《第70回》深層学習モデルを説明するための技術開発

渡　邊　千　紘

1.　はじめに

深層学習モデルはこれまで，画像，音声，自然言語などさまざまなデータに関する課題において高い性能を達成してきたが，その挙動を人間が理解できるかどうかという解釈性の観点において課題があった．本稿では，著者がこれまでに取り組んできた深層学習モデルの説明に関する研究について，一般の深層学習モデルに適用可能な技術（2章）と，音声処理の課題に特化した技術（3章）の両面から紹介する．

2.　深層学習モデルのモジュール分解

一般の深層学習モデルに対する説明手法を大まかに分類すると，(1) 深層学習モデル全体を1つのブラックボックス関数として捉えた時の入出力の関係を解析する大局的アプローチと，(2) 深層学習モデルを構成する部品（ユニットや層）ごとに役割を解析する局所的アプローチが存在する．著者はさらに，(1) と (2) の間をつなぐアプローチとして，学習済み深層学習モデルのユニットを互いに類似した役割をもつ者同士のグループ（モジュール）に分け，各モジュールの役割を明らかにするというアプローチを提案してきた（**図 1**）．

学習済み深層学習モデルにおけるモジュール構造の推定は，その各ユニットをそれぞれの特徴に基づいてクラスタリングする問題として捉えることができる．ここで，各ユニットの特徴量の定義や，クラスタリング手法を変えることにより，異なるモジュール構造が得られる．たとえば，学習済み深層学習モデルのネットワーク構造に着目し，隣り合う層のユニットとのつながり方が似ているユニット同士を同じモジュールに割り当てる方法を用いることで，各層のユニットをクラスタリングすることができる（**図 2**）[1]．

つぎに，抽出された各モジュールが予測において果たす役割を知るための手法として，各モジュールとモデルの入出力との関係を解析する手法を紹介する．たとえば図1のように，与えられた画像に写っているオブジェクトを認識する問題において，モデルへの入力は画像，出力は各オブジェクトの有無を表わしたものとなる．このとき，各モジュールに属するユニットの出力値が，入力画像のうちどの画素の値に大きく影響を受け，どのオブジェクトの認識結果に大きな影響を与えているかを定量評価することで，各モジュールの予測における役割を解釈することができる[2]．

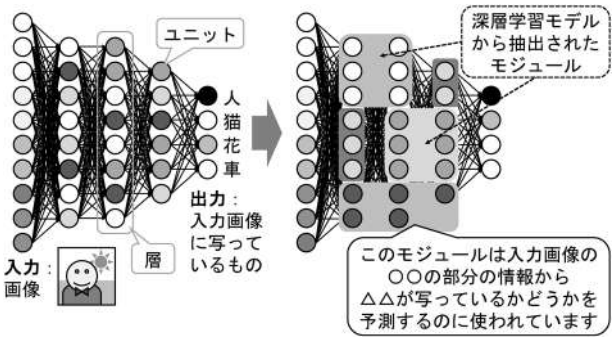

図 1　深層学習モデルのモジュール構造に基づく説明

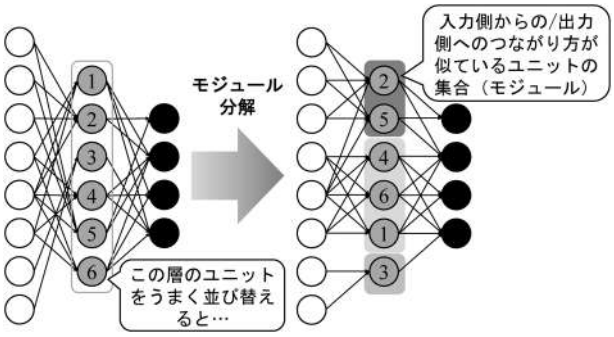

図 2　ネットワーク構造に基づくモジュール分解

JL 0011/24/6311–0665 ©2024 SICE
計測と制御　第 63 巻　第 11 号　2024 年 11 月号

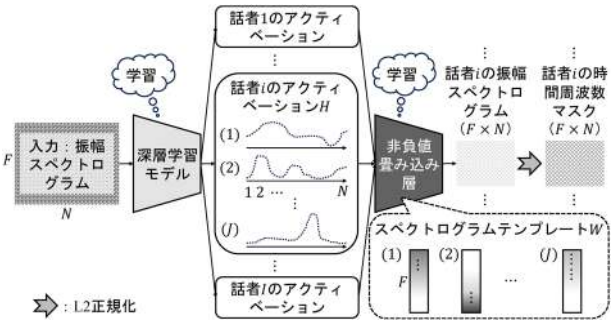

図 3　非負値スペクトログラムテンプレートの畳み込みとして解釈可能な深層音源分離モデルの構成

3. 音声分野における説明可能性

2 章では，一般の深層学習モデルに適用可能な技術を紹介したが，一方で扱うべきデータの種類や解くべき課題を一つに定めれば，その設定に特化した手法を開発するアプローチも考えられる．本章では，音声データに対する音源分離の課題において，解釈可能なモデルを構成し学習する手法を紹介する．

音源分離とは，複数人の話者からなる混合音声を単一の話者ごとの音声に分離する問題であり，具体的には以下のように定式化できる．まず，入力音声に短時間フーリエ変換を適用することで，音声の時間周波数表現（行方向が周波数，列方向が時間を表わす行列）を得る．この行列の周波数方向のサイズを F，時間方向のサイズを N とする．この $F \times N$ 個の各要素において，どの話者が支配的であるかを表わす時間周波数マスクを推定することを目指す．著者は，深層学習に基づく高精度な音源分離と解釈性を両立させるため，以下の構成を提案した（**図 3**）[3]．まず，混合音声の振幅スペクトログラムを，出力のサイズが $J \times N$（J は任意の自然数）かつ出力の全要素が非負値であるような任意の構成の深層学習モデルに入力し，特徴量 H を得る．一方，こちらも全要素が非負値であるような短時間のスペクトログラムのテンプレートを J 個用意しておく．このとき，H はこれらのテンプレート W が各時刻でどの程度使われているか（アクティベーション）を表わしたものであると捉えることができ，W と H の畳み込みに基づいて最終的な時間周波数マスクを得る構成をモデルの最終層に組み込んでおく．アクティベーション H を出力するモデルに加え，テンプレート W も学習により最適化することで，入力された混合音声を分離する過程を適切な非負値スペクトログラムテンプレートの組み合わせの形で表現し，解釈することができる．

4. おわりに

本稿では，深層学習モデルの挙動を説明するためのアプローチとして，一般の深層学習モデルにおけるモジュール構造を解析する手法と，音声データ処理に向けた解釈可能な深層学習モデルの構成法を紹介した．

そもそも解釈可能性の定義，つまり何が実現できれば解釈できたことになるのかということについては，満場一致の見解は得られていない．また，近年では説明手法自体の良し悪しを評価する方向性[4]や，説明手法が悪用される危険性を指摘した研究[5]も発表されている．そのため，説明手法を適用する際は，それがどのような場合に使える手法なのか，モデルについてどのような側面での情報が得られるものなのかということに注意しながら活用していく必要がある．発展を続ける解釈可能性の分野において，より広く・深く・安心して深層学習モデルの挙動を理解できるようになるための新手法の開発に，今後も取り組んでいきたい．

（2024 年 6 月 20 日受付）

参　考　文　献

1) C. Watanabe, K. Hiramatsu, and K. Kashino: Modular Representation of Layered Neural Networks, *Neural Networks*, **97**, 62/73 (2018)
2) C. Watanabe, K. Hiramatsu, and K. Kashino: Understanding Community Structure in Layered Neural Networks, *Neurocomputing*, **367**, 84/102 (2019)
3) C. Watanabe and H. Kameoka: X-DC: Explainable Deep Clustering based on Learnable Spectrogram Templates, *Neural Computation*, **33**–7, 1853/1885 (2021)
4) J. Adebayo, et al.: Sanity Checks for Saliency Maps, *NeurIPS*, **31**, 9505/9515 (2018)
5) U. Aivodji, et al.: Fairwashing: The Risk of Rationalization, *ICML*, **97**, 161/170 (2019)

［著　者　紹　介］

渡邊千紘（わたなべ ちひろ）君

2015 年東京大学大学院情報理工学系研究科システム情報学専攻修士課程修了．同年 NTT コミュニケーション科学基礎研究所入社，2022 年東京大学大学院情報理工学系研究科数理情報学専攻博士課程修了，現在に至る．機械学習，統計学の研究に従事．博士（情報理工学）．

E-mail: ch.watanabe@ntt.com

所属：NTT コミュニケーション科学基礎研究所
神奈川県厚木市森の里若宮 3–1