敵対的攻撃に対する堅牢性向上のための適応的マルチチャネル選択法

Adaptive multi-channel selection method for enhancing robustness against adversarial attacks

松井清修[1)]
指導教員　青木輝勝[2)]

1）東京工科大学大学院 バイオ・情報メディア研究科 コンピュータサイエンス専攻 青木研究室
2）東京工科大学 コンピュータサイエンス学部 コンピュータサイエンス学科 青木研究室

本研究では、深層学習モデルの敵対的攻撃に対する新しい防御手法「適応的マルチチャネル選択法（AMCS)」を提案する。AMCS は、小さいカーネルと大きいカーネルの畳み込み層を組み合わせることでモデルは多様な特徴を抽出し、頑健性を向上させることに成功した。

敵対的攻撃、適応的マルチチャネル選択法、深層学習モデル

１．はじめに

深層学習技術の発展により、自動運転車や医療画像診断システムなど、高度なセキュリティが要求されるアプリケーションの早期普及が期待されている。しかし、これらのシステムは一般に敵対的攻撃に対して脆弱であり、その影響は深刻である[1]。敵対的攻撃は、システムを悪用することを目的として行われる手法であり、特に AI モデルの認識を混乱させることを狙っている[2]。

本研究ではこの課題に対処するため、CNN（Convolutional Neural Network）の構造そのものに着目した新たなアプローチを提案する。

２．予備実験

本章では、CNN の構造が敵対的攻撃に対する堅牢性に与える影響を分析する 2 つの実験について述べる。

２.１チャネル影響分析

本実験では ResNet を使用し、Animals-10 データセット（10 のカテゴリに属する約 28,000 枚）に対して FGSM[3]、PGD[4]、LAP[5]の 3 種類の攻撃を適用した。敵対的サンプルの特定チャネルをクリーンサンプルのチャネルに最大3チャネル置換し、分類精度の評価を行った。

表 1 は、ResNet における入力層でのチャネル置換の影響を示しており、3 チャネルの置換で最大19.2%の精度向上が観察された。

表 1 ResNet18 におけるチャネル置換の影響

指標	FGSM	PGD	LAP
クリーンサンプル精度	81.3%	81.3%	81.3%
敵対的サンプル精度	70.0%	54.2%	45.3%
置き換え後の精度	77.3%	73.4%	60.4%
置き換えたチャネル	14, 18, 41	14, 18, 41	14, 18, 26

２.２カーネルサイズの影響分析

本実験では、ResNet の入力層のカーネルサイズのみを変更し、3×3（オリジナル)、5×5、7×7、9×9 の 4 種類のサイズで比較を行った。

表 3 に示している通り、カーネルサイズの増大に伴い敵対的サンプルに対する精度が向上する一方で、クリーンサンプルの精度が低下する傾向が観察され、7×7 のカーネルサイズが堅牢性と全体的な分類性能の最適なバランスを示した。

表 2 ResNet18 におけるカーネルサイズによる精度の変化

入力層のカーネルサイズ	クリーンサンプル精度	敵対的サンプル精度
3x3	81.3%	54.2%
5x5	78.9%	57.0%
7x7	75.4%	61.2%
9x9	70.1%	61.3%

２.３考察

これらの実験結果から、以下の重要な知見が得られた。

第 1 に、入力層での特徴抽出が敵対的攻撃に対

する堅牢性に決定的な役割を果たしていることが明らかになった。第 2 に、特定のチャネルが堅牢性に対して特に脆弱であり、これらのチャネルを適切に処理することで大幅な改善が見込めることがわかった。第 3 に、カーネルサイズの適切な選択により、堅牢性を向上させられる可能性が示された。

これらの知見は、CNN の構造自体を最適化することで、従来の防御手法を補完し、より効果的な敵対的攻撃対策を実現できる可能性を示唆している。

3．提案手法

本研究で提案する適応的マルチチャネル選択法（AMCS）は、2.3 で得られた知見に基づき、異なるサイズのカーネルを組み合わせた特徴抽出と、バッチサイズの動的調整を行うことにある。

AMCS は、デュアルカーネル入力層（3×3 と 7×7 の畳み込み層）、バッチサイズ調整機構、特徴統合層、そして標準的な CNN アーキテクチャから構成される。大きいバッチサイズは堅牢性の向上に、小さいカーネルサイズは精度向上に寄与する。

図 1 に示す AMCS の構造はこれらの要素を統合することで、従来のトレードオフを克服し、実用的な防御手法の確立が期待される。

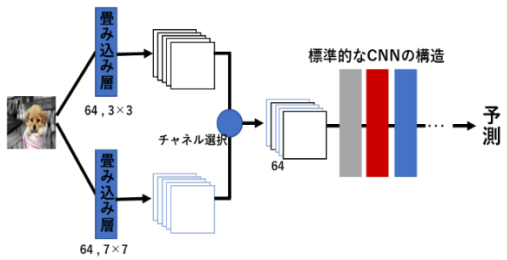

図 1 提案手法

4．実験

本章では、提案手法である AMCS の有効性を検証するために行った実験の結果について述べる。

4．1実験条件

実験では ResNet，EfficientNet，WideResNe の 3 つのモデルを使用し、従来のモデルと AMCS を加えた提案手法の比較を行う。攻撃には FGSM，PGD，LAP の 3 種類を使用し、Animals-10 データセットを使用して実験を行う。

4．2実験結果

表 4 に各モデルと攻撃手法の組み合わせにおける分類精度を示す。結果として AMCS を適用したモデルは、全ての攻撃手法に対して従来モデルよりも高い堅牢性を示した。クリーンサンプルに対する精度の変化は 2%程度に抑えられた。

表 3 チャネル交換による各攻撃手法の精度の変化と交換したチャネル

モデル	クリーンサンプル精度	FGSM精度	PGD精度	LAP精度
ResNet18	81.3%	53.1%	48.3%	49.5%
AMCS-ResNet18	**81.0%**	**69.8%**	**70.1%**	**65.2%**
EfficientNet-B0	75.2%	48.3%	40.4%	39.9%
AMCS-EfficientNet-B0	**77.8%**	**60.5%**	**51.4%**	**49.8%**
WideResNet-50-2	79.8%	59.7%	57.1%	61.4%
AMCS-WideResNet-50-2	**78.4%**	**66.2%**	**60.3%**	**68.9%**

5．結論

実験結果から、AMCS を適用したモデルは、従来のモデルと比較して敵対的攻撃に対する堅牢性が大幅に向上することが確認された。今後は、チャネル選択アルゴリズムの改良やカーネルサイズの最適化など、AMCS の性能を更に高めることが重要な課題である。

参考文献

[1] 田篭 照博. "AI セキュリティと敵対的サンプルの脅威." NRI セキュアブログ, NRI セキュア, 2021 年 8 月 25 日, https://www.nri-secure.co.jp/blog/hostile-sample.(2024 年 7 月 26 日参照)

[2] 松永, スキルアップ AI. "機械学習モデルへの敵対的攻撃とは〖スキルアップ AI キャンプ勉強記録〗." スキルアップ AI Journal, スキルアップ AI, 2023 年 3 月 28 日, https://www.skillupai.com/blog/tech/adversarial-attacks/.(2024 年 7 月 26 日参照)

[3] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." arXiv, 2014,https://arxiv.org/abs/1412.6572. (2024 年 7 月 26 日参照)

[4] Madry, Aleksander, et al. "Towards Deep Learning Models Resistant to Adversarial Attacks." ICLR, 2018, https://openreview.net/forum?id=rJzIBfZAb. (2024 年 7 月 26 日参照)

[5] Laidlaw, Cassie, and Soheil Feizi. "Laplacian Pyramid Adversarial Examples." (NeurIPS), 2019, https://arxiv.org/abs/1905.03345. (2024 年 7 月 26 日参照)