H-002

背景入替えデータ集計による XAI 結果評価方式の提案
Proposal of XAI result evaluation method based on background replacement

安井 雅彦† 浜 直史† 森 靖英† 和久井 一則‡
Masahiko Yasui Naofumi Hama Yasuhide Mori Katsunori Wakui

1. はじめに

機械学習の技術が発展する一方で，機械学習モデルやアルゴリズムのいわゆるブラックボックス化が加速している。ブラックボックス化とは，当該モデルがどのような理由でその結果を出力したのかについて直感的な理解を得ることが難しくなってきていることをいう。機械学習のブラックボックス化に対し，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。また，XAI を自らが開発する機械学習モデルに適用し，出力される判定根拠からその挙動の原因を解釈できるようにすることでモデルの精度向上に活用する動きも進められている。

XAI を以てモデルの精度向上を図る際，従来の方式ではあるデータに対するモデルの判定結果が正答か誤答かという観点と，判定根拠が妥当か不当かという観点とがあったが，いずれもこのように二値の分類となっていた。このためデータセットは，これらの組合せで得られる 4 象限への分類がなされていた。しかし，データセットに対するモデルの挙動を考察する際には，このような 4 象限への分類では実施する分析や操作に対してあまりに粒度が粗く不十分で，より精緻に分類することが求められていた。

本研究では，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に対して XAI を適用する場合を対象とする。中でも本報告では，人物が写り込んだ画像データの特徴を判定する機械学習モデルに対して XAI を適用するシーンを対象とする。我々は，画像の特徴判定 AI モデルにおいて，元画像に背景のセグメンテーションを行い，複数の背景入替え画像を作成しそれらの画像に対する XAI 適用結果を集計する評価方式を提案する。

2. 関連研究

データに含まれる対象の特徴を判定するモデルがあるとき，この特徴判定モデルの出力結果に対してデータ内の各構成要素に寄与を分配し，これを判定根拠類推の一助とすることができる XAI がこれまでも研究されてきている[1]。これら XAI を用いて画像データに特徴判定モデルを適用し，その際の寄与分布と対象領域との重なりを用いてその際の判定が妥当か否かについて評価する研究がある[2,3]。

しかし，先述のように従来の特徴判定モデルの結果の集計方式とこれに基づいた評価では，判定結果が正答か誤答かという観点と，判定根拠が妥当か不当かという観点とがあったが，いずれもこのように二値の分類となっていた。

3. 提案手法

従来の方式では，1 つの判定対象について 1 枚の画像にのみ XAI を適用し評価を行っていたために不十分だったと我々は考えた。そこで，Anchors[4]という XAI 手法から我々は提案手法の着想を得た。Anchors は機械学習モデルの推論結果を局所的な線形近似などで行うのでなく，if-then 形式のルールベースで行うものである。入力データが画像の場合，各画像に対し，入力データに含まれていれば他の部分がどのように変更を加えられても推論根拠は変わらないという必要十分な部分を提供する形での説明を行う。当該説明として必要十分な部分というのは判定対象となることが多く，それ以外の背景部分を他の画像と入替えてもモデルの判定結果は変わらないということを示す手法である。この手法から，背景入替えによって 1 つの対象データから複数の背景入替え画像を得る着想を得た。

提案手法では，予め入替え用に様々な背景画像を用意しておき，入力データに写り込んだ判定対象はそのままに，背景を入替えたときの XAI 結果の妥当性是非を集計する。背景部分に寄与が置かれる背景入替え画像がいくつかある場合においても，モデルが正しく機能している場合には背景入替え画像の大多数で判定対象に寄与が置かれることになるはずである。モデルが正しく機能していない場合には，背景入替え画像の多くで背景領域に寄与が置かれ，正しく判定対象を認識できていないと考えられる。このように，判定根拠が妥当か不当かではなく，用意した背景画像のうち何割で判定根拠が妥当となったかという評価が可能になることでデータセットのより精緻な分類が可能となる。

提案手法の処理過程を図 1 に示す。まず，図 1 左に示すように画像データから人物部分と背景部分を切り分ける処理を行い，それぞれのマスクを作成する。入替える背景画像は事前にデータを用意しておく。次に，得られたマスクを用いて人物部分を残し，図 1 中部に示すように背景部分を入替えたデータを作成する。得られた複数の背景入替え画像に対して，ピクセルごとに XAI を用いて寄与を計算し，図 1 右に示すような各画像についての寄与度マップを作成しておく。最後に，寄与度マップそれぞれに対して閾値に基づいた二値化処理を行い，人物領域との重なりがある場合には判定根拠妥当，そうでなければ判定根拠不当として集計を行う。

上記の閾値は，画像ごとの寄与分布の様態に基づいて決定した。寄与分布の傾向が，背景画像ごと，合成する人物画像ごとに大きく異なる事がわかっている。このため，すべての寄与度マップに対して単一の閾値を以て二値化処理を行うことは適切ではない。そこで我々は，背景の中に判定に寄与するべきではない空や地面といった領域を参照領域として定め，その内部の寄与分布の様態を以て予め定めた閾値への補正を行う手法を用いた。これは，上記寄与分布の平均と分散を用いる方式である。寄与分布の平均の正/負に応じて閾値に正/負の補正を行い，分散の大きさに応じて閾値に正の方向に補正をかける。

4. 評価

提案手法を用いて，データセットに対してクラスタリングを行い，この結果を以てデータ拡充を施すことでモデルの精度向上を評価した。

†(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.

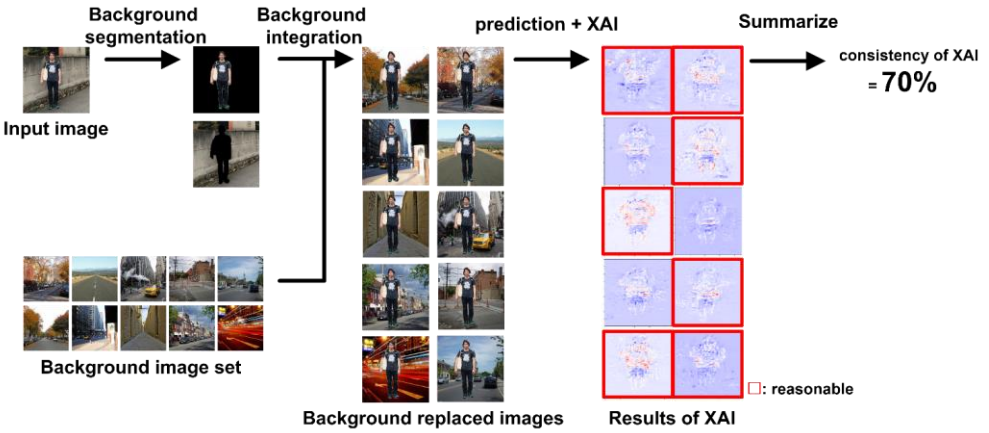

図 1　提案手法の処理過程

手順は以下 3 ステップで作成される 3 つのモデルを評価用データに適用し，その精度を評価対象とした。

STEP1(そのままの学習データ): 学習データでモデル(model_0)を作成。

STEP2(従来手法の学習データ拡充方式): model_0 と学習データを使って正答/誤答と妥当/不当の四象限分類を実施。正当かつ妥当の領域以外のデータに回転画像を追加した学習データを作成。これを学習データとしモデル(model_1)を作成。

STEP3(提案手法による学習データ拡充): model_0 と学習データ，提案手法を用いて，図 2 に示す 2D ヒートマップを作成。この 2D ヒートマップは横軸にモデルが正当する割合，縦軸に判断根拠が妥当である割合をとっている。正答割合が 0.9 以下で 1 マスの値が 30 以上のデータに背景入替え画像を追加した学習データを作成。これを学習データとしモデル(model_2)を作成。

モデルが判別する特徴は人物の性別とし，人物と背景の切り分けには OpenPose[5]を用いた。XAI には Integrated Gradients[6]を用いた。背景画像は kaggle 上で公開されている intel-image-classification[7]の中から屋外画像 10 枚を選出して使用した。学習データと評価データは MakeHuman[8]を用いて，それぞれ 3157 件，30161 件を用意した。

評価結果を表 1 に示す。提案手法を用いて，およそ 4%の精度向上を得た。また，従来の方式と比較してもおよそ 2%精度が高く，提案手法の有用性が示された。

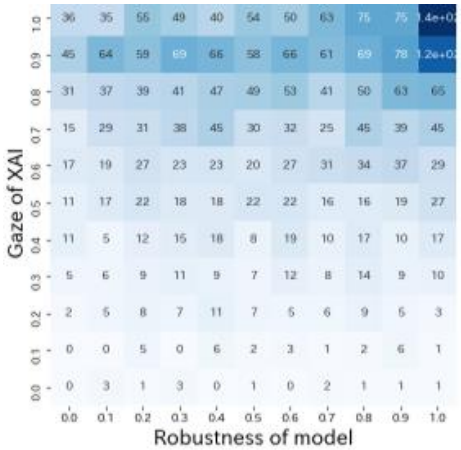

図 2　2D ヒートマップ

表 1 評価結果

モデル	学習データ	正答率
model_0	無操作	0.813
model_1	四象限による拡充	0.832
model_2	2DHM による拡充	0.852

5. まとめ

従来の XAI を用いた適用結果を集計する方式では，モデルの挙動を考察する際に不十分だった。そこで我々は，画像の特徴判定 AI モデルにおいて，元画像に背景のセグメンテーションを行い，複数の背景入替え画像を作成しそれらの画像に対する XAI 適用結果を集計する手法を提案した。提案手法を用いることで，XAI 結果をより細かな粒度で評価することができる。これによって，AI モデルの挙動をより精緻に解析することができ，実験では提案手法を活用した追加学習による精度向上を確認した。

参考文献

[1] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 4765-4774. 2017.
[2] Zhang, Jiajun, Pengyuan Ren, and Jianmin Li. "Deep Template Matching for Pedestrian Attribute Recognition with the Auxiliary Supervision of Attribute-wise Keypoints." arXiv preprint arXiv:2011.06798 2020.
[3] Chen, Lei, et al. "Adapting Grad-CAM for embedding networks." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.
[4] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.
[5] Cao, Zhe, et al. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields." IEEE transactions on pattern analysis and machine intelligence 43.1. 172-186. 2019.
[6] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." International Conference on Machine Learning. PMLR, 3319-3328. 2017.
[7] “Intel Image Classification - Image Scene Classification of Multiclass” https://www.kaggle.com/puneet6060/intel-image-classification/ (2021/06/01 確認).
[8] “MakeHuman” http://www.makehumancommunity.org/ (2021/06/01 確認).