AI 技術を活用した検査工程の省力化・効率化（第９報）

－AI の説明可能性の検討－

渡辺博己*、松原早苗*、内野義友輝*

A study on artificial intelligence for labor savings or efficiency improvements of inspection process（Ⅸ）

- Consideration of explainable artificial intelligence -

WATANABE Hiroki*, MATSUBARA Sanae* and UCHINO Yoshiyuki*

本研究では、深層学習モデルの推論過程の説明可能性を向上させるために、説明可能な人工知能（XAI）の一つの手法である Grad-CAM を使用し、欠陥画像分類モデルの予測結果に対する判断根拠を可視化する実験を行った。その結果、Grad-CAM による可視化画像を利用することで、モデルの推論過程が説明可能となり、モデルの信頼性向上や今後の学習データの選定、モデルの品質向上に寄与する可能性が示唆された。

1　はじめに

近年、深層学習により生成されたモデルの推論精度が高いことから、深層学習モデルが社会生活の様々な場面で使用されるようになりつつある。深層学習モデルが高い推論精度を持つ理由の一つとして、推論する上で非常に重要な特徴量を抽出する仕組みをモデルが獲得していることが挙げられる。しかし、このようなモデルの生成過程はブラックボックス性を有しており、モデルの結論や意思決定プロセスを明示的かつ理解可能にすることが課題となっている。

そこで、本研究では、深層学習モデルの予測結果に対する推論過程について、説明可能な人工知能（XAI）を用いた可視化を検討した。XAI は、人工知能モデルがなぜ特定の予測や意思決定を行ったのかを説明するための技術で、人工知能モデルの動作を理解しやすくすることを目的としている。本稿では、XAI の一つの手法である Grad-CAM[1]を用いて、欠陥画像分類モデル [2]の予測に対する判断根拠を可視化する実験を行ったので、その内容について報告する。

2　判断根拠の可視化

Grad-CAM は、画像分類や物体検出などのタスクで訓練されたモデルに適用される XAI で、モデルの最後の畳み込み層から出力された特徴量をもとに、入力画像のどの部分が予測に寄与しているかを可視化することが可能である。図 1 に Grad-CAM の概要を示す。

Grad-CAM による判断根拠の可視化画像は、以下の手順により生成される。

① モデルに画像を入力し、特定の層（通常は、最終の畳み込み層）の出力（特徴マップ）A^kを取得する。

② 各クラスに対するクラススコア（確信度）yを取得する。

③ バックプロパゲーションを実行し、特定のクラススコアy^cに関連する勾配を計算する。

＊ 情報技術部

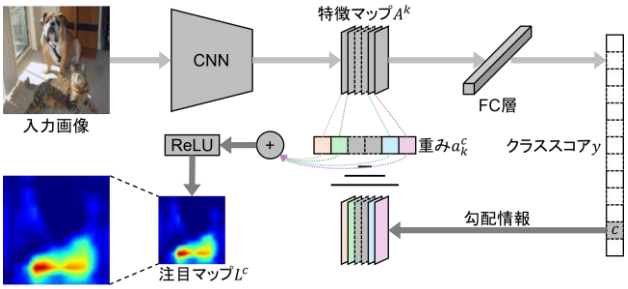

図 1　Grad-CAM の概要

④ 特定のクラスに対して、どの特徴がより影響を与えたかを示す重みa_k^cを、式(1)により特徴マップごとに計算する。

$$a_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

なお、Zは特徴マップのサイズである。

⑤ 重みa_k^cを特徴マップA^kに掛けて合成し、画像内のどの部分がモデルの判断に最も影響を与えたかを示す注目マップ（ヒートマップ）L^cを、式(2)により生成する。

$$L^c = \mathrm{ReLU}\left(\sum_k a_k^c\, A^k\right) \tag{2}$$

なお、ReLU 関数は負要素を 0 とするために適用される。

⑥ 注目マップL^cを入力画像の大きさにリサイズし、入力画像に重ね合わせた画像（可視化画像）を生成する。

3　可視化画像の生成実験

実験では、欠陥画像分類モデルに Grad-CAM を適用し、図 2 に示す欠陥画像データセット [2]における各クラスの検証データについて、どのクラスに該当するかを予測するとともに、正解クラスに対する判断根拠の可視化画像を生成した。図 3 に入力画像（左）と生成した可視化画

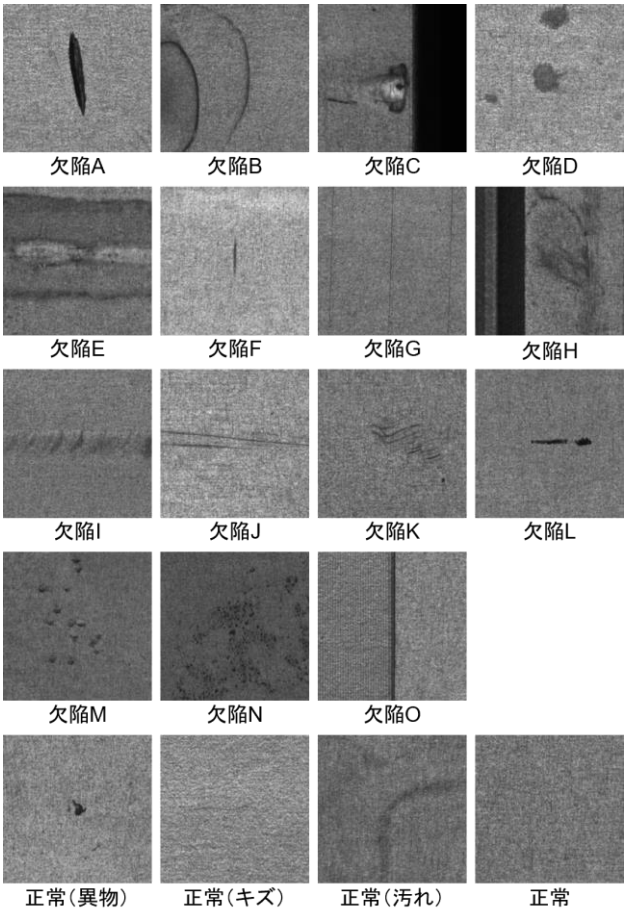

図２　欠陥画像データセットの画像例

像（右）の例を示す。可視化画像中の色は、寄与が高い画素ほど赤色が濃くなり、低い画素ほど青色が濃くなっている。

図 3(a)、(c)は予測が正しかった時の例であり、欠陥領域の画素の予測への寄与が高いことが、可視化画像から判断できる。一方で、予測に誤りが生じたのは、図 3(b)のように、欠陥領域以外の画素の寄与が高くなったり、図 3(d)のように、寄与の高い領域が分断され、欠陥領域の画素が予測に十分に寄与していなかったりした場合などであった。

これらの結果から、欠陥画像分類モデルの正誤に至った推論過程は、Grad-CAM により生成した判断根拠の可視化画像を利用することで、十分に説明可能であることが示唆された。

4　まとめ

本研究では、XAI を用いて、欠陥画像分類モデルの予測結果に対する推論過程の可視化を試みた。可視化にあたっては、画像分類タスクなどの判断根拠をヒートマップとして表示可能な Grad-CAM を利用した。その結果、欠陥画像分類モデルが正誤に至った推論過程を容易に説明することが可能となり、モデルの信頼性向上に貢献するだけでなく、今後の学習データの選定やモデルの品質

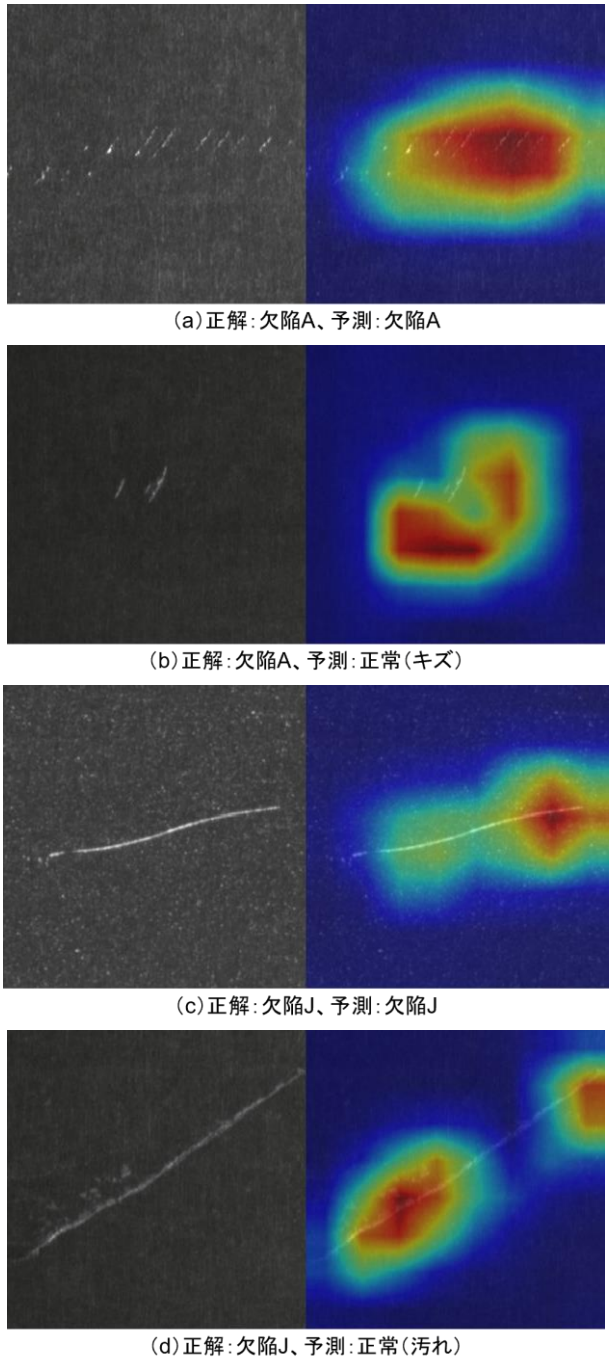

図３　入力画像(左)と可視化画像(右)の例

向上に有用であると考えられる。

【謝　　辞】

本研究を遂行するにあたり、欠陥画像データセットをご提供いただきました株式会社前田精工の皆様に深く感謝の意を表します。

【参考文献】

1)　R. R. Selvaraju, et al., arXiv: 1610.02391, 2016
2)　渡辺ら，岐阜県産業技術総合センター研究報告 No.4, pp69-72, 2023