特集
Special Feature

［AI 判断の根拠を説明する XAI を使いこなす］

5 説明可能な AI を身近にするためのディープラーニングツール

鈴木健二　ソニーグループ（株）R&D センター

概要

説明可能な AI を使いこなすためには，専門的な知識やプログラミング技術を必要とし敷居が高い．説明可能な AI とは AI の判断根拠を人間が理解できるように可視化する技術である．本稿では，AI の判断根拠の可視化や認識精度の向上を GUI 環境にて手軽に操作できるディープラーニングツール「Neural Network Console」を紹介する．

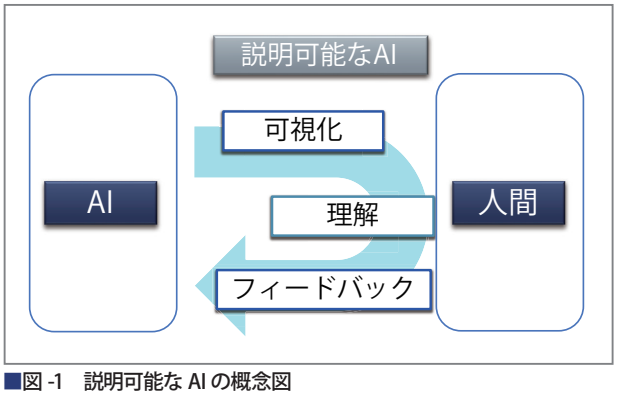
AI 倫理と説明可能な AI

社会の動向

AI の進展は目覚ましく，さまざまな分野において AI の活用が進んでいる．その中で，AI が人権を侵害する倫理的な問題も生じている．AI の判断結果は，不公平な予測を伴う場合もあり，またどのような理由で判断したのかが分からないこともある．大規模なディープラーニングによって作られた AI は，ブラックボックスと言われ，人間が AI の判断根拠を理解することが難しい．AI の判断根拠の説明は，AI を運用する人間にとっての重要な技術的課題であり，社会的にも重要視されている．あらゆる分野において AI の利活用が進む中で，AI の判断理由についての説明が求められている．

技術の位置づけ

このような背景のもと，説明可能な AI の研究開発が盛んになってきた．説明可能な AI とは，AI の判断根拠を人間に分かるように説明する技術である．図 -1 は，説明可能な AI の概念図を示す．人間が説明インタフェースを通じて，AI を理解することができる．AI の判断根拠は，説明可能な AI によって可視化され，人間が理解できるようになる．具体的には，AI がどの部分に着目して判断したのかを人間が知ることができる．また，AI が適切な判断根拠を示してない場合，人間が意図する方向へ AI に対してフィードバックを与えて改善することができる．説明可能な AI は，あらゆるモデルにおいて判断根拠の可視化のみならず，AI のデバッキング用途や，AI 倫理に沿った AI へと向けてフィードバックをすることができる．



■図 -1　説明可能な AI の概念図

ディープラーニングツール

身近なツールの必要性

あらゆる分野において AI の利活用が進む中で，AI の判断についての説明が求められている．しかしながら，説明可能な AI を使いこなすには , 従来のディープラーニングツールによるプログラミング技術や専門的なスキルが必要とされる．そのため産業界のさまざまな現場へ説明可能な AI を導入するには，敷居が非常に高いという問題がある． そこで，本稿では説明可能な AI を専門的なスキルやプログラミング技術を必要とせず，直観的に利用できるツールを紹介する．本ツールを使うことで，専門的なスキルやプログラミング技術を持たない人でも AI の判断根拠を理解することができる．

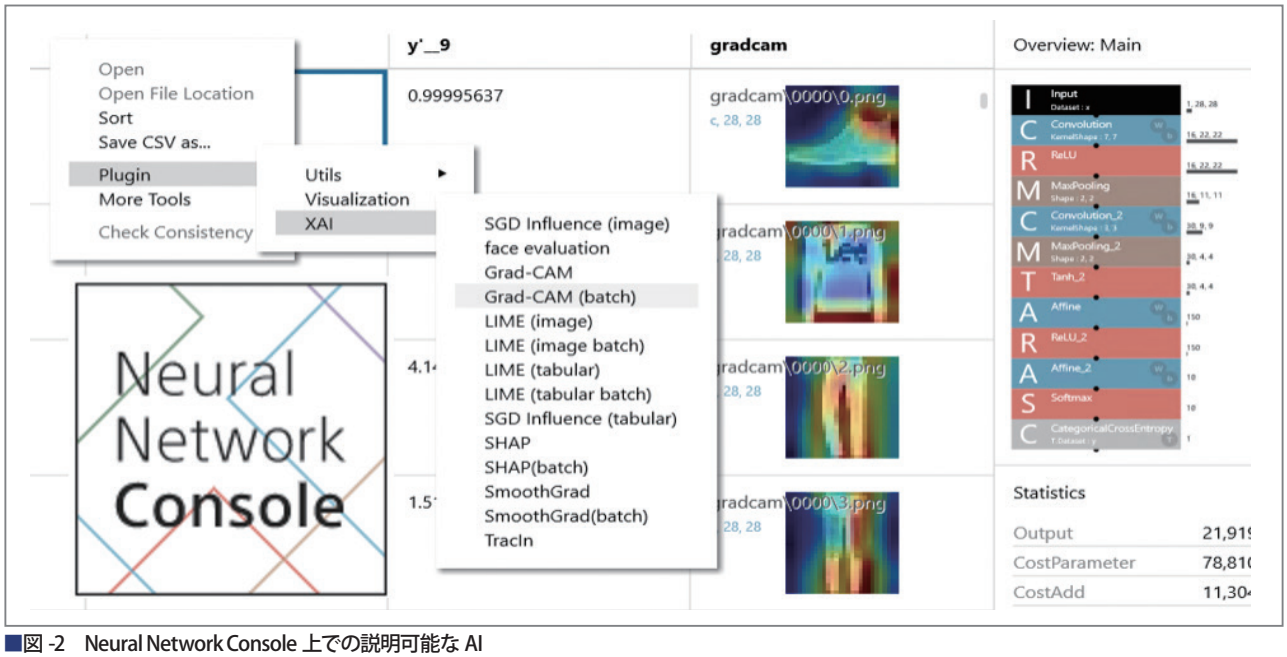
Neural Network Console

Neural Network Console（ニューラルネットワークコンソール）[☆1] は，効率的な AI 技術の開発を実現するために 2017 年 8 月にソニーがリリースした Windows 上で動作する無償のディープラーニングツールである．Neural Network Console は，ソニーが開発したディープラーニングフレームワーク Neural Network Libraries をコアとして，GUI 環境にて動作するラッパーツールである．この Neural Network Console は，ニューラルネットワークを視覚的に確認しながら簡単に設計できる特徴がある．また，設計したネットワークを GPU による高速学習や評価をすることができる．2018 年 5 月には，最大 8 台までのマルチ GPU での学習ができる有償の Neural Network Console のクラウドサービスの提供も開始された．

説明可能な AI プラグイン

Neural Network Console は，説明可能な AI をプラグインとして実装している．AI 開発者は，説明可能な AI を GUI 環境にて簡単に利用することができる．図 -2 は，Neural Network Console 上のスクリーンショットである．AI 開発者は，AI を設計し学習した後に，説明可能な AI プラグインを手軽に実行することができる．Neural Network Console の特徴は，設計，学習，説明可能な AI を含む評価

☆1 https://dl.sony.com/ja/



■図 -2 Neural Network Console 上での説明可能な AI

を一貫して同じソフトウェア上で実行できるところにある．図 -2 は，Neural Network Console にプラグインとして実装されている説明可能な AI によって，Convolution Neural Network（CNN）の画像分類での判断根拠をヒートマップで可視化した結果である．

最新版の Neural Network Console の説明可能な AI のプラグイン☆2 は，オープンソースソフトウェアとして公開されている．AI 開発者は，このオープンソースソフトウェアとして公開されているプラグインをダウンロードして，Neural Network Console へインストールすることで，最新の説明可能な AI の機能を使うことができる．また，AI 開発者自らが Neural Network Console のプラグインを作り，開発に参加することもできる．

判断根拠の可視化

代表的な手法

画像分類での AI の判断根拠の可視化を例に取り，代表的なアルゴリズムを解説していく．Grad-CAM[1] は，画像中の判断根拠となる個所をヒートマップ表示する技術である．Grad-CAM[1] の原理は，Convolution Neural Network（CNN）に対する入力勾配の大きな場所に注目することよって，判断根拠となる情報を可視化する方法である．Grad-CAM[1] は，予測クラスごとに関連の深い領域を可視化することができる．図 -2 は，この Grad-CAM[1] を適用し，画像分類での AI の判断根拠をヒートマップにて表示した例である．LIME[2] は，モデルを局所的に線形モデルで近似することによって判断根拠を示す方法である．その際に，LIME[2] は 入力に対して乱数を使いランダムな摂動を加えたデータを発生させる．SHAP[3] は，ゲーム理論の Shapley 値を求める手法を使って各特徴量の寄与度を計算する手法である．SmoothGrad[4] は，入力画像にガウシアンノイズを載せ，複数回の勾配計算をした後に平均を取ることによって，判断根拠となる個所の可視化画像を生成する手法である．このように，さまざまな説明可能な AI の手法が提案されている．各手法は，それぞれの論拠に基づいたアルゴリズムであり，説明の仕方にはいくつもの方法があることに留意する必要がある．

判断根拠の可視化への適用例

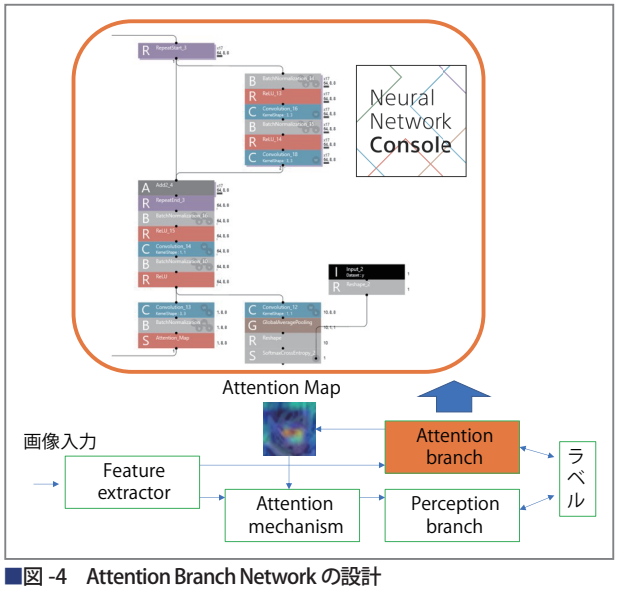
説明可能な AI を GUI 環境にて利用できる Neural Network Console を利用し，画像分類での AI の判断根拠へ適用した例を紹介する．この説明可能な AI を使い，図 -3 に示すように画像分類に おける判断根拠を可視化した．データセットは，訓練データ 60,000 枚，評価データ 10,000 枚からなる 10 クラスの白黒画像で構成されている Fashion-MNIST を利用している．AI は，LeNet にて訓練データを学習した．

図 -3（a）Original は，Ankle boot の評価データである．この AI において，AI の判断根拠は，どのようになっているだろうか．図 -3（b）は，Grad-CAM[1] による結果である．ヒートマップにて判断根拠が可視化され，Ankle boot の上部を主な根拠として画像分類がされたことが理解できる．このヒートマップの赤い部分は，判断根拠の重要な所を示している．次に，図 -3（c）は，LIME[2] による判断根拠の可視化である．LIME[2] は画像を分割し，その判断に最も寄与している画像の部分を表示する．Ankle boot の上部で，足の入る場所から左側のあたりを着目していることが分かる．次に，図 -3（d）

■図 -3　画像分類での AI モデルの判断根拠の可視化

☆2 https://github.com/sony/nnc-plugin

は，SHAP[3] による判断根拠の可視化である．赤いピクセル表示は，判断にプラスに寄与した部分であり，一方で，青い部分は判断へマイナスに寄与した部分である．LIME[2] と同様に，SHAP[3] もおおよそ足の入る場所から左側のあたりとその周辺を着目していることがうかがえる．最後に，図 -3（e）は Smooth Grad[4] による判断根拠の可視化である．レントゲン写真のような表示であり，白い部分が判断根拠を示している．Smooth Grad[4] も同様におおよそ足の入る場所から左側のあたりを根拠としている．それぞれの手法による判断根拠の可視化は，必ずしも一致しない．このように説明手法が異なれば，得られる説明も異なる．これらの 4 つの手法では，注目個所は Ankle boot の足の入る部分とその左側上部であった．判断根拠の可視化をした Ankle boot に似た画像には，ほかのクラスである Sandal が存在する．Ankle boot と Sandal の違いは，アキレス腱やくるぶしを覆うかどうかである．Ankle boot の特徴であるこの足の入る上部やその上部左側の辺りは，先ほどの 4 つの手法での説明にて可視化された部分である．このように，画像分類において AI がその特徴的な部分に着目して画像を識別したのではないかと考えられる．AI 開発者は，代表的な複数の手法を多角的に評価することも大切である．Neural Network Console は，Grad-CAM[1]，LIME[2]，SHAP[3]，Smooth Grad[4] を単一のソフトウェアで提供している．



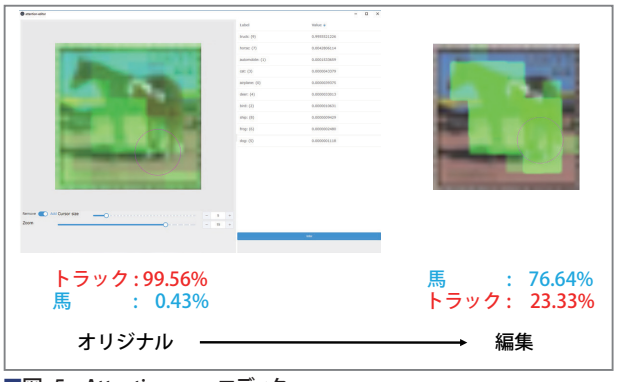
■図 -4　Attention Branch Network の設計

判断根拠の活用

先述の可視化手法は，AI の判断根拠を人間が知ることができ有用である．だが，その判断根拠から AI の精度を向上させるのは，困難を伴う問題がある．そこで，AI の判断根拠を可視化するだけでなく，その判断根拠を利用して AI の潜在的な性能を引き出し，AI の精度を向上させる手法として Attention Branch Network が提案されている[5]．

図 -4 に示すように，Attention Branch Network は，Feature extractor，Attention branch，Attention mechanism，Perception branch から構成される．Attention Branch Network を設計するためには，従来のディープラーニングフレームワークではプログラミング技術や専門的な知識を必要としていた．Neural Network Console は，この Attention Branch Network を図 -4 に示すように GUI 環境でブロックをつなげていくことで，設計することができる．

さらに，Attention map を変化させることによって，どのように推論値が変化するのかを人間が知ることができるツール Attention map エディターも Neural Network Console から利用できる．図 -5 の



■図 -5　Attention map エディター

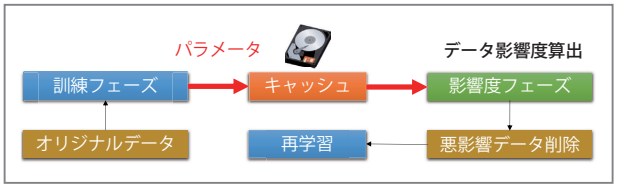
右側の画像は，Attention map エディターにより判断根拠を編集したものである．緑色に塗った部分は，判断根拠と思われる個所を編集した個所である．この Attention map エディターは，AI 開発者が Attention と思われる個所の付加や削除が自在にでき，その際の推論値をインタラクティブに理解することができる．図 -5 の画像の正解ラベルは，馬である．図 -5 の下の数値は，各ラベルに対する推論値 (%) を示す. 図 5 左は, トラック 99.56 %, 馬 0.43 % と誤認識したオリジナル画像の Attention map の出力画像である．図 -5 左の出力画像は，馬の中心部のみならず，その周りも全体的に注目領域としていることが判る．そこで，Attention map エディターによる注目領域を編集し，推論値がどのように変化するのかを観察した．図 -5 右は，馬と思われる個所を注目領域となるように編集した．その結果は，馬が 76.64 %，トラックが 23.33% へ推論値が変化した．このように，馬の画像をトラックと誤判断したのは，馬以外の周辺部分を判断根拠としていたことが起因していたと考えられる．つまり，AI 開発者は，画像中の AI の判断根拠の場所と推論値の関係をインタラクティブに理解することができる．Neural Network Console は，ネットワークの設計，学習，Attention map の可視化，Attention map エディターによる編集まで，同一のツール上で一貫して GUI 環境にて利用できる．

AI へのデータ影響度

ここからは，各データが AI へ与える影響について述べる．本章では，AI へのデータの影響度を明らかにし，データクレンジングをする手法を解説する．データ影響度とは，AI へ各訓練データがどのくらい影響を与えているのかを定量的に示す手法である . ここで言うデータクレンジングは，データ影響度の算出結果から AI へ悪影響を及ぼすデータを削除する手法である．データクレンジングによりデータの質を向上させることで，AI の性能を向上させることができる．データについての専門的な知識がなくとも，悪影響を与えている訓練データを特定して取り除くことで，AI の性能向上ができることが報告されている [6]．その提案されたデータクレンジングの方法は，確率的勾配降下法（SGD）で学習された AI において，SGD の学習ステップをたどることによって，影響力のある訓練データを推定する方法である．この方法を用いることで，AI 開発者にデータに関する広範な知識がなくても AI へ影響を与える訓練データを定量的に評価できる．図 -6 は，このデータクレンジング手法のワークフローを示す．まず，SGD を用いた訓練フェーズにて AI のパラメータを一時的にキャッシュに保存する．影響度フェーズにて，データ影響度を算出する．スコアリングされた各データから悪影響データを削除し，再学習をすることで AI を再構築することで，データクレンジングによる AI を作ることができる．

Neural Network Console へのデータ影響度演算での実装においての問題は，一時的に保存（キャッシュ）する AI のパラメータの量である．SGD を用いた訓練フェーズにおいて，キャッシュする必要がある AI のパラメータは，膨大である．そこで，Neural Network Console の実装は，データ影響度の計算に最後のエポックの最終パラメータだけを採用している．キャッシュサイズを縮小することによって，Neural Network Console へのデータ影響度演算の実装がされている☆3．そして，データクレンジングによる AI の精度も向上も確認されている．

☆3 https://arxiv.org/abs/2103.11807v2



■図 -6　データ影響度を利用したデータクレンジングのワークフロー

説明可能な AI の課題

説明可能な AI が身近になったとはいえ，さまざまな課題が残っている．

1 つ目は，説明手法の選択という点である．説明手法は，先に代表的な例を解説した通りいくつもの手法があり，ほかにも多くの説明手法が存在する．それぞれの説明手法は，同じモデルに対してそれぞれ異なる結果を示す場合がある．どの説明手法を使うのが適切であるのかは，人間が選ばなければならない．AI 開発者が説明可能な AI を利用するにあたっては，代表的な複数の手法を多角的に評価することが大切である．

2 つ目は，手法によって計算コストが多大な場合もあり，実際の運用において注意を要するという点である．計算量が大きいことは，開発時間をさらに要し，リードタイムできるだけ短縮したい AI 開発において重要な課題である．LIME[2] は，ランダムにデータ発生させ演算する手法のためある程度の計算時間を要するが，Grad-CAM[1] は，比較的高速に動作する利点がある．

3 つ目は，乱数パラメータを利用する説明手法では安定性に注意を要するという点である．先に解説した LIME[2] は，入力に対して乱数を使いランダムな摂動を加えたデータを発生させる手法であり，実行ごとに説明内容が変化することがある．

最後に，説明可能な AI は，AI 倫理の問題を技術的に直接解決するものではなく，あくまでも人間が AI 倫理を判断する際のツールに過ぎない．つまり，説明可能な AI が身近になったとはいえ，過剰な期待は禁物である．AI 倫理の観点で求められているものは，AI 開発者だけが AI を理解できるようになることではなく，AI 利用者が納得できるように AI を説明することである．説明可能な AI は，AI の説明責任を果たすために有用な一技術に過ぎない．

説明可能な AI が身近に

本稿は，説明可能な AI を専門的なスキルを必要とせず直観的に利用できるツールである Neural Network Console を紹介した．このツールは，画像分類において，AI の判断根拠の可視化のみならず，AI の判断根拠を活用した精度の向上や推論の変化，AI へ与えるデータの影響度も明らかにすることもできる．

今後，AI はさまざまな分野へ広がっていくことが想定される．あらゆるタスクにおいて，説明可能な AI が利活用されることが期待される．説明可能な AI を手軽に利用できることは，AI の加速的な普及に役立つと考えられる．

参考文献

1）Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. : Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, Proceedings of the IEEE International conference on computer vision, 618 (2017).
2）Riberio, M. T., Singh, S. and Guestrin, C. : "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, 1135 (2016).
3）Lundberg, S. M. and Lee, S.-I. : A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems 30, 4768 (2017).
4）Smilkov, D., Thorat, N. Kim, B., Viegas, F. and Wattenberg, M. : SmoothGrad : Removing Noise by Adding Noise, arXiv:1706.03825 (2017).
5）Fukui, H., Hirakawa, T., Yamashita, T. and Fujiyoshi, H. : Attention Branch Network : Learning of Attention Mechanism for Visual Explanation, Computer Vision and Pattern Recognition,10705 (2019).
6）Hara, S., Nitanda, A. and Maehara, T. : Data Cleansing for Models Trained with SGD, Advances in Neural Information Processing Systems 32, 4215 (2019).

（2022 年 4 月 28 日受付）

■鈴木健二（正会員） Kenji.B.Suzuki@sony.com

東京大学生産技術研究所，フランス IEMN を経て，ソニー（株）へ入社．現在，ソニーグループ（株）R&D センター．シニアマシンラーニングリサーチャー．博士（工学），学士（法学）．説明可能な AI，機械学習での公平性，AI 倫理，データ流通について，研究開発リーダーとして従事．