L-007

微小な再攻撃を用いた音声認識用ニューラルネットワークの敵対的事例の矯正に関する基礎検討

A Preliminary Study on Counter-attack-based Rectification of Adversarial Examples on Neural Networks for Speech Recognition

森田 龍斗[1)]　森本 文哉[1)]　小野 智司[1)]
Ryuto Morita　Fumiya Morimoto　Satoshi Ono

1　はじめに

深層ニューラルネットワーク（Deep Neural Networks: DNN）は画像分類や音声認識など様々な分野で応用され，高い性能を示している．一方，近年の研究により，人間に知覚できない程度の微弱で特殊な摂動を入力に加えた敵対的事例（Adversarial Examples: AE）を誤認識してしまう DNN の脆弱性の存在が明らかにされている [1]．この脆弱性は画像認識用 DNN において初めて報告されたが，音声認識用 DNN においても同様に脆弱性が存在することが示されている [2,3]．

このため，DNN を敵対的攻撃から防御する手法に関する研究も広く行われている．Zhao らは，AE の脆弱性，すなわち特徴空間において AE が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越えて分類結果が変化する特性に着目し，通常事例と AE に対する攻撃コストの差を用いて AE を検出する手法を提案した [4]．しかし，この手法は AE を検出することはできるものの，正しいクラスラベルを推定することは難しい．

森本らは，AE を検出する手法の後処理として，AE に対して再度攻撃を行うことで，AE を通常事例に矯正する手法を提案した [5,6]．これにより，画像認識用 DNN を対象として多様な攻撃手法で生成された AE を，正しいクラスラベルに矯正できることが示された．

一方，音声認識分野では，敵対的訓練や入力変換のような敵対的防御手法は提案されているものの，攻撃前の正しいラベルを推定する矯正手法についての研究は十分に行われていない [7,8]．森本らの手法は，DNN に入力するデータ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が特徴であるが，音声認識における有効性は検証されていない．

このため本研究では，画像認識用 DNN において有効である再攻撃を用いた AE の矯正手法を音声認識用 DNN に適用し，その有効性を検証することを目的とする．実験により，画像認識用 DNN と同様に，音声認識用の DNN に対しても多様な攻撃手法により生成された AE を適切に矯正できることを確認した．

2　関連研究

音声認識用ニューラルネットワークの敵対的防御手法として，AE の検出や敵対的訓練方式が提案されている．Samizade らは，敵対的防御を 2 値分類の問題として定式化し，入力された音声が AE であるかを畳み込みニューラルネットワークを用いて判別する検出手法を提案した [9]．また，Choi らは，異なる特徴抽出器である Mel-Spectrogram と Wav2vec から生成された敵対的な音声は異なる特徴を持つことに着目し，特徴抽出を多様化することで，入力音声が敵対的であるかを判別する検出手法を提案した [10]．これらの検出手法は，AE を検出し，検出された AE を除去することで DNN を防御することはできるものの，検出のみに焦点が当てられており，攻撃前の原音声の正しいクラスの識別は考慮していない．

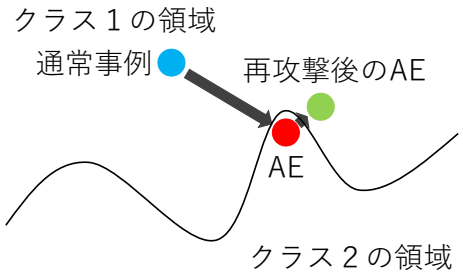

図 1: 再攻撃による AE の矯正

Pal らは，話者認識のタスクに敵対的訓練を適用し，生成した AE を学習データに追加して AE を原音声のラベルとして学習することで，AE を正しいクラスとして識別する手法を提案した [11]．敵対的訓練では，AE を学習データとして扱うため，通常事例のみを用いて学習を行う場合に比べ，通常事例の分類精度を低下させる可能性がある．

3　提案手法

本研究は，森本らによって提案された再攻撃による AE の矯正手法 [5,6] を音声認識用 DNN に適用し，その有効性を検証するとともに，本防御手法がモダリティにとらわれない手法であることを明らかにする．図1に示すように，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た位置に存在する．一方，人間が知覚することのできない微小な摂動を加えた AE は識別境界の近傍に存在する．このような AE の脆弱性は，様々な攻撃手法によって生成される AE に共通するため，攻撃に用いられる手法を防御に応用することが可能である．

本手法は FGSM や BIM などの任意のホワイトボックス攻撃手法を用いて AE を再攻撃することで，AE の矯正を行うことができる．ここでは，例として，FGSM を用いて再攻撃を行う場合について説明を行う．A^2D などの防御手法により検出された AE を $\boldsymbol{x}_{adv}$ とし，$\boldsymbol{x}_{adv}$ の原事例を $\boldsymbol{x}$ とする．FGSM は分類モデルの損失関数の勾配を利用して，入力されたデータの分類結果を変えるような摂動を加える方向を見つける手法である．損失関数の勾配を用いて，正解ラベルとの損失を大きくするように摂動を加えることで攻撃を行う．

$$x'_{adv} = x_{adv} + \epsilon \cdot \mathrm{sign}\left(\nabla_{x_{adv}} L\left(\theta, x_{adv}, C\left(x_{adv}\right)\right)\right) \quad (1)$$

ここで，ϵ は摂動量を調節するパラメータ，L は損失関数，θ は分類モデルのパラメータ，$C(\cdot)$ は DNN の出力

1)　鹿児島大学，Kagoshima University

表 1: 矯正成功率

DNN	矯正時の再攻撃手法	AE 生成時の攻撃手法 FGSM	BIM	DF	CW	JSMA
音声認識用	FGSM	0.979	0.998	0.997	1.000	1.000
	BIM	0.979	0.997	0.996	1.000	1.000
	DF	0.979	0.998	0.998	1.000	1.000
画像認識用	FGSM	0.992	1.000	1.000	1.000	0.994
	BIM	0.992	1.000	1.000	1.000	0.993
	DF	0.991	0.997	0.999	0.998	0.994

表 2: 再攻撃した AE の平均摂動量

矯正時の再攻撃手法	AE 生成時の攻撃手法 FGSM	BIM	DF	CW	JSMA
FGSM	0.158	0.135	0.316	0.082	0.240
BIM	0.114	0.077	0.222	0.018	0.195
DF	0.052	0.036	0.098	0.011	0.094

を表す．再攻撃の場合，損失関数の勾配を用いて，AE である $\boldsymbol{x}_{adv}$ の分類結果である $C(x_{adv})$ との損失を大きくするように AE に摂動を加える．本研究では，再攻撃により摂動を加えた AE が識別境界を越える，すなわち，$C(x_{adv}) \neq C(x'_{adv})$ となるまで ϵ を段階的に増加させる．これにより，誤ったクラスの領域から正しいクラスの領域に矯正されることを期待する．

4 評価実験

提案手法の有効性を検証するため，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，対象モデルとして一般的な音声分類用データセットに対して高い精度が報告されている，畳み込みベースのモデルである BC-ResNet-8 を用いた．また，AE を矯正する再攻撃手法として，モデルの内部情報を利用できるホワイトボックス攻撃を用いた．本実験では，AE を生成する際の攻撃手法として FGSM，BIM，DeepFool（DF），CW，JSMA の 5 種類，AE を矯正する際の攻撃手法として，FGSM，BIM，DF を採用した．本実験では，10 クラスの音声コマンドが含まれた Google Speech Commands をデータセットとして使用し，対象モデルが原音声を正しく識別でき，かつ，敵対的攻撃が成功したサンプルを使用した．評価指標は分類結果の矯正成功率，すなわち矯正後の AE を識別した結果が原音声と同じになる割合とした．

提案手法による AE の矯正成功率を表1に示す．参考までに，先行研究 [5] で検証された画像認識用 DNN を対象とした AE の矯正成功率についても併記する．また，表2は分類結果を矯正するために AE に加えた摂動の量を表す．音声認識用 DNN を対象とした AE の矯正成功率は，いずれの攻撃手法に対しても 97%以上であり，摂動量も微小であることから，音声認識用 DNN においても再攻撃を用いた AE の矯正が有効であることが示唆された．

次に，矯正に失敗した事例に着目する．矯正に失敗した全ての AE は，原音声のクラスとも，また AE のクラスとも異なるクラスに分類された．これは，図2に示すように，AE の近傍に原事例のクラス以外の識別境界が存在したためと考える．

5 結論

本研究では，音声認識用 DNN で検出された AE に対して再度攻撃を行うことで AE を矯正し，攻撃前の原音声の正しい分類結果を得る手法を提案した．実験結果から，音声認識用 DNN においても，多様な攻撃手法に

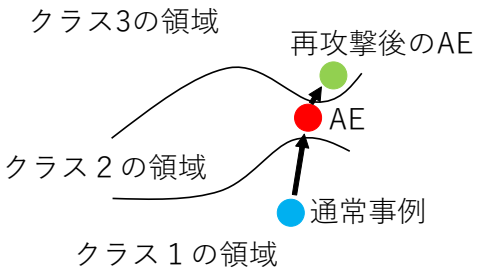

図 2: AE の矯正失敗例

よって生成された AE を高い精度で矯正できることが示された．本研究では，Mel-Spectrogram という比較的画像に近い音声の 2 次元表現での有効性を示すことができたが，音声信号に直接摂動を加える攻撃に対しての検証は行えていない．今後は，音声信号に摂動を加える攻撃に対する本手法の有効性を検証する．

参考文献

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX security symposium (USENIX security 16)*, pp. 513–530, 2016.

[3] Shoma Ishida and Satoshi Ono. Adjust-free adversarial example generation in speech recognition using evolutionary multi-objective optimization under black-box condition. *Artificial Life and Robotics*, Vol. 26, No. 2, pp. 243–249, 2021.

[4] Zhe Zhao, Guangke Chen, Jingyi Wang, Yiwei Yang, Fu Song, and Jun Sun. 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.

[5] 森本文哉, 赤垣敬吾, 小野智司. 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.

[6] Fumiya Morimoto and Satoshi Ono. Rectifying adversarial examples using their vulnerabilities. *Available at SSRN 4752243*.

[7] Xiaojiao Chen, et al. Adversarial attack and defense on deep neural network-based voice processing systems: An overview. *Applied Sciences*, Vol. 11, No. 18, p. 8450, 2021.

[8] Piotr Żelasko, et al. Adversarial attacks and defenses for speech recognition systems. *arXiv preprint arXiv:2103.17122*, 2021.

[9] Saeid Samizade, Zheng-Hua Tan, Chao Shen, and Xiaohong Guan. Adversarial example detection by classification for deep speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3102–3106. IEEE, 2020.

[10] Yujin Choi, Jinseong Park, Jaewook Lee, and Hoki Kim. Exploring diverse feature extractions for adversarial audio detection. *IEEE Access*, Vol. 11, pp. 2351–2360, 2023.

[11] Monisankha Pal, et al. Adversarial defense for deep speaker recognition using hybrid adversarial training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6164–6168. IEEE, 2021.