XAI（Explainable AI）

南 喜碩
Yoshimichi MINAMI

1　はじめに

近年，第 3 次 AI ブームが到来し，再び AI 技術に関心が高まっている．第 3 次 AI ブームを牽引しているのは機械学習（Machine Learning）の進化である．進化の背景には，ディープラーニング（Deep Learning）が用いられている．ディープラーニングは，自ら膨大なデータを学習し，自律的に答えを導き出すという特性上，内部構造を明らかにすることは困難である．そのため，結果に対する過程や根拠が説明不可能となり，ブラックボックス問題が生じている．ブラックボックス問題とは，学習の過程が不透明なために，結果に対する根拠が説明不可能な問題である．特に医療や金融，製造などの事業分野では，サービスに対して不信感や不安感をもたらすため，AI の利用において非常に重要な課題である．

そこでブラックボックス問題を解決するため，AI が導出した結果に根拠の説明能力を備えた AI が XAI（Explainable AI）である．

2　ブラックボックス問題

ブラックボックス問題とは，学習の過程が不透明なために，結果に対する根拠が説明不可能な問題である [1)]．ブラックボックス問題は，ディープラーニングの内部構造に隠れ層を持つことにより生じる．学習結果は，隠れ層におけるノードの重みにのみ反映されるため，学習の過程をユーザが理解できないためである．そこで，意思決定に責任が問われる場面や結果に対する根拠が必要となる場面では特に問題視されている．

3　XAI（Explainable AI）

3.1　XAI の概要

XAI とは，機械学習の精度と解釈性を両立する AI である．したがって XAI は，根拠の透明性や説得性が求められる分野で活躍可能な AI である．XAI の実現には，トレードオフの関係を持つ機械学習方式の精度と解釈性の改良が必要である．Fig. 1 に機械学習方式の精度と解釈性による関係図を示す．ディープラーニングは，解釈性が低い一方，AI により導かれる結果の精度が高い．そこで，Fig. 1 に示す A において解釈性の向上を行い XAI の実現を可能としている．また，線形回帰や決定木は，解釈性が高い一方，AI により導かれる結果の精度が低い．そこで，Fig. 1 に示す B において，結果の精度の向上を行い XAI の実現を可能としている．以上より，XAI の実現には様々なアプローチが存在する．

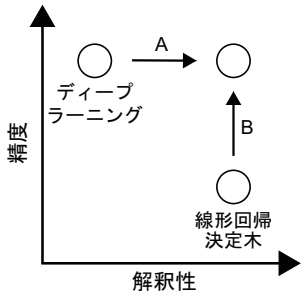

Fig.1　機械学習方式の精度と解釈性による関係図

3.2　XAI のアプローチ

3.2.1　Deep Explanation

Deep Explanation とは，ディープラーニングを状態解析し，解釈性を向上するためのアプローチである．Deep Explanation は主に図を用いたアテンションヒートマップや自然言語処理において用いられている．

Fig. 2 にアテンションヒートマップの生成手順を示す．はじめに，入力画像を畳み込みニューラルネットワーク（以後，CNN）を用いて画像解析を行い，画像の特徴量を算出する．次に，算出した特徴量を特徴マップに格納し，CNN の画像解析の結果から画像の対象物がどのクラスに属するのかを判別する．そして，判別した対象のクラスの特徴のみを特徴マップから抽出し，抽出された特徴量の多寡に応じて色を変化させ，アテンションヒートマップの生成を実現する．

以上の手順を行うことにより，色や濃淡で注目した部分が可視化され，結果に対する根拠が推定可能となる．

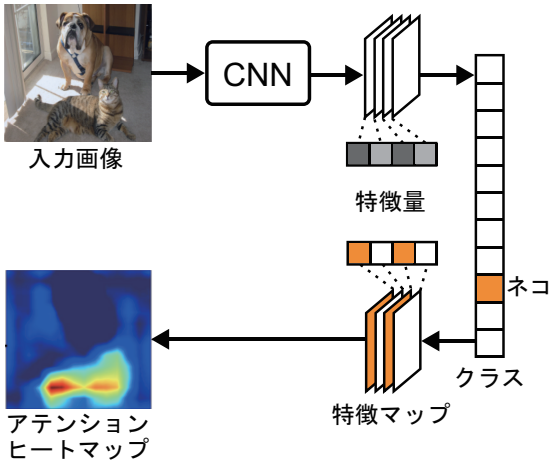

Fig.2　アテンションヒートマップの生成手順

3.2.2 Interpretable Models

Interpretable Models とは，精度が低く，解釈性の高いホワイトボックス型の機械学習に対し，適切な数学モデルを導出し適用することで精度を向上させるアプローチである．ホワイトボックス型の機械学習は線形回帰や決定木，Random Forest などが挙げられる．ホワイトボックス型機械学習の問題点は 2 点存在する．1 点目は入力データが複雑な場合，規則性の導出が容易でないことである．2 点目は，近似のモデルが線形近似であるため局所的な変化に対応が容易でないことである．

そこで，入力データが複雑な場合においても，正しい規則性を予測可能な XAI を実現するアプローチが Interpretable Model である．

4 XAI の適用例

4.1 Deep Explanation

Deep Explanation を用いた適用例として，がん組織片のアテンションヒートマップによる可視化が挙げられる．アテンションヒートマップの生成には，Grad-CAM と呼ばれるシステムが用いられている [2]．Fig. 3 に肺がん組織片の画像およびアテンションヒートマップを示す．肺がんの場合 Fig. 3 に示すがん組織片の画像を読み込み，アテンションヒートマップを生成する．アテンションヒートマップでは，腺がんは赤色，扁平上皮がんは青色など，注目した場所を色の違いにより可視化する．可視化することにより，解析した 2 つのがん組織片を総合的に判断して，肺がんと診断することが可能である．

以上により，どの部分に着目し，結果に至ったかを可視化することで，患者が納得して説明を受けることが可能となる．

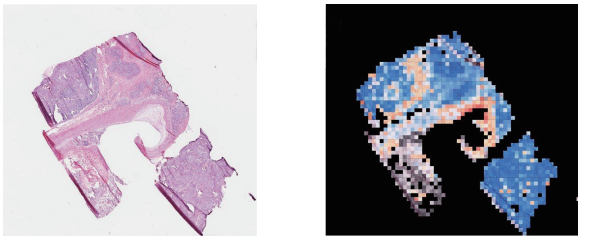

Fig.3 肺がん組織片の画像およびアテンションヒートマップ

4.2 Interpretable Models

Interpretable Models を用いた適用例として，ビルの電力消費の予測モデルに利用されている．予測モデルには，異種混合学習技術が使用されている．異種混合学習技術とは，複雑なデータにおいても高い精度で予測式の推定を可能にする技術である [3]．

$$p(x|\pi,\mu,\sigma^2,\lambda)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\sigma_k^2)+\pi_E f(x|\lambda) \quad (1)$$

予測式の推定は，(1) 式で示した新しい数学的モデルにより実現可能となっている．従来の数学的モデルとの違いは，(1) 式の第 2 項にある指数分布を加える点である．指数分布を加えることで，複雑なデータにおいても高い精度で予測式の推定を実現している．

Fig. 4 に単一モデルと Interpretable Models におけるビルの電力消費予測を示す．従来の単一モデルのみでは，予測するモデル式が 1 つで構成されている．そのため，電力消費の値に急激な変化が起こる土曜日や日曜日において実際の値との一致率の減少が見られる．一方，Interpretable Models では，状況に応じた複数のモデルを扱うことが可能である．Fig. 4 に示す平日はモデル A，土曜日と日曜日はモデル B のように状況に応じてモデルを適用する．適宜最適なモデルを適用することで，土曜日や日曜日においても高い一致率を実現している．

以上より，ビルの電力消費予測が精度の高い単純なピークカットではなく，知的に節電対策が可能となった．

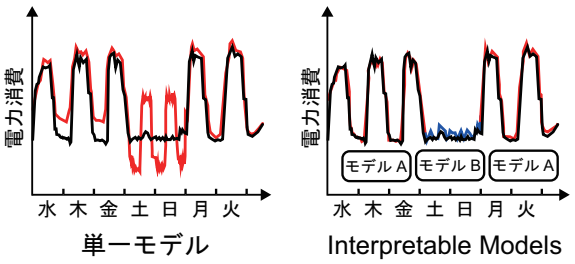

Fig.4 単一モデルと Interpretable Models におけるビルの電力消費予測

5 今後の展望

今後も 2 つのアプローチを用いて機械学習の精度と解釈性の両立に向けた XAI の研究が進むと考える．特に，ブラックボックス問題によって AI の浸透が阻まれている業界に XAI の浸透が期待される．医療業界では，脳の MRI 画像をもとにアテンションヒートマップを生成し，病名診断の根拠を推定することが実現できると考える．また，自動車業界では，自動運転時における AI 判断の説明をアテンションヒートマップで視覚的に示す研究が進められている．アテンションヒートマップを用いることで画像の着目点が可視化され，システム改善に役立つと考える．以上より，XAI の説明能力が向上することで，人が実施していた説明の機会を XAI が担うことができ，利用者が説明する負担の削減が可能となる．

参考文献

1) 福島俊一，藤巻遼平ら：ビッグデータ×機械学習の展望 最先端の技術的チャレンジと広がる応用，Vol.60, No.8, pp. 543 - 554（2017）.
2) CNN の可視化手法 Grad-CAM の紹介—SlideShare, https://www.slideshare.net/ToshinoriHanya/cnngradcamcnn，参照 Apr.16,2019
3) 異種混合学習技術—NEC, https://jpn.nec.com/press/201206/images/2202-01-01.pdf，参照 Apr.17,2019