私のブックマーク

説明可能 AI †1

原　　聡（大阪大学産業科学研究所）

1. は じ め に

2018 年に本誌 5 月号（Vol. 33, No. 3, pp. 366-369）の“私のブックマーク”に「機械学習における解釈性」[1] という記事を書いた．前記事の執筆から 1 年が経ち，機械学習モデルの解釈・説明技術を取り巻く社会的な情勢の変化や新たな研究の発展など，数多くの進展があった．本記事はこれら近年の変化・進展についてまとめた，上記の“私のブックマーク”の続編である．本記事を読む前に，まずは上記の前記事をご一読いただきたい．

用語について

本記事では，機械学習モデルの出力に加えて，その出力を補助する追加の情報（モデルの解釈，判断根拠の説明，など）を出力する技術一般および研究分野全体を指す用語として XAI（Explainable AI：説明可能 AI）を用いる．XAI はアメリカの国防高等研究計画局（DARPA）が主導している研究プロジェクト [2] で使われている略称である．

2. 研究を取り巻く情勢

2·1　社会情勢

2018 年は日本国内において XAI への注目が大きく高まった一年であった．機械学習技術の利用の拡大とともに，将来的にもたらされるだろう社会的な問題への関心・懸念がより高まってきた．このような社会的な関心の高まりを受けて，行政を中心に各種の団体が XAI を重要な研究領域の一つとして取り上げている（2019 年 4 月時点）．

- **人工知能技術戦略会議（内閣府）**
  「人工知能技術戦略実行計画（案）」[3] の中に「信頼できる AI」へのアプローチの一つとして XAI が盛り込まれている．
- **AI ネットワーク社会推進会議（総務省）**
  2017 年の「AI 開発ガイドライン案」[4] に続いて，報告書 2018[5] の中で「AI 利活用原則案」[6] が公開されている．その中に「透明性の原則」や「アカウンタビリティの原則」など XAI に関連する項目が盛り込まれている．
- **経済産業省**
  「次世代人工知能・ロボット中核技術開発」[7] の資料の中に，AI の安全性の担保などへ向けたアプローチの一つとして XAI があげられている．
- **科学技術振興機構（JST）**
  研究開発戦略センターから「（戦略プロポーザル）AI 応用システムの安全性・信頼性を確保する新世代ソフトウェア工学の確立」[8] が公開された．この中で XAI が重要技術領域の一つとして取り上げられている．
- **日本経済団体連合会**
  「AI 活用戦略」[9] の中で AI のブラックボックス性に言及しており，XAI の研究開発の必要性が指摘されている．

2·2　研究界の動向

2018 年度は前年に引き続き，さまざまな XAI 技術が提案された．ここでは情勢把握に有用な情報を紹介する．個別の研究については本記事後半で紹介する．

---

†1 http://www.ai-gakkai.or.jp/my-bookmark_vol34-no4

・ Peeking Inside the Black-Box：A Survey on Explainable Artificial Intelligence（XAI）[10]
XAI 技術のサーベイ論文．Fig.6 にて，XAI 関連論文が 2016 年以降に急激に増加していることが報告されている．
・ DARPA の資料 [11]
アメリカの DARPA が主導する XAI プロジェクトの資料．近年のさまざまな研究の概要がまとめられている．

3. まとめ・解説資料

3·1 日本語資料

2018 年度は「機械学習における解釈性」[1] 以外にも，XAI 技術について複数の日本語のまとめ・解説資料が公開された．

・ 機械学習モデルの判断根拠の説明 [12]，［講演動画］[13]
第 20 回ステアラボセミナー [14] の講演資料．「機械学習における解釈性」[1] をベースに，いくつかの代表技術について紹介している．
・ 機械学習と解釈可能性 [15]
ソフトウェアジャパン 2019 にて LINE の吉永尊洸さんが講演された際の資料．代表的な手法のいくつかを実際に動かしてみた様子を紹介している．
・ モデルを跨いでデータを見たい [16]，［コード］[17]
第 76 回 R 勉強会 @ 東京で @kato_kohaku[18] さんが R のライブラリを紹介された際の資料．R のライブラリを元に，各種手法の概要と適用例，利点・欠点についてまとめられている．
・ ディープラーニングの判断根拠を理解する手法 [19]
@icoxfog417[20] さんによるディープラーニングの説明法に特化したまとめ記事．各種の説明法の基本的なアイディアやその評価方法について紹介している．
・ もうブラックボックスとは呼ばせない〜機械学習を支援する情報可視化技術 [21]
お茶の水女子大学の伊藤貴之先生の講演資料．機械学習モデルを理解するための可視化に関する資料．近年の可視化研究について紹介している．

3·2 英語資料

英語でもまとめ資料が複数公開されて注目を集めている．

・ Interpretable Machine Learning：A Guide for Making Black Box Models Explainable [22]
2018 年から無料公開されている XAI に関する教科書的な資料．ついに完成したとのことで，PDF 版も販売されている．
・ AAAI 2019Tutorial On Explainable AI：From Theory to Motivation, Applications and Limitations [23]
AAAI’19 で行われた XAI に関するチュートリアルの資料．XAI を複数の異なる側面から解説したいろいろな研究の概説．
・ Use Cases for Model Insights[24]
Kaggle にて公開されたチュートリアル資料．
・ Visualization for Machine Learning [25]
NeurIPS’18 の可視化に関するチュートリアルのスライドおよび動画．グラフの色の選び方などの可視化の基礎から，データの俯瞰，深層学習モデルの可視化などの最近の成果について幅広く紹介している．

4. ライブラリ

Python や R で XAI の代表的な手法が使えるようにツールの整理も進んでいる．各論文の著者が個別に公開している実装以外にも，各種手法を集めて整理したライブラリが公開されている．

・ ELI5 [26]
Python のライブラリ．各種の説明法・可視化法が実装されている．
・ iml [27]
R のパッケージ．Interpretable Machine Learning：A Guide for Making Black Box Models Explainable [28] の著者により種々の説明法がまとめられている．詳細は @kato_kohaku [18] さんのモデルを跨いでデータを見たい [29] を参照．

・ DALEX [30]

R のパッケージ．詳細は @kato_kohaku [18] さんのモデルを跨いでデータを見たい [31] を参照．なお, パッケージ開発者の別リポジトリ [32] では論文をはじめさまざまな情報がまとめられている．

5. いろいろな説明法

前記事「機械学習における解釈性」[1] で紹介した研究のほかにも，さまざまな手法が研究･提案されている．特に，「局所的な説明」（特定の入力に対するブラックボックスモデルの予測の根拠を提示することで説明とする方法）に関する研究が活発に行われている．

5･1 局所的な説明

・ Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking [33], ［Python 実装 ml-feature-tweaking [34]；R 実装 featureTweakR [35]］

KDD'17 論文．LIME などのように「判断根拠になった特徴量」を説明するのでなく，判断結果を反転させる「“有用”な特徴量」を出力する方法を提案した論文．日本語での解説記事には [36] や [37] などがある．

・ Anchors：High-Precision Model-Agnostic Explanations [38]［Python 実装 Anchor [39]］

LIME の作者による AAAI'18 論文．LIME では説明対象データの周りで線形モデルをつくるが，Anchor では領域モデル（決定木の葉ノードのようなもの）をつくる．

・ Human-in-the-Loop Interpretability Prior [40]

NeurIPS'18 論文．局所的な説明を“わかりやすい”モデルの学習に応用する方法を提案した論文．人間がモデルの学習ループの中に入ることで，“わかりやすい”モデルを学習する．学習の過程で人間に局所的な説明を与えて，人間の反応時間（わかりづらい説明ほど反応が遅くなると想定）をもとに現在のモデルの“わかりやすさ”を評価する．学習を繰り返すことで，人間の反応時間が短くなるように（モデルがわかりやすくなるように）モデルを学習する．

5･2 深層学習モデルの説明

深層学習モデルの説明法としても，従来の「画像の注目領域のハイライト」だけでなくいろいろな説明法が考案されている．

・ Interpretability Beyond Feature Attribution：Quantitative Testing with Concept Activation Vectors（TCAV）[41], ［著者スライド [42]］, ［Tensorflow 実装 tcav[43]］

ICML'18 論文．CNN などの画像認識モデルにおいて，モデルが特定の“コンセプト”に基づいて対象を認識しているかを測る方法を提案した論文．特定のコンセプトの画像（e.g. ストライプ柄の画像）を集めて，コンセプトを代表するベクトルを学習する．学習されたコンセプトベクトルとモデルの認識過程との関係を定量化することで，認識結果への“コンセプト”の影響度合いを調べる．

・ This Looks Like That：Deep Learning for Interpretable Image Recognition [44]

画像認識モデルにおいて，類似の学習データを認識の根拠として提示する方法を提案した論文．事後的に認識対象と学習データのひも付けをするのでなく，モデル内部に訓練データとの類似度を計算するレイヤを明示的に埋め込んでいる．これにより，モデルが訓練データとの類似度に基づいて画像認識をするようになる．

・ Generating Visual Explanations [45], ［Caffe 実装 [46]］

ECCV'16 論文．画像認識モデルの認識根拠を文章で説明する方法を提案した論文．CNN の中間層の特徴を使って，説明文を生成するモデル（RNN + LSTM）を学習する．このような画像＋文章の研究は説明以外のさまざまな文脈でも研究されている．これらの研究についてはオムロンサイニックエックスの牛久祥孝 [47] さんの資料（[48] や [49]）に詳しい．

5･3 説明の学習

従来の説明法は，説明獲得のための計算プロセス（e.g. 最適化問題を解く，微分を計算する，など）は人間が設計していた．これに対し，説明法そのものを学習しよう，という試みが提案されている．

・ Real Time Image Saliency for Black Box Classifiers [50], ［PyTorch 実装 pytorch-saliency [51]］

NIPS'17 論文．深層学習モデルの「画像の注目領域のハイライトの仕方」を学習する方法を提案している．ResNet を使い画像の潜在表現を抽出し，潜在表現をもとにハイライトを生成するネットワークを構築・学習する．

・Learning How to Explain Neural Networks：PatternNet and PatternAttribution [52], ［Tensorflow 実装 innvestigate [53]］

ICLR’18 論文．上の論文と同様に，深層学習モデルの「画像の注目領域のハイライトの仕方」を学習する方法を提案している．誤差逆伝播の枠組みの中で，各層に学習可能なフィルタを入れることでハイライトの仕方を学習する．

・Learning to Explain：An Information-Theoretic Perspective on Model Interpretation [54], ［Tensorflow 実装 L2X [55]］

ICML’18 論文．LIME のように「判断根拠になった特徴」を出力する説明モデルを学習する方法を提案している．説明モデルは入力データに応じた少数の入力特徴を自動的に選び，そこから元モデルの出力を近似できるように学習させる．一度説明モデルを学習してしまえば説明の生成に複雑な計算が不要なため，高速に説明が生成できることが報告されている．

6. 説明法の見直し

ここまでに紹介したように，さまざまな説明法が提案されている．しかし，あまりにも多くの説明法が提案された結果，どの説明法が本当に良いのかがわからなくなってきた．2018 年には特に「画像の注目領域のハイライト」について，理論および実用の面でどの説明法が良いか・悪いか，といった“見直し”が始まった．以下ではこれら“見直し”に関する研究を紹介する．

・A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations [56]

ICML’18 論文．CNN の画像注目領域をハイライトする方法である Guided-Backprop および DeconvNet が，実際にはモデルの認識とは関係ない領域をハイライトしている可能性を理論的な観点から指摘した論文．適当な仮定のもとで，これらの手法は単に入力画像の復元を行っていると解釈できる．結果的に，見た目では画像を“説明”しているように見えるが，これらの方法による説明はモデルの判断根拠を反映していない．

・Sanity Checks for Saliency Maps [57], ［実装 sanity_checks_saliency [58]］

NeurIPS’18 論文．どのハイライト法が妥当かを確認する方法を提案した論文．具体的には，モデル内部のパラメータを意図的に乱数に置き換えたり，データのラベルをランダムに付け替えて学習することで，“明らかにダメなモデル”を構築する．このダメなモデルでのハイライトと，真っ当なモデルでのハイライトを比較する．このとき，両方のモデルで似たハイライトを生成する方法は，モデルやデータの破壊を見ていない（モデルやデータの実態を反映していない）という意味で，良いハイライト法ではないと判断できる．

・Evaluating Feature Importance Estimates [59]

ハイライト法の良さを定量的に評価する指標を提案した論文．訓練画像についてハイライトを計算しハイライト箇所をマスクすることで，マスク付きの訓練画像を用意する．このマスク付き訓練画像で再度モデルを学習した際に，精度が元のモデルからどれだけ低下するか，でハイライト法の良さを測る．重要領域を適切に特定できる良いハイライト法では，マスク後の画像に認識に必要な情報が残らないため，再学習後のモデルの精度が大きく低下する．

7. 説明の信頼性

XAI の大きな目的の一つは「ブラックボックスな AI は信頼できない」という不信の払拭である．しかし，以下の論文では説明そのものの信頼性に問題がある可能性が指摘されている．

・Interpretation of Neural Networks is Fragile [60], ［Tensorflow 実装 InterpretationFragility [61]］

AAAI’19 論文．深層学習モデルの説明法（画像ハイライト，影響関数）に対して adversarial example が生成可能なことを示した論文．画像のハイライト部位を変えるような微小ノイズを実際に生成できることを実証した．

・Attention is not Explanation [62], ［PyTorch 実装 AttentionExplanation [63]］

NAACL’19 論文．自然言語処理で広く使われている Attention 機構が，単語間の関係性を適切に説明していない可能性を指摘した論文．具体的には，Attention の重みが勾配などの他の説明結果とあまり相関しないことを報告している．加えて，同じ予測をしながらも全く異なる Attention の重みをもつモデルが存在することを実証した．

・Fairwashing：The Risk of Rationalization [64]

ICML’19 論文．利己的なユーザが自身の主張を正当化するために恣意的な説明を生成できる可能性を指摘した論文．実際は性別や人種などに基づく差別的な意思決定をするモデルが，あたかも公平であるかのような

説明を生成できることを実証した．

8. お わ り に

XAI に関する近年の進展および研究トピックの広がりについて紹介した．最後に，今後の研究の展望について筆者の意見を述べて本記事を締めくくりたい．

- **知見の蓄積**

  数多くの説明法が提案されてきているが，どのような問題やデータにはどの手法を使うのが良いか，といった知見の蓄積はいまだ十分とはいえない．今後，XAI 技術の実問題への適用が広がることで，これら知見の蓄積が進むことを期待したい．現状では，試しに A という問題やモデルに B という手法を適用してみた，という研究レベルの報告が中心である．今後は，実運用まで見越したうえでの手法の開発や検証が重要になると考える．

- **説明の“良さ”の評価法の確立**

  本記事で述べたとおり，どの説明法が良いか，そもそも説明は信頼できるのか，といった基本的な問いへの関心が高まっている．今後は，説明の“良さ”をどのように評価・担保するか，といった方向の研究の重要性が増すと考える．

[1] https://www.ai-gakkai.or.jp/my-bookmark_vol33-no3/
[2] https://www.darpa.mil/program/explainable-artificial-intelligence
[3] https://www8.cao.go.jp/cstp/tyousakai/jinkochino/7kai/siryo3.pdf
[4] http://www.soumu.go.jp/main_content/000499625.pdf
[5] http://www.soumu.go.jp/menu_news/s-news/01iicp01_02000072.html
[6] http://www.soumu.go.jp/main_content/000564147.pdf
[7] http://www.meti.go.jp/main/yosangaisan/fy2019/pr/ip/sangi_06.pdf
[8] http://www.jst.go.jp/crds/report/report01/CRDS-FY2018-SP-03.html
[9] http://www.keidanren.or.jp/policy/2019/013_honbun.pdf
[10] https://ieeexplore.ieee.org/document/8466590
[11] https://asd.gsfc.nasa.gov/conferences/ai/program/003-XAIforNASA.pdf
[12] https://www.slideshare.net/SatoshiHara3/ss-126157179
[13] https://www.youtube.com/watch?v=Fgza_C6KphU&feature=youtu.be
[14] https://stair.center/archives/1858
[15] https://speakerdeck.com/line_developers/machine-learning-and-interpretability
[16] https://www.slideshare.net/kato_kohaku/how-to-use-in-r-modelagnostic-data-explanation-with-dalex-iml
[17] https://github.com/katokohaku/compareModels_with_MLR-IML
[18] https://twitter.com/kato_kohaku
[19] https://qiita.com/icoxfog417/items/8689f943fd1225e24358
[20] https://twitter.com/icoxfog417
[21] https://www.slideshare.net/iTooooooooooooT/six-abeja
[22] https://christophm.github.io/interpretable-ml-book/index.html
[23] https://xaitutorial2019.github.io/
[24] https://www.kaggle.com/dansbecker/use-cases-for-model-insights
[25] https://nips.cc/Conferences/2018/Schedule?showEvent=10986
[26] https://eli5.readthedocs.io/en/latest/
[27] https://github.com/christophM/iml
[28] https://christophm.github.io/interpretable-ml-book/index.html
[29] https://www.slideshare.net/kato_kohaku/how-to-use-in-r-modelagnostic-data-explanation-with-dalex-iml
[30] https://github.com/pbiecek/DALEX

[31] https://www.slideshare.net/kato_kohaku/how-to-use-in-r-modelagnostic-data-explanation-with-dalex-iml
[32] https://github.com/pbiecek/xai_resources
[33] https://www.kdd.org/kdd2017/papers/view/interpretable-predictions-of-tree-based-ensembles-via-actionable-feature-tw
[34] https://github.com/gtolomei/ml-feature-tweaking
[35] https://github.com/katokohaku/featureTweakR
[36] http://setten-qb.hatenablog.com/entry/2017/10/22/232016
[37] http://kato-kohaku-0.hatenablog.com/entry/2018/01/22/212205
[38] https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982
[39] https://github.com/marcotcr/anchor
[40] https://papers.nips.cc/paper/8219-human-in-the-loop-interpretability-prior
[41] http://proceedings.mlr.press/v80/kim18d.html
[42] https://beenkim.github.io/slides/TCAV_ICML_pdf.pdf
[43] https://github.com/tensorflow/tcav
[44] https://arxiv.org/abs/1806.10574
[45] https://link.springer.com/chapter/10.1007/978-3-319-46493-0_1
[46] https://github.com/LisaAnne/ECCV2016
[47] https://yoshitakaushiku.net/index_ja.html
[48] https://www.ai-gakkai.or.jp/my-bookmark_vol32-no1
[49] https://www.slideshare.net/YoshitakaUshiku/deep-learning-73499744
[50] https://papers.nips.cc/paper/7272-real-time-image-saliency-for-black-box-classifiers
[51] https://github.com/PiotrDabkowski/pytorch-saliency
[52] https://openreview.net/forum?id=Hkn7CBaTW
[53] https://github.com/albermax/innvestigate
[54] http://proceedings.mlr.press/v80/chen18j.html
[55] https://github.com/Jianbo-Lab/L2X
[56] http://proceedings.mlr.press/v80/nie18a.html
[57] http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps
[58] https://github.com/adebayoj/sanity_checks_saliency
[59] https://arxiv.org/abs/1806.10758
[60] https://arxiv.org/abs/1710.10547
[61] https://github.com/amiratag/InterpretationFragility
[62] https://arxiv.org/abs/1902.10186
[63] https://github.com/successar/AttentionExplanation
[64] https://arxiv.org/abs/1901.09749