6ZM-01

画像特徴量のロバスト性に着目した教師なし敵対的サンプル検知

神田 悠斗 †　　波多野 賢治 †
† 同志社大学文化情報学部

1 はじめに

Deep Neural Network (DNN) は，画像処理や自然言語処理をはじめとする様々なタスクにおいて近年大きな成果を上げている機械学習手法である．一般的に， DNN モデルはランダムに発生するノイズに対して頑健であるように設計される．しかし，ある特定のパターンを持ったノイズに対しては強く反応し，予測結果が変化することがある．そのようなノイズのうち，人間には知覚できないような小ささで意図的に作成されたものを敵対的摂動と呼び，それが加えられた入力データのことを敵対的サンプルと呼ぶ．人間には正常な入力と敵対的サンプルを区別することは難しいため，利用者には気づかれずにモデルの誤分類を引き起こすことが可能となる．このような敵対的サンプルの存在は DNN モデルの信頼性を低下させ，実社会における DNN 活用の足枷となる可能性がある．

敵対的サンプルに対する防衛策の一つは，モデルに入力されるデータから敵対的サンプルを検知することである．本研究では画像処理分野における敵対的サンプル検知の新たなアプローチとして，入力画像特徴量のロバスト性に着目した敵対的サンプルの検知手法を提案する．

2 関連研究

敵対的サンプルを解釈するにあたり，Ilyas らはロバスト特徴量と非ロバスト特徴量と呼ばれる二つの概念を導入している [1]．ロバスト特徴量とは，高い予測性を持ちながら，ノイズに対して頑健である特徴量を指す．一方で非ロバスト特徴量は，高い予測性を持つものの，ノイズに弱く壊れやすい特徴量を意味する．正常な入力に対してごく小さな敵対的摂動を付与することによって敵対的サンプルが作成されることから，敵対的サンプルは真のクラス以外の非ロバスト特徴量が強められることで生成されると Ilyas らは主張している．

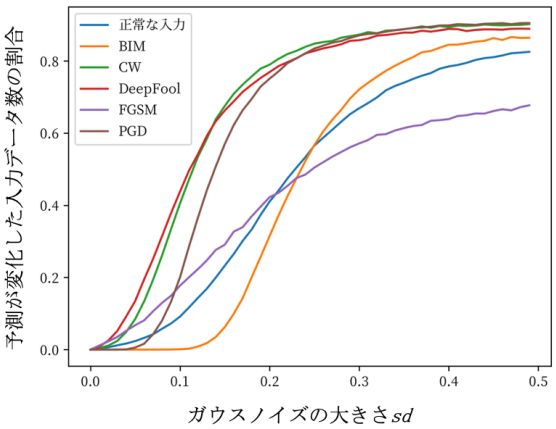

図 1　大きさ sd のガウスノイズを加えた際に予測が変化した入力データ数の割合の推移

3 提案手法

Ilyas らの仮説に従えば，正常な入力と敵対的サンプルでは，モデルが予測の根拠とする特徴量のロバスト性が異なる可能性がある．

この仮説を実証するために，正常な例と敵対的サンプルの画像特徴量のロバスト性を検証する基礎分析を行った．具体的には，正常な入力と評価実験で用いた 5 種類の敵対的サンプルに対してある大きさのガウスノイズを付与した際の，全入力データ数に対する予測が変化した入力データの数の割合を計測した．ガウスノイズは $N(0, sd)$ に従うものとし，sd は 0 から 0.01 刻みで 50 パターンを計測したところ，図 1 のようになった．

CW, DeepFool, PGD の 3 種類の敵対的サンプルについては，ほぼ全ての大きさのガウスノイズで正常な入力よりも予測が変化した入力の割合が大きく，正常な入力よりもロバスト性が

An Unsupervised Detecting Method for Adversarial Examples based on The Robustness of Image Features
† KANDA Yuto and HATANO Kenji, Faculty of Culture and Information Science, Doshisha University

低いことが読み取れる．また，BIM と FGSM では，それぞれある sd を境に正常な入力とロバスト性が逆転しているものの，正常な入力とは異なるロバスト性の推移を辿っている．

この結果から，いずれの敵対的サンプルも正常な入力とは異なるロバスト性を持つことが確認できた．つまり，特徴量のロバスト性に着目することで正常な入力と敵対的サンプルを見分けられる可能性がある．そこで，特徴量のロバスト性を反映させた出力値であるロバスト出力値を提案し，敵対的サンプルの検知に利用する方法を提案する．

ロバスト出力値は，モデルの出力値からロバスト性の低い特徴量の影響を取り除くことで求められる．クラス i $(i = 1, 2, ..., N)$ におけるロバスト出力値 Z_{r_i} は，次の式で定義される．

$$Z_{r_i} = Z_{c_i} - |Z_{c_i} - Z_{n_i}|$$

ここで Z_{c_i} は，入力画像を N クラスに分類する DNN モデルの，クラス i に対するモデルの出力値である．また Z_{n_i} は，ある大きさのガウスノイズを加えた入力画像をモデルへ入力した時のクラス i の出力値とする．

非ロバスト特徴量は壊れやすいため，ガウスノイズを付与することで簡単にモデルの出力値が変化する．したがって，ガウスノイズを加えた際のモデルの出力値の変化量 $|Z_{c_i} - Z_{n_i}|$ は，非ロバスト特徴量の影響量と捉えることができる．この考えから，Z_{r_i} は，モデルの出力値 Z_{c_i} から非ロバスト特徴量の影響量 $|Z_{c_i} - Z_{n_i}|$ を取り除くことで求められる．

敵対的サンプルと正常な入力では特徴量のロバスト性が異なるため，ロバスト出力値の分布も異なることが予想される．これを利用し，Z_{r_i} の集合 Z_r を特徴量として用いた敵対的サンプルの教師なし検知を行う．

4 評価実験

提案するロバスト予測値が敵対的サンプル検知に有効であることを示すために，敵対的サンプル検知に関する比較実験を行った．比較対象には敵対的サンプルの教師なし検知で最も高い精度を有している研究 [2] を採用した．画像分類モデルのアーキテクチャには DenseNet-BC，Resnet-34 を，モデルの学習に用いるデータセットには，CIFAR-10，CIFAR-100，SVHN を使用した．また，敵対的サンプルの生成には，文献 [2] で用いられていた五つの手法を採用した．ロバスト出力値を算出する際に付与するガウスノイズは $N(0, 0.01)$ ，$N(0, 0.1)$，$N(0, 0.2)$ の 3 パターンでロバスト出力値を算出し，これらを結合したものを敵対的サンプル検知の特徴量として利用した．敵対的サンプルの検知器には One-Class SVM を用い，画像分類モデルの学習に用いたデータで学習を行った．

表 1　各敵対的サンプル生成法に対する AUROC (%)

モデル	データセット	手法	BIM	CW	DeepFool	FGSM	PGD
DenseNet	CIFAR-10	既存手法	97.07	69.49	66.46	77.27	93.85
		提案手法	**99.68**	**76.91**	**79.09**	**90.50**	**99.23**
	CIFAR-100	既存手法	96.08	61.81	61.52	**97.58**	88.33
		提案手法	**99.48**	**65.61**	**61.65**	74.42	**97.56**
	SVHN	既存手法	94.69	86.17	85.83	**96.92**	94.56
		提案手法	**95.45**	**89.90**	**90.68**	89.19	**98.42**
ResNet	CIFAR-10	既存手法	**95.88**	**74.94**	78.98	**97.79**	79.17
		提案手法	94.60	72.07	**88.14**	92.54	**96.25**
	CIFAR-100	既存手法	79.30	72.70	70.65	**96.56**	69.72
		提案手法	**88.83**	**86.03**	**95.19**	90.09	**96.36**
	SVHN	既存手法	**95.60**	**90.06**	90.96	**98.92**	84.03
		提案手法	88.83	86.03	**95.19**	90.09	**94.88**

各データセットの評価データに対して敵対的サンプルを作成し，正常な評価データと敵対的サンプルを混在させた検知評価データセットに対して検知の精度評価を Area Under the Receiver Operating Characteristic Curve (AUROC) を用いて行ったところ，表 1 のようになった．FGSM ではやや劣るものの，その他の敵対的サンプルでは平均して大幅な精度向上が確認できた．

5 おわりに

本研究では，DNN の脆弱性である敵対的サンプルに対して，ロバスト出力値に基づく検知法を提案した．

今後の課題としては，最新のモデルアーキテクチャや敵対的サンプルの生成法に対する検知の精度を検証することである．

参考文献

[1] Andrew Ilyas, Shibani Santurkar, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. In *Advances in Neural Information Processing Systems*, Vol. 32 of *NeurIPS Proceedings*, 2019.

[2] Bartosz Wójcik, Paweł Morawiecki, Marek Śmieja, Tomasz Krzyżek, Przemysław Spurek, and Jacek Tabor. Adversarial Examples Detection and Analysis with Layer-wise Autoencoders. In *Proceedings of 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence*, pp. 1322–1326. IEEE, 2021.