6ZC-01

微小な再攻撃による敵対的事例の矯正に関する基礎検討

森本　文哉　　小野　智司 †
鹿児島大学 †

概要

深層ニューラルネットワークの誤認識を引き起こす敵対的事例（Adversarial Examples: AE）に対する防御手法が広く研究されている．本研究では，AEと正常入力を区別することなく極めて微小な摂動を加えることで，AEと正常入力の双方から正しい分類結果を出力する手法を提案する．

1　はじめに

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事例（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．

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．

A Preliminary Study on Rectification of Adversarial Examples by Slight Re-attacks
† Fumiya Morimoto, Satoshi Ono, Kagoshima University

2　関連研究

敵対的防御手法の一つである検出手法は，入力の特徴からAEであるかを判別する手法であり，正常入力の識別精度を保つことが可能である．Attack as Defense（A^2D）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 [2]．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 [3]．

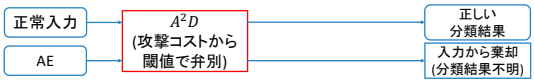
3　提案手法

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1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）[4] を使用する．FGSMの式を以下に示す．

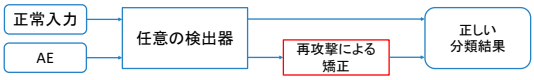
$$\boldsymbol{x}' = \boldsymbol{x} + \epsilon \cdot sign(\nabla_{\boldsymbol{x}} L(\theta, \boldsymbol{x}, y))$$

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．

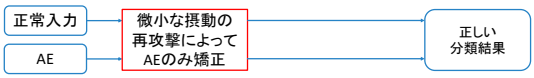
提案手法が1ステップで正しい分類結果を得るためには，再攻撃の摂動パラメータ ϵ を適切に決定することが重要となる．本手法は，正常データのみから構成される訓練データを用い，Z-scoreに基づく異常検知の閾値によって設定する．このときに用いる



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



(b) 従来手法 [3]



(c) 提案手法

図 1　提案手法の処理手順

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 値によって閾値 h が得られ，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　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．AE を生成する際の攻撃の種類は，FGSM [4]，BIM [5]，JSMA [6]，CW [7] を採用した．本実験では，CIFAR-10，ImageNet 使用し，正常入力 1,000 事例と，敵対的攻撃が成功した AE 各 1,000 事例を使用した．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 [2] 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 事例使用し，s を [1e-4, 1e-5]，p 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，多くの事例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 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

表 1　パラメータ間の防御性能の比較

(a) CIFAR-10

s	p	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	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

攻撃の FGSM や特定画素を攻撃する JSMA は，必要な摂動量が大きくなる．p が小さくなるに従い閾値も小さくなるため，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．

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．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，敵対的に頑健なモデルに対する有効性を検証する．

参考文献

[1] Szegedy, et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] 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.

[3] 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.

[4] Goodfellow, et al. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[5] Kurakin, et al. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. Chapman and Hall/CRC, 2018.

[6] 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.

[7] Carlini, et al. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.