

進化型多目的最適化と2次元離散コサイン変換を用いた 敵対的サンプルの生成

Adversarial Example Generation using Evolutionary Multi-objective Optimization and
Two-dimensional Discrete Cosine Transform

鈴木 崇大^{*1} 竹下 真悟^{*1} 小野 智司^{*1}
Takahiro Suzuki Shingo Takeshita Satoshi Ono

^{*1}鹿児島大学大学院 理工学研究科 情報生体システム工学専攻

Department of Information Science and Biomedical Engineering, Graduate School of Science and Engineering, Kagoshima University

This paper proposes Evolutionary Multi-objective Optimization (EMO)-based Adversarial Example (AE) design method that performs under black-box setting. Previous gradientbased methods produce AEs by changing all pixels of a target image, while previous EC-based method changes small number of pixels to produce AEs. Thanks to EMO's property of population based-search, the proposed method produces various types of AEs involving ones locating between AEs generated by the previous two approaches, which helps to know the characteristics of a target model or to know unknown attack patterns. Experimental results showed the potential of the proposed method, e.g., it can generate robust AEs and, with the aid of DCT-based perturbation pattern generation, AEs for high resolution images.

1. はじめに

画像分類や音声合成などの分野において深層学習は高い性能を示しており、実社会への応用が期待されている。畳み込みニューラルネットワーク (Convolutional Neural Network: CNN) の急速な進歩による物体の画像認識性能の向上は、その代表例といえる。一方、近年の研究により、ニューラルネットワーク (Neural Network: NN) にもとづく分類器は、攻撃者が意図的にモデルが誤認識するように設計した敵対的サンプル (Adversarial Example: AE) の影響を受け、正しく物体認識を行えないことが明らかにされている [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]。

AEの生成方式の多くは、NN内部の情報である損失関数の勾配を必要とする方式が多い [1]。一方で、例えば市販のソフトウェアやサービスを対象としてAEを生成する場合、上記の勾配情報が利用できない場合が多い。このため、攻撃者が対象モデルの勾配を利用できないブラックボックス条件下でAEの生成を試みる研究が行われている [2, 3, 4, 5, 6]。ブラックボックス条件下では、目的関数の勾配を必要としない進化計算 (Evolutionary Computation: EC) を利用することでAEの生成が行えると考える。実際に、先行研究の1つは差分進化 (Differential Evolution: DE) を利用した1画素攻撃を提案している。しかし、この手法では、摂動を加える対象の画素および摂動幅を設計変数とした定式化を行っているため、より多くの画素に摂動を加えるようなAEの生成は困難である。逆に、勾配に基づいたAE生成方式は、対象画像のほぼ全ての画素に変更を加えることでAEを生成する。未知の種類AEを生成すること、および、対象となるNNの特性をより深く理解することの双方の観点から、従来のECによって生成されたAEと勾配法の間に位置するAEなどの様々なAEを包括的に生成することは重要である。

一方、AEを生成する問題は本質的に、分類精度と摂動量のようなトレードオフの関係にある複数の目的関数を含む。先行研究の多くは線形結合によって複数の目的関数を単一の目

的関数にまとめており、目的関数を統合せずに多目的最適化 (Multi-Objective Optimization: MOO) を行うことでAEを生成する研究は存在しない。

このため、本研究は、進化型多目的最適化 (Evolutionary Multi-objective Optimization: EMO) を用いるAE生成法を提案する。提案手法は、多点探索により複数の目的関数を同時に最適化するEMOの特性を活かし、ブラックボックス条件下で多様なAEを同時に生成することができる。

また、本研究では、高解像度画像に対しても多数の画素を修正するAEを生成できるように、2次元離散コサイン変換 (two dimensional Discrete Cosine Transform: 2D-DCT) を利用する定式化を提案する。実験により、提案手法が高解像度の画像に対して多様なAEを生成できることを示す。

2. 関連研究

AEを生成する最も一般的なアプローチは、対象となるNNモデルの内部情報である損失関数の勾配を利用する方法である。すなわち、損失関数の勾配に従って対象画像の全ての画素に微小な摂動を加えることでAEを生成する。また、近年、任意の画像に摂動を与えることで、複数のNNモデルに対して有効な普遍的AEの生成が可能であることも示されている [9]。複数のNNモデルにおいて誤認識を誘発する点が興味深いものの、パターンが普遍的であるために一度そのパターンが既知となると容易に検出されてしまう。

一方で、ブラックボックス条件下でAEを生成する手法も提案されている。Papernotらは、対象となるNNモデルの代替モデルを構築することでAEを生成する手法を提案した [5]。また、Suらは、DEを用いた1画素攻撃方法を提案し [6]、Ninaらは、ネットワークの勾配を近似する局所探索法を提案した [4]。

3. 提案手法

3.1 基本アイデア

本研究では、ブラックボックス条件下で多様なAEを同時に生成する手法を提案する。提案する方式の基本アイデアを以下に示す。

連絡先: 鈴木崇大, 鹿児島大学大学院理工学研究科情報生体システム工学専攻, 〒890-0065 鹿児島市郡元 1-21-40, sc115029@ibe.kagoshima-u.ac.jp

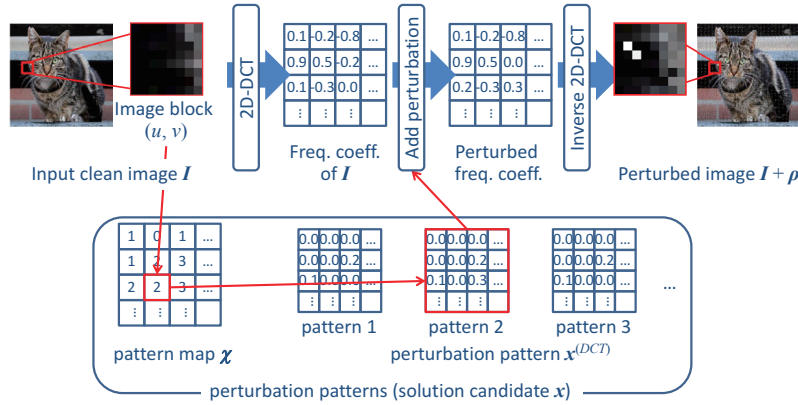


図 1: DCT を利用した AE の生成

1. 多目的最適化としての AE 設計問題の定式化:

AE を生成する問題は、分類精度と摂動強度のように互いに競合する複数の目的関数を本質的に含んでいる。したがって、これらの目的関数を単一の目的関数に統合せずに、多目的最適化問題としてモデル化し、パレート解集合の発見を試みることは妥当なアプローチであると考えられる。提案手法は、目的関数を統合するためのパラメータを必要しない点、および、微分不可能かつ非凸形状の目的関数を最適化できる点に利点がある。例えば、摂動画素数（摂動パターン ρ の l_0 ノルム）と摂動強度（ l_1 ノルム）の 2 つの目的関数を分離して導入すると、目的関数間のトレードオフの関係を明確にすることができる。一度の最適化によりパレート解集合を発見することで、対象画像の特性に応じた、かつ、最も用途に合致した AE を選択できる。

2. 進化的多目的最適化 (EMO) アルゴリズムの適用:

提案手法は、多目的最適化を実行するために EMO アルゴリズムを採用する。代替モデルを訓練するアプローチと比較すると、提案手法は代替モデルを訓練する必要はなく、また、NN 以外に対しても適用が可能な点に利点がある。さらに、EMO は多点探索であるため、対象画像に適した AE を網羅的に生成できる。なお、提案手法は先行研究よりも優れた AE を生成するとは限らないが、提案手法を用いて様々な AE を見つけることで、対象となる NN モデルの特性をより深く知ることや未知の攻撃パターンを理解することが可能となる。

特に、EMO を適用することで、提案手法では、微分不可能、多峰性の関数を含め、様々な種類の目的関数および制約条件を使用できる。例えば、提案手法は、分類精度の期待値に加えて分類精度の標準偏差を目的関数に加えることによって、画像変換に対して頑健な AE を生成することができる。

3. 離散コサイン変換 (DCT) を利用した摂動の表現:

摂動パターンを最適化により直接発見する場合、画像の解像度に応じて問題サイズが増大し、膨大な設計変数を決定する必要が生じる。このため、次元数の増加を抑えるために、本研究では、DCT に基づく摂動表現方法を提案する。摂動を与えるパターンを DCT の周波数成分として表現することで、対象画像の解像度が大きい場合も AE を生成できる。

3.2 定式化

3.2.1 設計変数

提案手法では、対象画像に対する摂動を DCT の周波数成分に対して加える。ここで、対象画像の画素値を直接変更する方法（直接法）を以下のように定義する。直接法では、解候補 $\mathbf{x}$

は、以下のように変数 $x_{u,v,c}^{(Dir)}$ から構成される。

$$\mathbf{x} = \left\{ x_{u,v,c}^{(Dir)} \right\}_{(u,v,c) \in \bar{I}} \quad (1)$$

ここで、 (u, v) は $\mathbf{I}$ における $N_w \times N_w$ 画素のブロックの位置を表し、 c は色成分を表す。

$\mathbf{I}$ の解像度を $W_I \times H_I$ 画素とすると、 $\mathbf{I}$ は $\lceil \frac{W_I}{N_w} \rceil \times \lceil \frac{H_I}{N_w} \rceil$ ブロックに分解される。

これに対して、本研究で提案する方法 (DCT 法) は、以下の 2 種類の変数を含む。

$$\mathbf{x} = \chi \cup \mathbf{x}^{(DCT)} \quad (2)$$

$$\chi = \left\{ \chi_{u,v}^{(PS)} \right\}_{(u,v) \in \bar{I}} \quad (3)$$

$$\mathbf{x}^{(DCT)} = \left\{ \mathbf{x}_r^{(DCT)} \right\}_{1 \leq r \leq N_{AP}} \quad (4)$$

$$\mathbf{x}_r^{(DCT)} = \left\{ x_{p,q,r}^{(DCT)} \right\}_{1 \leq p \leq N_{DCT}, 1 \leq q \leq N_{DCT}} \quad (5)$$

ここで、 $\mathbf{x}_{p,q,r}^{(DCT)}$ は、周波数帯 (p, q) の 2D-DCT 係数を表す。DCT 法では、画像ブロックの特徴に従って入力画像 $\mathbf{I}$ を適応的に摂動させるために、 N_{AP} 種類の摂動パターンを持つこととし、 r はパターンの番号を表す (図 1)。 $\chi_{u,v}^{(PS)}$ は、入力画像 $\mathbf{I}$ 内の画像ブロック (u, v) に適用する 2D-DCT 係数の摂動パターンを表す。すなわち、 $\chi_{u,v}^{(PS)} \in \{0, 1, \dots, N_{AP}\}$ となる。 $\chi_{u,v}^{(PS)} > 0$ の場合、対応する摂動パターンを画像ブロック (u, v) に適用し、 $\chi_{u,v}^{(PS)} = 0$ の場合は、当該ブロックの周波数係数は変化しない。

3.2.2 目的関数

提案手法は EMO を適用するため、目的関数や制約条件を柔軟に設定でき、かつ、複数の目的関数を同時に最適化することができる。ここでは、対象画像が正しく認識される際の信頼度と、対象画像に加える摂動の量とを 2 つの目的関数とする。これにより、対象画像における誤認識率と摂動量との関係性を明確化することができ、また、多様な AE を一度の最適化で生成することができる。

$$\begin{aligned} \text{minimize} \quad & f_1 = P(C(\mathbf{I} + \rho) = C(\mathbf{I})) \\ \text{minimize} \quad & f_2 = \|\rho\|_e \end{aligned} \quad (6)$$

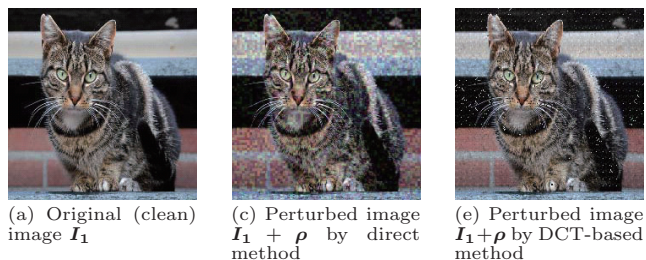


図 2: 実験 1: 生成された AE の例

表 1: 実験 1: 生成された各 AE の認識結果と信頼度

順位	$C(I_1)$		$C(I_1 + \rho)$	
			直説法	DCT 法
1st	Tabby:	60.8%	Envelope: 13.6%	Coyote: 35.2%
2nd	Tiger_cat:	30.4%	Jigsaw_puzzle: 10.0%	Wallaby: 16.0%
3rd	Egyptian_cat:	7.4%	Carton:	9.7%
4th	Doormat:	0.4%	Wallet:	9.7%
5th	Radiator:	0.2%	Door_mat:	9.2%
				German_shepherd: 4.5%

ここで、 $C(\cdot)$ は分類結果を表し、第 1 目的関数 f_1 は、対象となる分類器が摂動画像 $I + \rho$ を正しいクラス $C(I)$ に分類する信頼度を示す。第 2 目的関数は、摂動 ρ の量を l_e ノルムによって表す。

3.3 処理手順

本研究では EMO のアルゴリズムとして MOEA/D[12] を用いる。他の多くの EMO と同様、MOEA/D は解候補の生成と評価を繰り返すことで非劣解集合を発見する。解候補を評価する際は、対象となる NN モデルを用いて摂動を加えた画像の認識を試み、その結果に基づいて目的関数の値を算出する。

4. 評価実験

4.1 実験設定

提案する方式の有効性を検証するために、Keras フレームワークに実装された事前学習済みの VGG16 モデルを対象として AE の生成を試みた。MOEA/D では、多目的最適化問題をスカラー最適化問題の集合に変換するために、チェビシェフ法を選択した。近傍サイズ N_n は 10、 $\delta = 0.8$ 、 $n_r = 1$ に設定した。

4.2 実験 1: DCT 法の有効性の検証

高解像度の画像における DCT 法の有効性を検証するために、図 2(a) に示す ImageNet-1000 の画像に対して AE を生成することで、直接法と DCT 法の比較を行った。本実験では、ImageNet-1000 で割り当てられたラベルよりも一般的なクラスを考慮する。すなわち、正しいラベル “Tabby” を持つ I_1 の AE を生成する場合、他の品種の猫のラベル (“Egyptian_cat”, “lynx”, “Persian_cat”, “Siamese_cat”, “tiger_cat”) も正しいラベルとみなした。第 2 目的関数は、原画像と AE 画像間の二乗平均平方根誤差 (RMSE) とした。なお、本実験では、誤認識率が低い個体を淘汰するために、 $P(C(I + \rho)) \leq 0.4$ という制約を追加した。

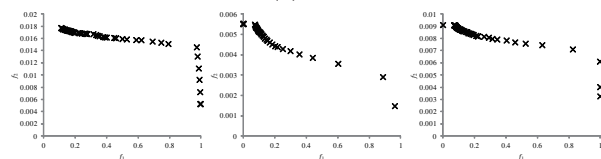
入力画像の解像度は 224×224 画素にリサイズすることとした。直接法を使用する場合、 $N_w = 3$ と設定し、摂動を I の輝度成分に追加したため、設計変数の総数は 5,625 である。DCT 法を使用する場合、 8×8 画素のブロック単位で適用するものとし、 $N_{DCT} = 10$ パターンの摂動を用意することとしたため、設計変数の総数は 1,424 となった。

表 2: 実験 2 において正しいとみなしたクラスラベル

画像	元ラベル	正しいとみなしたラベル
I_2	electric_guitar	acoustic_guitar, Violin, Banjo, cello
I_3	Plastic_bag	mailbag, sleeping_bag
I_4	Promontory	Seashore, Lakeside, Cliff, cliff_dwelling, Valley, Breakwater



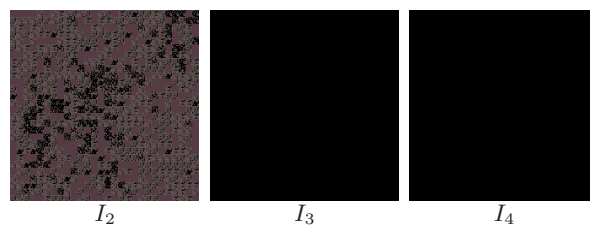
(a) 原画像



(b) 得られた非劣解集合



(c) 代表的な AE の例



(d) (c) の摂動パターン

図 3: 実験 2: 対象画像および実験結果

図 2 に、生成された代表的な AE を示し、表 1 に生成された AE の認識結果と信頼度を示す。直接法および DCT 法の双方とも、識別器を騙す AE を生成することに成功したこと、および、それぞれ異なる特性の摂動パターンを生成できたことがわかる。

4.3 実験 2: 他の AE の生成例

実験 2 では、DCT 法を用いて、ImageNet-1000 の他の画像に対して AE の生成を試みた。本実験では、すべての変数が 0 に設定された解候補 x_0 を初期集団に追加した。その他の実験条件は実験 1 と同様である。図 3(a) は対象とする原画像を示す。なお、実験 1 と同様、表 2 に示すように、元ラベルと同様のクラスラベルを正しいラベルとみなすこととした。

得られた非劣解の分布を図 3 (b) に示す。 x_0 を追加することで、提案した手法は分類精度と摂動量間のトレードオフの関係を明確にすることができた。なお、DCT と逆 DCT 変換の

表 3: 実験 2: 生成された各 AE の認識結果と信頼度

(a) I_2				
順位	認識結果と信頼度			
	$C(I_2)$		$C(I_2 + \rho)$	
1st	electric_guitar:	96.7%	Eft:	19.7%
2nd	acoustic_guitar:	2.7%	Banded_gecko:	11.3%
3rd	pick:	0.4%	European_fire_salamander:	10.2%
4th	violin:	0.1%	Common_newt:	10.1%
5th	banjo:	0.0%	alligator_lizard:	9.7%

(b) I_3				
順位	認識結果と信頼度			
	$C(I_3)$		$C(I_3 + \rho)$	
1st	Plastic_bag:	96.2%	sock:	22.5%
2nd	brassiere:	1.0%	brassiere:	8.6%
3rd	Toilet_tissue:	0.2%	pillow:	7.8%
4th	diaper:	0.2%	diaper:	7.8%
5th	sulphur-crested_cockatoo:	0.2%	handkerchief:	7.8%

(c) I_4				
順位	認識結果と信頼度			
	$C(I_4)$		$C(I_4 + \rho)$	
1st	Promontory:	96.6%	alp:	17.8%
2nd	seashore:	1.7%	Irish_wolfhound:	8.8%
3rd	cliff :	1.4%	marmot:	7.5%
4th	bacon:	0.2%	timber_wolf:	7.4%
5th	lakeside:	0.0%	bighorn:	7.4%

影響により、 x_0 の RMSE は 0 とならない点に留意されたい。

生成された AE とその摂動パターンの例をそれぞれ図 3(c) および (d) に示す。様々な種類の摂動パターンが見られ、提案手法が目標原画像特性に従って適応的に AE を生成できることがわかる。

表 3 に認識結果と信頼度を示す。 I_2 および I_4 は、 I_3 と比較して特徴的な色を含むため、様々な動物を含む多様なクラスへと誤認識されていることがわかる。

5. まとめ

本研究では、機械学習モデルの誤認識を誘発する AE を進化型多目的最適化により生成する手法を提案した。提案手法は、対象モデルの内部情報を必要としないブラックボックス手法であり、トレードオフの関係にある複数の目的関数を同時に最適化することで多様な AE を生成することができる。また、高解像度の画像に対する AE を生成するために DCT を利用して摂動を加える方法を提案した。実験により、提案手法が高解像度の画像を対象として多様な AE を生成できることを確認した。なお、問題の次元をさらなる削減や局所探索の併用は本研究の重要な課題である。

参考文献

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [2] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” in *AISec@CCS*, 2017.
- [3] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” *arXiv preprint arXiv:1804.08598*, 2018.
- [4] N. Narodytska and S. P. Kasiviswanathan, “Simple black-box adversarial attacks on deep neural networks,” in *CVPR Workshops*, vol. 2, 2017.
- [5] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017, pp. 506–519.
- [6] J. Su, D. V. Vargas, and K. Sakurai, “One pixel attack for fooling deep neural networks,” *CoRR*, vol. abs/1710.08864, 2017. [Online]. Available: <http://arxiv.org/abs/1710.08864>
- [7] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” *arXiv preprint arXiv:1802.00420*, 2018.
- [8] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, and T. Kohno, “Physical adversarial examples for object detectors,” in *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.
- [9] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 86–94.
- [10] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, “Simple black-box adversarial attacks,” 2019. [Online]. Available: <https://openreview.net/forum?id=rJeZS3RcYm>
- [11] R. Shin and D. Song, “Jpeg-resistant adversarial images,” in *NIPS 2017 Workshop on Machine Learning and Computer Security*, 2017.
- [12] Q. Zhang and H. Li, “MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition,” *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, December 2007.