6T-03

XAIを用いたノイズに頑健なモデル構築手法の提案

川端 祐也† 市川 嘉裕‡ 山口 智浩‡

奈良工業高等専門学校 専攻科 システム創成工学専攻† 奈良工業高等専門学校 情報工学科‡

1 研究背景と目的

機械学習モデルの判断根拠を説明する手法として説明可能なAI（eXplainable AI: XAI）が注目されている[1]．しかし，モデルが根拠としている部分にノイズを加えることで容易に予測結果を変えられることや，モデルの特性を利用した攻撃の対象になることが問題として挙げられる．この問題に対して，本研究ではXAIのSHAP[2]を用いてモデルを欺くのに効果的なノイズ画像を生成し，それを学習データとして学習させるモデル構築手法を提案する．

関連手法であるAdversarial Training（以下，AT法）は，モデルの特性を利用して敵対的サンプルと呼ばれるノイズがかった画像を生成し，再学習させることによりモデルを頑健にする手法である．しかし，AT法で作成したモデルは敵対的サンプルに対しての認識精度を大きく向上させるが，通常のCNNであればあまり問題にならない程度のランダムノイズに対する精度は下がることが問題として挙げられている[3]．

その点，提案手法により生成されるノイズ画像はモデルの根拠としている部分にピンポイントでノイズを加えるため，AT法により生成されるノイズ画像よりも元画像に近い状態でノイズを付加することができる．そのため，AT法よりもランダムノイズに対する精度が維持できると考えられる．

本稿では，提案手法の評価のために，提案手法が利用するノイズ画像とAT法が利用するノイズ画像（敵対的サンプル）をそれぞれ学習データとした際に，認識精度に与える影響を調査する．

2 モデル構築手法

提案するモデル構築手法の概要を図1に示す．本研究では手書き文字認識を対象とした関係で，それに用いたデータセット（MNIST）を例として説明する．まず，MNISTの一部を用いてCNN(Convolutional Neural Network)によってモデルを作成する．これを従来モデルと称する．このモデルに対してSHAPを適用し，判断の根拠を算出する．SHAPでは画像の平面領域に対してヒートマップで判断根拠を表現することが可能である．例えば，肯定的な根拠とされる部分は赤，否定的な根拠とされる部分は青となる．これを基に，自作のアルゴリズムによって画像にノイズを付加する．紙面の都合により詳細は割愛するが，SHPA値によって重み付けをして，各画素の白黒を反転させるような変更を加える．作成したノイズ画像を従来モデルで予測し，予測結果がノイズを付加する前のラベルと異なる画像だけを抽出する．作成したノイズ画像は元のラベルから大きく見た目が変わってしまっている画像や，読み取り不可能な画像が混在するため，ノイズ画像に対して人手により改めてラベル付けを行う．このノイズ画像とラベルを学習データとしたモデル構築手法が本研究の提案手法である．ここで，従来モデルを作成する画像群，ノイズ画像の元となる画像群，提案モデルの作成に用いる画像群は全て違う画像を用いている点に注意されたい．

3 実験

実験では，提案手法が生成するノイズ画像と敵対的サンプルが，それぞれ学習データとして有効かどうかを検証する．以下で実験に使用する学習データ①～④と評価データ Ⅰ～Ⅳについて整理する．

① MNISTのトレーニング画像の一部
② 提案手法により作成したSHAPノイズ画像
③ 従来モデルで作成したモデルを用いて生成した敵対的サンプル
④ ③とは別の学習データで作成した従来モデルを用いて生成した敵対的サンプル

Ⅰ. MNISTのテスト画像
Ⅱ. 提案手法によりテスト用に作成したSHAPノイズ画像
Ⅲ. ③と同じ敵対的サンプル
Ⅳ. ③と④とは異なる従来モデルで生成した敵対的サンプル

ここで，提案手法では①と②を併せてモデル構築し，AT法では①と③を組み合わせてモデル構築される．また，純粋なCNNでは①でモデル構

A Construction Method of Noise-Robust Learning Model with XAI

†Yuya Kawabata・Department of Systems Innovation, Nara National Institute of Technology, Nara College

‡Yoshihiro Ichikawa, Tomohiro Yamaguchi・Department of Information Engineering, Nara National Institute of Technology, Nara College

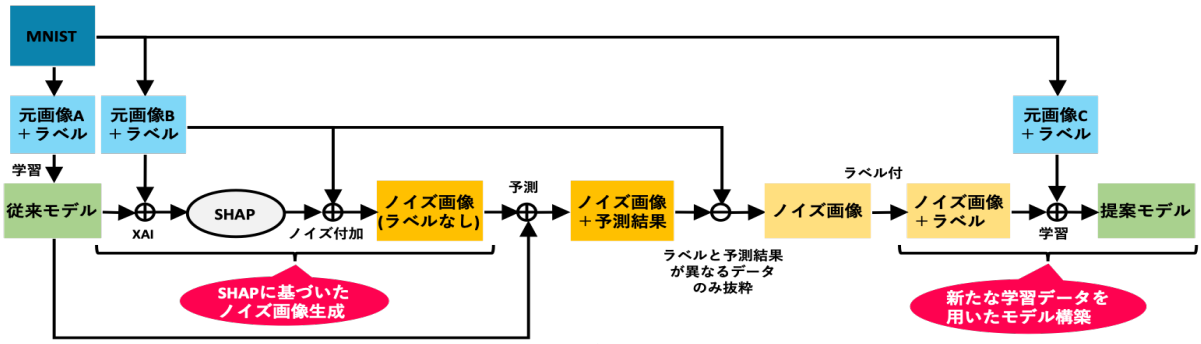

図 1 提案モデルの作成方法

築し，Ⅰで評価されることになる．AT 法はⅢに対して一定の認識精度が確保されており，提案モデルがⅡに対して一定の認識精度が確保されていることはほぼ自明である．

今回の実験では，①③で構築したモデルがⅡに対する精度と①②③や①②④で構築したモデルがⅡに対する精度を比較することで，SHAP ノイズ画像の有効性を確認する．提案モデルは AT 法の弱点を改善できたことになる．また，②③で構築したモデルについて，全ての評価データに対する精度が向上すれば，より少ない学習データで頑健なモデルを構築できることになる．ここで，これらのモデルはⅠ，Ⅱ，Ⅲに対する認識精度が落ちていないかどうかも検証する．実験で作成したモデルの構成条件を表 1 に示す．

表 1 各モデルを構築する学習データの内訳[枚]

	①	②	③	④
モデル 0	10000	0	0	0
モデル 1	10000	0	4130	0
モデル 2	10000	4130	4130	0
モデル 3	10000	4130	0	4130
モデル 4	0	4130	4130	0

4 実験結果・考察

前述の条件で構築したモデル 1～4 それぞれが評価データⅠ～Ⅳに対する正解率を表 2 に示す．

表 2 各モデルの正解率[%]

	Ⅰ	Ⅱ	Ⅲ	Ⅳ
モデル 0	96.70	54.02	69.61	72.57
モデル 1	97.02	70.26	98.09	85.93
モデル 2	97.20	87.42	97.87	90.07
モデル 3	97.35	87.42	87.24	91.91
モデル 4	93.85	85.57	97.92	87.31

従来の AT 法は通常のランダムノイズに対して精度が低いことが知られているが，以上の実験結果から SHAP ノイズ画像に対しての認識精度も低いことがわかった．そこで，モデル 2 やモデル 3 のように SHAP ノイズ画像を学習させることによって，敵対的サンプルに対する精度を落とすことなくノイズ画像に対しての認識精度も向上させることができた．今回の実験で，モデル 4 のようなより少ない学習データでモデルを頑健にするという課題に関しては，SHAP ノイズ画像や敵対的サンプルに対する認識精度は良いが，元画像に対する正解率が減少している点から元画像を一緒に学習させることが頑健なモデルを作成する最善策であると考えられる．

5 まとめと課題

本稿では AT 法の弱点として通常のランダムノイズに対する認識精度の低下という点に着目して研究を行った．そのために XAI の SHAP を用いてノイズ画像を生成し，ノイズに対して頑健なモデルを作成する手法を提案し，従来の AT 法と提案モデルを組み合わせることでノイズに対して頑健なモデルを作成可能であることを実験で検証した．

今後は SHAP ノイズ画像のデータ数を増やしてそれぞれのデータ数の割合による認識精度の変化を検証する．また，本実験では学習データとテストデータの SHAP ノイズ画像は同アルゴリズムで生成したものであった．今後はアルゴリズムを変えた場合の認識精度の変化を検証する．

参考文献

[1] 恵木正史：XAI(eXplainable AI)技術の研究動向．日本セキュリティ・マネジメント学会誌，Vol. 34, No. 1, pp. 20-27, 2020.
[2] Lundberg, Scott M.; Lee, Su-In.: A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, Vol. 2017, No. 20, pp. 1-10, 2017.
[3] 先崎 佑弥，大畑 幸矢，松浦 幹太：深層学習における Adversarial Training による副作用とその緩和策．コンピュータセキュリティシンポジウム 2017 論文集，Vol. 2017, No. 2, pp. 385-392, 2017.