4ZD-02

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　　玉城　大生　　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）に対する防御手法の多くは AE の検知に焦点をおいている．一方で，自動運転における標識認識などのタスクでは，攻撃前の原画像における正しいラベルを認識することが求められている．このため本研究では，検出された AE に対して再度攻撃を加えることで，原画像の正しいラベルを推定する手法を提案する．

1　はじめに

深層ニューラルネットワーク（Deep Neural Network: DNN）は，画像分類や音声認識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，DNN に基づく学習器は入力データに対して，人間の知覚が困難な程度に微小で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を有することが明らかにされている．

このため，AE に対する防御手法を有する DNN の研究も広く行われている．例えば，入力事例の特徴から AE を判別する検出手法 [1] が提案されており，これらは通常事例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，攻撃前の画像における正しいカテゴリの認識までを考慮しない．このため，例えば自動運転における標識認識において，一時停止の標識に対して攻撃が加えられた際にそれを AE として検出はできるものの，防御手法のみでは一時停止の標識であることを認識することができず，何らかの後処理が必要となる．

このため本研究では，防御手法により検出された AE に対するラベルの矯正手法，すなわち，攻撃前の原画像における正しいラベルを推定する手法を提案する．本手法は，AE に対して再度攻撃を行うことで，誤分類されていた分類結果を原画像の分類結果に矯正する．本手法は，DNN の入力信号の種別やタスクに依存せずに適用できる汎用性の高い手法であることに特徴がある．実験により，比較手法より高い矯正性能を有し，広範な攻撃手法に対して適用可能であることを示す．

2　関連研究

敵対的攻撃からの防御手法は，主に敵対的訓練，入力変換，検出手法に大別される．敵対的訓練を行った DNN は AE に対して過剰に適合し，通常事例の分類精度を低下させる可能性がある．入力変換は，すべてのサンプルに同様の変換を適用するため，通常事例が変換によって歪み，分類精度が低下するほか，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 [1] は，入力事例の特徴から AE であるかを判別する手法であり，敵対的訓練や入力変換と異なり通常事例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 は識別境界付近に位置し，再度攻撃を受けると容易に識別境界を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 [1]．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．

Kao らは説明可能 AI（eXplainable AI: XAI）を用いた矯正手法を提案し，XAI により推定された注視領域を修正または削除することで，AE を正しい事例に戻すことが可能であることを示した [2].

3　提案手法

本研究は，防御手法により検出された AE に対して再度攻撃を加えることで，原画像の正しいラベルを推定する手法を提案する．すなわち，Kao らによる AE の矯正と同様の目的を，A^2D と同様に AE の脆弱性に着目した再攻撃により実現する．本手法は

A Preliminary Study on Counter-attack-based Rectification of Adversarial Inputs
† Fumiya Morimoto, Daiki Tamashiro, Satoshi Ono, Kagoshima University

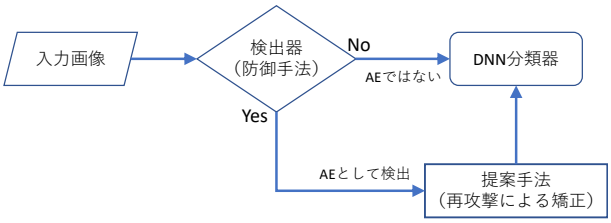

図 1　提案手法の位置付け

A^{2}D の後処理として位置づけることもできる．タスクに特化した後処理（画像処理用 DNN における平滑化やノイズ付与／除去等）により AE を矯正することも考えられるが，提案手法は DNN の入力データの種別を問わない点に特徴がある．

提案手法と従来の防御手法との関係を図 1 に示す．本手法は「再攻撃による矯正」の処理を担う．再攻撃に利用可能な手法には特に制限はなく一般的な手法が利用可能である．

一般に，汎化性能を高めるように学習された分類モデルにおいて，通常事例は識別境界から比較的離れている．一方で，人間が知覚できない程度の摂動を付加された AE は識別境界を越えた近傍に位置する．この脆弱性は様々な手法によって生成される AE に共通するため，本手法は多様な攻撃手法に対して適用可能である．

4　実験

提案手法の有効性を検証するため，防御手法により AE が検出されたという前提のもとで，AE の矯正を試みる実験を行った．本実験では先行研究 [2] に倣い，AE の矯正後に利用する DNN モデルの内部情報を利用できると仮定し，FGSM [3]，BIM [4]，DeepFool（DF）[5] の 3 手法を再攻撃に用いることとした．AE を生成する際の攻撃の種類は，FGSM [3]，BIM（L_2, L_∞） [4]，CW [6] を採用した．本実験では，MNIST，CIFAR-10 の 2 種類のデータセットを使用することとし，各データセットにおいて，分類モデルが原画像を正しく識別でき，かつ，敵対的攻撃が成功した 1,000 事例を使用した．また，矯正後の AE を識別した結果が原画像と同じになる，すなわち矯正が成功した割合を評価指標とした．分類モデルは，先行研究 [2] をもとに実装した．

XAI を用いた矯正手法 [2] と提案手法の比較を行った結果を表 1 に示す[*1]．Kao らの手法 [2] では，

[*1] 比較手法の結果は，論文に示された結果のうち条件に合致する結果のなかで最良の結果を引用した．提案手法とは実験に利用した事例および分類モデルが異なる可能性があることから，厳密な比較ではない点に留意されたい．

表 1　先行研究 [2] との矯正成功率の比較

データセット	矯正手法	AE 生成時の攻撃手法			
		FGSM (L_∞)	BIM (L_2)	BIM (L_∞)	CW (L_2)
MNIST	先行研究 [2]	0.889	0.949	0.905	0.972
	提案手法（FGSM）	0.999	0.996	0.999	1.000
	提案手法（BIM）	0.998	0.996	0.999	1.000
	提案手法（DF）	0.993	0.992	0.998	1.000
CIFAR-10	先行研究 [2]	0.581	0.616	0.729	0.936
	提案手法（FGSM）	0.992	0.997	1.000	1.000
	提案手法（BIM）	0.992	0.997	1.000	1.000
	提案手法（DF）	0.991	0.995	0.997	0.998

CIFAR-10 において，CW により生成された AE と比較して，FGSM や BIM により生成された AE の矯正の成功率が低下する傾向がみられた．これに対して提案手法では，データセットや攻撃手法の組み合わせによって成功率が大きく変化することなく，すべての攻撃に対して高い矯正性能を示した．

5　結論

本研究では，検出された AE に対して再度攻撃を行うことで AE を矯正し，攻撃前の原画像の正しい分類結果を得る手法を提案した．実験結果から，提案手法は従来手法と比較して，多様な攻撃方法によって生成された AE をより安定的に矯正できることが示された．今後は，より高解像の事例や，より多様な攻撃に対する本手法の有効性を検証する．

参考文献

[1] Z Zhao, et al. Attack as defense: Characterizing adversarial examples using robustness. *Proc. Int'l Symp. Software Testing and Analysis*, pp. 42–55, 2021.

[2] CY Kao, et al. Rectifying adversarial inputs using xai techniques. *Europ. Signal Processing Conf.*, pp. 573–577, 2022.

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

[4] A Kurakin, et al. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pp. 99–112. 2018.

[5] SM Moosavi-Dezfooli, et al. Deepfool: a simple and accurate method to fool deep neural networks. *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

[6] N Carlini, et al. Towards evaluating the robustness of neural networks. *Symp. Security and Privacy*, pp. 39–57, 2017.