画像分類における効率的なバックドア防御のためのアクティベーションクリッピング

徐 煜凱 [1,*]　顧 玉杰 [2]　櫻井　幸一 [2]

概要：バックドア攻撃は，特に画像分類において，深層ニューラルネットワークに対する重大な脅威である．バックドア攻撃は，わずかな汚染された訓練データで被害者モデルを汚染することが可能である．その結果，被害者モデルはバックドアパターンを含むテストサンプルを攻撃者の指定したクラスに誤分類し，バックドアのないサンプルは正しく分類する．攻撃アルゴリズムはますます目立たない方法でバックドアパターンを埋め込むことができるため，検出および緩和が困難となる．本研究では，訓練データセットのみを用いて，訓練過程中にアクティベーション値をクリップし制約する新しい防御メカニズムを提案する．提案手法は，訓練データのみを使用し，訓練過程でアクティベーション値をクリップおよび制限する新しい防御メカニズムを提案する．包括的な実験により，提案手法が最新技術に比べて高い精度と低い攻撃成功率（ASR）を達成し，さらに時間消費も少ないことが示された．これらの結果は，提案されたアクティベーションクリッピング手法が，バックドア攻撃に対する画像分類モデルのロバスト性を向上させるための効果的かつ効率的な解決策であることを示した．

キーワード：深層ニューラルネットワーク，バックドア攻撃，バックドア防御

Efficient Activation Clipping for Backdoor Defense in Image Classification

Yukai Xu[1,*]　Yujie Gu[2]　Kouichi Sakurai[2]

Abstract: Backdoor attacks represent a significant threat to deep neural networks, particularly in image classification. A backdoor attack can poison the victim model with only a few contaminated training samples. Consequently, the victim model misclassifies test samples containing the backdoor patterns into the attacker's target class, while classifying those without backdoors correctly. Advanced attack algorithms can embed backdoor patterns in increasingly subtle ways, making them difficult to detect and mitigate. In this paper, we propose a novel defense mechanism that clips and bounds activation values during the training process using only the training data. Comprehensive experiments demonstrate that the proposed method achieves higher accuracy and lower attack success rates (ASR) than state-of-the-art approaches, while consuming less time. These results suggest that the proposed activation clipping method is an effective and efficient solution for enhancing the robustness of image classification models against backdoor attacks.

Keywords: Deep Neural Networks, Backdoor Attack, Backdoor Defense

1. はじめに

深層学習, 特に深層ニューラルネットワーク（DNN）は，多くの分野で顕著な成功を収めており[1][4], 特にその高い効果により画像分類で広く活用されている．しかし[9][13], 大量の訓練データの必要性[3]から，信頼性の低い第三者リソースが関与することが多く，これが深刻なセキュリティリスクをもたらす. バックドア攻撃のメカニズムは, 通常，モデルの訓練データセットを一つまたは複数のソースクラスからのサンプルで汚染し，特別に設計されたバックドアパターンを埋め込み，ラベルを特定のターゲットクラスに変更することで達成される．DNN モデルは，汚染された訓練データセットを学習することで，バックドアパターンとターゲットクラスを強く関連付ける[5]．その結果，検証過程において, DNN モデルは同じバックドアパターンを持つサンプルをターゲットクラスとして分類するが，正常なサンプルに対しては通常通りの分類を維持する．バックドア攻撃の気づきにくい性質，実際の運用においてその検出や防止を困難にしている．

本研究では，汚染されたモデルを浄化するための訓練後手法として，アクティベーションクリッピングによるバックドア防御メカニズムを提案する．この方法は，アクティベーション層の出力に着目している．先行研究に基づき，汚染されたモデルは特にアクティベーション層において，バックドアパターンに対して強く反応することが確認されている．いくつかのノードは，入力サンプルにバックドアが含まれている場合にのみ活性化され，その値は通常のアクティベーション値よりも高くなるため，最終的な分類結

1 九州大学大学院　マス・フォア・イノベーション連携学府
Joint Graduate School of Mathematics for Innovation, Kyushu University
2 九州大学大学院 システム情報科学府
Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University
* yukaiiu@yeah.net

果が歪められる．

2. 関連研究

2.1 バックドア攻撃

BadNets[5]は，最も代表的なバックドア攻撃である．具体的には，元の正常なデータセットからいくつかのサンプルをランダムに選択し，バックドアパターンを正常なサンプルにスタンプし，ラベルを攻撃者が指定するターゲットラベルに変更して，汚染サンプルを生成した．これらの汚染サンプルは残りの正常サンプルと組み合わせて，ユーザーに提供される汚染された訓練データセットを形成する．その後，Chen ら[2]は，汚染サンプルがステルス性を高めるためにその正常バージョンに似ているべきだと提案し，これに基づいてブレンド攻撃を提案した．Nguyen ら[12]は，元のサンプルを最大限に保持しながらバックドアパターンを埋め込むために，サンプルの内容を保ちつつ微妙に歪めるイメージワーピングをバックドアパターンとして使用する WaNet を提案した．

2.2 バックドア防御

現在，バックドアの脅威を軽減するためのいくつかの防御手法が存在する．バックドア防御は，モデルの訓練段階または訓練後に展開することができ，それぞれのシナリオにおいて防御者の役割や能力が異なると仮定されている．

Hendrycks ら[7]は，自己教師あり学習メカニズムが，画像破損やラベル破損など，さまざまな攻撃に対してよりロバストなモデルを訓練するための効果的なアプローチであると主張した．これに基づき，Huang ら[8]は，訓練過程を自己教師ありの事前訓練，教師ありの訓練，そして半教師ありのファインチューニングに分けるデカップリングベースのバックドア防御（DBD）を提案した．しかし，この方法は，防御者が汚染サンプルの存在を認識しているというシナリオに基づいており，実際の応用においては必ずしも現実的ではない．また，訓練過程が複雑であり，時間と計算負荷が増加する．

Li ら[11]は，クリーンデータセットで訓練された教師モデルを用いて，汚染された生徒モデルに対する知識蒸留を実施した．しかし，この手法は別のモデルを使用し，さらに別途正常な訓練データセットが必要となるため，現実のシナリオでは実用的ではないことが多い．

Wang ら[14]は，訓練後過程に着目し，最大マージンに基づくバックドア検出手法（MM-BD）を提案した．この方法では，防御者が小規模なクリーンデータセットを所有していることを前提としており，そのデータセットを用いてモデルが汚染されているかどうかを検出し，さらにクリーンデータセット上でアクティベーションクリップの境界を最適化して，汚染されたモデルを浄化する．しかし，現実のシナリオでは，信頼できるクリーンデータセットが常に入手可能とは限らず，その最適化過程も計算コストを増加させる．本研究では，追加のデータセットや最適化プロセスを必要とせず，汚染されたモデルを浄化するための訓練後アクティベーションクリッピング防御メカニズムを提案する．この手法は，実用的なアプリケーションにおいて効率的である．

3. 提案手法

3.1 問題設置

典型的なバックドア攻撃は次のように定式化される．元のクリーンな訓練データセットを$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ とし，ここでx_i は訓練サンプル，y_i はその対応するラベルである．バックドア攻撃では，攻撃者がクリーンデータセットの部分集合$\mathcal{D}_p \subset \mathcal{D}$を選択し，各サンプル $x_i \in \mathcal{D}_P$ にバックドアパターンδ を加えて汚染サンプル$\tilde{x}_i = x_i + \delta$ を作成する．これらの汚染サンプルのラベルはターゲットラベルtに変更され，$\tilde{y} = t$ と表される．したがって．汚染されたデータセットは次のように表される：

$$\widetilde{\mathcal{D}} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{|\mathcal{D}_p|} \cup \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|-|\mathcal{D}_p|} \quad (1)$$

汚染されたモデルの訓練目標は，汚染データに対する損失を最小化することになる：$\ell = \sum_{(x_i,y_i)\in\widetilde{\mathcal{D}}} \ell(f'_\theta(x_i), y_i)$．ここで$\ell$ は損失関数，f'_θ は汚染されたモデルである．

3.2 アクティベーションクリッピング

DNN モデルf'_θ が複数の層から構成されるとし，入力 x に対する層l のアクティベーション出力を $\boldsymbol{a}^l(x)$ とする．訓練中，モデルはバックドアパターン δ を特定の層でのアクティベーション値と関連付けることを学習する．バックドア効果を軽減するために，アクティベーションを特定の層で制約するアクティベーションクリッピングメカニズムを導入する．層l におけるアクティベーションのクリップ境界を$\boldsymbol{c}^l$とする．層のアクティベーションは，この境界を超えないようにクリップされる．入力 x に対して，層 l のクリップされたアクティベーションは次のように表される：$\widehat{\boldsymbol{a}}^l(x) = min\,(\boldsymbol{a}^l(x), \boldsymbol{c}^l)$．これにより，特にバックドアパターン$\delta$ によって引き起こされる過剰なアクティベーションが抑制され，入力が攻撃者のターゲットクラスt に誤分類されることを防ぐ．

アクティベーション層のクリップ境界を設定するために，各層の各ノードにおけるアクティベーションの平均値と標準偏差を計算した．具体的には，層 lll におけるアクティベーションの平均を$\boldsymbol{\mu}^l$，標準偏差を $\boldsymbol{\sigma}^l$ とし，クリップ境界$\boldsymbol{c}^l$ は次の式に従って設定した：$\boldsymbol{c}^l = \boldsymbol{\mu}^l + \lambda \cdot \boldsymbol{\sigma}^l$，ここで λ はクリッピングの厳しさを調整するためのスケーリング係数である．

4. 実験と結果

4.1 実験設置

提案手法の有効性を評価するために，実験を行った．CIFAR-10 データセット[10]と ResNet18[6]という広く使用されている深層ニューラルネットワークアーキテクチャを使用して実験を行った．

CIFAR-10 データセットは，10 クラスに分類された 60,000 枚の 32x32 カラー画像で構成されており，そのうち 50,000 枚を訓練用，10,000 枚をテスト用として使用した．バックドア攻撃は, 50,000 枚の訓練画像のうち 1%にあたる 500 枚のサンプルに特定のパターンを埋め込み，それらをターゲットクラスに割り当てる設定で行った．モデルの性能は，正常データに対する分類精度と，汚染データに対する攻撃成功率（ASR）で評価し，防御手法の適用前後の結果を比較した．

4.2 実験結果

表 1Table 1 は, 3 つの異なるバックドア防御手法（NAD, MM-BD，そして本論文で提案した手法）における分類精度（ACC）と攻撃成功率（ASR）の比較分析を示している，3 つの攻撃シナリオ（BadNet，Blend，WaNet）に対して比較している．BadNet 攻撃では，提案手法は最も高い分類精度（90.2%）を達成し，ASR も 1.1%と MM-BD と同等で，NAD よりも優れている．WaNet 攻撃では，提案手法は他の手法をわずかに上回り，分類精度 90.9%と ASR 9.8%を達成し，NAD および MM-BD よりも良好な結果を示している．実験結果は，提案されたアクティベーションクリッピング防御の有効性を示した．

表 1　ResNet18 モデルにおける CIFAR-10 の分類精度と攻撃成功率（ASR）の比較

Table 1　Comparison of Accuracy and Attack Success Rate (ASR) on CIFAR-10 with ResNet18.

攻撃手法		NAD	MM-BD	This paper
BadNet	ACC	88.7%	89.1%	**90.2%**
	ASR	2.2%	**1.1%**	**1.1%**
Blend	ACC	**87.6%**	83.8%	85.9%
	ASR	10.3%	11.7%	9.8%
WaNet	ACC	90.7%	89.77%	90.9%
	ASR	9.9%	10.1%	9.8%

5. おわりに

本研究では，画像分類モデルにおけるバックドア攻撃を軽減するためのアクティベーションクリッピングに基づく防御メカニズムを提案した．この手法は，追加のデータセットや複雑な最適化プロセスを必要としない訓練後手法として機能し，実用的な応用において効率的であることを特徴としている．CIFAR-10 データセットと ResNet アーキテクチャを使用した典型的な実験を通じて，提案された手法が BadNet，Blend，WaNet を含むさまざまなバックドア攻撃シナリオにおいて，攻撃成功率（ASR）を効果的に低減しながら，高い分類精度を維持することを示した．NAD および MM-BD との比較結果も，堅牢性と精度のバランスにおいて提案手法の有効性を強調している．これらの結果から，アクティベーションクリッピングは, DNN のセキュリティと信頼性をバックドアの脅威から強化するための，有望かつ実用的な解決策であることが示唆される. 今後の研究では，さらなる最適化や他のモデルやデータセットへの拡張が検討されるべきである．

参考文献

[1] Baevski, A. et al.. wav2vec 2.0: A framework for self-supervised learning of speech representations. NIPS. 2020, p.12449–12460.

[2] Chen, X. et al.. Targeted backdoor attacks on deep learning systems using data poisoning. 2017.

[3] Deng, J. et al.. Imagenet: A large-scale hierarchical image database. CVPR. 2009, p.248–255.

[4] Devlin, J. et al.. Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT. 2019, p. 4171–4186.

[5] Gu, T. et al.. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 2019, vol 7, p.47230-47244.

[6] He, K. et al.. Deep residual learning for image recognition. CVPR, 2016. p.770–778.350.

[7] Hendrycks, D. et al.. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. NIPS. 2019.

[8] Huang, K. et al.. Backdoor defense via decoupling the training process. ICLR. 2022.

[9] Kim, H. E. et al.. Transfer learning for medical image classification: A literature review. BMC Medical Imaging. 2022, 22(1), p. 69.

[10] Krizhevsky, A. et al.. Learning multiple layers of features from tiny images. Technical Report. 2009.

[11] Li Y. et al.. Neural attention distillation: Erasing backdoor triggers from deep neural networks. ICLR. 2021.

[12] Nguyen, A., Tran, A. WaNet--imperceptible warping-based backdoor attack. ICLR. 2021.

[13] Redmon, J. and Farhadi, A. Yolov3: An incremental improvement. CoRR abs/1804.02767. 2018.

[14] Wang, H. et al.. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. IEEE Symposium on Security and Privacy. 2023.