敵対的手法による医用画像分類モデルの説明性に関する研究

1275107　中嶋 大雅　【 知能情報学研究室 】

A Study on Explainability of Medical Image Classification Models Using Adversarial Techniques

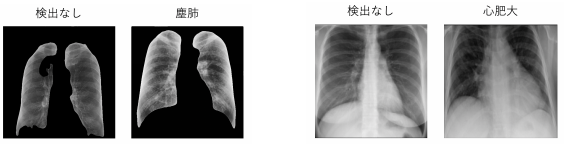
1275107　NAKAJIMA, Taiga　【 Intelligent Informatics Lab. 】

1　はじめに

医師不足や患者数の増加により, 医療現場の負担は深刻化している. この問題の解決策として, 医用画像診断支援システム (CAD) の活用が注目されている [1]. 特に, Convolutional Neural Network（CNN）を用いた画像認識が広く採用されており, 病気を高精度に識別することが出来ている. しかし, ニューラルネットのブラックボックス性が課題となり, クラス活性化マップ (CAM) 等の説明手法の活用が試みられているものの十分な解決にはなっていない. 本研究では, 敵対的サンプル [2] を活用した新たな説明性向上手法を提案する. 本手法では, 敵対的サンプルを用いて CNN の分類に寄与する領域やその形状・パターンを特定することで, 説明性の向上を目指す. また, 既存の説明性手法との比較を行い, 提案手法の有効性を検証する. 実験には, 塵肺および心肥大のラベル付き胸部 X 線画像を対象とした CNN モデルを用いる. ここで, 敵対的サンプルとは機械学習モデルに誤った予測をさせるために, 意図的に小さな摂動 (ノイズ) を持たせた画像などのことである. また, CycleGAN は, ペア画像がない非対整データを用いて, 一つの画像スタイルを別のスタイルへ変換するための深層学習モデルである. 塵肺症を例に挙げると, 「塵肺」ならば「検出なし」に, 「検出なし」ならば「塵肺」に変換が可能である.

2　提案手法

本研究では, CNN が認識する分類に寄与する領域とその領域内の形状やパターンの違いの獲得を目指して敵対的サンプルを用いた手法を提案する. 本実験では敵対的サンプルを作成するにあたり, 敵対的攻撃手法として Projected Gradient Descent(PGD, 射影勾配降下) を採用する. PGD は画像に複数回摂動を加える手法であり, 他の手法と比較してより強力な攻撃が可能である. PGD の数式を式 (1) に示す. 作成された敵対的サンプルの結果が分類に寄与する領域とその中のパターンや形状を反映していると仮定し, 敵対的サンプルを解析する. 式 (1) の漸化式により，敵対的サンプル x^{n+1} がサンプル x^n から生成される. これを，入力画像 x^1 から始める.



(a) 塵肺データセット　　**(b)** 心肥大データセット

図 1: データセットの構成

$$x^{n+1} = P(x^n + \alpha * \mathsf{sign}(\nabla_{x^n} J(\theta, x^n, y))) \quad (1)$$

θ: パラメータ, x^n: 入力, y: ラベル, J: 損失関数
α: 摂動の大きさ, sign: 勾配の符号, P: 範囲の制限

3　実験内容

3.1　データセット

本研究では塵肺と心肥大の 2 種類の胸部 X 線画像データセットを用いる.

3.1.1　塵肺データセット

塵肺は, 長期間にわたる粉塵や微粒子の吸入・蓄積によって引き起こされる肺疾患である. 用いる画像数は 237 枚であり, クラスは「塵肺」と「検出なし」の 2 クラスである. また, 塵肺画像の分類精度を向上させるためにデータセットに対して U-Net を用いたセグメンテーションによる肺野領域抽出を実施した.

3.2　心肥大データセット

心肥大の学習には 718 枚の胸部 X 線画像を使用する. クラスは「心肥大」と「検出なし」の 2 クラスである.

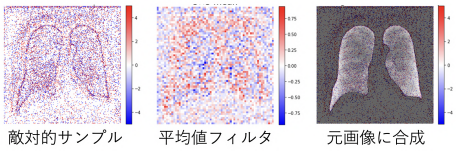
3.3　実験手順

塵肺と心肥大を学習したそれぞれの分類モデルに対して, 誤分類を引き起こすような敵対的サンプルを作成し, 傾向を分析する. 作成した敵対的サンプルと Grad-CAM, Guided Backpropagation, CycleGAN の注意マップを比較して, 注目領域や領域内の形状やパターンの違いが得られるかを検証する.

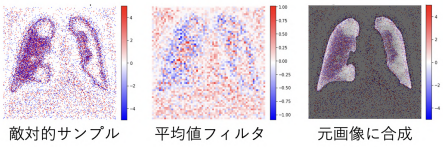
4　結果・考察

4.1　分類モデルの学習結果

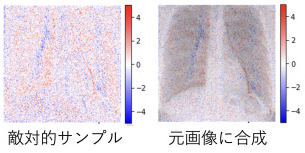
塵肺の分類精度は 97.14%, 心肥大の分類精度は 97.22% となった.



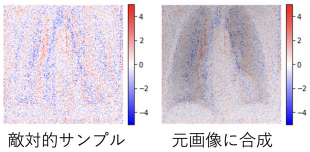
(a) 検出なし



(b) 塵肺

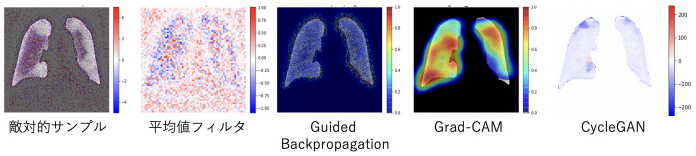


(c) 検出なし

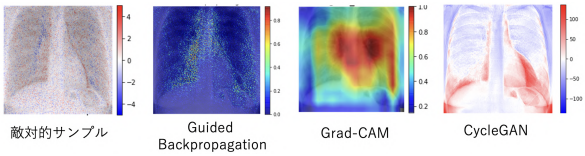


(d) 心肥大

図 2: 敵対的サンプルの適用



(a) 塵肺



(b) 心肥大

図 3: 敵対的サンプルと既存の説明性手法との比較

4.2　敵対的サンプルの生成

学習済みの分類モデルを用いて塵肺と心肥大それぞれの敵対的サンプルを作成した.

4.2.1　塵肺への適用

図 2(a) より, 検出なし画像から生成された敵対的サンプルでは肺野領域の境界の外側の領域の画素値を増加させる傾向が見られた. 一方で, 図 2(b) より, 塵肺画像から生成された敵対的サンプルは肺野領域の境界線の内側の領域の画素値を減少させる傾向が見られた. 二つの結果から, 敵対的サンプルは塵肺かそうでないかを肺野領域の境界の輝度の大きさと差で判断していると考えられる.

4.2.2　心肥大への適用

図 2(c) より, 検出なし画像から生成された敵対的サンプルでは心臓部分の側面の画素値を増加させる傾向が見られた. 一方で, 図 2(d) より, 心肥大画像から生成された敵対的サンプルは心臓部分の側面の画素値を減少させる傾向が見られた. 二つの結果から, CNN は心肥大かどうかを心臓部分の側面の領域の広さで識別していることが示唆された.

4.3　既存の説明性手法との比較

敵対的サンプルによる注目領域と Grad-CAM, Guided Backpropagation による注目領域は大まかに一致し, 注目領域に一貫性が認められた. さらに, 敵対的サンプルの解析結果から, 塵肺では肺野境界の画素値が低いと「検出なし」, 高いと「塵肺」と分類される傾向が, 心肥大では心臓側面の画素値が低いと「検出なし」, 高いと「心肥大」と分類される傾向が明らかになった. これにより, Grad-CAM や Guided Backpropagation が注目領域の特定に留まるのに対し, 敵対的サンプルは分類に寄与する具体的な特徴を詳細に解析できることが示された. また, CycleGAN と比較した場合, CycleGAN も分類に必要な特徴を視覚化できるものの, 変換後の画像に変換前の画像とは逆のラベルを付与してモデルに予測させた際の精度は塵肺で 42.85%, 心肥大で 77.02%に留まった. 一方, 敵対的サンプルを加えた画像は, その性質上, 本来とは逆のクラスに確実に誤分類させることが可能であり, 予測確率も平均 99.9%と高い精度で判定を反転させている. 以上の結果から, 敵対的サンプルを用いたアプローチは, 変換された画像の判定が変化した上でその差を見ていることから, CycleGAN よりも高い信頼性を持つ説明になると考えている.

5　おわりに

本研究では, 敵対的サンプルを活用して CNN 分類モデルの説明性を向上させる手法を提案した. 提案手法は, 従来手法と比較して CNN の注目領域だけでなく, その領域内における分類に寄与する詳細な特徴を視覚化することに成功した. 実験の結果, 塵肺では肺野領域の境界周辺, 心肥大では心臓部分の側面領域が分類に高く寄与していることが明らかとなった. さらに, 敵対的サンプルから得られる特徴が分類に寄与していることを示す根拠の信頼性も高いことが確認された.

参考文献

[1] Gao, Jun, et al. “Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: An overview.” *Math. Biosci. and Eng.* 16.6 (2019): 6536-6561.

[2] Goodfellow et al. “Explaining and harnessing adversarial examples.” *arXiv:1412.6572* (2014).