The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

脳内情報表現の融合による深層学習ネットワークの認識能力向上

Improving recognition performance of a deep neural network via integration with information representation in the brain

西田知史*1,2
Satoshi Nishida

西本伸志*1,2
Shinji Nishimoto

*1 情報通信研究機構
National Institute of Information and Communications Technology

*2 大阪大学
Osaka University

Abstract: Deep learning has recently shown splendid performance in pattern-recognition tasks, such as object identification. However, even using the state-of-the-art deep neural network, it is still difficult to predict human subjective judgements, such as preferences or impressions, from sensory patterns. Here we investigate whether the performance of a deep neural network in such pattern recognition improves by integrating brain representations into deep-learning feature representations. The feature representations of visual inputs in a deep neural network are transformed into those in the brain via their association pre-learned from measured brain response. Then, the transformed representations are used to estimate human cognition induced by the visual inputs. We demonstrate that the estimation performance improves when the brain representations are integrated. Thus, brain data integration can provide an effective way to extend the general applicability of deep learning in the estimation of human subjective judgements.

1. はじめに

機械学習は、脳の構造や振る舞いからヒントを得て、発展を遂げてきた。特に近年、脳の階層的情報処理を模倣する深層学習は、物体認識のような特定の問題において、人間を凌駕する性能を示している [He 16]。深層学習の強みは、未知の問題に対して、大量の教師データから、問題に適した特徴空間を自動的に獲得可能な点にある [LeCun 15]。この利点が、従来の機械学習手法を性能で上回る大きな要因だと考えられている。しかし、そのような脳を模倣した深層学習であっても、人間の感性や嗜好といった主観性の強い認知情報を、感覚入力パターンと結びつける問題においては、未だ高い性能を発揮できていない。この要因の一つとして、脳を模倣する深層学習といえども、現状の枠組みでは、そのような主観情報の特徴表現の獲得が難しい点が挙げられる。

一方で、人間の脳内にはそのような特徴表現が内在しており、感性や嗜好といった主観的判断に利用されている。近年、脳神経科学の研究分野において、脳計測データに基づき個人の脳内における情報表現空間をモデル化して、理解する試みが行われている [Güçlü 15、Nishida 15、Nishimoto 11]。また、その枠組みを用いて、計測脳応答から、感性のような主観的認知内容を解読する技術の開発も行われている [Nishida 18]。モデル化された情報表現空間は定量的に扱うことが可能で、他の工学システムに導入して利用することは難しくない。

そのような脳内情報表現のモデル化手法を応用して、我々は近年、深層学習ネットワークの特徴表現から脳内情報表現への写像を獲得および利用したうえで、任意の感覚入力に誘発される人間の知覚内容を推定するための手法を提案した [西田 18]。この提案手法で作成したモデルにより、深層学習ネットワークに含まれない特徴表現を、脳内情報表現への変換を介して補うことで、人間の主観的判断が強く影響する推定問題において、推定性能の向上がもたらされると予想する。本研究では、この予想を、実データを用いた2種類の推定問題において、提案手法と既存手法の推定性能の比較に基づき検証する。

2. 提案手法

2.1 深層学習特徴表現からの脳応答予測

本研究では、機能的磁気共鳴画像法（fMRI）により計測した脳応答パターンを脳内の情報表現とみなした。そして、任意の映像入力から生じる深層学習ネットワーク16層 VGG [Simonyan 15] の中間層活性化パターンから、同じ映像刺激により生じる脳応答パターンを予測するモデル（脳応答予測モデル）を構築した（図 1A）。なお、深層学習の中間層活性化パターンは、視覚入力に誘発される脳応答を精度良く予想することが、先行研究で報告されている [Güçlü 15]。

モデル構築には、符号化・復号化モデリング手法 [Naselaris 11] を利用した。符号化モデリングでは、任意の特徴空間から脳応答空間への写像を学習する。fMRI の計測単位であるボクセルごとの脳応答の系列を $\mathbf{R}$、感覚入力の系列を $\mathbf{S}$、その特徴表現の系列を $f(\mathbf{S})$とすると、以下の数式により表現されるモデル重み $\mathbf{W_e}$をリッジ線形回帰により推定する。

$$\mathbf{R} = f(\mathbf{S})\mathbf{W_e}$$

本研究における $f(\mathbf{S})$は、映像入力 $\mathbf{S}$ に対するVGGの中間層活性化パターンに相当する。これにより、一旦 $\mathbf{W_e}$ が学習された後は、新たな映像入力が得られれば、誘起される脳応答 $\mathbf{R}$ の予測が可能なモデルとして利用できる。

モデル学習のために、特徴表現の系列 $f(\mathbf{S})$として、fMRI 実験で被験者が視聴した映像の各フレームに対する、VGG の中間層活性化パターンの 1 秒ごとの最大値を算出する。モデル学習は各被験者の脳応答データを別々に用いて行う。また、VGG の中間層のうち 8 層（5 つのプーリング層、3 つの全結合層）のそれぞれを用いて、別々にモデルの学習を行い、各被験者で 8 個のモデルを作成する。そして、新しい映像入力に対する 8 個のモデルの脳応答予測結果を、モデル学習時に算出した予測精度（実測脳応答と予測脳応答の時系列に対して算出した相関係数）に基づき重み付けを行って平均し、最終的な脳応答予測結果を得る。

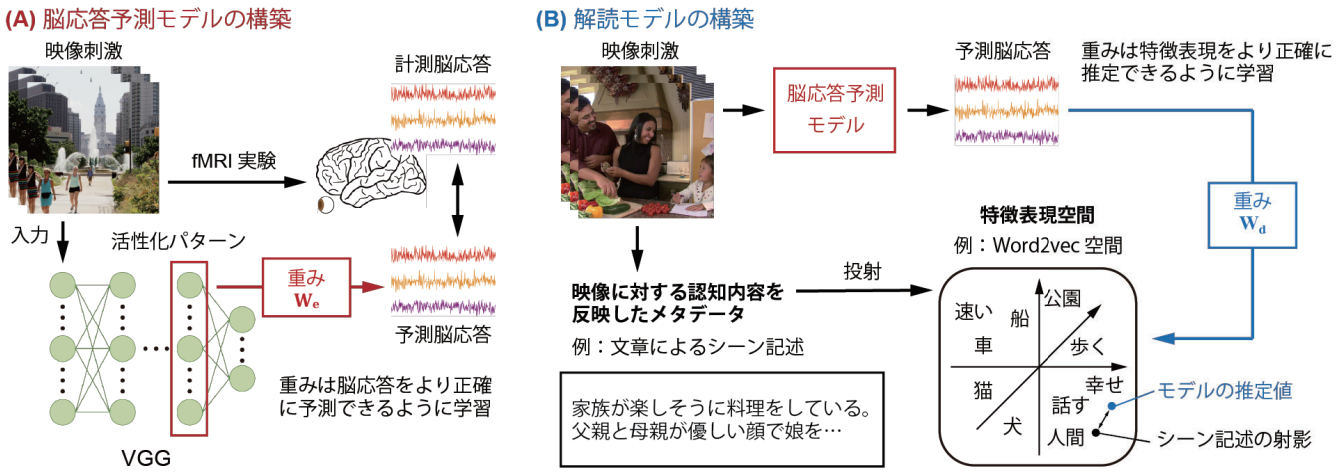

図 1：脳応答予測モデルと解読モデルの概要

2.2 予測脳応答からの認知内容解読

続いて、脳応答から映像に対する人間の認知内容を推定するモデル（解読モデル）の構築を行う（図 1 B）。これは復号化モデリングに相当する。このモデリングでは、符号化モデリングとは逆に、脳応答空間から特徴表現空間の写像を学習する。つまり、以下の数式により表現されるモデル重み $\mathbf{W}_d$ をリッジ回帰により推定する。

$$g(\mathbf{S}) = \mathbf{R}\mathbf{W}_d$$

本研究では、$\mathbf{R}$ として脳応答予測モデルが出力した予想脳応答を用いた。また、$g(\mathbf{S})$として映像に紐付いた、認知内容を反映するメタデータの特徴表現を用いた。そのうえで、$\mathbf{W}_d$ の学習を行った。この解読モデルは、一旦 $\mathbf{W}_d$の学習が完了すると、任意の新たな $\mathbf{R}$ に対して、認知内容と結びついた特徴表現 $g(\mathbf{S})$ を推定できるようになる。

本研究では、2 種類のメタデータの推定を目的として 2 種類の解読モデルを構築した。メタデータの 1 つは、映像に対して人手で付与した、1 シーンあたり 50 文字以上の日本語のシーン記述である。シーン記述は、1 シーン（1 秒）あたり 5 名から取得し、記述内にはシーンを表現する多様な客観的・主観的記述が含まれていた。この記述を単語に分解し、Wikipedia コーパスから学習した word2vec 空間 [Mikolov 13] に投射して、100 次元のベクター表現を得た。これを特徴表現 $g(\mathbf{S})$)として用い、解読モデルの構築と評価を行った。

もう 1 つのメタデータは、Web 上で集計された、映像に対する視聴者の嗜好を反映する行動指標である。本研究で用いた映像は、Web 上で公開された広告映像であり、映像にアクセスした視聴者のうち、映像をクリックして広告元のサイトに移動した割合（クリック率）と、映像を最後まで視聴完了した割合（視聴完了率）のデータが各広告に紐付いている。これらのマスデータを特徴表現 $g(\mathbf{S})$として、解読モデルの構築と評価を行った。

3. MRI 実験

3.1 被験者

40 名の被験者（男性 25 名、女性 15 名、20～61 歳）が fMRI を用いた脳計測実験に参加した。全被験者から実験前に書面で同意を得た。また、実験プロトコルは情報通信研究機構の倫理審査委員会および安全審査委員会から承認を得た。

3.2 MRI 計測

被験者の脳機能画像を Siemens 社の 3T MRI MAGNETOM Prisma を用いて取得した。撮像パラメータは次の通りである：64ch receiver coil、multiband gradient echo-EPI sequence [multiband factor = 6]、TR = 1000 ms、TE = 30 ms、flip angle = 60°、voxel size = 2×2×2 mm、matrix size = 96×96、number of slices = 72。

3.3 映像視聴課題

被験者には、MRI スキャナー内のスクリーン（視角 28.0°×15.5°）に写し出される 1 スキャンあたり 10 分 10 秒間（最初 10 秒分の脳活動データは使用しない）の映像刺激を、20 スキャンに分けて呈示した。

映像刺激として、株式会社 NTT データの協力により、Web 広告映像を入手した。また、映像と紐付いた視聴者のクリック率、視聴完了率のマスデータも、同社の協力により入手した。各広告は 15 秒または 30 秒の長さを持ち、それらをランダムな順番でつなげて、10 分 10 秒×14 本の映像刺激を作成した。

20 スキャンのうち 12 スキャンで取得した脳応答は、モデル学習用に用いるデータ（学習データ）である。この 12 スキャンにおいては、12 本の異なる映像を呈示した。残りの 8 スキャンで取得した脳応答は、手法の評価に用いるデータ（評価データ）である。評価データのスキャンでは、データの SN 比を上げるため、2 本の映像をそれぞれ 4 スキャンで繰り返し呈示し、各映像に対する応答の平均値を評価データとして利用した。脳応答は 1 秒 1 サンプルとして取得しており、最終的に被験者 1 名あたり、7200 サンプルの訓練データと 1200 サンプルの評価データを得た。

4. 結果

4.1 映像シーン記述推定

推定精度の検証では、提案手法に加え、2 種類の既存手法の精度を評価した。1 つの既存手法は、脳活動空間を介さずに、VGG の中間層活性化パターンから直接 word2vec ベクターを回帰して推定するモデルである（以下、深層学習手法と呼ぶ）。これは、深層学習を用いた転移学習 [Bengio 12] の一つの形であり、既存研究でよく用いられる手法である。もう 1 つの既存手法

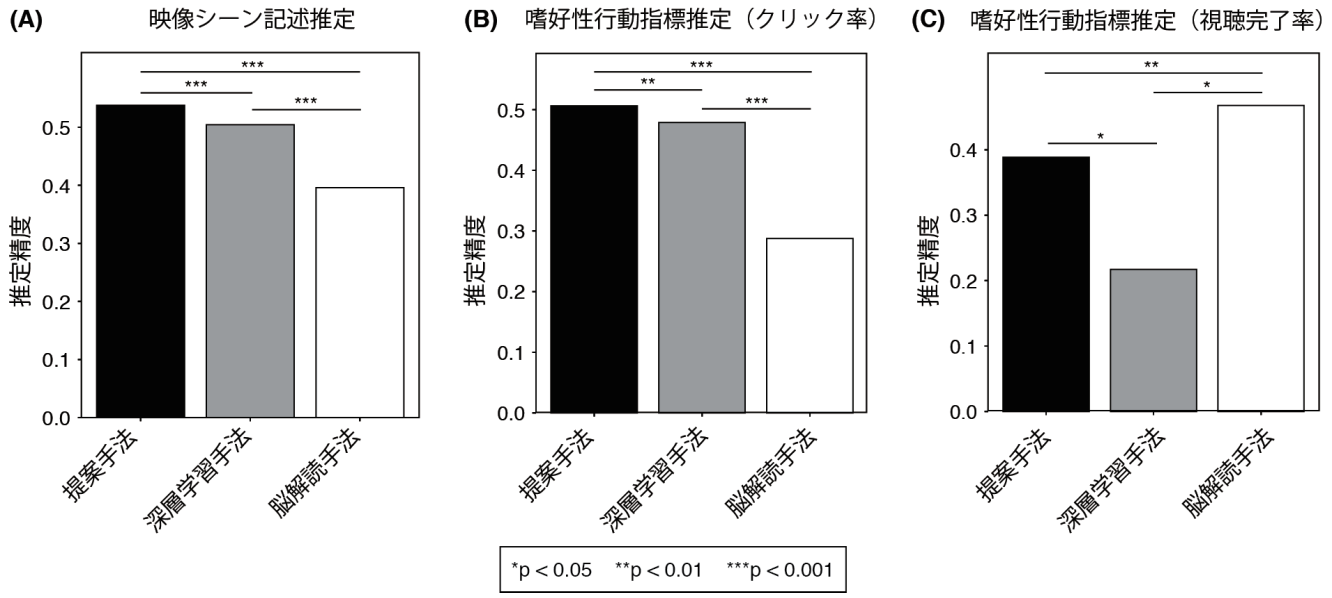

図 2：推定精度における提案手法と既存手法の比較

は、実測した脳応答から word2vec ベクターを解読する復号化モデルである [Nishida 18]（以下、脳解読手法と呼ぶ）。

図 2A に推定精度の比較結果を示す。推定精度は、評価データの各シーンにおいて、各手法を用いて推定した word2vec ベクターと、同シーンの実記述データから算出した word2vec ベクターのピアソン相関を、全シーンについて平均した値で評価した。なおここでは、提案手法と脳解読手法の推定には、全被験者のモデルの推定を平均した値を用いている。推定精度は、推定手法が最も高く、次いで深層学習手法、脳解読手法という順になった。深層学習手法と提案手法の間で 6.6%の精度向上が見られた（bootstrap test、$p < 0.0001$）。

4.2 嗜好性行動指標推定

映像に対する嗜好性を反映する行動指標（クリック率、視聴完了率）の推定精度についても、評価データを用いて、提案手法と既存手法の比較を行った。推定精度は、各手法を用いて推定した指標の時系列と、実測値としての行動指標の時系列におけるピアソン相関係数を用いて評価した。図 2B および C に、クリック率および視聴完了率の推定精度の比較結果を示す。クリック率推定は、映像シーン記述推定と同じく、提案手法、深層学習手法、脳解読手法の順で高くなった。深層学習手法と提案手法の間で 5.8%の精度向上が見られた（bootstrap test、$p < 0.0033$）。一方で、視聴完了率の推定においては、脳解読手法、提案手法、深層学習手法の順で高くなった。深層学習手法と提案手法の間で 79%の精度向上が見られた（bootstrap test、$p < 0.017$）。

5. 考察

本研究では、深層学習ネットワークの特徴表現に脳情報表現を融合することで、人間の主観的判断が強く影響する 2 種類の推定課題において、深層学習ネットワークの性能が向上するか検証を行った。そして、深層学習のみを用いた既存の推定方法と比較して、提案手法が高い推定精度を示すことを確認した。本研究の成果は、脳を模倣するのではなく、脳情報を機械学習手法と融合することで、認識性能の向上ならびに応用範囲の拡大が可能になることを示唆しており、機械学習研究に新たなパラダイムシフトをもたらす可能性を秘めているといえる。

提案手法は、映像シーン記述推定において、既存手法より高い精度を示した（図 2A）。シーン記述の中には、物体や動作を表す名詞、動詞（例：男性、走る）だけでなく、印象を表す形容詞（例：格好いい）も多様に含まれている。前者のみであれば既存の深層学習でも高い精度で推定を行うが [He 16]、後者を含んだことにより、脳情報を取り入れた提案手法が高い精度を示したと考えられる。

一方で、嗜好性行動指標推定においては、2 つの推定項目のいずれでも、提案手法が深層学習手法を上回る精度を示したが（図 2B、C）、視聴完了率の推定では、脳解読手法が最も精度が高くなった（図 2C）。これは、後者の推定の方が、人間の脳内でのみ表現される情報が、より重要であることを示している。そして、提案手法は、脳解読手法には及ばなかったが、深層学習手法より高い精度を示した。以上のことから、提案手法は深層学習の特徴表現と脳の情報表現をうまく融合し、いずれの表現も推定に利用することで、精度向上をもたらしていると推測される。

提案手法は、個々人の脳から脳情報表現をモデル化し、深層学習と融合することができる。本研究での推定問題では、全被験者のモデルの平均を推定に利用したが、個々のモデルは個人の情報表現の特性を反映すると考えられ、認知内容の個人差を推定できる可能性がある。もし提案手法が認知内容の個人差を推定可能であれば、個性を取り入れたパターン認識システムとして、特定の人物の代役をつとめるエージェントの開発や、個性のデジタル・アーカイブ化などを実現する、多様な社会応用の可能性を秘めた技術となることが期待される。

参考文献

[Bengio 12] Bengio Y (2012) Deep Learning of Representations for Unsupervised and Transfer Learning. In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, pp 17–37.

[Güçlü 15] Güçlü U, van Gerven MAJ (2015) Deep neural networks reveal a gradient in the complexity of neural

representations across the ventral stream. J Neurosci 35:10005–10014.

[He 16] He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778.

[LeCun 15] LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444.

[Mikolov 13] Mikolov T, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26:3111–3119.

[Naselaris 11] Naselaris T, Kay KN, Nishimoto S, Gallant JL (2011) Encoding and decoding in fMRI. Neuroimage 56:400–410.

[Nishida 15] Nishida S, Huth AG, Gallant JL, Nishimoto S (2015) Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions. Soc Neurosci Abstr 45:333.13.

[Nishida 18] Nishida S, Nishimoto S (2018) Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage 180:232–242.

[Nishimoto 11] Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL (2011) Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol 21:1641–1646.

[Simonyan 14] Simonyan K, Zisserman A (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[西田 18] 西田知史, 西本伸志 (2018) 脳表象モデルを用いた任意の視覚入力に対する知覚内容推定システム. 第 32 回人工知能学会全国大会 4Pin1-37.