画像認識における説明可能な AI に関する研究

（令和 5～6 年度）

産業システム部　○全　　慶樹、本間　稔規
企画調整部　　近藤　正一

１．はじめに

深層学習などの AI モデルは膨大な量のパラメータから構成されているため、モデルの内部構造から認識や予測の根拠を説明することが困難である。予測根拠を説明できない場合、「目的外の対象を学習する」等のモデルの問題の発見が困難になることから、近年、AI モデルの予測根拠を説明するための様々な研究が進められている。AI を利用する道内企業から当場に対して、画像認識モデルの認識や予測の根拠に関する相談が寄せられているなど、AI モデルの予測根拠を説明する技術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。

２．画像認識モデルの予測根拠の可視化手法

画像認識モデルの予測根拠を可視化する手法には、モデルが予測の際に重視した画像内の領域を示す手法やモデルが予測の際に利用した概念（concept、代表的な画像の集合で表現される）を示す手法などが知られている。最新の手法である CRAFT（Fel et al., 2023）は、これらの手法を組み合わせることでモデルの予測根拠の詳細な可視化を可能とした。本研究では、Python の深層学習フレームワークである TensorFlow を使用して当該手法のプログラムを実装し、画像認識モデルへ適用した。

３．農作物不良品判別モデルへの可視化手法の適用

開発したプログラムをブロッコリーの画像（図 1 左）から不良品（腐敗）を判別する画像認識モデルへ適用し、開発中のモデルが不良品の判別において、どのような概念を、画像内のどの領域で利用しているか、を可視化することでモデルの妥当性を評価した。

図１　ブロッコリー画像（左）と可視化結果（右）

可視化手法により抽出された 5 つの概念を図 2 に、各概念の重要度を図 3 に示す。また、画像認識モデルが判別の際に各概念を利用した画像内の領域を色分けして図 1 右に示す。図 2 と図 3 からモデルは①の茎の概念を利用して不良品（腐敗）を判別していることがわかるが、腐敗は花蕾が部分的に黒色に変化する現象であり、茎の概念を重視したとする可視化手法の分析結果と整合しない。このことから開発中のモデルは学習に失敗したということが結論づけられた。以上より、予測根拠の可視化手法はこれまで困難だったモデルの問題発見に有用であることがわかった。

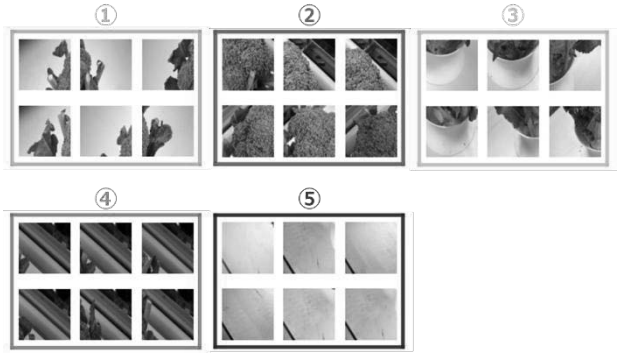

図２　可視化手法により抽出された概念

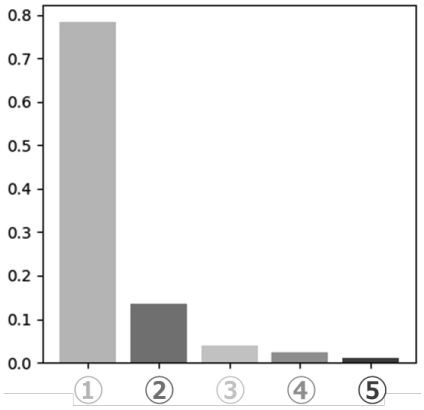

図３　各概念の重要度

４．おわりに

本研究では、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ブロッコリーの不良品判別モデルへ適用することでその有用性を検証した。その結果、これまで困難だった、モデルの問題の発見に有用であることがわかり、より信頼性の高い AI の開発が可能となった。

（連絡先：zen-keiki@hro.or.jp）