微小な再攻撃による敵対的サンプル矯正の試み

森本 文哉[1,a)]　小野 智司[1,b)]

概要：深層ニューラルネットワークには，入力に対して人間に知覚できない特殊な摂動が加えられた敵対的サンプル（Adversarial Example: AE）を誤認識してしまう脆弱性が存在する．この脆弱性はセキュリティが重要なタスクにおいて深刻な問題であり，AE に対する防御手法が研究されている．既存の研究では，検出器により検出された AE の分類結果を矯正する手法があるが，2 段階の処理が必要となる．本研究では，AE と正常な入力の区別なく，正しい分類結果を出力する手法を提案する．本手法は，極めて微小な摂動を設計することで，正常な入力に対しては分類結果を維持し，AE の分類結果のみを矯正することが可能である．

An Attempt to Rectify Adversarial Examples by Minimal Re-attacks

Abstract: Deep neural networks have a vulnerability of misrecognizing Adversarial Examples (AEs), in which special perturbations are applied to the inputs that are imperceptible to humans. This vulnerability is a serious problem in security-critical tasks, and defenses against AEs have been studied. Previous study has shown the method of rectifying classification results of AEs detected by detectors, but it requires two-step processing. In this study, we propose method that outputs correct classification results without distinguishing between AEs and normal inputs. By designing extremely small perturbations, this method can maintain the classification results for normal inputs and rectify only the AE classification results.

1. はじめに

深層ニューラルネットワーク（Deep Neural Network: DNN）は，画像や音声など様々な分野で高い性能を示しており，実応用が進んでいる．一方，近年の研究により，深層ニューラルネットワーク（Deep Neural Network: DNN）に基づく分類器は，入力に対して人間の知覚が困難な程度に微小かつ特殊な摂動が加えられた敵対的サンプル（Adversarial Examples: AE）を誤認識してしまう脆弱性が存在する [1].

このため，AE に対する防御手法である敵対的防御の研究が広く行われている．敵対的防御手法には，入力の特徴から AE を判別する検出手法がある．これらは正常入力の認識精度を保証できるものの，AE を検知することに留まっており，入力本来の正しいカテゴリの認識までを考慮していない．タスクによっては攻撃前の入力の識別が必要であり，例えば自律走行車における標識認識では，攻撃が加えられたことを検出するのみでは不十分であり，自動運転を継続するために標識を正しく認識することが求められる．

このため著者らは，検出器により検出された AE の分類結果を矯正する手法を提案した [2]．しかし，この手法では防御性能が検出器に依存してしまう．

本研究では，AE と正常入力とを区別することなく，正しい分類結果を出力する手法を提案する．本手法は，入力に対して極めて微小な摂動で攻撃を行うことで，正常入力に対しては分類結果を維持し，AE である場合にのみ分類結果を矯正することが可能である．様々な攻撃手法を用いた実験により，提案手法は，正常入力の分類結果を維持し，AE を正しい分類結果に矯正可能であることを示す．

2. 関連研究

2.1 敵対的攻撃

敵対的攻撃は，攻撃者が意図的に AE を生成する手法である．AE の一例として Goodfellow らによって生成された AE を図 1 に示す [3]．AE $\boldsymbol{x}'$ は入力画像 $\boldsymbol{x}$ に対して微

[1] 鹿児島大学
Korimoto, Kagoshima, Kagoshima 8900065, Japan
[a)] k5801545@kadai.jp
[b)] ono@ibe.kagoshima-u.ac.jp

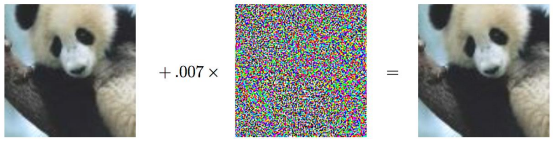


図 1 AE の一例 [3]

小な摂動 $\boldsymbol{\delta}$ を加えることで生成され，以下の式で表される．

$$\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}, \quad s.t.\ C(\boldsymbol{x}') \neq C(\boldsymbol{x})$$

ここで $C(\cdot)$ は分類器の分類結果である．また，摂動 $\boldsymbol{\delta}$ は L_p ノルムが ε を下回ることとする（$\| \boldsymbol{\delta} \|_p < \varepsilon$）．$\| \boldsymbol{\delta} \|_p$ は以下の式で表される．

$$\| \boldsymbol{\delta} \|_p = (|\delta_1|^p + |\delta_2|^p + \cdots + |\delta_n|^p)^{\frac{1}{p}}$$

ここで $\delta_1, \delta_2, \cdots, \delta_n$ は摂動 $\boldsymbol{\delta}$ の各要素である．

敵対者の知識は，ホワイトボックス（White-Box: WB）とブラックボックス（Black-Box: BB）に大別される．WB ではモデルの勾配やパラメータに関する完全な知識を有する．一方，BB ではモデルに関する知識が無く，得られる情報には様々な程度がある．

2.2 敵対的防御

DNN モデルを実世界に応用する場合，敵対的攻撃により損害を与えることでインセンティブが得られる限り，敵対者からの攻撃を受ける危険性がある．このような攻撃からシステムを保護するため，敵対的攻撃からの防御手法である敵対的防御の研究がなされている．また実環境の多くは予測が困難なランダム性があり，敵対的サンプルを実環境における最悪のケースとして防御することで，システムの頑健性を検証できる．

敵対的防御手法は，主に敵対的訓練 [4–6]，入力変換 [7–9]，検出手法 [10–13] の 3 種類に大別される．

敵対的訓練は，敵対的攻撃を防ぐ最も一般的なアプローチであり，AE を学習データに含めることで AE に対する頑健性を向上させる．しかし，通常サンプルの精度が低下することや計算コストが大きくなってしまうことが問題に上げられる．

入力変換は，前処理によって入力データに変換を加えることで，AE の影響を弱める手法である．画像分類タスクにおいては，R&P 変換 [7] や JPEG 変換 [8] などによって，入力画像を変換する手法がある．R&P 変換は，入力画像をランダムな画像サイズに変更し，画像の周囲にゼロパディングを行う手法である．また，JPEG 変換では，JPEG 圧縮を通して画像変換を行うことで，AE の影響を弱めることを期待する．しかし，これらの入力変換手法は，すべてのサンプルに同様の変換を適用するため，通常サンプルが変換によって歪み，分類精度が低下する可能性がある．また，画像や音声といった DNN の入力データの種別に応じた処理が必要となる．

検出手法は，入力サンプルの特徴から AE であるかを判別し，入力から除外する手法であり，敵対的訓練や入力変換と異なり通常サンプルの識別精度を保つことが可能である．しかし，自動運転における標識認識などの入力が必要なタスクにおいての適用が困難であるという問題点がある．

2.3 先行研究

Attack as Defense（A^{2}D）は，AE の脆弱性，すなわち特徴空間において AE は識別境界の近傍に位置し，再度攻撃を受けると容易に識別境界を超えて分類結果が変わってしまう特性に着目して検出を行う [13]．一方，上記のような検出手法は，AE の検出にのみ焦点を置いており，攻撃前の原画像の正しいクラスの識別等は考慮していない．

このような問題点に着目し，著者らは AE の脆弱性を用いた矯正手法を提案し，検出された AE に対して再度敵対的攻撃を適用することで，AE を正しい分類結果に戻すことが可能であることを示した [2]．

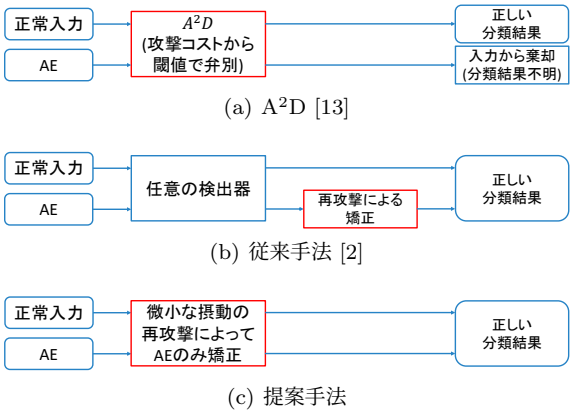
3. 提案手法

本研究では，AE と正常入力とを区別することなく，正しい分類結果を出力する手法を提案する．我々のアイデアは AE の脆弱性に基づいており，本手法は入力を区別することなく攻撃を行うが，正常入力に対しては摂動が小さいことで攻撃が失敗し，分類結果を維持する．一方，AE に対しては，AE の脆弱性によって極めて微小な摂動でも攻撃が成功し，分類結果を矯正する．提案手法の処理手順を図 2 に示す．本手法における再攻撃は，非反復型の攻撃手法である Fast Gradient Sign Method（FGSM）[3] を使用する．FGSM の式を以下に示す．

$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\boldsymbol{x}} L(\theta, \boldsymbol{x}, y))$$

ここで，$\boldsymbol{x}$ は入力画像，$\boldsymbol{x}'$ は攻撃後の画像，y は $\boldsymbol{x}$ の分類ラベル，ϵ は摂動量パラメータ，θ はモデルパラメータ，L は勾配損失，$sign$ は符号関数である．

提案手法が 1 ステップで正しい分類結果を得るためには，再攻撃の摂動パラメータ ϵ を適切に決定することが重要となる．本手法は，正常データのみから構成される訓練データを用い，Z-score に基づく異常検知の閾値によって設定する．このときに用いる分布は，正常な訓練データに対する攻撃に必要な最小摂動量の集合である．この最小摂動量は訓練データに対して，FGSM の ϵ をステップサイズ s ずつ増加して適用し，攻撃に成功する最小の ϵ を求める．オムニバス検定により上記の分布が正規分布に従うことを確認し，正規分布に従わない場合は Box-Cox 変換を適用する．片側検定の p 値があらかじめ定められた値 p_D となる



(a) A^{2}D [13]

(b) 従来手法 [2]

(c) 提案手法

図 2 提案手法の処理手順

際の摂動量を h とする．この h を再攻撃摂動 ϵ として適用する．

このように正常データのみから構成される訓練データから再攻撃摂動 ϵ を調整することで，特定の攻撃手法に対する過剰な適合を抑制できる．

4. 評価実験

提案手法の有効性を検証するため，正常入力と AE の双方に対して本手法を適用した．実験 1 では，先行研究 [13] と同様の実験条件において，提案手法の有効性を検証し，各パラメータ間での防御性能の比較を行った．実験 2 では，敵対的頑健性を評価可能なベンチマークである RobustBench [14] において，適応的な敵対的攻撃である AutoAttack [15] によって生成された AE に対する提案手法の有効性を検証した．

4.1 実験 1: パラメータと防御性能の関係

提案手法は，再攻撃に利用する FGSM のステップサイズの増加幅 s と，入力を正常か AE であるかを識別する際の閾値 h を決定する p_D 値とをパラメータとして含む．提案手法が効果的に作用するパラメータ値を求めるための実験を行った．本実験において，AE を生成する際の攻撃の種類は，FGSM [3]，BIM [4]，JSMA [16]，CW [17] を採用した．本実験では，CIFAR-10，ImageNet（ILSVRC2012）を使用し，正常入力 1,000 サンプルと，敵対的攻撃が成功した AE 各 1,000 サンプルを使用した．また，正常入力は分類結果を維持した割合，AE は分類結果の矯正が成功した割合を評価指標とした．CIFAR-10 を対象とした分類器は先行研究 [13] をもとに実装し，ImageNet を対象とした分類器は事前学習された ResNet-101 を用いた．提案手法は，上記の正常入力とは異なる訓練データを 1,000 サンプル使用し，s を [1e-4, 1e-5]，p_D を [0.1, 0.05, 0.01] で変更し比較した．

実験結果を表 1 に示す．提案手法は，多くのサンプルで正常入力の分類結果を維持し，AE の分類結果の矯正に成功した．s が小さい場合は AE に対して，大きい場合は正常入力に対して，有効に機能することが示唆される．p_D を 0.01 とした場合は，CIFAR-10 の FGSM と JSMA で性能が著しく低下した．CIFAR-10 は特徴空間が ImageNet より小さく，1 ステップ攻撃の FGSM や特定画素に対する貪欲手法である JSMA は，必要な摂動量が大きくなる．p_D が小さくなるに従い閾値 h も小さくなるため，再攻撃に使用する ϵ が，AE の矯正が困難なほどに小さくなったと考えられる．

表 1 実験 1: 防御性能（正解率）へのパラメータの影響

(a) CIFAR-10

s	p_D	FGSM	BIM	JSMA	CW	AE 平均	正常入力
1e-4	0.1	0.615	0.958	0.940	0.980	0.873	0.870
	0.05	0.502	0.964	0.862	0.990	0.830	0.926
	0.01	0.248	0.972	0.543	0.988	0.688	0.971
1e-5	0.1	0.599	0.960	0.935	0.980	0.869	0.877
	0.05	0.461	0.966	0.828	0.991	0.812	0.937
	0.01	0.200	0.968	0.459	0.987	0.654	0.977

(b) ImageNet

s	p_D	FGSM	BIM	JSMA	CW	AE 平均	正常入力
1e-4	0.1	0.923	0.980	0.999	0.974	0.969	0.840
	0.05	0.933	0.986	0.999	0.966	0.971	0.891
	0.01	0.940	0.989	1.000	0.948	0.969	0.941
1e-5	0.1	0.932	0.986	0.999	0.966	0.971	0.889
	0.05	0.940	0.989	1.000	0.954	0.971	0.936
	0.01	0.818	0.989	0.998	0.907	0.928	0.972

4.2 実験 2: AutoAttack に対する有効性の検証

Croce らは，防御評価のためのパラメータ調整が不要な攻撃が存在しないことを，敵対的防御手法の適切な評価が得られない原因として，AutoAttack を提案した [15]．AutoAttack では，ステップサイズを選択する必要のない Auto-PGD と，代替損失関数である Difference of Logits Ratio Loss を開発し，これら 2 つの手法による攻撃と WB 攻撃の FAB，BB 攻撃の Square Attack のアンサンブルによって，防御手法に依存しない適応的な敵対的攻撃を実現した．AutoAttack は 50 以上の防御モデルにおいて，頑健性を 10%以上低下させることに成功した強力な敵対的攻撃である．また，末神らの研究によると AutoAttack に対する頑健性の向上は 10%未満であり，AutoAttack の攻撃性能を課題としていた [18]．我々は，実験 1 で検証した 4 つの攻撃手法より強力な攻撃として，AutoAttack に対する提案手法の防御性能を検証し，提案手法による極めて少ない計算量での防御が AutoAttack に対してどの程度有効か評価する．

本実験では，CIFAR-10，ImageNet（ILSVRC2012）を使用し，各データセットにおける分類器は RobustBench で提供されている標準モデルを使用した．CIFAR-10 を対象とした分類器は WideResNet-28-10，ImageNet を対象とし

表 2 実験 2: AutoAttack に対する提案手法の性能比較 [%]

(a) CIFAR-10

	standard accuracy	robust accuracy
Standard model	94.38	0.00
Ours($p_D = 0.1$)	73.96	3.78
Ours($p_D = 0.05$)	79.98	3.56
Ours($p_D = 0.01$)	87.18	3.06

(b) ImageNet

	standard accuracy	robust accuracy
Standard model	76.68	0.00
Ours($p_D = 0.1$)	65.44	4.74
Ours($p_D = 0.05$)	68.60	3.94
Ours($p_D = 0.01$)	72.42	2.60

た分類器は ResNet-50 を用いた．AutoAttack の脅威モデルはどちらのデータセットにおいても L_∞ としており，摂動パラメータは CIFAR-10 では $\epsilon = 8/255$，ImageNet では $\epsilon = 4/255$ とした．提案手法は，訓練データ 1,000 サンプルを使用し，$s = 1e-4$，p_D を [0.1, 0.05, 0.01] で変更し比較した．また，正常入力と AE の各 5,000 サンプルに対して正常入力の分類精度（clean accuracy）と AE の分類精度（robust accuracy）を評価指標とした．ここで，評価に使用されるサンプルは，モデルが正しく分類できたか考慮されていない点に注意されたい．

実験結果を表 2 に示す．CIFAR-10 では 3.78%，ImageNet では 4.74%の頑健性が改善された．しかし，パラメータ p_D による standard accuracy 低下に見合った頑健性は得られなかった．実験 1 で検証した AE は，AE として摂動の視認性を下げるために摂動量の削減に注力したアルゴリズムによって生成されているため矯正が可能であった．一方，AutoAttack による AE は誤認識を引き起こすために，動的に摂動を調整することができるため矯正が容易ではないことが示唆された．対象とした分類モデルは敵対的訓練がなされていないため，AE に対して脆弱であることも原因の一つとして考えられる．敵対的頑健性を有するモデルでは，我々の提案手法によって分類結果が変わってしまう可能性が少ないことが想定されるため，敵対的訓練によって clean accuracy を維持し，robust accuracy の向上が可能か検証する必要がある．

5. 結論

本研究では，AE の脆弱性を用いて，AE と正常入力の区別なく，正しい分類結果を出力できる手法を提案した．提案手法は既存の防御アイデアより極めて少ない計算量で防御が可能であり，実験結果から，適切なパラメータを設定することで，正常入力の分類結果を維持し，AE を矯正可能であることを示した．今後の課題として，AutoAttack に対する防御性能の向上のため，敵対的に頑健なモデルに対する有効性を検証する．

謝辞

本研究の一部は JSPS 科研費 JP22K12196 の助成による．

参考文献

[1] Szegedy, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
[2] 森本文哉ほか. 敵対的事例の脆弱性を用いた分類結果矯正の試み. 人工知能学会全国大会論文集 第 37 回 (2023), pp. 2K5GS201–2K5GS201. 一般社団法人 人工知能学会, 2023.
[3] Goodfellow, et al. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
[4] Kurakin, et al. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.
[5] Madry, et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
[6] Zhang, et al. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.
[7] Xie, et al. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.
[8] Dziugaite, et al. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.
[9] Meng, et al. Athena: A framework based on diverse weak defenses for building adversarial defense. *arXiv preprint arXiv:2001.00308*, 2020.
[10] Feinman, et al. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.
[11] Wang, et al. Dissector: Input validation for deep learning applications by crossing-layer dissection. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pp. 727–738. IEEE, 2020.
[12] Wang, et al. Adversarial sample detection for deep neural network through model mutation testing. In *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, pp. 1245–1256. IEEE, 2019.
[13] Zhao, et al. Attack as defense: Characterizing adversarial examples using robustness. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pp. 42–55, 2021.
[14] Croce, et al. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
[15] Croce, et al. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.
[16] Papernot, et al. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pp. 372–387. IEEE, 2016.
[17] Carlini, et al. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.
[18] 末神奏宙ほか. Self-examination mechanism: 説明可能 ai を用いた敵対的攻撃に対する軽量な防御機構. 人工知能学会全国大会論文集 第 37 回 (2023), pp. 2A1GS203–2A1GS203. 一般社団法人 人工知能学会, 2023.