ブラックボックス条件下における画像解釈器の脆弱性検証の試み

廣瀬 雄大[†1,a)]　梶浦 梨央[†1,b)]　小野 智司[†1,c)]

概要：深層ニューラルネットワーク (DNN) の普及により，DNN モデルの判断根拠を明らかにするための研究の重要性が高まっている．一方，DNN には画像に対して特殊な摂動を加えることで誤分類を引き起こす敵対的事例 (Adversarial Examples： AE) と呼ばれる脆弱性が存在する．このような脆弱性は画像解釈器にも存在するため，画像解釈器を安全に使用するために脆弱性の調査が不可欠である．本研究では，DNN モデルや解釈器の内部構造が未知であるブラックボックス条件下で解釈結果の注視領域を誤誘導する敵対的事例を生成する手法を提案する．実験により提案手法は予測ラベルを変えずに解釈器の解釈結果のみを誤誘導することに成功した．

Attempt to verify vulnerability of image interpreters under black box conditions

Abstract: With the spread of deep neural networks (DNNs), it has become increasingly important to clarify the basis for DNNs' decisions. On the other hand, DNNs have vulnerabilities called Adversarial Examples (AEs), which can cause misclassification by adding special perturbations to images. Since such vulnerabilities can also exist in image interpreters, it is essential to investigate them to ensure safe use of image interpreters. In this study, we propose a method to generate adversarial cases that misguide the gazing region of the interpretation result under black box conditions where the internal structures of the DNN model and the interpreter are unknown. Experimental results showed that the proposed method successfully misled only the interpretation result of the interpreter without changing the prediction label.

1. はじめに

深層ニューラルネットワーク（Deep Neural Network: DNN）は画像分類，医療画像診断，物体検出などの幅広い分野で高い性能を示しており，実問題への応用が進んでいる．また近年，人の意思決定に関わる業務を DNN に代替させる取り組みが増えている．しかしこのような場面では，出力に公平性や倫理面などの妥当性，モデルの不透明性などが生じることが問題となる．これを軽減するために DNN の推論根拠を説明する説明可能 AI（eXplainable AI: XAI）の研究が活発に行われている．

一方，DNN に基づくモデルは，人間には知覚できないような特殊な摂動を入力データに加えることで誤った判断を引き起こす敵対的事例（Adversarial Examples: AEs ）と呼ばれる脆弱性が存在することが明らかにされている [1]．このような脆弱性はセキュリティが重要な応用サービスにとって深刻な問題である．例えば，自動運転における道路標識の誤認識や顔認証システムでのなりすましなど重大なインシデントにつながる可能性があり，脆弱性の検証が必要である．

このような脆弱性は画像分類での XAI である画像解釈器でも存在することが明らかになっている [2,3]．例えば医療画像診断において誤った病変部位と異なる場所が注視領域として提示されることにより，誤診による医療事故を招く恐れや，予測ラベルと関連しない領域を注視領域とすることで利用ユーザの DNN や画像解釈器に対する信頼性の低下につながるため脆弱性の検証を行うことは重要である．

本論文では分散共分散行列適応進化戦略（Covariance Matrix Adaptation Evolution Strategy: CMA-ES）におけ

[1] 鹿児島大学
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[†1] 現在，鹿児島大学
Presently with Kagoshima University
[a)] k4648853@kadai.jp
[b)] k0301280@kadai.jp
[c)] ono@ibe.kagoshima-u.ac.jp

る分散共分散行列を対角成分のみに制限することで高次元の最適化問題への適用を可能にした Sep-CMA-ES を用いた画像解釈器への敵対的攻撃手法を提案する．目的関数が未知であるブラックボックス (BlackBox: BB) 最適化問題に対して最も優れたアルゴリズムとして知られる CMA-ES を用いることでモデルの内部情報を利用せずに敵対的攻撃を可能にする．また，画像分類器の誤認識を誘因する攻撃については広く研究されていること，画像解釈器の脆弱性を検証することが目的であることから画像の予測ラベルを変えることなく，画像解釈器の解釈結果のみを誤誘導させることのできる脆弱性を検証する．評価実験により提案手法は画像の予測ラベルを原画像から変化せずに解釈結果のみを変化させることができること，また注視領域を特定の箇所に向けることができる脆弱性を確認した．

2. 関連研究

2.1 画像解釈器

近年，DNN が発達し幅広い分野で高い性能を示しており，実問題への応用が進んでいく中で，出力に公平性や倫理面などの妥当性，モデルの不透明性などが生じることが問題となっている．例えば医療画像診断において「癌」と予測された画像において医療画像分類モデルが画像のどの領域を見て「癌」であると判断したのかが不明であるとき，ユーザはそのモデルを信頼することができなくなることから特にこのような繊細な分野においては予測根拠の可視化が非常に重要である．以下に画像分類における予測根拠の解釈手法について述べる．

Zhou らは畳み込みニューラルネットワーク（Convolutional Neural Network: CNN）の予測根拠の可視化手法である，クラス活性化マップ（Class Activation Map: CAM）を提案した [4]．この手法は ResNet や VGG などの Global Average Pooling（GAP）を使用しているモデルに適用できる手法である．GAP 層，全結合層，softmax 層の順番からなるモデルを考えたとき，k 番目の特徴マップに GAP 層を適用したものを F_c^k，あるクラス c の softmax 層を通す前のスコアを S_c とし，w_k^c はクラス c と接続する k 番目の重み，$f_k(x,y)$ は最終畳み込み層の k 番目の特徴マップの x 座標，y 座標とする．この時，F_c^k は式（1），S_c は式（2）で表される．

$$F_c^k = \frac{1}{xy}\sum_{x,y} f_k(x,y) \tag{1}$$

$$\begin{aligned} S_c &= \sum_k w_k^c F_c^k \\ &= \frac{1}{xy}\sum_{x,y}\sum_k w_k^c f_k(x,y) \end{aligned} \tag{2}$$

そして最終的に得られるクラス c に対する解釈ヒートマップ $M_c(x,y)$ は式 (3) で表される．CAM を用いたヒートマップ M_{CAM}^c は最終畳み込み層の特徴マップに対して対応する重みを乗算することで得ることができる．

$$M_{CAM}^c = \sum_k w_k^c f_k(x,y) \tag{3}$$

CAM は特徴マップに乗算する重みにおいて GAP 層の結果を利用していたため，GAP 層を含むモデルでないと利用できないという問題があった．この問題を解決するために Selevaraju らは重みの部分を逆伝播時の勾配で代用することで様々なモデルで予測根拠の可視化を行うことのできる手法 Grad-CAM を提案した [5]．解釈ヒートマップ $M_{gradcam}^c$ 算出の式は (4)，(5) で表される．

$$\alpha_k^c = \frac{1}{xy}\sum_{xy}\frac{\partial S_c}{\partial f_k(x,y)} \tag{4}$$

$$M_{gradcam}^c = ReLU(\sum_k \alpha_k^c f_k(x,y)) \tag{5}$$

この時，クラス c に対する k 番目の重みを表す α_k^c と，k 番目の特徴マップ $f_k(x,y)$ の線形結合において，対象クラス c に正の影響を与える部分が特に重要であることから ReLU 関数を適用している．

2.2 画像解釈器の脆弱性

Ghorbani らは DNN の決定境界は区分的線形性という特徴を持ち，入力空間において微小な摂動 δ を加えることにより損失関数に対する勾配の方向を容易に変更できることから，複数の勾配に基づく解釈手法が微小な摂動に敏感であり，画像の予測ラベルを変えずに解釈結果のみを変えることができる脆弱性を示した [6]．

Zhan らはブラックボックス条件下において，従来手法で作成された敵対的事例に対して，1 ピクセルずつランダムに摂動を加え，ヒートマップ間の類似度を高くするように最適化を行うことで解釈ヒートマップが原画像と変化しない手法 (Dual Black-box Adversarial Attack: DBAA) を提案した [7]．

Heo らは画像分類モデルのパラメータを変更することで分類精度を損なうことなく画像解釈器の注視領域を変更する攻撃手法を提案した [3]．中止させたい部分を 1 に，それ以外の場所を 0 に設定した 2 値マスクを用いて，通常の損失関数に加えて，注視領域に関する項を追加し学習を行うことで，注視領域が本来とは異なる部分を注視する脆弱性を発見した．

2.3 先行研究: 画像解釈器に対するホワイトボックス攻撃

Subramanya らは画像分類器と画像解釈器の両方を誤認させる敵対的パッチを生成することに成功した [8]．原画像の一部にパッチを載せることで予測ラベルが変化した

とき，その原因はパッチ部分にあると考えられるため画像解釈器はパッチ部分を注視領域として示すはずであるが，パッチ部分が注視領域として示されないようにするため，以下の式 (6) と式 (7) の損失関数を最適化することでパッチ部分を注視しないような敵対的パッチの生成に成功し，画像解釈器の脆弱性を示した．

$$L = \underset{z}{argmin}[l_{ce}(\tilde{x}, t) + \lambda \sum_{i,j}(\tilde{G}^t(\tilde{x}) \odot m)] \tag{6}$$

$$z^{n+1} = z^n - \eta Sign(\frac{\partial L}{\partial z}) \tag{7}$$

ここで，z は最適化したいパッチ，$\tilde{x}$ は原画像にパッチを載せた画像，t は予測ラベルの標的カテゴリ，λ はハイパーパラメータ，$\tilde{G}^t(\tilde{x})$ は画像 $\tilde{x}$ のカテゴリ t に対する解釈結果，m はパッチを載せたい部分を 1，それ以外を 0 とする 2 値マスク，$\odot$ はアダマール積を表す．

Song らは画像の予測ラベルを変えることなく画像解釈器の注視領域を誤誘導する攻撃手法を提案した [9]．注視を向けたい領域を 1 に，それ以外の領域を 0 にした 2 値マスクを作成し，2 値マスクが 1 の部分を注視するような以下の式 (9) の位置重要度損失関数 $l_{loim}(h, m)$ とクロスエントロピーからなる以下の式 (10) の目的関数を用いて，摂動を最適化することで予測ラベルを変えずに画像解釈器の注視領域のみを誤誘導させることに成功した．

$$\tilde{x} = x + z \odot m \tag{8}$$

$$l_{loim}(h, m) = ||h(\tilde{x}) - m||_2 \tag{9}$$

$$L = l_{ce}(\tilde{x}; c) + \lambda l_{loim}(h, m) \tag{10}$$

ここで，x は入力画像，$\tilde{x}$ は摂動を載せた画像，z は最適化する摂動，m は 2 値マスク，h は解釈手法，c は原画像の予測ラベル，λ は位置重要度損失の影響度合いを考慮するハイパーパラメータを表す．

これらの手法は目的関数の勾配を利用する必要があることから先ほどと同様にホワイトボックス手法であるため，商用モデルやサービスの脆弱性を検証するという点で現実的ではない．また，これらの手法はパッチや摂動，最適化後の注視領域が矩形であることから知覚されやすくなる可能性がある．

3. 提案手法

3.1 基本アイデア

本論文では，画像解釈器の示す注視領域を誤誘導させる敵対的攻撃手法を提案する．提案手法は特に，識別モデルおよび画像解釈器の内部情報を利用せずに敵対的事例の生成を行うブラックボックス攻撃を行う点に特徴がある．このため，対象とする識別モデルや画像解釈器を限定せず，特にソースコードが公開されていない商用のフレームワークやサービスに対しても応用が可能である．

提案手法はブラックボックス条件下での敵対的攻撃を行うために，下記の方針にもとづいて最適化を行う．

方針 1: 高次元の最適化問題への適用が可能な Sep-CMA-ES を使用する．提案手法では分散共分散行列適応進化戦略（Covariance Matrix Adaptation Evolution Strategy: CMA-ES）を用いることにより，目的関数の勾配を用いることなく大域的最適化を行う．ただし，$(\mu/\mu_w, \lambda)$-CMA-ES などの一般的な CMA-ES は，通常 100 次元程度までの規模の最適化問題を対象としており，設計変数の次元数 n に対して，時間計算量が $O(n^3)$，空間計算量が $O(n^2)$ であることから，本論文のような高次元の最適化問題への適用が困難である．このため，分散共分散行列を対角成分のみに制限し，学習率を変更することで，時間計算量と空間計算量を $O(n)$ に低減する Sep-CMA-ES [10] を使用する．

方針 2: 画質評価指標として使用される SSIM を用いて解候補を評価する．先行研究で利用される式 (8) から式 (10) の目的関数は画素単位でヒートマップの評価を行っているため，解釈器が出力する注視領域の空間的な特徴を十分に考慮していない．このため，提案手法では，画像の空間的特徴を考慮し，人間の主観に近い評価が可能な SSIM を目的関数として利用する．

なお，本論文では，先行研究 [8] とは異なり，画像分類器の予測ラベルを原画像から変化させることなく，画像解釈器の出力する注視領域のみを誤誘導する敵対的事例を生成する．分類器の予測ラベルが変化する場合に解釈器が注視する領域が変化することは自然であるため，分類器の予測ラベルを変化させない制限はより挑戦的な問題設定となる．

3.2 定式化

3.2.1 設計変数

提案手法は，画像認識における敵対的攻撃の一般的な手法と同様に，画像に加える摂動を最適化することにより求める．画像に加える摂動の次元数は赤，緑，青の 3 チャンネルそれぞれの画素数の和によって決まる．つまり設計変数の総数 N は，入力画像 x の解像度 $w \times h$ に対してチャンネル数である 3 を乗じた数（$N = 3wh$）となる．本論文では $w = h = 224$ とするため，最適化の次元数は $N = 224 \times 224 \times 3$ となり，約 15 万次元の最適化となる．また，画像類似度算出に用いる解釈ヒートマップは全て $[0, 1]$ で正規化して使用する．

3.2.2 目的関数

ここでは画像解釈器に対する非標的型攻撃の目的関数について述べる．目的関数は，予測ラベルを原画像から変化させることなく，注視領域のみを誤誘導させるために，AE 候補の解釈ヒートマップを原画像の解釈ヒートマップから遠ざける．つまり解釈ヒートマップ間の類似度を最小化する必要がある．また予測ラベルが原画像から変化しないようにするため，AE 候補の中で予測ラベルが変わっている

ものがあれば何らかのペナルティを与える必要がある．一般的に SSIM は画像に対して様々な処理を加えるなどして劣化した際にどの程度劣化したのかを評価する画質評価指標である．

提案手法では，原画像解釈ヒートマップと摂動を載せた画像の解釈ヒートマップの 2 枚の画像の類似度計算に使用し，最小化する最適化を行うことで予測ラベルを変えずに解釈ヒートマップの注視領域のみを誤誘導させる AE を作成する．SSIM は，輝度，コントラスト，構造の 3 つの要素から計算される SSIM を使用することで空間的特徴に関しても考慮することでより効果的にヒートマップ間の類似度を下げることができると考える．

原画像の解釈ヒートマップと原画像に摂動を載せた画像の解釈ヒートマップの 2 枚の画像の間の画像類似度を SSIM により算出する．また，予測が原画像から変化した際に加えるペナルティ関数，摂動量が一定量を超えた際に加えるペナルティ関数を組み合わせることで目的関数としこれを最小化する．

$$\begin{aligned} minimize f(\tilde{x}) = \alpha_s s(h(\tilde{x}), h(x)) \\ + \alpha_{p1} p_1(\tilde{x}) + \alpha_{p2} p_2(\tilde{x}) \end{aligned} \tag{11}$$

$$p_1(\tilde{x}) = \begin{cases} ||\tilde{x} - x||_2 & \text{if } ||\tilde{x} - x||_2 \geq 30 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

$$p_2(\tilde{x}) = \begin{cases} ||\tilde{x} - x||_2 & \text{if } g(\tilde{x}) \neq g(x) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

ここで $\tilde{x}$ は原画像 x_{ori} に摂動を載せた画像，$s(\cdot,\cdot)$ は 2 枚の画像の SSIM の算出，$h(\cdot)$ は画像を入力したときの解釈ヒートマップ，$p_1(\cdot)$ は摂動量に関するペナルティ関数，$p_2(\cdot)$ は予測に関するペナルティ関数，α_s，α_{p1}，α_{p2} はそれぞれの項の影響度合いを調節するパラメータを表す．

3.3 処理手順

提案手法により，非標的型攻撃で AE を生成するアルゴリズムを図 1，2 と以下に示す．

STEP 1： 解候補群の生成（3 – 4 行目）
多変量正規分布に基づいて $\lambda = 4 + \lfloor 3ln(n) \rfloor$ 個（n は次元数）の敵対的事例候補をサンプリングし，$\tilde{x}_1^{gen}, \tilde{x}_2^{gen} \ldots \tilde{x}_\lambda^{gen}$ を生成する．

STEP 2： 解候補群の評価（5 行目）
生成した敵対的事例候補 $\tilde{x}_1^{gen}, \tilde{x}_2^{gen} \ldots \tilde{x}_\lambda^{gen}$ を画像解釈器 $h(\cdot)$ に入力し解釈結果のグレースケールヒートマップ間の画像類似度 $s(h(\tilde{x}_k^{gen}), h(x_{ori}))$ などから目的関数値を算出する．

STEP 3： 最良個体の更新（6 – 9 行目）
各解候補の SSIM の値がその時点での最良解の SSIM よりも低く，摂動量が 30 以下である，かつ予測ラベルが原画像の予測ラベルと同一の敵対的事例候補が存在する場合，最良個体 $\tilde{x}_{best}$ を更新する．

STEP 4： Sep-CMA-ES のパラメータ更新（10 行目）
生成した解候補群のなかで，評価値が良い上位 $\frac{\lambda}{2}$ 個体をもとに Sep-CMA-ES アルゴリズムの平均ベクトル $\boldsymbol{m}$ などのパラメータを更新する．

STEP 5： 新しい解候補を生成（3 行目）
終了条件を満たしていないとき，更新したパラメータをもとに新しい分布から新しい解候補を生成する．**STEP 1** から **STEP 4** の操作を終了条件を満たすまで繰り返す．

STEP 6： 最良解の出力（12 行目）
設定した世代数に達したときにその時点での最良個体 $\tilde{x}_{best}$ と最良個体のラベル $g(x_{best})$ を出力する．

Algorithm 1 提案手法のアルゴリズム

Input: 原画像 x_{ori}，画像分類器 $g(\cdot)$，画像解釈器 $h(\cdot)$
Output: 最良敵対的画像ヒートマップ $h(x_{best})$，最良敵対的画像予測ラベル $g(x_{best})$

1: $gen = 0$
2: **while** gen $\leq$ gen_max **do**
3: /*敵対的事例候補の生成*/
4: $\tilde{x}_k^{gen} = \mathcal{N}(\boldsymbol{m}^{(gen)}, {\sigma^{(gen)}}^2 \boldsymbol{C}^{(gen)})$ $(k = 1, 2, \ldots, \lambda)$
5: 敵対的事例候補 $\tilde{x}_k^{gen}$ $(k = 1, 2, \ldots, \lambda)$ の評価
6: /*最良解の更新*/
7: **if** $g(\tilde{x}_k^{gen}) = g(x_{ori})$ *and* $s(h(x_{ori}), h(\tilde{x}_k^{gen})) < s(h(x_{ori}), h(\tilde{x}_{best}))$ *and* $||\tilde{x}_k^{gen} - x_{ori}||_2 \leq 30$ **then**
8: $\tilde{x}_{best} = \tilde{x}_k^{gen}$
9: **end if**
10: Sep-CMA-ES のパラメータを更新
11: **end while**
12: $\tilde{x}_{best}$，$g(\tilde{x}_{best})$ の出力

図 1: 提案手法での AE 生成手法

3.4 CMA-ES

分散共分散行列適応進化戦略（Covariance Matrix Adaptation Evolution Strategy: CMA-ES）は，目的関数が未知である BB 最適化問題に対して優れたアルゴリズムの一つとして知られている [11].

CMA-ES は，進化戦略（Evolution Strategy: ES）に分散共分散行列を導入したアルゴリズムであり，多変量正規分布 $\mathcal{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{C})$ に従って解候補群を生成し，より良い評価値を得た解候補を用いて分布の変数 (平均 $\boldsymbol{m} \in \mathbb{R}^n$，分散共分散行列 $\boldsymbol{C}$ およびステップサイズ $\sigma > 0$) の更新を行う．そして，分布の更新と個体の生成を繰り返して分布

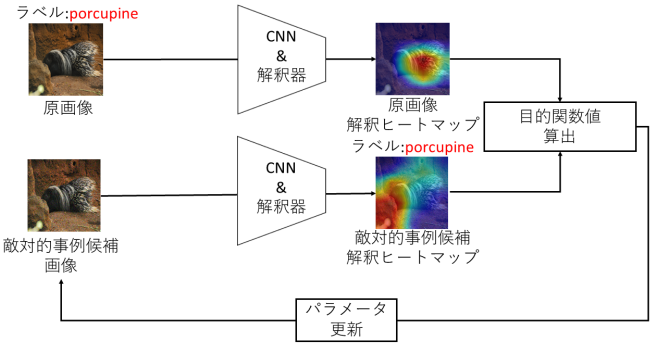


図 2: 提案手法の概要図

の最適化を行うことにより解を探索する．CMA-ES の特徴として，分散共分散行列の効果的な更新方法により，変数分離不能かつ悪スケールな単峰性関数を目的関数にもつ最適化問題において，優れた探索が可能となる．また，平均の初期値のように最適化問題の探索領域に依存して決めなければならないパラメータを除けば，各内部パラメータに推奨値が設けられている [12]．推奨される解候補数は $4+\lfloor 3\ln(n) \rfloor$（$n$ は次元数）と少なく，単峰性関数における解への収束が高速である．

CMA-ES のパラメータの更新方法はいくつか存在する．現在広く用いられている更新方法は，生成された λ 個の解候補の内，評価値の良い上位 μ 個体を使用して分散共分散行列 $\boldsymbol{C}$ とステップサイズ σ の進化パス $\boldsymbol{p}_\sigma$，$\boldsymbol{p}_c$ を学習し，パラメータの更新を行う $(\mu/\mu_w, \lambda)$-CMA-ES である [13]．

3.5 Sep-CMA-ES

本研究では，分散共分散行列を対角成分に限定することで時間・空間計算量を低減した Sep-CMA-ES を用いる [10]．CMA-ES の問題として，設計変数の次元 n に対し，時間計算量が $O(n^3)$，空間計算量が $O(n^2)$ であることから，高次元な最適化問題への適用が困難な点が挙げられる．Sep-CMA-ES は，目的関数が高次元で CMA-ES の適用が困難である問題を解決するために，CMA-ES における分散共分散行列を対角成分に限定したアルゴリズムである．Sep-CMA-ES では，変数間の依存関係を考慮せず各次元独立してサンプリングする代わりに，時間計算量・空間計算量をともに $O(n)$ に低減した．

Sep-CMA-ES は，$(\mu/\mu_w, \lambda)$-CMA-ES と比較して，(1) 分散共分散行列 $\boldsymbol{C}$ を対角成分に限定すること，(2) 分散共分散行列の学習率を増加すること，の 2 つの変更が行われた．

(2) の変更は，分散共分散行列 $\boldsymbol{C}$ が対角に制限されたことで，分散共分散行列の学習率を再調節するために行われた．Ros ら [10] は，設計変数の次元を n としたとき，rank-one 更新の学習率と rank-μ 更新の学習率に $\frac{(n+2)}{3}$ を係数に乗じることで調節を行った．

3.6 Structural Similarity: SSIM

Structural Similarity（SSIM）とは画像構造の類似度が人間の画質劣化の知覚に影響を与えるという仮説をもとに Wang らによって提案された画質評価指標である [14]．従来の画質評価指標である Peak Signal to Noise Ratio（PSNR）は以下の式 (14) で算出することができる．ここで，MAX は画像が取りうる最大画素値（8bit 画像であれば 255），MSE は平均二乗誤差表す．PSNR では主観評価において画像の一部が集中的に劣化していた場合と画像全体が少しずつ劣化していた場合の PSNR 値は同じになってしまう．これは人間の主観評価と差が生まれている．このような問題を解決するために提案された．SSIM の式を以下の (15) から (17) に示す．(15) は輝度に関する式，(16) はコントラストに関する式，(17) は構造の比較に関する式である．この 3 つの式を乗算することで (18) の SSIM 値を求めることができる．ここで μ は輝度の平均値，σ は標準偏差，σ_{xy} が共分散を表し，$C_1 = (K_1 L)^2, C_2 = (K_2 L)^2, C_3 = \frac{C_2}{2}$，$K_1 = 0.01, K_2 = 0.03$，$L$ は画像が取りうる最大画素値とする．

$$PSNR = 10 \log_{10} \frac{MAX^2}{MSE^2} \tag{14}$$

$$l(x,y) = \frac{2\mu_x + \mu_y + C_1}{\mu_x^2 \mu_y^2 + C_1} \tag{15}$$

$$c(x,y) = \frac{2\sigma_x \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{16}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{17}$$

$$s(x,y) = \frac{(2\mu_x + \mu_y + C_1)(\sigma_{xy} + C_2)}{(\mu_x^2 \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{18}$$

4. 評価実験

4.1 実験設定

提案手法の有効性を検証するために複数の画像認識モデル，画像解釈手法を対象として AE 生成を試みた．自然画像では VGG19 [15]，ResNet50 [16]，DenseNet121 [17] の 3 つの画像認識モデルを使用した．画像解釈器は，Grad-CAM [5]，GradCAM++ [18]，Eigen-CAM [19]，Layer-CAM [20] の 4 つを使用した．Sep-CAM-ES の学習率などのパラメータは Hansen ら [12] の推奨値を使用し，$\sigma = 0.01$，世代数の上限 gen は 500 とした．

使用する画像データは，Imagenet（ILSVRC2012）[21] の検証データにおいて，Top-1 クラスラベルが攻撃対象モデルによって正しく分類される画像のうち 4 枚を使用する．

4.2 実験結果

図 3 に，提案手法の最適化において，各世代の最良解における SSIM の値と，初期個体，100 世代，200 世代，および 500 世代における敵対的画像ヒートマップを示す．初期解ではハリネズミ全体を注視しているが，得られた最良解

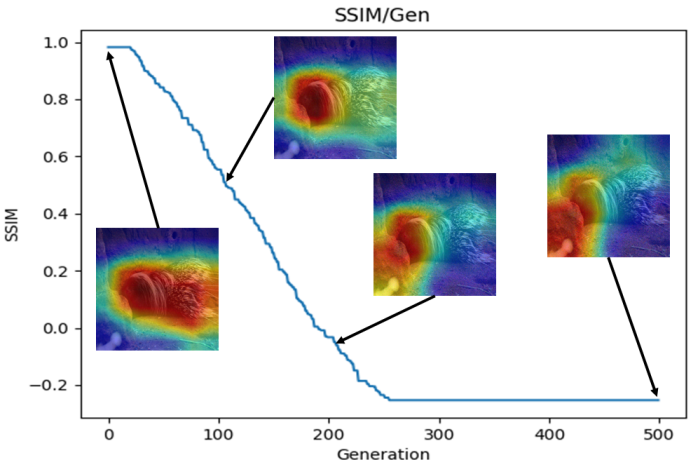


図 3: 世代ごとの SSIM の変化

ではハリネズミ以外の関連しない箇所を注視していることがわかる．

図 4 に DenseNet121 の GradCAM に対する提案手法の結果の一例を，図 5 に ResNet50 の GradCAM++に対する提案手法の結果の一例を示す．表 4，表 5 より，提案手法が様々な画像分類器，画像解釈器において，敵対的画像解釈ヒートマップが原画像の解釈ヒートマップから大きく変化していることがわかる．

5. 結論

本論文では，Sep-CMA-ES を用いることで，ブラックボックス条件下で予測ラベルを変えずに解釈結果のみを誤らせる画像解釈器への敵対的攻撃手法を提案した．実験により画像分類器と画像解釈器の多様な組み合わせに対して提案手法が有効であることを確認した．今後は摂動量の削減や指定した箇所を注視する標的型攻撃についての手法を検討する．

参考文献

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[2] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. *Explanations Can Be Manipulated and Geometry is to Blame*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[3] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society.

[5] Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[6] Amirata Ghorbani, Abubakar Abid, and James Zou. INTERPRETATION OF NEURAL NETWORK IS FRAGILE, 2018.

[7] Yike Zhan, Baolin Zheng, Qian Wang, Ningping Mou, Binqing Guo, Qi Li, Chao Shen, and Cong Wang. Towards black-box adversarial attacks on interpretable deep learning systems. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022.

[8] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification, 2018.

[9] Qianqian Song, Xiangwei Kong, and Ziming Wang. Fooling neural network interpretations: Adversarial noise to attack images. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part II*, page 39–51, Berlin, Heidelberg, 2021. Springer-Verlag.

[10] Ros et al. A simple modification in cma-es achieving linear time and space complexity. In *Proceedings of the 10th International Conference on Parallel Problem Solving from Nature — PPSN X - Volume 5199*, page 296–305, Berlin, Heidelberg, 2008. Springer-Verlag.

[11] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 312–317, 1996.

[12] Nikolaus Hansen and Anne Auger. Principled design of continuous stochastic search: From theory to practice. In *Theory and principled methods for the design of metaheuristics*, pages 145–180. Springer, 2014.

[13] Nikolaus Hansen. The cma evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.

[14] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[18] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM*plusplus*: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, mar 2018.

[19] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-CAM: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2020.

[20] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-

Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

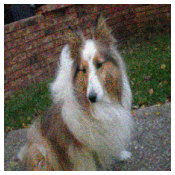
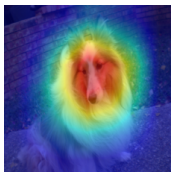
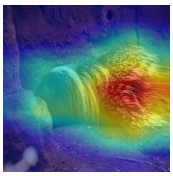
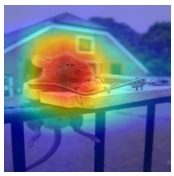
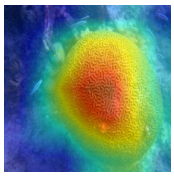
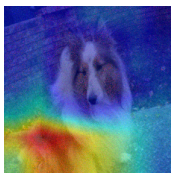
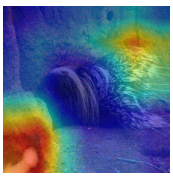
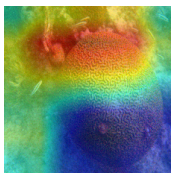
画像分類モデル	DenseNet121			
画像解釈手法	GradCAM			
原画像				
敵対的画像				
原画像予測ラベル	Shetland sheepdog	porcupine	mousetrap	brain coral
敵対的画像予測ラベル	Shetland sheepdog	porcupine	mousetrap	brain coral
原画像解釈 ヒートマップ				
敵対的画像解釈 ヒートマップ				

図 4: 提案手法の実行結果の例 (1)

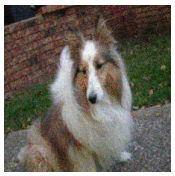
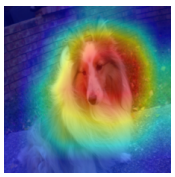
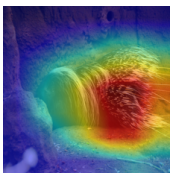
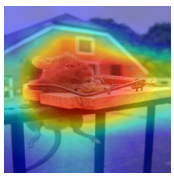
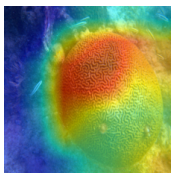
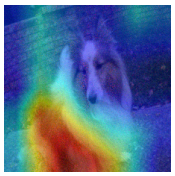
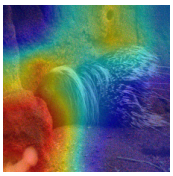
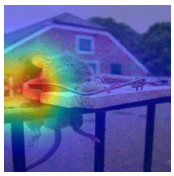
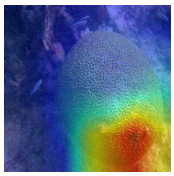
画像分類モデル	ResNet50			
画像解釈手法	GradCAM++			
原画像				
敵対的画像				
原画像予測ラベル	Shetland sheepdog	porcupine	mousetrap	brain coral
敵対的画像予測ラベル	Shetland sheepdog	porcupine	mousetrap	brain coral
原画像解釈 ヒートマップ				
敵対的画像解釈 ヒートマップ				

図 5: 提案手法の実行結果の例 (2)