AI モデル説明 Grad-CAM に対する敵対的攻撃の提案と評価

寶木　隆正　　菊池　浩明

明治大学総合数理学部

1　はじめに

深層学習モデル [2] は画像認識などの分野で高い性能を発揮している一方で, その判断プロセスが不明確であることが課題になっている. この課題に対し, モデルの判断根拠を可視化する説明可能 AI（XAI）が注目されている．例えば,Gradient-weighted Class Activation Mapping（Grad-CAM）[1] は画像分類モデルの可視化に用いられる代表的な XAI である.

しかし,Grad-CAM による可視化結果には不確性があり, しばしば人間の判断基準と一致しないことがある. 加えて, 近年研究が著しい Adversarial Example Attack（AE）敵対的攻撃 [3] により, 人には知覚できない微小なノイズを加えるだけで, 分数モデルの結果を誤らせることが可能であることが知られている. 従って,XAI の不確性を用いた新しい攻撃の潜在性があると考える.

そこで, 本研究は,Grad-CAM による可視化された注目領域は変更することなく, 分離モデルを誤らせることができるかを検証する. 本編で, 既存の AE である FGSM[3] をベースにして,XAI に影響を与えない新しい AE 手法を提案する. 提案方法の効果をオープン画像データセット [5] を用いて実験評価を行い, その攻撃成功率などの評価結果を報告する.

2　準備

2.1　Grad-CAM

Grad-CAM[1] は CNN の最後の畳み込み層における勾配情報を用いて，クラスの判断に画像のどの部分が影響したかを可視化する.

2.2　FGSM

Fast Gradient Sign Method（FGSM）[3] は, ニューラルネットワークの勾配を利用して，入力画像に摂動を加えることで誤分類を引き起こす敵対的攻撃手法である.

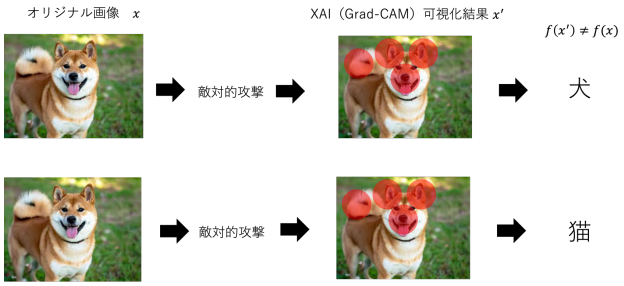


図 1　敵対的攻撃下における Grad-CAM の安全性評価の概念図

標準的な FGSM では，交差エントロピー損失を最大化することで，正解クラスの予測確率を低下させる.

3　提案手法

3.1　攻撃の概要

本研究は, 代表的な画像分類モデル ResNet50[2] を対象として攻撃者が標的分類モデルを知ることができるホワイトボックス攻撃を行う. 提案する攻撃概要を図 1 に図視する．入力画像 $\mathbf{x}$ に対し,FGSM を用いて摂動を加えた敵対的画像 $\mathbf{x}_{\text{adv}}$ は [4] に従って,

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \varepsilon \,\text{sign}\left(\nabla_{\mathbf{x}} L(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y})\right) \tag{1}$$

で算出される. ここで,θ は小さな定数である.

3.2　提案手法

本研究では,Goodfellow らの研究 [3] を基にする．この研究に基づき, 本研究における FGSM の ε の値は, 人間には知覚困難なノイズを生成するように十分小さくする. モデルが出力する予測値 z の順位に注目して攻撃を仕掛ける次の 2 つの損失関数を提案する.

1. FGSM_{L1}：損失関数 $L_1(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) = z_{c2} - z_{c1}$ を (1) 式に基づき最大化する．ここで z_{c1}, z_{c2} はそれぞれ攻撃前の画像を分類したときの予測順位 1 位と 2 位の予測値である.
2. FGSM_{L2} : 損失関数 $L_2(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) = (z_{c2} - z_{c1}) + (z_{c2} - z_{c3})$ を最大化する．z_{c3} は,3 位のラベル.

Evaluation and Proposal of Adversarial Attacks to Explainable AI Grad-CAM,

Ryusei Takaragi and Hiroaki Kikuchi, School of Interdisciplinary Mathematical Science, Meiji University.

表 1　攻撃成功率と XAI モデル説明変化

攻撃手法	ASR (%)	MOD
FGSM_CE (従来)	78.03	0.1101
FGSM_L1 (提案)	80.00	0.099
FGSM_L2 (提案)	74.34	0.086

表 2　分類成否と可視化変化の割合（%）

分類 XAI	成功 不変	成功 変化	失敗 不変	失敗 変化
FGSM_CE (従来)	19.54	2.43	34.80	43.24
FGSM_L1 (提案)	18.38	1.62	42.66	37.34
FGSM_L2 (提案)	24.62	1.04	44.28	30.06

4　実験

4.1　データセット

ResNet50 の学習に使用された ImageNet[6] のサンプリングの画像データセット 1000 枚を使用する．

4.2　実験方法

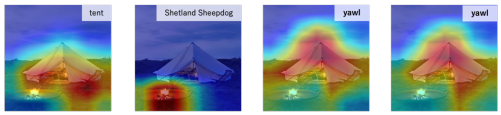
本実験は，Grad-CAM によるモデルの判断根拠の可視化と敵対的攻撃の生成から攻撃を実行する．各画像，各攻撃について，予測クラス，次の各予測指標を評価する．ここで, モデルの 1 位の出力がオリジナルから変更したことを, 攻撃成功と定める. なお, 攻撃前に正しく分類された画像（本実験では 865 枚）のみを対象として評価を行った．

- Attack Success Rate（ASR）：攻撃により 1 位予測が変化する確率.
- Mean Observed Dissimilarity（MOD）[4] : XAI のモデル説明の変化量．1 – SSIM の平均値で定める．ここで，SSIM は，攻撃前後の Grad-CAM で可視化したヒートマップ間の構造的類似度指標 SSIM（Structural Similarity Index Measure）.

4.3　結果

表 1 に従来方式と提案 2 方式の攻撃実験結果を示す．L2 が全手法の中で最も MOD 低い．これは、XAI の注目点を最小限に抑えたまま、誤分類を引き起こしたことを意味する．

表 2 に AI の分類成否と 1−SSIM（閾値 0.1）による可視化結果の変化を示す．表 2 より, 提案方式 L1,L2 は従来方式 CE と比較して、Grad-CAM の可視化結果を維持したまま分類を誤らせる効果が高い（44.28）ことが示された．図 2 に，XAI の説明を変化させずに分類結果だけを誤らせた「テント」の例を示す．実験の結果、提案手法が AI の判断における重要なポイントを操作することで確信度を高く保ったまま誤分類を引き起こしていることを示している．



Original　FGSM_CE　FGSM_L1　FGSM_L2

図 2　攻撃前後の Grad-CAM の比較の例：tent

4.4　考察

表 2 の結果は, ヒートマップの分布がある程度維持された状態で AI が誤分類していることを示している．提案手法は予測順位 2 位のものに誤分類させることのみ狙った手法であるため, ヒートマップの分布に大きな影響を与えず誤分類させることができたと考察する．

5　おわりに

本研究は，予測順位を操作することを目的とした提案方式が，従来攻撃よりも注目点を動かすことなく高い確信度で誤分類を引き起こすことを実証した．今後は，PGD などより強力な反復攻撃への適用や，ResNet 以外への検証などが挙げられる．

参考文献

[1] Selvaraju, R. R., et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” ICCV, pp.4-5,2017.

[2] He, K., et al., “Deep Residual Learning for Image Recognition.,” CVPR, pp. 6-7, 2016.

[3] Goodfellow, I. J., et al. “Explaining and Harnessing Adversarial Examples,” ICLR, pp.2-3,2015.

[4] Chakraborty, S., et al., “Generalizing Adversarial Explanations with Grad-CAM,” CVPRW,pp.3, 2022.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” CVPR, 2009.