

画像データ向け local XAI 結果の集計方式の提案 Summarize method of XAI results for image data

浜 直史[†] 森 靖英[†] 和久井 一則[‡]

Naofumi Hama Yasuhide Mori Katsunori Wakui

1. はじめに

近年、機械学習モデルが出力する予測値の精度向上を目的とした技術開発の進展によって、モデルやアルゴリズムのブラックボックス化が加速している。これによって機械学習モデルを開発するユーザは、自分の作成したモデルが必要な精度に達しない場合や意図しない挙動を示す場合に、どのような改善手法を施せばよいかが分からず、実際の案件に適用できるモデルを作成できないという課題があった。

この問題に対し、モデルがなぜそのような判断結果を出力したのかを説明できるようにする、XAI (eXplainable AI, 説明可能な AI) という技術が研究されている。XAI 技術を自らが開発する機械学習モデルに適用し、出力根拠からその挙動の原因を解釈できるようにしておくような、説明可能性を機械学習モデルの品質保証の 1 項目とする動きも進められている[1]。本研究は、XAI 技術によって機械学習モデルの精度不足や予期せぬ挙動といった問題の原因を突き止めることで、適切な改善手法を立案可能にすることを目的とする。以下では特に、画像を入力データとした分類タスクに係る機械学習モデルを取り扱う。

2. XAI 技術

既存の AI モデルを対象に適用する XAI は、大きく Model Explanation, Outcome Explanation, Model Inspection の 3 つに分類される[2]。本研究ではこのうち、主に Outcome Explanation と Model Inspection を取り扱う。Outcome Explanation は Local XAI と呼ばれ、入力データごとに異なる出力根拠を説明とするものである。これは現在最も研究が盛んで、特に対象とする機械学習モデルを特定せずに、既存の学習済み機械学習モデルへアドオンのように適用することができる XAI 技術として有用とされる。一方 Model Inspection は、ブラックボックスである機械学習モデルが、データを入力され推論結果を出力するまでのプロセスを可視化するなど、人間が理解できる形で表現することによって機械学習モデル全体の説明とするものである。入力データが画像の場合に Model Inspection 型の説明を与える有力な手法は我々の知る範囲ではまだ存在しない。そして本研究は、機械学習モデル全体の精度を改善するために必要な情報としての Model Inspection 型の説明を、Outcome Explanation の結果を抽出および集計することから得ることを目的としている。

図 1 は Local XAI の一つである SHAP[3] を画像分類モデルに適用した例である。鳥類の画像[4](図 1(a)) から種名を判定する分類モデルにおいて、図 1(b) のように区切られた画像パッチごとに、種名の推定に寄与した大きさを SHAP によって求め、赤色の濃さによって表現したのが図 1(c) である。本例では、種名判定に無関係な道路標識は全く推定に寄与せず、鳥の体側面の茶色い部分から種名を推定しているという妥当な結果が得られている。

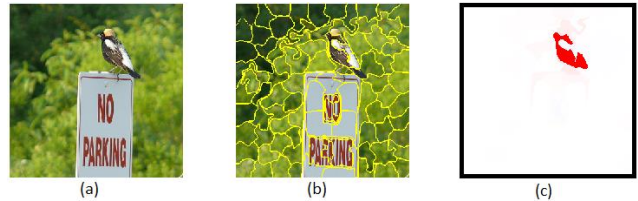


図 1 Local XAI 結果の例

3. 機械学習モデルの改善に用いる際の課題

Local XAI として機械学習モデルの出力結果に対する各説明変数の寄与の大きさを出力根拠として算出することで、機械学習モデルの推定プロセスについて理解できるケースが多くなった。画像のような非構造データで特徴量ベースの説明を行う際には、ハイライト表現のように出力根拠を表現した場合に元の画像と見比べ、注視すべき点を機械学習モデルが適切にハイライトしているかを以て妥当性を検討することができる。一方で、Local XAI の結果を集計して機械学習モデルの改善を図る際に、テストデータセット全体についてこのような目視による比較検討を行うことは困難である。機械学習モデルの改善プロセスに必要なのは、どのような特徴を持つ画像に対しての精度が悪く、どのような特徴は適切に捉えられているか、などの機械学習モデル全体の傾向である Model Inspection 型の説明である。各データに Local XAI の結果を適用した結果を、テストデータセット全体に亘って機械的に集計することでこの Model Inspection 型の説明を得ることが必要となっており、特にその集計の観点を表現し Local XAI 適用結果をフィルタ出来るような抽出基準を作成することが、XAI 技術を機械学習モデルの改善プロセスに用いる際の課題である。

4. XAI パーツパターン作成手法の提案

前章までに述べた課題に対し以下では、Local XAI の結果を抽出(フィルタリング)するための、集計の観点を表現する XAI パーツパターンを定義する。ここで XAI パーツパターンとは、分配される寄与値が大きく、機械学習モデルが推論に用いている特徴的なパーツを複数個組み合わせたものである。このような XAI パーツパターンを持っている画像データを抽出し、それらのみで機械学習モデルの精度や Local XAI 結果などを集計すると、画像データセットに対して機械学習モデルがどのような挙動をする傾向があるかという Model Inspection 型の説明が得られる。更にこれを半自動的に作成する手法として、一部の XAI 適用結果から XAI パーツパターンの候補を作成する手法を提案する。

XAI パーツパターンの作成手法ではまず、Local XAI の結果が「分かりやすい」、即ちハイライトされた複数箇所が局所的あるいは特徴的であるデータを抽出し、そのハイライトされた箇所を半自動的に複数箇所抽出したものの組み合わせを XAI パーツパターンとして採用する。その上でそのよ

うな XAI パーツパターンがデータセットの他の(XAI 結果が比較的局所的もしくは特徴的でない) データの Local XAI の結果にマッチングするかどうかを検索し, XAI パーツパターンごとにマッチングした画像群について機械学習モデルの精度や Local XAI 結果などを集計する手法である。

5. 評価実験

Inception V3[5]の最終層に fine-tuning を施し、鳥類の画像データセット[4]に対し種別を精度 76%で判定するモデルを作成した。以下ではこのモデルに対し XAI を適用する。

SLIC[6]を用いて画像を 50 個前後に分割し、それぞれを特徴量として kernel SHAP により寄与値を分配している。今回は, superpixel ごとに算出された寄与値が、特に 2 つの superpixel に局在する画像を Local XAI の結果が「分かりやすい」画像の候補とし、その上位から目視によって当該局在箇所に分かりやすさを基準に XAI パーツパターンを定義する画像(図 2 左)を選択した。以下、当該画像から定義した XAI パーツパターンによる抽出結果を報告する。



図 2 左: 元のテストデータ画像 中: Local XAI の結果 右: 上位 2 superpixel を用いて定義した XAI パーツパターン

上で定義した XAI パーツパターンを用いたマッチングによる抽出は以下のように行った。まず、検索対象の画像へ Local XAI を適用し、出された寄与値が高い 3 つの superpixel を当該画像の判断根拠に用いられた superpixel と定義する。次に、XAI パーツパターンのうち寄与値の高い方の superpixel(図 2 右図の白い箇所)を用いて判断根拠に用いられた全ての superpixel との組み合わせとテンプレートマッチングを行う。この際、同時に両 superpixel に対し、回転角(-30° ~ +30°)および拡大率(80%~120%)についてパラメータサーチを行いながら superpixel の変形を施し、最も高いマッチングスコアを算出する組み合わせを得る。その際の回転角および拡大率のパラメータを用いて XAI パーツパターンのうち寄与率の低い方の superpixel を変形し、検索対象画像の判断根拠に用いられた superpixel とテンプレートマッチングにおけるマッチングスコアの積を本検索のスコアとするものである。テスト画像 4642 枚中、最も高いスコアを与えた 2 枚の画像が図 3,4 である。いずれも鳥の右側の羽を抽出できている。またここでは、Local XAI 結果を用いたマッチングを行っているため、XAI パーツパターンと色が異なっているにも関わらず、判断根拠として同じ箇所を用いた画像を抽出することができている。

また、特にこのスコアでソートした上位 20 枚の画像についてはモデルの正答率が 50%となっていた。これは同じく上位 100 枚の画像に対する正答率 70%、データセット全体に対する正答率 76%と比較して著しく低く、Local XAI 結果を基準とした誤答データの特徴を見つけれられているとも考えられる。今回の抽出結果を精査することでモデルの精度を向上させられる可能性がある。

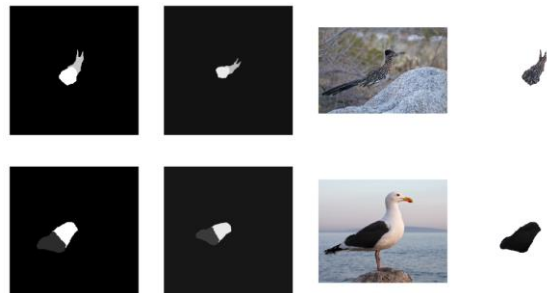


図 3 上段: XAI パーツパターン定義に用いた画像 下段: マッチングスコア最上位の画像。左から順に、マッチングに用いた superpixel (色は XAI による寄与値を参照), 変形結果, 元の画像データ, 判断根拠箇所

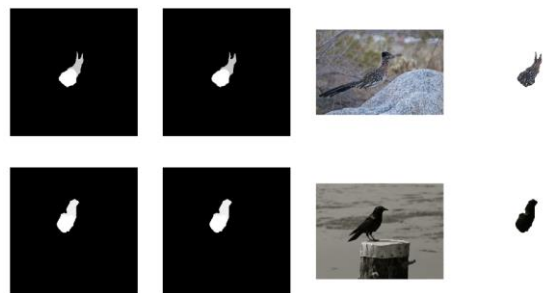


図 4 図 3 と同じくマッチングスコアが次点となったもの

6. まとめ

機械学習モデルの精度向上施策の立案に必要な情報を得るための、Local XAI の抽出基準の作成方法を検討した。モデルが持つ潜在的な誤判定リスクを検出するためには、Local XAI の結果を参照し適切な抽出基準を定める必要がある。そこで我々は XAI パーツパターン作成手法を提案し、原理検証のための初期実験として所与のデータセット及びモデルで提案手法の妥当性を検証した。その結果、Local XAI 結果を基準とした検索により誤答データを抽出可能な基準を作成できた。これによって XAI パーツパターン作成手法を用いた Local XAI 結果の集計を用いて、モデルの精度向上に係る施策立案に必要な情報を得る見込みを得た。

参考文献

- [1] AI プロダクト品質保証コンソーシアム『AI プロダクト品質保証ガイドライン 2019.05 版』,
- [2] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42..
- [3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).
- [4] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.
- [5] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
- [6] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & S. S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11), 2274-2282.

† (株)日立製作所 Hitachi, Ltd.

‡ (株)日立産業制御ソリューションズ Hitachi Industry & Control Solutions, Ltd.