Adversarial examples に対する畳み込みニューラルネットワークの挙動分析と検出法の提案

EP18075 土松千紗　　指導教授：山下隆義

1. はじめに

Adversarial Examples(AEs) は，入力画像に摂動を付与することで，畳み込みニューラルネットワーク (CNN) の誤認識を誘発する．AEs の防御方法として，様々な Adversarial detection 手法が提案されている．それらは，入出力の関係のみに着目しており，AEs がネットワーク内部にどのような影響を及ぼしているか十分に調査されていない．そこで本研究では，AEs を入力した際のネットワークの内部状態を把握するために特徴マップを分析し，分析結果に基づいた AEs の検出法を提案する．

2. Adversarial detection

Adversarial detection は，AEs の特徴に注目して AEs かどうか検出する防御手法である．代表的な手法である Feature squeezing[1] は，入力画像に対して色深度の削減や平滑化をした際の出力値の違いから AEs を検出する．色深度の削減は画素に使用するビットを削減する処理で，表現できる色数を減らしている．AEs に対して色深度の削減や平滑化を行うと摂動の影響が弱まり，推測クラスが処理前後で変化するため，AEs を検出できる．

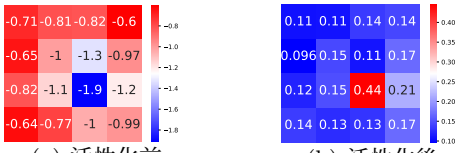
3. 提案手法

AEs の摂動は出力確率を低くするように入力画像に対する勾配から作成するため，防御する際に CNN の内部状態を考慮することが重要だと考える．しかし，Feature squeezing は入出力の関係のみに着目しており，内部状態に対する分析が十分にされていない．本研究では，入力画像に様々な幾何変化を施した時の内部状態に着目して，まず通常のサンプル (Clean) と AEs の傾向を調査する．

3.1. CNN の挙動分析

本分析では，CIFAR-10 データセットを 80 エポック学習した ResNet-18 モデルを対象とする．画像に対して左右反転と {90°, 180°, 270°} の回転の幾何変換を加えた画像を入力とする．本分析では特徴マップをチャネル方向に平均し，活性化前後の各 Res-block の平均特徴マップを比較する．

Clean に対する，活性化前と活性化後の特徴マップのチャネル方向の平均を図 1 に示す．図 1 より，活性化後の最大値の座標と活性化前の最小値の座標が一致することが確認できる．したがって，活性化後の最大値と活性化前の最小値の座標が一致する割合で AEs を検出できると考える．



(a) 活性化前　　(b) 活性化後

図 1: Clean 入力時の特徴マップの平均

次に，各学習サンプルに対して座標の一致率と閾値を比較し AEs を検出した場合の AUROC を表 1 に示す．ここで真陽性率は AEs を検出した確率，偽陽性率は Clean を AEs と検出した確率とする．表 1 より，5 番目の Res-block の AUROC が最も優れている．したがって，5 番目の Res-block が AEs の検出に最適であると考える．

表 1: 検出に利用する特徴マップに対する AUROC

	Res-block		
	3 番目	5 番目	7 番目
AUROC	0.4946	**0.5378**	0.4993

3.2. AEs の検出

3.1 節で得た挙動分析の結果をもとに AEs の検出法を提案する．提案する AEs 検出の流れを図 2 に示す．まず Step1 として事後確率を用いて AEs の候補を求める．ここでは，各幾何変換画像の事後確率と元画像の事後確率の KL ダイバージェンスの中央値を閾値と比較して検出する．そして Step2 では，特徴マップを用いて AEs の最終判定を行う．分析結果をもとに AEs の候補は，5 番目の Res-block のチャネル方向に平均した特徴マップを取得し，活性化後の最大値と活性化前の最小値の座標が一致する割合を閾値と比較して検出する．

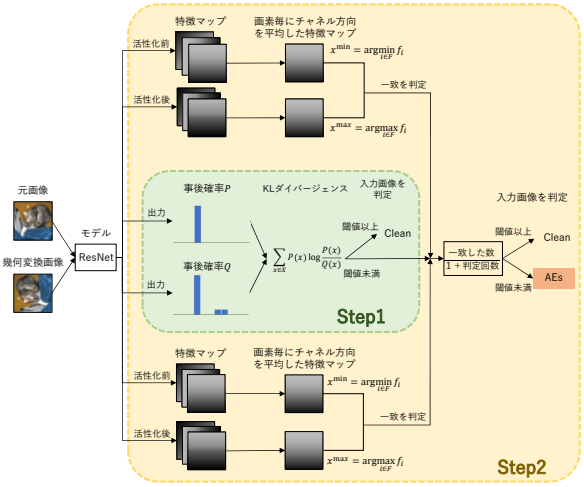


図 2: 提案手法による検出の流れ

4. 評価実験

提案手法と Feature squeezing の検出精度を比較し，提案手法の有効性を示す．

4.1. 実験概要

本実験ではデータセットに CIFAR-10，ベースモデルに ResNet-18 を用いる．学習回数は 300 エポックとする．攻撃手法は FGSM，PGD を使用する．ここで，摂動の強さ ϵ は 0.031，PGD のステップサイズ α は 0.003，PGD の摂動計算の反復回数は 20 とする．また，Step1 における閾値は 0.1，Step2 における閾値は 0.125 とする．

4.2. 実験結果

提案手法および Step1 のみ，Step2 のみ，Feature squeezing を用いた場合の認識率を表 2 に示す．表 2 より，提案手法は FGSM に対する認識率が Feature squeezing より低い一方で，PGD に対する認識率は高い．また，提案手法の FGSM，PGD に対する認識率は Step1 のみを用いた手法，Step2 のみを用いた手法よりも低下した一方で，Clean に対する認識率は向上した．以上より提案手法は，事後確率と特徴マップによる検出を組み合わせることで Clean に対する認識精度を維持しつつ AEs に対する堅牢性を向上させることができた．

表 2: 認識率の比較 [%]

	Clean	FGSM	PGD
FS	**86.50**	**61.54**	2.97
提案手法 (Step1 のみ)	83.34	43.52	15.07
提案手法 (Step2 のみ)	78.15	47.27	**22.39**
提案手法	83.44	43.49	14.56

5. おわりに

本研究では，特徴マップと事後確率の観点から AEs が与える影響を分析し，AEs の影響を考慮した検出手法を提案し，完全に AEs を検出することは困難であることを確認した．今後は，特徴マップを取得する箇所を変更して分析を行う予定である．

参考文献

[1] W. Xu, *et al.*, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks”, NDSS, 2018.