画像認識における説明可能なAIに関する研究

全　　慶樹、近藤　正一、本間　稔規

Research on Explainable AI in Image Recognition

Keiki ZEN, Shouichi KONDOU, Toshinori HONMA

抄　録

近年、深層学習をはじめとする人工知能（AI）技術が急速に発展し、様々な分野でAIを活用した研究が進められている。しかし、近年のAIは膨大な量のパラメータから構成される巨大なモデルであり、モデルの内部構造が極めて複雑なため、なぜその認識や予測の結果が得られたのかという根拠を明確に説明できないという課題を抱えている。このような背景からAIの予測根拠を人間が解釈可能な形で説明するための技術である説明可能なAI（XAI: Explainable AI）に関する研究が注目されており、道内においてもAIを利用する企業から画像認識AIの予測の根拠に関する相談が当場に寄せられている。そこで本研究では、農作物の画像から良品と不良品を判別するAIモデルに対して画像認識AIの予測根拠を可視化して説明する、説明可能なAIの最新手法を適用することでその有用性を検証したので報告する。

キーワード：説明可能なAI、画像認識、可視化、概念抽出

Abstract

In recent years, artificial intelligence (AI) technologies, including deep learning, have developed rapidly, and research utilizing AI has been advancing across a wide range of fields. However, modern AI models are enormous systems composed of vast numbers of parameters, and their internal structures are extremely complex, which makes it difficult to clearly explain the rationale behind their recognition or prediction results. Against this background, research on explainable AI (XAI)—a technology that enables AI predictions to be explained in a human-interpretable manner—has been attracting increasing attention. In Hokkaido as well, our institute has received inquiries from companies using AI regarding the rationale behind predictions made by image recognition systems. In this study, we report on the usefulness of explainable AI by applying a state-of-the-art visualization method to elucidate the basis of image recognition AI predictions in a model that classifies agricultural products as either good or defective.

KEY-WORDS : Explainable AI, Image recognition, Visualization, Concept extraction

1. はじめに

近年、深層学習をはじめとする人工知能（AI）技術が急速に発展し、AIを活用した自動化等の研究が様々な分野で進められている。しかし、近年のAIは膨大な量のパラメータから構成される巨大なモデルであり、モデルの内部構造が極めて複雑なため、なぜその認識や予測の結果が得られたのかという根拠を明確に説明することができない。これは「ブラックボックス問題」と呼ばれており、AIの信頼性の評価を困難にすることから、企業や自治体におけるAIの導入を妨げる要因になっている。

AIの開発では、収集したデータをAIに学習させてモデルを構築する。この学習プロセスでは、認識や予測の対象に関する「特徴」を入力データから自動的に学習するが、開発者の意図していない特徴を学習してしまうことがあり、その場合、期待した結果を出力しないなどAIモデルの性能は著し

事業名：経常研究
課題名：画像認識における説明可能なAIに関する研究

く低下する。このような現象は、特に実運用現場で取得されるデータにはない手がかりとなる特徴が学習用データに存在する場合に発生しやすい。例えば、農作物の良品と不良品を画像から判別するAIの開発において、学習用データの撮影に良品と不良品で異なる撮影台を用いるとAIは撮影台の違いを学習してしまい、農作物の一部が腐敗しているかなどの良品と不良品を判別するために必要な特徴を学習しない可能性がある。この場合、実運用現場では学習時の撮影台の情報を利用できないためAIの判別性能が極めて低くなる。AIの予測の根拠を説明できない場合、開発プロセスにおいてこのようなモデル内部に潜む問題を発見することは容易ではない。

このような背景からAIの予測根拠を人間が解釈可能な形で説明するための技術である、説明可能なAI（XAI: Explainable AI）に関する研究が注目されている。XAIは、AIが意図どおりに機能しているかを確認する上で重要な技術であり、認識や予測の妥当性の確認や誤判断の原因特定による性能改善など、AIの信頼性向上を可能にする。XAIは、開発者にとって有用であるだけでなく、AIの導入にあたってそのリスクと効果のより正確な評価を可能にすることから、企業等がAIを自社の業務に導入する際の意思決定支援にも有用である。ここ数年、AIを利用する企業から画像認識AIの予測の根拠に関する相談が当場に寄せられているなど、AIの認識や予測の根拠を説明する技術開発の重要性が高まっている。

そこで本研究では、農作物の画像から良品と不良品を判別するAIモデルに対し、画像認識AIの予測根拠を可視化して説明する最新手法を適用し、その有用性を検証したので報告する。

本稿では、まず説明可能なAIの概要と本研究で使用したアルゴリズムの詳細について述べる。次に当該アルゴリズムの実装および具体的なAIモデルへの適用事例として、ブロッコリーの良品と不良品を判別するAIモデルへアルゴリズムを適用した結果について述べる。最後に得られた予測根拠にもとづくAIモデルの妥当性評価について述べ、本研究で使用したアルゴリズムの有用性についてまとめる。

2. 画像認識モデルの予測根拠の可視化手法

本章では、まず説明可能なAIの概要と画像認識モデルの予測根拠を可視化する一般的な手法について述べる。次に本研究で使用した可視化アルゴリズムであるCRAFT[1]（Concept Recursive Activation Factorization for explainability）の仕組みと特徴について説明する。

2.1 説明可能なAI

説明可能なAIは、AIモデルが出力した認識や予測の根拠を人間が理解できる形で提示する技術であり、AIシステムの透明性や信頼性の向上において重要な役割を果たしている。画像認識モデルに対して説明可能なAIを適用する場合、予測根拠を視覚的に可視化して提示する手法が広く知られており、大きく分けて二種類のアプローチが存在する。

一つは、AIモデルが予測の際に画像内のどの領域を特に重視したかを示す手法であり、通常はヒートマップ等による表現を用いてモデルが着目したピクセルレベルの情報を視覚的に提示する。代表的な手法としてGrad-CAM[2]（Gradient-weighted Class Activation Mapping）があり、これは畳み込みニューラルネットワーク（CNN: Convolutional Neural Network）の最後の畳み込み層の勾配情報を用いることで、特定のクラス予測に最も影響を与えた画像領域をヒートマップとして可視化するアルゴリズムである。図1にGrad-CAMによる可視化の例を示す。図1左の入力画像に対してAIモデルが犬と認識した際に重視したと考えられる画像領域を図1右のヒートマップの赤い領域として可視化している。

また、モデルの構造に依存せずに入力の重要度を推定できるLIME[3]やSHAP[4]のような汎用的手法を画像認識モデルに適用することも可能であり、Grad-CAMと同様にクラスの予測に最も影響を与えた画像領域を可視化することができる。

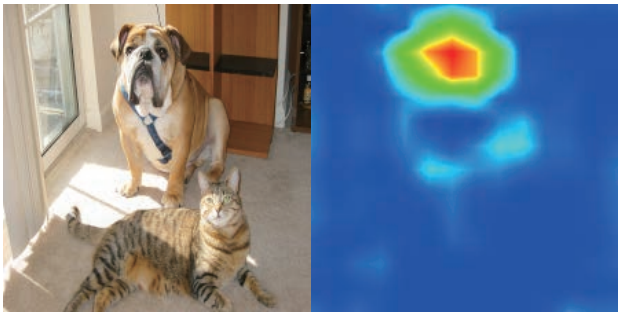

図1　Grad-CAMによる可視化

もう一つのアプローチは、モデルが予測の際に利用した「概念」（concept）を示す手法である。ここでの概念とは、モデルが学習により獲得した特定の視覚的パターンや特徴の集合である。概念にもとづく手法では、これらの概念が予測にどの程度寄与したかを示すことで、画像の「どこを重視したのか」ではなく「何を重視したのか」という観点からモデルの予測根拠を説明する。TCAV[5]等の初期の手法では、開発者があらかじめ定義した概念（例えば「縞模様」や「尖った耳」等の特徴を持つ画像の集合）を用いて、特定の概念が予測にどの程度寄与しているかを計算するが、次節で説明するCRAFTのような新しい手法では、モデルの内部構造から自動的に概念を抽出することが可能である。これにより開発者が想定していない潜在的な概念の発見とそれにもとづく説明が可能になった。

2.2 概念に基づく可視化手法

本研究では、画像認識モデルの予測根拠を可視化する手法としてFelらによって提案されたCRAFTを使用した。CRAFTは、モデルが予測の際に重視した画像内の領域と概念を同時に提示できる手法であり、従来の二種類のアプローチを統合することにより詳細な可視化を実現している。

CRAFTの基本的な考え方は、深層学習モデル内部の活性化された中間層の出力を分析し、それらを人間が理解しやすい概念へと分解することにある。具体的には、画像認識モデル内部の最後の畳み込み層などの出力に対して多変量解析手法の一種である非負値行列因子分解（NMF: Non-negative Matrix Factorization）を適用することで、モデルの内部に構築されている潜在的な概念を自動的に抽出する。このプロセスは再帰的に実行されるため、抽出された概念はさらに細分化され、階層的な構造を持つ。例えば、動物の画像を分類するモデルにおいて「動物」という上位概念は、「犬」「猫」といった中位概念、さらに「犬の耳」「猫のひげ」といった下位概念へと分解できる。このような階層的概念の利用は、モデルが予測において重視した視覚的特徴に関する正確な理解に有用である。

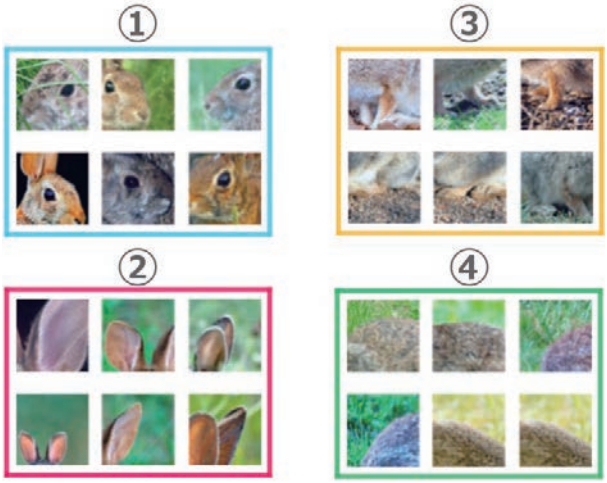

図 2　CRAFTにより抽出された概念
（予測に重要なものから順に①～④）

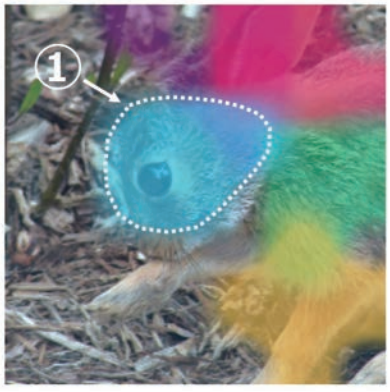

図 3　予測において各概念が利用された画像領域

抽出された概念は、それぞれの概念に強く反応する代表的な画像の集合として可視化される。これにより数値的な指標ではなく、視覚的な情報からモデルが何を重視しているのかを把握することができる。また、CRAFTはどの概念が予測にどの程度寄与したのかを定量的に示すことができる。これらの寄与度の情報を用いることで、自動的に抽出された複数の概念から重要なものを選択的に提示することが可能になる。さらにそれぞれの概念に関してモデルが着目した画像領域を提示することが可能であり、以上をまとめると「予測で重要なこれらの概念を画像のこの領域で利用した」という詳細な説明を得ることができる。

図 2 および図 3 にCRAFTによる可視化の例を示す。画像認識モデルが入力画像をウサギと認識した結果を分析した例であり、図 2 では抽出された概念のうち予測に重要なものを 4 つ示し、図 3 では各概念が入力画像のどの領域で利用されたかをそれぞれ示している。最も重要とされた①の概念は、ウサギの顔の概念と考えられ、ウサギの認識に利用することは妥当であると考えられる。また、①の概念は図 3 においてウサギの顔の領域で利用されており、この点からもこの画像認識モデルの妥当性を確認することができる。

本研究では、Pythonの深層学習フレームワークであるTensorFlowと説明可能なAIのライブラリであるXplique[6]を使用してAIモデルへCRAFTを適用するプログラムを実装し、農作物の良品と不良品を判別する画像認識モデルに対して適用した。

3. 農作物不良品判別モデルへの可視化手法の適用

3.1 農作物不良品判別モデル

実装したプログラムの適用対象は当場で開発を進めている、ブロッコリーの良品と不良品を判別する画像認識モデルとした。本モデルは、ブロッコリーの画像から良品か不良品かを判別する畳み込みニューラルネットワークモデルである。モデルの学習には、選果場で判別された良品および不良品を同一の撮像条件下で撮影した画像をそれぞれ約1,300枚用いた（図 4 ）。不良品は主として腐敗により花蕾（食用とされるつぼみ）の一部が黒等に変色していた。学習後のモデ

図 4　ブロッコリー画像
左：良品、右：不良品（腐敗）

ルはテストデータにおいて正解率90%以上の良好な性能を示した。

3.2 可視化手法の適用

学習後のモデルの妥当性を評価するため、テストデータの画像を不良品（腐敗）であると正しく判別した際の予測根拠を、CRAFTを適用して可視化した。結果を図5および図6に示す。図5に抽出された概念を予測に重要なものから順に①～④で示し、図6に各概念が入力画像のどの領域で利用されたかを示す。また、予測における概念の寄与率は、①の概念が約80%となり、最も重要であるとされた。

図5の各概念を代表する画像から、①はブロッコリーの茎および葉、②はブロッコリーの花蕾、③は容器、④は搬送台のフレームの概念であると考えられる。今回の腐敗ブロッコリーは花蕾に変色が生じているものであり、本来であればモデルは②の概念を重視して腐敗の有無を判別すると想定される。しかし、予測における寄与度が最も高い概念は①の茎および葉であり、想定と異なる。この不整合は、本来意図している腐敗ブロッコリーの特徴をモデルが学習できていない可能性を示唆している。

3.3 モデルの改善

予測根拠の可視化結果から、モデルが本来注目すべき腐敗の特徴が現れる花蕾ではなく、茎や葉といった部位に注目して腐敗を判別していることが示唆された。この原因について調査した結果、学習に使用した腐敗ブロッコリーの多くが撮影日程の都合により収穫から日数の経過した個体であり、葉や茎のしおれや乾燥など腐敗（花蕾の一部黒色化）とは直接関係ない視覚的特徴を併せ持っていたことがわかった。これにより、モデルは腐敗そのものではなく、しおれや乾燥といった異なる特徴を学習してしまっていた可能性が高いと推測された。

この問題に対応するため、学習データを見直し、しおれや乾燥等の特徴が強く現れていた画像約200枚をデータセットから削除し、モデルの再学習を実施した。再学習後のモデルに対して同様にCRAFTを用いた予測根拠の可視化を行った。結果を図7および図8に示す。

図7は、再学習後のモデルから抽出された概念のうち予測において重要とされた上位4つを示している。また、図8は各概念が入力画像のどの領域で利用されたのかを可視化したものである。今回の結果では、最も寄与率の高かった概念①がブロッコリーの花蕾に関する特徴を反映しており、図8においても当該概念が花蕾の部位で利用されていることが確認

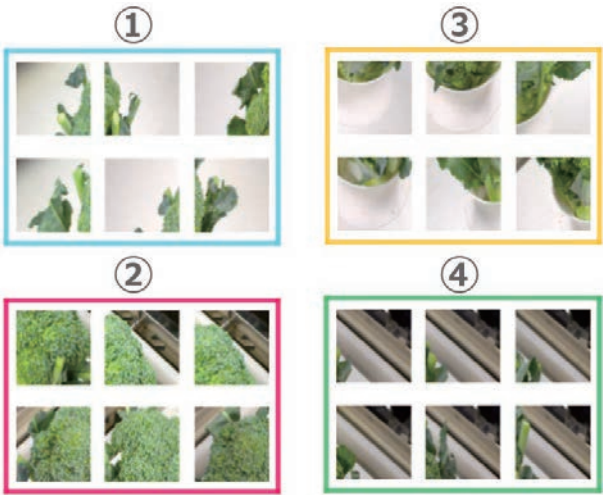

図5　抽出された概念
（予測に重要なものから順に①～④）

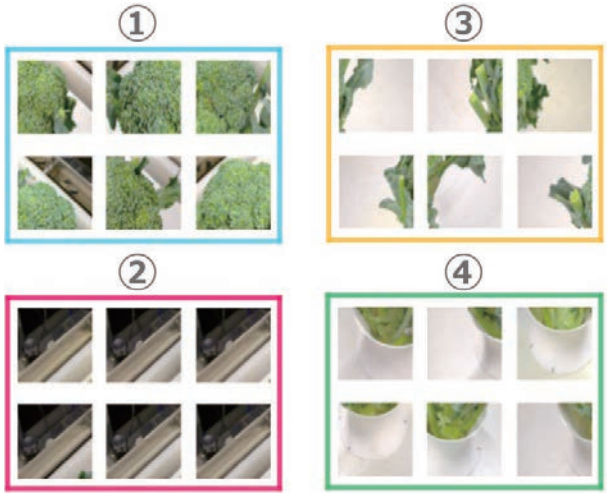

図7　抽出された概念
（予測に重要なものから順に①～④）

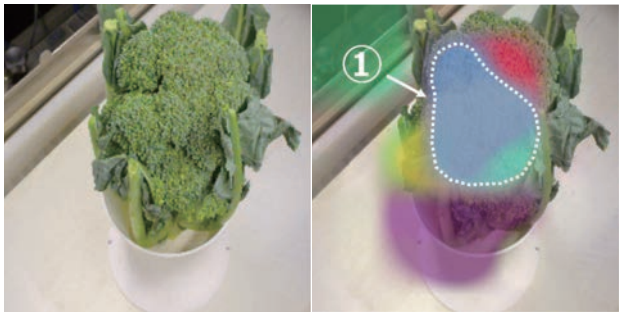

図6　予測において各概念が利用された画像領域

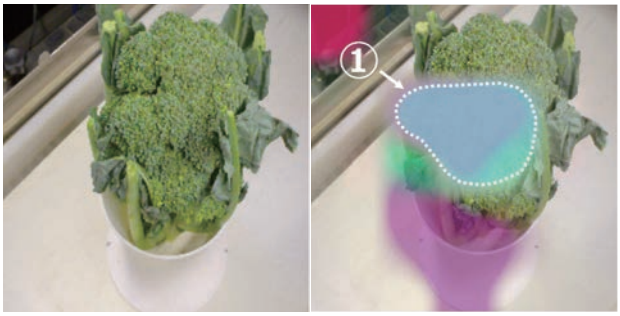

図8　予測において各概念が利用された画像領域

できた。このことから、再学習後のモデルは従来よりも適切な視覚的特徴にもとづいて不良品を判別していると判断できる。

一方で、モデルの予測性能を示す正解率は再学習の前後で大きな変化は見られず、いずれも90%以上であった。このことは、モデルが誤った特徴に依存して予測していたとしても、正解率といった性能指標からは学習の偏りに気づくことが難しいことを示している。特に学習データのみに存在する特徴が予測性能の向上に有利にはたらく場合は、見かけ上の性能が高くなるため、正解率等の性能指標ではモデルの内部状態を適切に評価することができない。

今回のように、概念にもとづく可視化手法を用いることで、モデルが予測において「何を重視したのか」を詳細に分析することが可能となり、正解率などの数値指標では見逃されやすいモデル内部の問題やデータセットの偏りの発見に有効であることが確認された。

4. おわりに

本研究では、農作物の画像から良品と不良品を判別するAIモデルに対し、画像認識AIの予測根拠を可視化して説明する最新手法を適用し、その有用性を検証した。具体的にはモデルが判別に利用した視覚的「概念」と画像内における注目領域を同時に可視化できるアルゴリズムであるCRAFTを農作物判別モデルへ適用し、開発中のモデルが予測において意図していない茎と葉の部位に着目していることを明らかとした。その結果、データの見直しによる再学習を通じて、モデルが本来注目すべき花蕾の部位に基づいて判断するようモデルを改善することが可能となった。これは、予測根拠を可視化することでモデル内部に潜在していた学習の不整合を把握し、改善につなげることが可能であることを示している。

また、モデルの正解率は再学習の前後で大きく変化しなかったことから、一般的な性能指標のみではモデルの妥当性を評価しきれない場合があることも確認された。このことは、モデル開発において予測性能だけでなく、判断過程を可視化・分析する重要性を示すものである。

一方、今回使用したCRAFTには技術的な制限も存在する。特に、概念の抽出に非負値行列因子分解（NMF）を使用しているため、対象となるニューラルネットワークの出力が非負値である必要がある。したがって、活性化関数としてReLU等の非負性を持つモデルには適用できるが、Swish等の負の値を取る関数を用いる最新のモデルにはそのまま適用することができない。今後、NMFの代替として非負制約を緩和した行列分解アルゴリズムを組み込むなど、より広範なモデル構造へ適用可能な可視化手法の開発も検討する必要がある。

今後は、AIを使用する様々な研究開発において本研究の知見を広く活用することで、より信頼性の高いAIモデルの開発を図る予定である。

参考文献

1）Thomas Fel, Agustin Picard, et al.：*CVPR 2023*, pp. 2711-2721,（2023）

2）Ramprasaath R. Selvaraju, Michael Cogswell, et al.：*ICCV 2017*, pp. 618-626,（2017）

3）Marco Tulio Ribeiro, Sameer Singh, et al.：*KDD 2016*, pp. 1135-1144,（2016）

4）Scott M Lundberg, Su-In Lee.：*NIPS 2017*,（2017）

5）Been Kim, Martin Wattenberg, et al.：*ICML 2018*, pp. 2668-2677,（2018）

6）Thomas Fel, Lucas Hervier, et al.：XAI4CV Workshop at CVPR 2022,（2022）