IS1-51：特徴マップの幾何変換前後に着目した敵対的サンプルの検出

土松千紗，足立浩規，平川翼，山下隆義，藤吉弘亘
中部大学

PDF

研究背景

Adversarial Examplesの防御
- Adversarial Examples (AEs) : 畳み込みニューラルネットワーク (CNN) の誤認識を誘発
- Adversarial detection : AEsの特徴や挙動に注目した検出器をモデルの前に配置する防御法

従来手法の問題点
- モデルの内部状態に対する分析が不十分
  - AEsの多くがモデルの勾配をもとに摂動を導出

  → モデル内部の変化を考慮することで更なる性能向上を期待

研究目的

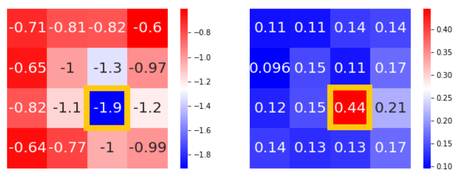
AEsを入力した際のネットワークの内部状態の把握
- 幾何変換を施した画像の活性化前後の特徴マップの調査

内部状態の分析から検出手法を提案
- AEsに対する検出率，認識率の向上

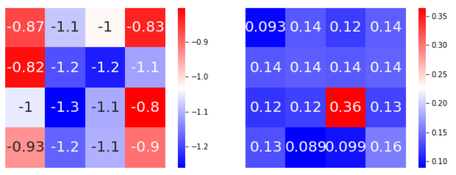
分析

設定
- データセット
  - CIFAR-10
- モデル
  - ResNet-18
- 攻撃手法
  - PGD ($\epsilon = 0.031$，$\alpha = 0.003$)
- 幾何変換
  - 左右反転+$\{90°, 180°, 270°\}$の回転

1サンプルを用いた分析結果



通常の画像 (Clean)

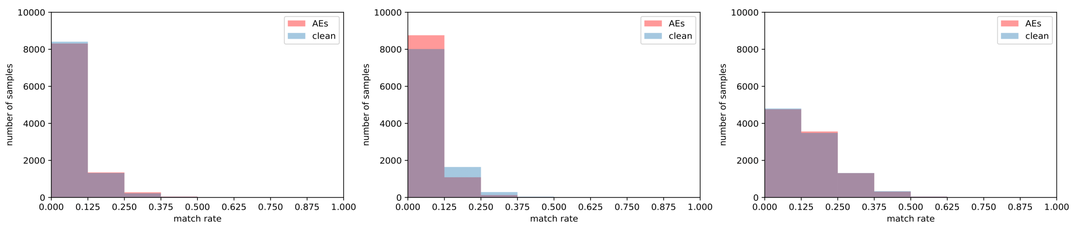


AEs
特徴マップのチャネル方向の平均

- Clean : 座標の一致率が高い
- AEs : 座標の一致率が低い

活性化前の最小値，活性化後の最大値の座標の一致率を用いたAEsの検出が可能

複数サンプルを用いた分析結果



3層目　5層目　7層目

各層における座標一致率の分布

特徴マップに対するAUROC

	3層目	5層目	7層目
AUROC	0.4946	**0.5378**	0.4993

- 5層目
  - CleanとAEsの分布の違いが最も大きい
  - AUROCが最も高い値

5層目の特徴マップを用いた検出が有効

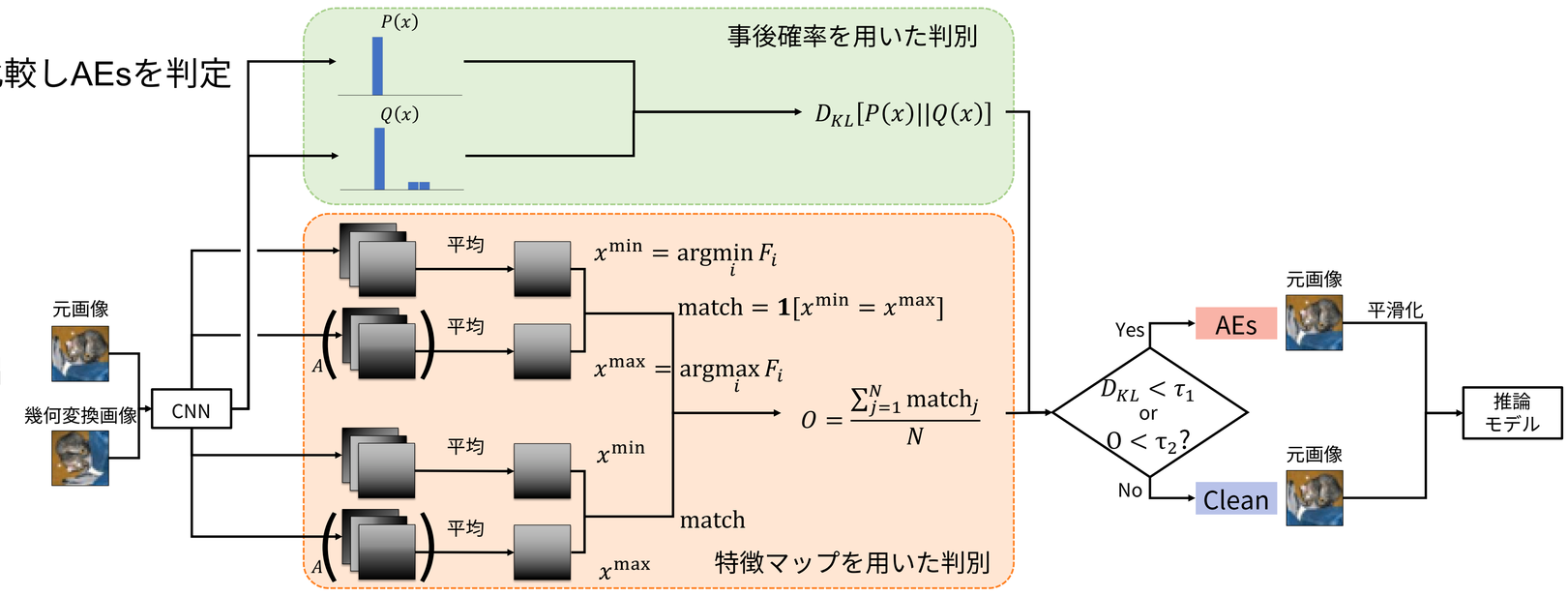
提案手法

特徴マップと事後確率を利用した検出器
- 特徴マップと事後確率を用いた計算結果と閾値を比較しAEsを判定

事後確率を用いた判別
- 各幾何変換画像と元画像の事後確率を取得
  - KLダイバージェンスを導出し，閾値と比較

特徴マップを用いた判別
- 活性化前の最小値，活性化後の最大値の座標を導出
  - 座標の一致率を導出し，閾値と比較

検出後の処理
- AEsと判定した画像に対する平滑化
  - 2×2の平均フィルタを使用



実験

実験設定
- データセット : CIFAR-10
- モデル : ResNet-18
- 攻撃手法
  - PGD ($\epsilon = 0.031$，$\alpha = 0.003$)
  - FGSM ($\epsilon = 0.031$)
- 幾何変換
  - 左右反転+$\{90°, 180°, 270°\}$の回転
- 閾値
  - $\tau_1 = \{0.1,1,10,15\}$
  - $\tau_2 = 0.125$
- 比較手法
  - DLA [P. Sperl+, EuroS&P, 2020]
  - Feature Squeezing [W. Xu+, NDSS, 2018]
  - PixelDefend [Y. Song+, ICLR, 2018]

検出性能の評価
- 検出率のF値を比較

従来手法と提案手法の検出率のF値の比較

	FGSM	PGD
DLA	**0.815**	**0.833**
Feature Squeezing	0.667	0.667
PixelDefend	0.571	0.571
提案手法($\tau_1 = 0.1$)	0.572	0.581
提案手法($\tau_1 = 1$)	0.666	0.660
提案手法($\tau_1 = 10$)	0.667	0.667
提案手法($\tau_1 = 15$)	0.667	0.667

- 従来手法と同程度，または低下

修正性能の評価
- 認識率を比較

従来手法と提案手法の認識率の比較 [%]

	Clean	FGSM	PGD
防御なし	**92.39**	17.51	0.01
Feature Squeezing	86.50	61.54	2.97
PixelDefend	85.00	46.00	46.00
提案手法($\tau_1 = 0.1$)	88.08	28.62	1.10
提案手法($\tau_1 = 1$)	86.19	33.44	1.52
提案手法($\tau_1 = 10$)	86.20	**61.93**	72.89
提案手法($\tau_1 = 15$)	86.20	**61.93**	**72.90**

検出の組み合わせごとの認識率の比較 [%]

	Clean	FGSM	PGD
防御なし	**92.39**	17.51	0.01
事後確率のみ	92.30	17.82	0.01
特徴マップのみ	88.11	28.48	**1.10**
提案手法	88.08	**28.62**	**1.10**

- AEsに対する認識率
  - 従来手法より向上
  - 特徴マップと事後確率を組み合わせることで向上

まとめ・今後の予定

まとめ
- 提案手法はAEsに対する修正性能が高い

今後の予定
- 特徴マップの取得位置の変更 : 畳み込み前後より取得
- 統計手法の変更 : バイスペクトル，コサイン類似度の導出