敵対的サンプルを用いた胸部 X 線画像分類モデルの説明性向上手法

椎 葉　啓 介　【 知能情報学研究室 】

Explainability of Chest X-ray Image Classification Models Using Adversarial Examples

SHIIBA, Keisuke　【 Intelligent Informatics Lab. 】

1　はじめに

深層学習による医用画像解析の説明性に関して，Grad-CAM よりも説明性のある手法として，中嶋らは CNN に対し CycleGAN や敵対的サンプル（Adversarial Examples）を用い，分類に重要な領域や形状及びパターンの特定で説明性の向上を図っている [1]．敵対的サンプルとは，モデルに誤分類させることを目的として小さな摂動（ノイズ）を画像などに加えたものである．ただし中嶋らは，PGD のみの敵対的サンプルから得られた情報を扱っており，その他の手法は，検証されていなかった．さらに，複数手法の敵対的サンプルで比較した説明性の検討はされていない．そこで本研究は，FGAM，FAB，DeepFool，C&W の手法についても敵対的サンプルを作成し，新たな説明性が得られるか検証する．CNN および ViT を対象とし，敵対的摂動が予測結果に与える影響を比較・分析することで，各モデルが利用している判別根拠を明らかにする．本手法により，モデル構造に依存しない説明性の向上を図り，医用画像分類モデルの信頼性向上に貢献することを目指す．

2　提案手法

本研究では，学習したモデルそれぞれに，FGSM，FAB，DeepFool，C&W の 4 つの手法で敵対的サンプルを生成することで，生成した摂動をもとにモデルがどのような特徴を学習しているのかを明らかにする手法を提案する．また，複数手法の敵対的サンプルを生成し比較することで，手法間の差を明確にしつつ，モデルが学習した特徴の分析を行う．

3　実験

3.1　データセット

本研究では，Kaggle で公開されている Chest X-Ray Images (Pneumonia) を使用した．本データセットは，肺炎と検出なしの胸部 X 線画像で構成されている．さらに，肺炎には細菌性とウイルス性が存在し，「細菌性肺炎（bacteria）」「ウイルス性肺炎（virus）」「検出なし（normal）」の胸部 X 線画像をそれぞれ 1583 枚，1493 枚，1583 枚使用し，合計で 4659 枚を使用した．また，それぞれにおいて train，validation，test が 6:2:2 となるようランダムに分割した．

表 1: モデルの学習結果

モデル	データセット	validation	test
VGG16	bacteria-normal	98.42	96.38
VGG16	virus-normal	95.47	94.50
VGG16	bacteria-virus	72.32	72.67
Vit	bacteria-normal	95.89	94.18
Vit	virus-normal	94.13	93.00
Vit	bacteria-virus	75.00	74.00

3.2　モデルの学習

すべてのモデルにおいて二値分類を行った．「細菌性肺炎と検出なし」,「ウイルス性肺炎と検出なし」,「細菌性肺炎とウイルス性肺炎」の 3 組でそれぞれ VGG16 と Vit で学習を行った．また，一方の画像がもう一方の画像よりも多い場合，少ない方の画像と同数になるよう，ランダムに抽出し使用した．そして，VGG16 と Vit はどの組であっても train，validation，test において全く同じデータセットを使用した．

3.3　敵対的サンプルの生成

本実験では，4 種類の方法で敵対的サンプルを生成した．各条件で学習したモデルそれぞれに FGSM，FAB，DeepFool，C&W を適用した．

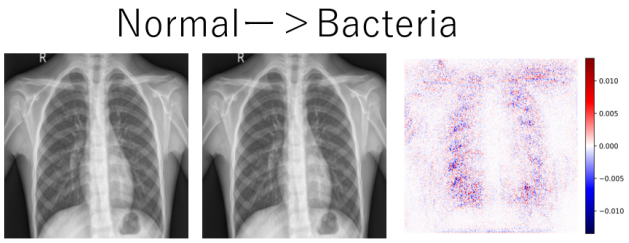
4　結果

4.1　モデルの学習

表 1 に各条件下での学習結果を示す．検出なしと肺炎の胸部 X 線画像を使用したモデルでは，すべてのモデルで validation と test 共に 90%以上の精度を示し，Vit よりも VGG16 の方がやや精度が高い．また，細菌性肺炎とウイルス性肺炎の胸部 X 線画像を使用したモデルでは，どちらのモデルにおいても validation と test で 70%から 75%の精度を示し，VGG16 よりも Vit の方がやや精度が高い．

表 2: 元画像と敵対的サンプル画像の精度比較

モデル	Adv	orig_acc	adv_acc
VGG16(b-n)	FGSM	96.38	7.55
VGG16(b-n)	FAB	96.38	96.38
VGG16(b-n)	DeepFool	96.38	3.62
VGG16(b-n)	C&W	96.38	49.69
Vit(v-n)	FGSM	93.00	2.33
Vit(v-n)	FAB	93.00	45.67
Vit(v-n)	DeepFool	93.00	7.00
Vit(v-n)	C&W	93.00	49.67
Vit(b-v)	FGSM	74.00	0.83
Vit(b-v)	FAB	74.00	47.17
Vit(b-v)	DeepFool	74.00	26.00
Vit(b-v)	C&W	74.00	47.33



(a)元画像　(b)敵対的サンプル　(c)摂動

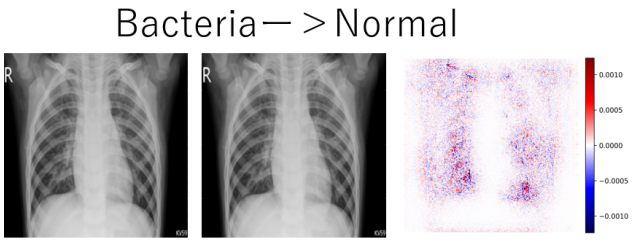
図 1: VGG16(b-n) DeepFool の N から B への変換

4.2　敵対的サンプルの生成

表 2 に各条件下での元画像と敵対的サンプル画像の精度例を示す．左列から，モデル（データセット），敵対的サンプル，元画像の正解率，敵対的サンプル画像の正解率を表す．生成した敵対的サンプルの精度において，最も精度が低いのは Vit の「細菌性肺炎とウイルス性肺炎」における FGSM の 0.83%で，最も精度が下がったのは VGG16 の「細菌性肺炎と検出なし」における DeepFool の 92.76%であった．敵対的サンプルの例として，図 1 および図 2 に示す．また，VGG16 の FAB では，どのデータセットでも精度の変化はなかった．

4.3　摂動の分析

今回最も精度が落ちた VGG16 の「細菌性肺炎と検出なし」における DeepFool において，求めた摂動がモデルの学習したどのような特徴に攻撃を加えているのか確認する．具体的には，モデルの各中間層を取得し，元画像と敵対的サンプル画像を入力しそれらに対する中間表現の差分を取る．ラベル別の平均を求めると，いずれのラベルにおいても最後の中間層である block5 の層で最も差が大きく，bacteria では 2.0044×10^{-1}，normal では 1.7853×10^{-1} となった．



(a)元画像　(b)敵対的サンプル　(c)摂動

図 2: VGG16(b-n) DeepFool の B から N への変換

5　考察

DeepFool では，どのモデルや画像においても，元画像の予測ラベルと敵対的サンプルの予測ラベルが反転した．図 1 および図 2 では肺の背骨寄りの領域に強い摂動を加えていることがわかる．VGG16 は、浅い層において一般的なエッジやテクスチャを学習する一方で、深い層においては肺炎に関連する意味的特徴表現に強く依存する。敵対的摂動はこれら深層表現を集中的に変化させており、モデルの判別が高次概念に基づいていると考える。

FGSM において，ほとんどの場合で精度が 10%以下へと変化した．これらは勾配情報を利用し効率的に決定境界の外側へ誘導するため，モデルが勾配方向に強く依存した特徴を使っていると考える．FAB においては，VGG16 では精度の変化がなかったが，Vit では精度が 50%以下になった．そのため，VGG の勾配が FAB の探索アルゴリズムと相性が悪い可能性があると考える．

モデルの構造に着目すると，敵対的サンプルでの精度は，C&W 攻撃では VGG16 の方が Vit よりも精度が低く，逆に FGSM では Vit の方が VGG16 よりも精度が低かった．ViT は広域的な特徴を捉えるのが得意であるので，入力画像全体に散布される微細なノイズに対し，予測の根拠となる「Attention」機構が乱されやすい性質を持つ可能性があると考える．

6　まとめ

本研究では，複数の敵対的サンプルを生成し，生成した摂動をもとにモデルがどのような特徴を学習しているのかを明らかにする手法を提案した．また，モデルの中間層を取り出し元画像と敵対的サンプル画像の表現差を確認することで，モデルが肺炎に関してエッジやテクスチャよりも意味的特徴を学習していることを明らかにした．

参考文献

[1] T. Nakajima, et al.,“Explainability of CNN Classification Models Using CycleGAN and Their Application to Medical Imaging,” in *Computational Intelligence and Industrial Applications*, 2025.