともに，画像の低周波領域にのみ摂動を加えることが可能な低周波摂動を応用することで，視認性の高さを維持した摂動除去耐性を持つ A.E. を提案する．

提案手法には 2 つの要件がある．1 つ目は，機械学習分類器による解読に耐性があること．2 つ目は，人間に

* 静岡大学　大学院総合科学技術研究科, 静岡県浜松市中区城北 3 丁目 5 − 1 , Shizuoka University Graduate School of Integrated Science and Technology, 3-5-1 Johoku, Naka-ku, Hamamatsu City, Shizuoka Prefecture

対する視認性を維持していることである.1 つ目の要件は，機械学習モデルを用いた分類によって検証する．2 つ目は，国際電気通信連合の無線通信部門が定める，画像・映像の主観的品質評価手法の 1 つである，二重刺激方式を用いて評価する．

2 つの評価手法によって，低周波摂動を利用した視認性の高い A.E.CAPTCHA が，人間と計算機械とを区別する機能を取り戻す手法として有効であることを示す．

2 関連研究

2.1 Adversarial Example

A.E. は，人に知覚できない程度の小さい摂動を入力に加えることによって，機械学習分類器を誤分類させることができる手法として Szegedy らにより提案された [5]. Szegedy らの手法が提案された後，Goodfellow らにより，A.E. を高速に生成する手法として，Fast Gradient Sign Method（以下，FGSM）が提案された [6]．FGSM で作成される A.E.$\boldsymbol{X}^{adv}$ は，元となる入力 $\boldsymbol{X}$ とそのラベル L，モデルパラメータを $\boldsymbol{\theta}$，更新する量を示すパラメータを ϵ として式 (1) で表される．

$$\boldsymbol{X}^{adv} = \boldsymbol{X} + \epsilon \ \mathrm{sign}\left(\nabla_{\boldsymbol{X}} J(\boldsymbol{\theta}, \boldsymbol{X}, L)\right) \qquad (1)$$

また，入力画像をピクセル単位で摂動を制限して，反復的に FGSM をかける手法 Basic Iterative Method（以下，BIM）が提案されている [7].

2.2 Adversarial Example の除去手法

A.E. は微小な摂動を入力に加えることで機械学習分類器を誤分類させる攻撃手法である．一方で，微小な摂動を除去，または無効化することで，機械学習分類器の誤分類をなくし，本来出力すべき結果を出力可能とする手法が提案されている [8][9]．摂動除去の手法としては，たとえば，A.E. 作成時に加える摂動が微小であることに着目し，メディアンフィルタなどの画像にボケを加える処理が用いられる．本稿では，このようにして微小な摂動を除去，または無効化することを単に摂動を除去する，または摂動除去と呼ぶこととする．

2.3 Adversarial CAPTCHA

Adversarial CAPTCHA は，A.E. の技術を応用することで，機械学習分類器による解読耐性を CAPTCHA に付与することを提案する手法である．ここでは代表的な手法として Osadchy らが提案した DeepCAPTCHA を例に挙げて説明する [10]．DeepCAPTCHA は機械学習分類器に対する解読耐性だけでなく，2.2 節で述べた，メディアンフィルタを用いた摂動除去に対する耐性を併せ持つ A.E. 作成手法を用いる．著者らはこの手法を Immutable Adversarial Noise（以下，IAN）と呼んでおり，本稿でもこれに従う．DeepCAPTCHA は次の要件を満たすアルゴリズムとして定義される．

1. 敵対性: 追加する摂動は，入力の少なくとも 98.5 % のケースにおいて，対象システムに対し誤分類を引き起こすことに成功すること
2. 堅牢性: 追加する摂動は，機械学習やフィルタ操作といった，効率的な手段による除去が難しいこと
3. 微小性: 追加する摂動は，人間による画像コンテンツの良好な視認を妨げない程度に，十分小さくなければならないこと
4. 機能性: 追加する摂動は，CAPTCHA チャレンジを毎秒数百万個生成できる程度には，生成の効率が良いこと

以上の要件を満たす，摂動除去耐性をもつ A.E. 作成アルゴリズム，IAN を Algorithm1 に示す．IAN アルゴリズムは，FGSM（式 (1)）による A.E. 作成と，メディアンフィルタによる摂動除去耐性の評価を，IAN アルゴリズム中のパラメーターを更新しながら繰り返す手法である．また，δ_ϵ は ϵ の更新量を示すパラメーターである．Algorithm1 中でメディアンフィルタが評価対象として使用されているのは，同論文中で行われた摂動除去を行うフィルタ操作検証の中で，A.E. の摂動を最も効率的に除去できたからだとしている．

Algorithm 1 IAN を用いた A.E. 作成手法

Require: Net は深層学習ネットワーク; I は原画像; C_i は真のクラス; C_d は分類クラス; p はネットワークの信頼度; M_f はメディアンフィルタを表す.

$adv(I, C_d, p) \leftarrow I$; ▷ $adv(I, C_d, p)$ は A.E. を示す.
$\eta \leftarrow 0$;
while $Net(M_f(adv(I, C_d, p))) = C_i$ **do**
  while $Net(adv(I, C_d, p)) \neq C_d \vee confidence < p$ **do**
    $\eta = -\epsilon \cdot sign(\nabla_I Net(I, C_d))$;
    $adv(I, C_d, p) \leftarrow adv(I, C_d, p) + \eta$;
  end while
  $\epsilon = \epsilon + \delta_\epsilon$; ▷ 摂動を一定量増やして更新する.
end while
Output: η

本稿では，IAN アルゴリズムを基とし，メディアンフィルタによる摂動除去手法への耐性を持ち，かつ複数回のフィルタ適用による画質低下を抑えることが可能な CAPTCHA 作成手法を提案する．なお，DeepCAPTCHA

は，チューリングテストの問題にあたる Adversarial CAPTCHA 作成に止まらず，作成した問題を用いてユーザーに解答を行わせることまでを CAPTCHA と定義しているが，本稿では問題の作成に止め，ユーザーによる解答は行わないものとする．本稿の提案手法は，チューリングテストに使用する，機械学習分類器に対して，誤分類を引き起こすことができる問題作成の手法である．

3 提案手法

本節では，2 節で述べた問題点を解決するために，摂動による画像の変更量を制限する手法である BIM[7] と，画像の低周波数領域に限定的に摂動を加える低周波摂動方式を組み合わせた方式を提案する．2.3 節でも述べたように，DeepCAPTCHA 作成アルゴリズムにおける画質劣化の原因は，摂動除去耐性，つまり機械学習解読耐性を与えることを目的に複数回の摂動を加えることである．この問題を解決するためには，(1) 複数回の摂動を加えた A.E. を作成する際に，原画像から変化可能な量に制限を設けること，(2) 人間の視覚的に原画像からの劣化が少ないと感じられる摂動であること，の 2 点を満たすアルゴリズムを提案する必要があるが，BIM を用いることで (1) を，低周波摂動を用いることで (1) および (2) を解決することが期待できる．

3.1 BIM による摂動量の制限

BIM における入力画像をピクセル単位で摂動を制限する関数 Clip は，A.E.$\boldsymbol{X}^{adv}$ の値とパラメータ α を用いて式 (2) で示される．

$$\begin{aligned}\mathrm{Clip}_{\boldsymbol{X},\epsilon}\left\{\boldsymbol{X}^{adv}\right\} &= \min\left\{255, \boldsymbol{X} \right. \\ &\left. +\alpha, \max\left\{0, \boldsymbol{X}-\alpha, \boldsymbol{X}^{adv}\right\}\right\}\end{aligned} \tag{2}$$

したがって，BIM における A.E. の更新式は式 (3) で表すことができる．

$$\begin{aligned}\boldsymbol{X}_0^{adv} &= \boldsymbol{X}, \\ \boldsymbol{X}_{N+1}^{adv} &= \mathrm{Clip}_{\boldsymbol{X},\epsilon}\{\boldsymbol{X}_N^{adv} \\ &\quad +\alpha\ \mathrm{sign}(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true}))\}\end{aligned} \tag{3}$$

ここで，N は更新回数，$\boldsymbol{X}_N^{adv}$ は N 回目の更新後の入力 $\boldsymbol{X}$ の値，y_{true} は入力 $\boldsymbol{X}$ に対応する正解ラベルを示す．式 (3) を用いることで，$N+1$ 回の更新後に，入力 $\boldsymbol{X}_{N+1}^{adv}$ の取りうる値の範囲を，式 (2) で示される範囲に制限することができる．これにより，仮に複数回の摂動が加えられた場合にも，人間による認識が難しくなる段階まで画像が破壊されてしまうことを防ぐことができる．本稿では，BIM を用いることで，機械学習分類器による解読耐性の維持と，摂動の量の制限を行う．

3.2 低周波摂動を用いた視認性の高い A.E. の作成

既存手法では，摂動の作成を式 (1) の FGSM で行い，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摂動を加算する．このため，加算された摂動は一様な高周波摂動に近くなるため，メディアンフィルタ等の高周波摂動を除去するフィルタによる，摂動除去耐性の低下を引き起こす．そこで，本稿では A.E. 作成手法を BIM に変更し，さらに低周波成分に限定して摂動を加える低周波摂動を用いることで，効率的に摂動除去耐性を与える．また，低周波成分に限定して摂動を加えることは画像のぼかしに近い処理となり，高周波ノイズと比較して画質劣化が抑えられることから，視認性の向上に寄与することが予想される．

提案手法の詳細を以下に示す．

1. 入力画像 $\boldsymbol{x}$ から BIM(式 (3)) を用いて A.E. を作成する．
2. 作成に成功した A.E. の摂動除去耐性を，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. メディアンフィルタで摂動が除去できた A.E. を入力として，BIM(式 (3)) を用いて摂動 η を作成する．
4. 摂動 η をフーリエ変換し，フーリエ画像 $\mathcal{F}(\eta)$ を作成する．
5. フーリエ画像 $\mathcal{F}(\eta)$ の高周波を，カット量 c でカットする．
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摂動 $\tilde{\eta}$ を加算して新たな A.E. にする．
8. 摂動除去耐性を再評価し，摂動除去耐性が存在しなければ，手順 3 から手順 7 を繰り返す．繰り返しは，画像の更新回数が n 回を超える，またはメディアンフィルタによって式（3）における摂動 α の更新回数が m 回を超えるまでを条件として終了する．

Algorithm2 に上記 1. から 8. の手順を擬似コードにより示す．DeepCAPTCHA の摂動作成アルゴリズム Algorithm1 を，提案手法の摂動作成アルゴリズム Algorithm2 へと変更することで，視認性の高い機械学習耐性を付与することが可能になる．δ_α は，手順 8 における α の更新量を示すパラメーターである．

Algorithm 2 低周波摂動を利用した視認性の高い A.E. の作成

Require: Net は深層学習ネットワーク; I は原画像; C_i は真のクラス; C_d は分類クラス; $confidence$ はネットワークの信頼度; p はネットワークの信頼度に関するしきい値; M_f はメディアンフィルタ; $\mathcal{F}$ フーリエ変換; $\mathcal{C}$ は高周波領域制限; L_i は画像更新回数上限; L_α は摂動更新回数上限を表す.

$adv(I, C_d, p) \leftarrow I$; ▷ $adv(I, C_d, p)$ は A.E. を示す.
$\eta \leftarrow 0$;
while $Net(M_f(adv(I, C_d, p))) = C_i \vee m < L_\alpha$ **do** ▷ m は摂動の更新回数を示す.
  while $Net(adv(I, C_d, p)) \neq C_d \vee confidence < p \vee n < L_i$ **do** ▷ n は画像の更新回数を示す.
    $\eta \leftarrow run\ BIM\ with\ noise\ magnitude\ \alpha$;
    $\tilde{\eta} \leftarrow \mathcal{F}^{-1}(\mathcal{C}(\mathcal{F}(\eta)))$;
    $adv(I, C_d, p) \leftarrow adv(I, C_d, p) + \tilde{\eta}$;
  end while
  $\alpha = \alpha + \delta_\alpha$; ▷ 摂動を一定量増やして更新する.
end while
Output: η

4 実験手法

4.1 前提事項

2.3 節でも述べたように，本提案における評価では CAPTCHA に用いる問題の作成のみを行い，ユーザーにおける解答が可能であるかの評価は行わない．攻撃者の目的は，システムによって提示された CAPTCHA 画像に対する正解ラベルを機械学習識別器を用いて推定することとし，この際，防御側によって提示される画像のサブセットに関する一定の知識を用いて，攻撃用のネットワークを学習可能であると仮定する．なお，メディアンフィルタを用いて A.E. を除去する処理も機械学習を用いた識別処理の一部に含まれるとする．これに対し，防御側の目的は機械学習識別器による解読が困難な CAPTCHA 画像を A.E.，IAN，提案手法等を用いて生成することである．ここで，防御側は攻撃側が用いる攻撃用機械学習識別器の構造に関する知識（たとえば，VGGNet を利用している，など）を有するとし，攻撃側がラベルの推定に用いる機械学習識別器と防御側が A.E. の生成に用いる機械学習識別器の構造は同一のものを用いる．

4.2 データセットおよびネットワーク

本実験ではデータセットとして MNIST，Caltech-256 を使用する [11][12]．MNIST は縦 28 横 28 の白黒二値で構成された一桁の手書き数字 10 クラスを分類するデータセットである．また，Caltech-256 は，サイズの異なる画像で構成された，256 クラスの一般物体認識用データセットである．

使用するネットワークは（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および VGG-Net[2] を用いる．多層パーセプトロンは MNIST，VGG-Net は Caltech-256 を用いた A.E. 作成に使用する．

4.3 評価実験手法

本稿では，機械学習分類器による解読耐性評価と主観評価による画質評価を行う．機械学習分類器による解読耐性評価実験は，機械学習モデルを用いた分類の結果から，耐性保持率を算出し，比較することによって行う．耐性保持率は，全入力中におけるメディアンフィルタで攻撃が不可能な入力の割合と定義する．

ここで，本稿での入力の分類を図 1 に示す．入力の総数 N のうち，原画像の段階で機械学習分類器による解読に失敗している入力の数を分類失敗件数 n_{cf} とする．そして，分類が成功した入力のうち，BIM による A.E. の作成に失敗し，機械学習分類器に対する解読に対する対策ができない入力の数を作成失敗件数 n_{mf} とする．さらに，A.E. の作成に成功した入力のうち，メディアンフィルタによる摂動の除去に失敗している入力の数を攻撃失敗件数 n_{af} とする．また，メディアンフィルタによる摂動除去に成功した入力のうち，IAN または提案手法による攻撃への対策によって，機械学習分類器を誤分類させる A.E. の機能を取り戻すことができない入力の数を対策失敗件数 n_{sf}，取り戻すことができる入力の数を対策成功件数 n_{ss} とする．このうち，本提案手法を適用した場合であっても，機械学習分類器に解読されてしまう入力は作成失敗件数と対策失敗件数に集計される入力である．

以上のことから，入力の総数 N，作成失敗件数 n_{mf}，対策失敗件数 n_{sf} を用いて耐性保持率 P_d を式で表すと式 (4) となり，機械学習分類器による解読に対する耐性は耐性保持率 P_d の大小によって評価する．

$$P_d = \frac{N - n_{mf} - n_{sf}}{N} \tag{4}$$

主観評価による画質評価実験は，二重刺激方式を用いて行う [13]．二重刺激方式は，国際電気通信連合の無線通信部門（ITU-R）が発行する BT.500 : Methodologies for the subjective assessment of the quality of television images で定められている，画像・映像の主観的品質評価手法の 1 つである. 二重刺激方式は，一組の原画像と摂動が加わった画像の問題を，どちらに摂動が加わっているか，出題される順番からは被験者にわからないよう，無作為に提示し，画質を評価する手法である．また，問

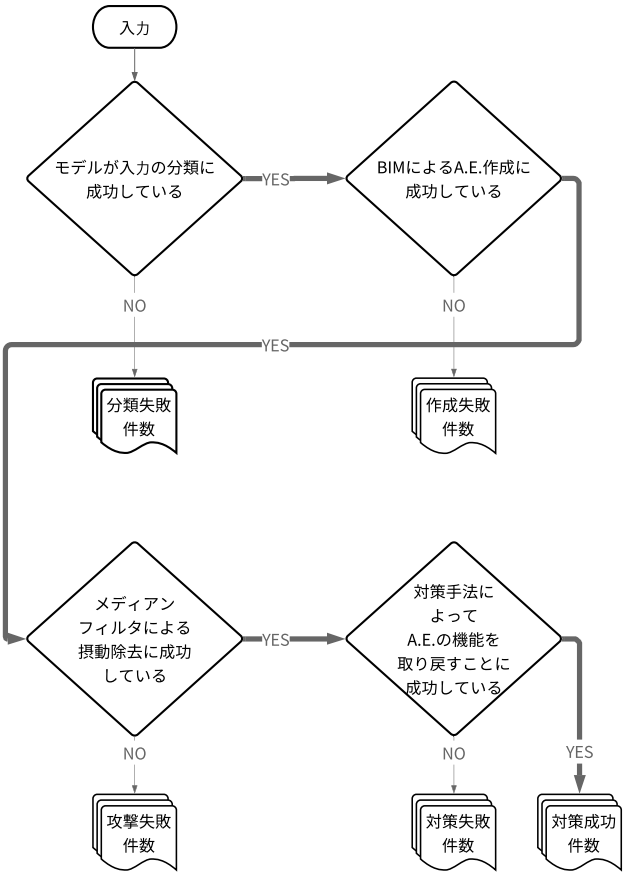

図 1: 実験システムの入力遷移図．A.E. の機能とは機械学習分類器を誤分類させる働きのことを指す．

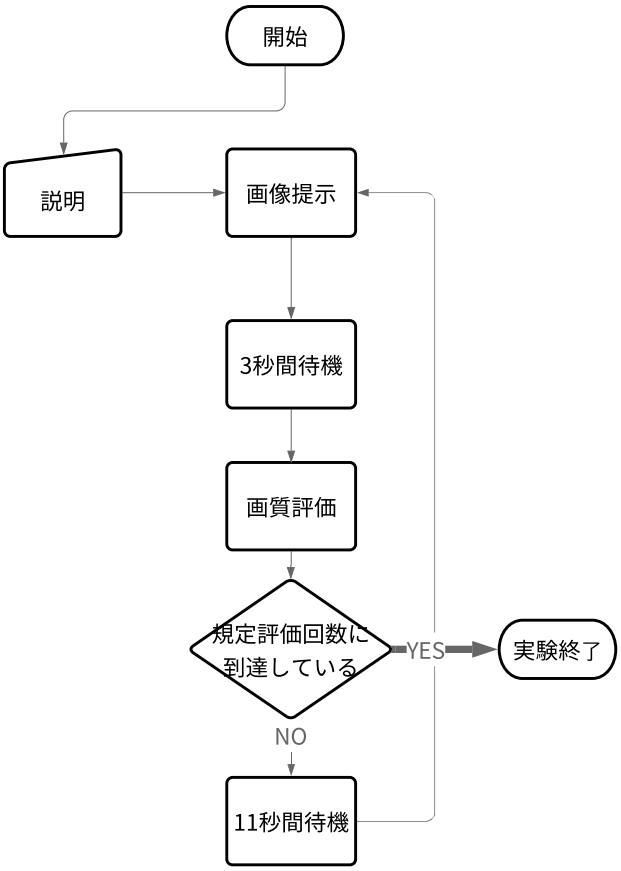

図 2: 主観評価実験の手順．

題の組は，データセットから無作為に選ばれる．今回の二重刺激方式を用いた実験は，被験者 1 人に対して行う実験を，複数回それぞれ異なる被験者に対して行う．被験者が 1 人の場合の二重刺激方式は，次の手順 1 から手順 6 で行われる．

1. 被験者に対して，実験に関する同意事項と実験手順の説明を行う．
2. 被験者は，データセットから無作為に抽出される 1 組の問題を，10 秒間の間，自由に入れ替えながら，画質についての主観的評価を持つ，この時，どちらが摂動が入っている画像であるかは，被験者にはわからない．
3. 3 秒間，灰色の画面で待機する．
4. 1 組の問題それぞれについて，主観的評価を行う．被験者は 0 から 100 までの連続した整数値の得点を用いて，評価を行う．数字が高いほど画質が良いと評価したことになる．
5. 11 秒間，灰色の画面で待機する．
6. 以上，手順 1 から手順 5 までを規定回数に達するまで繰り返す．ただし，説明を含めた合計実験時間が三十分を超えないように注意する．

加えて，画質の主観評価実験の手順を図 2 に示す．説明のブロックでは，同意事項と実験手順の説明に加え，被験者に対し，実験に全く関係しない画像を用いて予備実験を行い，実験手順の確認を実施する．なお，予備実験の評価内容は，結果に考慮しないものとする．

画質の主観評価実験で使用する，データセットから無作為に選ばれる問題は，MNIST を使用した IAN アルゴリムを用いた A.E. と原画像，MNIST を使用した提案手法を用いた A.E. と原画像，Caltech-256 を使用した IAN アルゴリズムを用いた A.E. と原画像，Caltech-256 を使用した提案手法を用いた A.E. と原画像の 4 種類，各 10 問で，それぞれの問題は種類ごとに連続して出題される．

被験者は静岡大学大木研究室及び西垣研究室に所属している，21 歳から 24 歳の学生計 16 名に協力を募り，4 種類の問題の組，各 10 問，合計 40 問を解答してもらうことによって，画像評価の得点を採取する．採取した得点は，各問題の組ごとに，摂動のない画像の得点から摂動のある画像の得点の差を取った平均を算出し，95% 信頼区間とともに示す．また，外れ値は ITU-R が発行

している BT.500-14 中の，Annex 1 to Part 1 Analysis and presentation of results の A1-2.3.1 Screening for DSIS, DSCQS and alternative methods except SSCQE method に従い、検知を行う [13].

5 実験結果

4.3 節で示した，機械学習分類器による解読耐性評価と主観による画質評価を行った結果を示す．MNIST と多層パーセプトロンを用いたシナリオにおいては，メディアンフィルタ適用回数 n を 6，カーネルサイズ k を 5，提案手法にのみ使用するローパスフィルタのカット量 c を 9 として実験を行った．また，Caltech-256 と VGG-Net を用いたシナリオにおいては，メディアンフィルタ適用回数 n を 1，カーネルサイズ k を 7，ローパスフィルタのカット量を $c = 102$ として実験を行った．なお，IAN および提案手法において，A.E. 作成判定に用いる信頼度のしきい値は $p = 0.8$ とする．なお，MNIST を使用した A.E. の作成の時に用いる，メディアンフィルタの適用回数とカーネルサイズ，信頼度の値は DeepCAPTCHA で評価の際に用いられていた値である [10].

まず，機械学習分類器による解読耐性評価実験の結果を示す．IAN と提案手法の二つの対策手法を，MNIST と Caltech-256 にそれぞれ適用したときの耐性保持率を表 1 に示す．表 1 から，IAN と提案手法を比較したとき，提案手法の方が耐性保持率が高いことがわかる．

次に，IAN と提案手法それぞれの手法で作成された A.E の例を図 3 に示す．図 3 は実際に出力された画像の中から無作為に抽出した，原画像と対策手法適用後の画像の組である．後述する，画質の主観評価実験において，被験者は図 3 のような画像の組を交互に入れ替えながら，画質評価を行う．

最後に，主観評価による画質評価実験の結果を示す．エラーバーは 95%信頼区間を表している．MNIST における差の平均をみると，僅差ではあるが，提案手法の方が既存手法に比べ値が小さい．加えて，Caltech-256 における差の平均をみると，提案手法の方が既存手法に比べ値が小さい．差の平均の値が小さいことは，原画像からの画質劣化が小さいことを示す．よって，IAN アルゴリズムに比べ提案手法の方が画質劣化を抑えることに成功しているといえる．なお，4.3 節で示した，ITU-R の文書で定められた手法を用いて外れ値検知を行ったが，外れ値に該当する解答をしている被験者は存在しなかった．

対策手法 & データセット	耐性保持率
IAN & MNIST	79.72%
Proposed & MNIST	81.29%
IAN & Caltech-256	81.62%
Proposed & Caltech-256	82.10%

表 1: 機械学習分類器に対する解読耐性評価実験の結果．ネットワークの詳細は 4.2 節，パラメーターの詳細は 5 節をそれぞれ参照のこと．

6 議論

機械学習分類器による解読耐性に対する評価では，IAN による A.E. 作成手法に比べ，提案手法による A.E. 作成手法の方が解読耐性が高いことが表 1 からわかる．ここで，2.3 節では DeepCAPTCHA の要件の 1 つとして，対象システムに対し 98.5%のケースで誤分類を引き起こせることが示されている．今回の耐性保持率を用いた評価では提案手法を用いた二つの結果のいずれも DeepCAPTCHA の要件を満たしていないが，ユーザーに Adverasarial CAPTCHA を複数回解答してもらうシステムの設計を想定すると，機械学習分類器による解読に耐性を持たない入力を選ぶ割合が小さくなり，結果として DeepCAPTCHA の要件を満たすことができると考える．

画質の主観評価では，IAN による A.E. 作成手法に比べ，提案手法による A.E. 作成手法の方が画質が維持されているという評価が，図 4 から得られたことがわかる．特に，MNIST と Caltech-256 両データセットにおいて，既存手法に比べ，提案手法の方が画質劣化が少ないという結果が得られたことで，本稿が目的とする画質維持がデータセットに依存しないことがわかる．一方で，Caltech-256 と提案手法を用いた A.E. の画質評価では，差の平均の値が極めて小さく，95%信頼区間の一部が負の値を含んでいることが確認できる．差の平均は 4.3 節で定義したとおり，摂動のない画像の評価の値から，摂動のある画像の評価の値の差をとって平均したものである．つまり，差の平均の値が負の値をとるということは，原画像より提案手法を用いて作成した A.E. の方が画質が良い，と評価した被験者が存在したことを示唆している．本稿の提案手法が意図した主観評価とは異なる評価の可能性があることに対する検証は，今後の課題である．また，MNIST を用いた評価に比べ Caltech-256 を用いた評価の数値が低いことから，被験者に画質の差が感じ取られにくかったことがわかる．これは，MNIST が Caltech-256 と比較して画像サイズが小さく，より局所的に摂動が適用された結果，画質劣化が被験者に認知されやすかったのではないかと考えられる．ただし，本実験では被験者に対して画像を提示する順番を MNIST から Caltech-256 という順番に統一して実験を行ったため，摂動がある状態に慣れてしまい，画質劣化が少ないという評価になった可能性も考えられる．実験実施順序の順序効果による結果の偏りに関する検証は今後の課題である．

また，本稿は A.E. 作成までに止まったが，提案手法

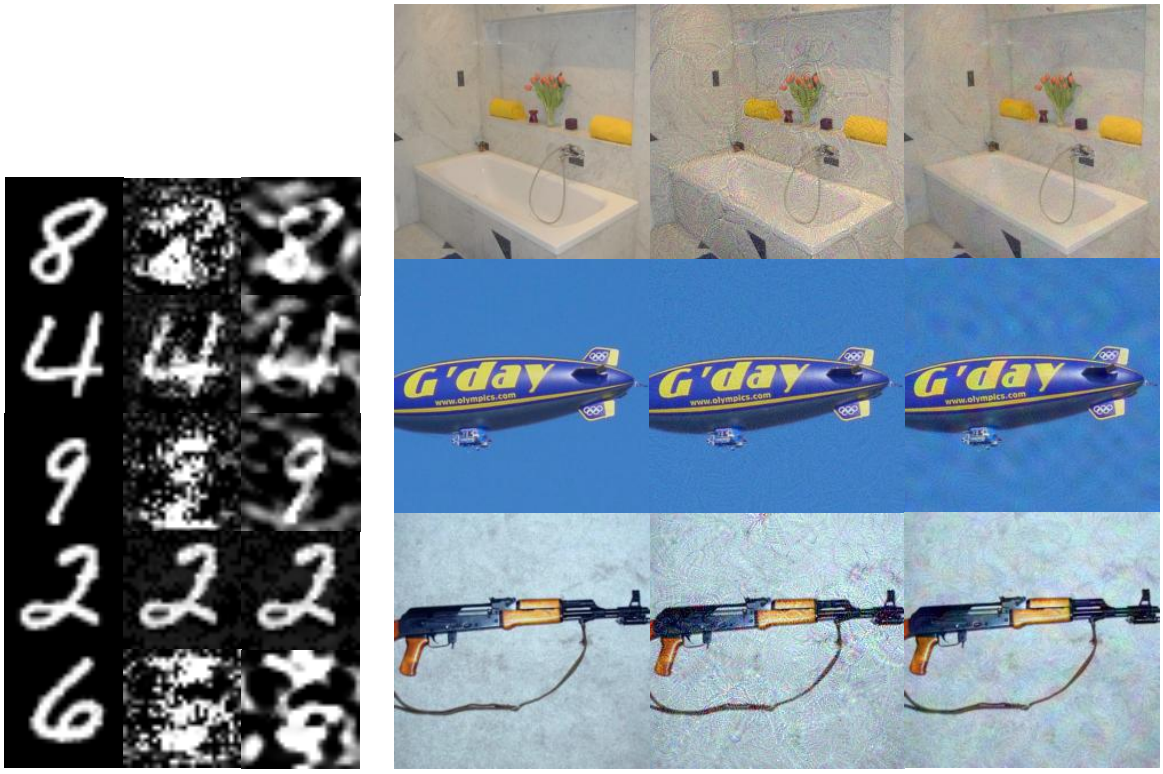

図 3: 原画像と対策手法適用後の出力画像の例．左が MNIST 右が Caltech-256 を用いた出力画像で，[I] が原画像，[Ⅱ] が IAN アルゴリズムを用いたときに出力された A.E.，[Ⅲ] が提案手法を用いたときに出力された A.E. である．図中の画像はデータセットごと同じ倍率で，拡大・縮小を行っているが，主観評価実験の際は，それぞれ原寸大の画像を用いて評価実験を行う．

を CAPTCHA に適用したときに，機械学習による解読耐性評価と画質の主観評価の検証も同様に今後の課題である．

7 結論

本稿では，機械学習分類器による解読攻撃に対する耐性を持ち，かつ人間による視認性を維持した CAPTCHA を実現するために，BIM および低周波摂動を利用する Adversarial CAPTCHA の提案を行った．また，本提案手法の有効性を検証するために，MNIST と Caltech-256 の 2 種類のデータベースによる機械学習解読耐性，および主観評価実験を行った．実験により，提案手法は機械学習分類器に対する解読耐性を有し，視認性を維持できる Adversarial CAPTCHA の問題を作成できることが示された．

今回の提案手法を CAPTCHA へ応用したとき，機械学習分類器に対する解読耐性と画質の主観評価がどのように変化するのか検証することが今後の課題である．

参考文献

[1] Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford. Captcha: Using hard ai problems for security. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 294–311. Springer, 2003.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[4] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan,

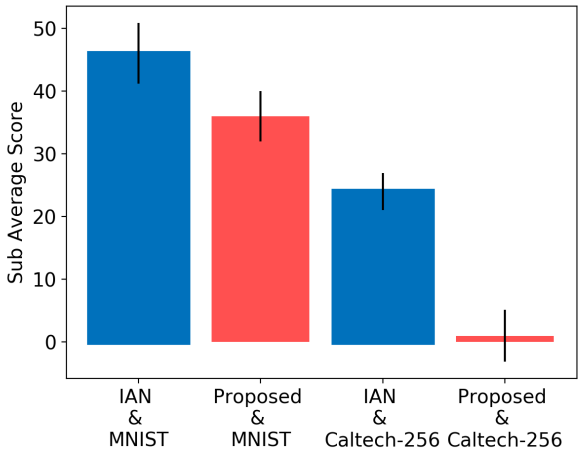

図 4: 主観評価実験によって得られた，各手法・各画像ごとの原画像と対策手法適用後の得点における差の平均をとった結果．エラーバーは 95%信頼区間を示す．既存手法である IAN より，提案手法による A.E. 作成手法の方が，原画像と A.E. との画質評価の差の平均が小さいことから，画質劣化が抑えられていることがわかる．

Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, Vol. 358, No. 6368, p. eaag2612, 2017.

[5] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[7] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14. ACM, 2017.

[9] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 2019.

[10] Margarita Osadchy, Julio Hernandez-Castro, Stuart Gibson, Orr Dunkelman, and Daniel 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.

[11] L. Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 141–142, Nov 2012.

[12] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[13] Recommendation ITU-R BT. 500-14: Methodology for the subjective assessment of the quality of television pictures, 2019.