GAN を用いた CNN の分類結果の説明性に関する研究

1245127　筒井 康行　【 知能情報学研究室 】

A Study on Explainability of Convolutional Neural Networks Using Generative Adversarial Networks

1245127　Yasuyuki Tsutsui　【 Intelligent Informatics Laboratory 】

1　はじめに

近年，CAD(Computer Aided Diagnosis) は重要な研究分野の 1 つとなっており，Convolutional Neural Network(CNN) を用いた画像認識による診断が広く使われる．一方で，医学では判定結果の理由も重要である．これまで CNN の分類過程の解析には，分類に寄与する領域を特定する Class Activation Mapping(CAM) ベースの手法が用いられてきた．しかし結果の説明には分類に寄与する領域だけでなく，領域内の形状やパターンの違いも必要になると考えられる．

そこで本研究では，データ生成の学習からデータ分布の獲得が期待できる Generative Adversarial Network (GAN) を用いて，分類に寄与する領域と領域内の形状やパターンの違いを取得する分析方法を提案する．本研究では GAN モデルとして CycleGAN に注意機構を導入した Attention-Guided CycleGAN(AG-CycleGAN) を用い，得られる注意マップと変換結果から分類に寄与する領域と領域内の形状やパターンの違いの獲得を目指す．変換結果，Grad-CAM++，学習済み CNN を用いて分析結果を比較し，妥当性を検証する．

2　関連研究

本研究で用いる AG-CycleGAN と CNN の説明性に広く使われる CAM ベースの手法について説明する．

2.1　AG-CycleGAN

AG-CycleGAN[1] は Mejjati らによって提案され，生成モデル G_x, G_y，識別モデル D_x, D_y，注意モデル A_x, A_y の計 6 つのニューラルネットワークモデルから構成される．また生成モデルと注意モデルによる変換機構を変換モジュール M_x, M_y と表す．AG-CycleGAN では，教師なし学習でデータドメイン X, Y 間における相互変換 $M_x : Y \to X, M_y : X \to Y$ を学習する．また，変換過程で注意モデルは生成モデルによる変換の領域を制限する $[0, 1]$ の値を持つ注意マップを出力し，生成モデルとともに識別モデルが区別できないデータの生成を学習する．本研究では変換領域の制限のために注意モデルから出力される注意マップがドメイン間の違いとなる領域を選択していると捉えられることに注目し，

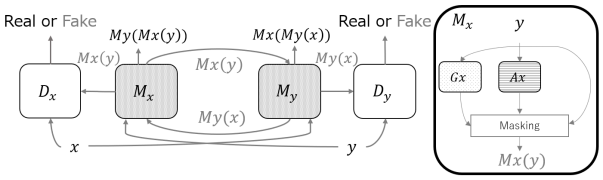


図 1: AG-CycleGAN 上でのデータ遷移

注意マップを変換に寄与する領域の分析に用い，生成モデルによる変換結果を領域内の形状やパターンの違いとして用いる．Mejjati らが提案した AG-CycleGAN の損失関数は Adversarial Loss と Cycle Consistency Loss によって構成される．Adversarial Loss は一方のデータをもう一方のドメインのデータに近づける学習に寄与し，Cycle Consistency Loss は変換後のデータを元のドメインに変換した際に元のデータに戻る様な一貫性のある変換の学習に寄与する．M_x では，G_x による変換結果を A_x から得られる注意マップによって制限する．式 (1) は，M_x による変換を表し，式中の演算子 $\odot$ はアダマール積を表す．M_y についても同様である．図 1 は AG-CycleGAN 上でのデータ遷移を表したものである．

$$M_x(y) = A_x(y) \odot G_x(y) + (1 - A_x(y)) \odot y \qquad (1)$$

2.2　CAM ベースの手法

CAM ベースの手法では，学習済みモデルにおける最終の畳み込み層の特徴マップと出力層の重みや勾配から分類に寄与する領域を表す顕著性マップを算出する．しかし，CNN モデルでは物体の位置に対するロバスト性を確保するためにモデル内部で特徴マップの縮小が行われ，高次の層では特徴が持つ空間情報が失われる．そのため，CAM ベースの手法から得られる顕著性マップは詳細な領域を選択することができない．また，得られる情報が領域のみでモデルが認識する具体的なクラス間の違いを得ることが困難であり，それらが CNN の説明可能性における問題として挙げられる．

3　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領域内の形状やパターンの違いの獲得を目指して AG-

CycleGAN を用いた分析手法を提案する．分析手順としては，まず CNN と共通の特徴マップを用いた変換を行うために共通のバックボーンを用いた AG-CycleGAN による相互変換を学習する．そして得られた注意マップを分類に寄与する領域，変換結果を領域内の形状やパターンの違いと捉え，手法の有効性を評価するために変換結果の評価，注意マップと CAM ベース手法の比較，学習済み CNN を用いた変換前後での CNN の分類結果の違いを検証する．

4　実験

本稿では，NIH Clinical Center が提供する胸部 X 線画像データセット [2] における心肥大と検出なし間での提案手法の検証ついて記述する．また本研究では，CAM ベースの手法との比較に Grad-CAM++[3] を用いた．

5　実験結果

CNN に対して提案手法を検証した結果を記述する．

5.1　AG-CycleGAN による変換結果

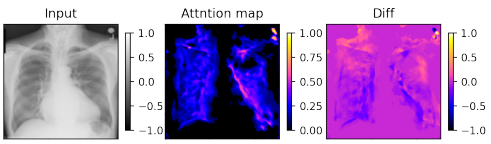
AG-CycleGAN による心肥大と検出なし間の相互変換結果として図 2 が得られた．図 2 の Attention map(注意マップ) では心臓の外側が分類に寄与する領域として得られ，変換画像から入力画像を引いた差分 (Diff) を見ると図 2(a) の心肥大から検出なしへの変換では心臓を小さく，図 2(b) の検出なしから心肥大への変換では心臓を大きくするような心肥大の症状に沿う領域内の形状やパターンの違いが得られた．

5.2　注意マップと Grad-CAM++の比較

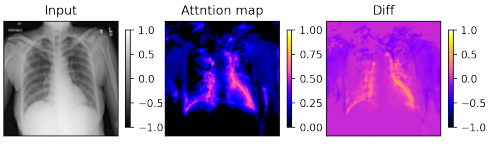
学習済み CNN から Grad-CAM++，AG-CycleGAN から注意マップを出力させた結果として図 3 に示す結果が得られた．結果より，Attention map(注意マップ) は Grad-CAM++と比べてより詳細な領域が得られた．

5.3　変換前後での CNN の分類結果の違い

学習済み CNN を用いて変換前後データセット間の分類結果の違いを検証した結果として，表 1 の結果が得られた．結果より，変換後のデータが CNN にもう一方



(a) 心肥大 (Cardiomegaly) → 検出なし (No Finding)



(b) 検出なし (No Finding) → 心肥大 (Cardiomegaly)

図 2: 相互変換の結果

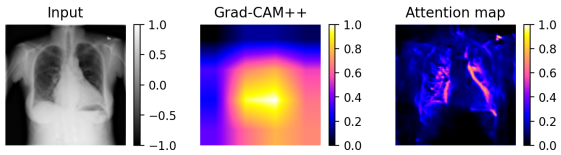


図 3: Grad-CAM++と注意マップの比較

表 1: 変換前後の混同行列

		分類結果			
		変換前		変換後	
		検出なし	心肥大	検出なし	心肥大
ラベル	検出なし	8682	3391	5108	6965
	心肥大	152	404	340	216

のクラスと認識される傾向が見られ，CNN と GAN で共通の違いを認識した可能性が高いと考えられる．

6　考察

本研究では変換の有効性の評価に変換前後の混同行列を用いたが，変換後も元クラスと認識されるデータが見られた．これらにおいて AG-CycleGAN による変換と Backbone に用いた CNN による分類で苦手とするデータが一致する可能性がある．また混同行列の傾向の変化で変換の有効性を見ているが，変化が小さい場合に変換が有効かを判断できない可能性がある．

7　まとめ

本研究では CNN の説明可能性の向上を目指し，GAN を用いた分析手法の提案，検証を行った．その結果として，Grad-CAM++と比べ詳細な領域が得られ，変換前後の差分から領域内の形状やパターンの違いを確認できた．しかし変換が有効に働かないデータもあり，今後は分析手法としての妥当性の判断の為に CNN と GAN で苦手とするデータ傾向の類似性の検証，変換の有効性について定量的な判断基準が必要である．

参考文献

[1] Youssef Alami Mejjati, et al. Unsupervised attention-guided image-to-image translation. In *NeurIPS2018*, 2018.

[2] Xiaosong Wang., et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR2017*, 2017.

[3] Aditya Chattopadhay, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE WACV2018*, 2018.