

切羽評価 AI モデルの説明可能性に関する検討

株式会社大林組 正会員 ○松崎 晃 畑 浩二 稲川 雄宣

1. はじめに

昨今、機械学習の建設分野への適用は増加しており、自動・自律運転のみならず品質評価に関する AI も開発されている。しかし、AI の性能は向上したものの、複雑化した機械学習の出力結果に対して、「何故そう判断したか」に関しての説明がつかずブラックボックス化してしまうという問題がある。この問題は説明可能性問題と呼ばれる。この問題の解決策として、機械学習の判断の根拠を可視化する手法（以下、説明可能手法と称す）の開発が進んでいる。本稿では既存の切羽評価 AI に対して、説明可能手法を適用した結果とその知見について述べる。



図-1 切羽 AI の出力

2. 概要

大林組では、山岳トンネルの掘削に携わる未熟練技術者の切羽評価の補助ツールとして、AI を用いた切羽評価手法（以下、切羽評価 AI）の開発を行っている¹⁾。切羽評価 AI は切羽の画像から風化変質、割目間隔、割目状態、圧縮強度、走向傾斜、湧水量および劣化度合いの計 7 項目の評価をそれぞれ推定し、図-1 に示すように単位部分ごとに評価点を出力する。切羽画像と実際の切羽観察結果を教師データとして、深層学習した AI であり、図-2 に示すように Alex-Net²⁾をベースとして、分類器を SVM（Support Vector Machine）に変更する工夫を施した。これは全結合ニューラルネットワークを使用するよりも SVM を使用する方が、分類精度が高いためである。文献 1)に示すように、風化変質の正解率は 73%となっている。

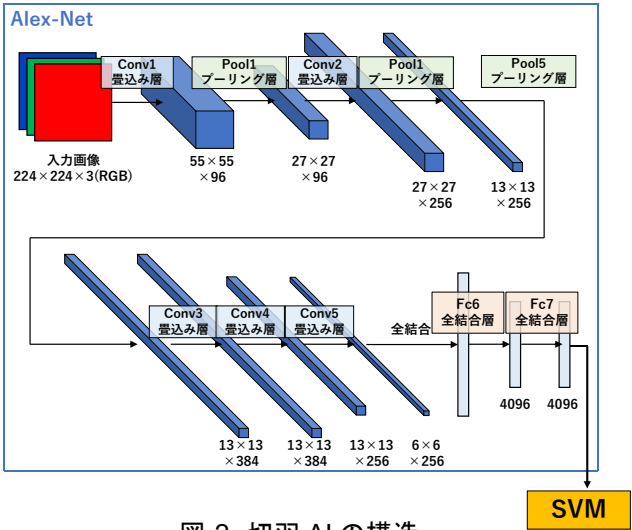


図-2 切羽 AI の構造

表-1 Grad-CAM の
ヒートマップ値の総和

LIMEによる 寄与度の順位	Grad-CAMの ヒートマップ値の総和 () は順位
Top1	6581703 (2)
Top 2	1001963 (10)
Top 3	2456185 (9)
Top 4	5507794 (5)
Top 5	4254828 (6)
Top 6	5965901 (3)
Top 7	3062075 (8)
Top 8	5755119 (4)
Top 9	3796592 (7)
Top 10	7792375 (1)

本検討では、試行として風化変質の評価を対象に、説明可能手法を適用し、評価を実施した。図-2 中の Fc7 層において、分類結果への寄与度が高い特徴量を抽出し、その特徴量が獲得している特徴を可視化した。分類結果への寄与度の算出には LIME（Local Interpretable Model-agnostic Explanations）³⁾を、特徴量の可視化には Grad-CAM（Gradient-weighted Class Activation Mapping）を使用した。LIME は局所的なデータをもちいて近似式を作成し、AI モデルと比較することでその寄与度を算出するアルゴリズムである。Grad-CAM は強く認識している部分をヒートマップとして出力する。なお、入力画像は、強風化した岩盤と新鮮な岩盤が含まれる画像とし、説明可能手法は、切羽評価 AI の出力結果の単位部分ごとに適用した。

3. Fc7 層の特徴量の順位

表-1 に LIME による特徴量の寄与度の順位と Grad-CAM のヒートマップ値の総和を、図-3 に Top10 特徴量の Pool1 層、Pool2 層、Pool5 層での Grad-CAM 適用結果を示す。Grad-CAM は、図-3 に示すように分類結果への寄与度が大きい部分を強く活性化する。LIME による寄与度の順位が上位の特徴量においては、全体的に分類への寄与度が大きく、Grad-CAM によっても全体として強く活性化されると予測された。しかし、表-1 に示すように双方

キーワード 説明可能、LIME、Grad-CAM、切羽評価

連絡先 〒204-8558 東京都清瀬市下清戸 4-640 株式会社大林組 技術研究所 TEL050-3829-2163

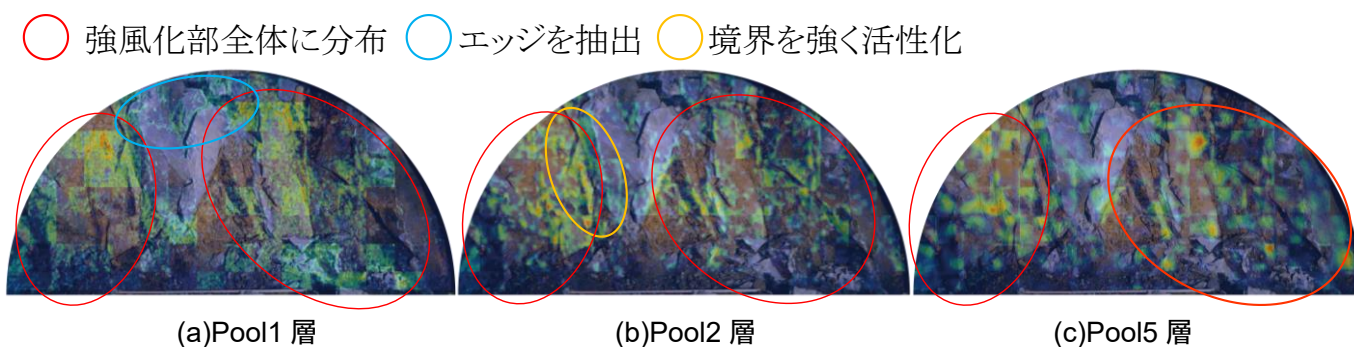


図-3 各層での Grad-CAM 適用結果(Top10 の特徴量)

の評価の順位は一致していない。例えば LIME による寄与度の順位が 10 番目の特徴量（以下 Top10 特徴量と称す）が Grad-CAM のヒートマップ値の総和が最も高い。これは寄与度算出と可視化でアルゴリズムが異なるため違いが生じたと考えられるが今後も検討が必要である。

4. 各層での可視化の検討

一般に畳込みを有する画像分類 AI モデルは、比較的浅い層で色やエッジ等の人間が理解しやすい低次元の特徴量を抽出し、深い層に向かうにしたがって高次元の特徴量を抽出する。そこで、切羽評価 AI に前述の説明可能手法を適用し、各層での説明性に関して検討を実施した。

Pool1 層では、強風化した部分に活性化が高い部分が分布していることに加え、新鮮な部分ではエッジを抽出している。すなわち、風化変質している岩の判断基準として、エッジの特徴を含めて判断していると考えられる。Pool2 層では、Pool1 層と同様に強風化した部分に活性化が広がっていることに加え、強風化した岩と新鮮な岩の境界を強く活性化しており、岩種の境界に関する特徴を獲得している可能性がある。Pool5 層では、強風化した岩の範囲に活性化が部分的に分布しているが、人間が視覚的に認識できる特徴を捉えきれてはいない。このように深い層になると、その判断の根拠を人間的に理解するのは困難であり、説明が難しくなる傾向にある。浅層では人間が視覚的にわかりやすい活性化をしているが、Pool1 層よりも Pool2 層の方がキーポイントである境界部やエッジを捉えている。どの層が最も適切に説明できるかは、現段階では不明であるが、分類結果に対して説明可能手法を適用することで、その根拠に対して説明がつく可能性があることが示唆された。

5. まとめ

既存の切羽評価 AI に対して、説明可能手法を適用した結果について述べた。得られた知見を以下に示す。

- 1) LIME+Grad-CAM を適用した結果、LIME による寄与度の順位と Grad-CAM の可視化には直接的な関連性はなく今後も検討が必要である。
- 2) どの層が最も適切に説明できるかは、現段階では不明であるが、切羽評価 AI の分類結果に対して、説明可能手法を適用することで、その根拠に対して説明がつく可能性があることが示された。
- 3) 今後、本検討では Fc7 層のみを対象にしているが、その他の層での特徴量の寄与度算出や浅い層での可視化を実施することで、人間が視覚的にわかりやすい説明ができる可能性がある

謝 辞

本研究に関し、MathWorks Japan の縣 亮氏、萩谷 俊幸氏にご協力いただいた。ここに記して感謝の意を表す。

参考文献

- 1) 畑浩二：山岳トンネル切羽評価における DNN の適用、地盤工学会誌、JGS2022-7、Vol.70、No.7、Ser.No.774、pp33～36、2022
- 2) Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hilton：ImageNet Classification with Deep Convolutional Neural Networks, Conference and Workshop on Neural Information Processing Systems(NIPS)2012.
- 3) Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin：Why should I trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.