The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 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

A computational system estimating perception evoked by arbitrary visual inputs on the basis of modeling the perceptual 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

Although brain decoding techniques using functional magnetic resonance imaging (fMRI) have many potential real-world applications, the measuring cost of fMRI makes it difficult to realize many of such applications. Here, we propose a new decoding framework for estimating perceptual experiences evoked by visual materials with no additional fMRI measurement after model construction. Our framework consists of brain-activity prediction and perceptual decoding models constructed from individual brain data. Once the training of these models using movie-evoked fMRI data has been done, the framework combines these models and estimates each person's perceptual experiences regarding novel scenes without any additional fMRI measurements. Our results showed that our framework well estimated perceptual experiences that were evoked by novel scenes, and the estimated contents varied across the models for individual persons. Thus, our framework may provide a new computational system for estimating personal perceptual experiences, evoked by arbitrary visual inputs, via the perceptual representation in the brain.

1. はじめに

機能的磁気共鳴画像法（fMRI）で計測した脳活動から知覚内容や意図を解読する脳情報デコーディング技術では、近年、自然な映像により誘発される知覚内容が、映像や単語の形で解読可能になっている [Huth 16、Nishida 17、Nishimoto 11]。このような技術は、例えば CM 映像の印象評価サービスなどの形で、社会実装も進められている（参考：NeM sweets DONUTs®, http://www.nem-sweets.com/）。しかし、新たな視覚入力に対する脳活動解読を行うためには、多大な金銭的および時間的コストを必要とする fMRI 計測が必要であり、これが技術の社会実装を困難にする一因となっている。

一方で、深層学習 [LeCun 15]のような最新の人工知能システムは、任意の視覚入力から物体カテゴリのような知覚ラベルを、高い精度で推定することが可能で、様々な形での社会実装が進められている。しかし、人間並み、あるいはそれ以上に正確なレベルで画像や映像に客観的な知覚ラベルを付与することが可能な手法であっても、個人差を含むような主観的な知覚内容を推定することはまだ難しい。

そこで本研究では、任意の視覚入力により誘発される知覚内容を、少量の脳計測データからモデル化した脳内の知覚表象空間を介して推定する手法を提案する。これにより、追加の fMRI 計測を要さずに、任意の視覚入力を扱える脳情報デコーディング技術が実現できる。また提案手法を別の観点で捉えると、個人の脳内知覚表象空間のモデルを取り入れることで、個人差を含む人間らしい知覚推定を行うことが可能となりうる、人工知能システムとみなすこともできる。本研究では、この提案手法を、映像により誘発する知覚内容を単語の形で推定する解読モデル [Nishida 17] に適用し、従来のデコーディング技術である脳活動からの知覚推定と性能の比較を行ったうえで、提案手法の推定結果が知覚の個人差も反映しうることを検証した。

連絡先：西田知史，情報通信研究機構，s-nishida@nict.go.jp

2. 提案手法

2.1 脳活動予測モデル

脳活動予測モデルの構築（図 1A）は、符号化・復号化モデリング手法 [Naselaris 11] に準ずる。符号化モデリングは、任意の特徴空間を介した、知覚入力空間から脳活動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活動の系列を **R**、知覚入力の系列を **S**、その特徴表現の系列を $f(\boldsymbol{S})$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。

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

これにより、一旦 $\mathbf{W_e}$ が学習された後は、新たな $f(\boldsymbol{S})$が得られれば、誘起される脳活動 **R** の予測が可能なモデルとして機能する。

本研究では、特徴空間として、深層学習ネットワークの一種である 16 層 VGG [Simonyan 15] の中間層活性化パターンを利用した。深層学習ネットワークの中間層活性化パターンを利用すると、任意の視覚入力に対して活性化パターンが算出できる。また、同様の活性化パターンは、視覚入力に誘発されるヒト脳活動を精度良く予想することが報告されている [Güçlü 15]。知覚刺激として用いた映像の各フレームに対して、活性化パターンを計算したうえで 1 秒ごとに平均し、特徴表現の系列 $f(\boldsymbol{S})$としてモデル学習に用いた。モデル学習は被験者ごとのデータを別々に用いて行った。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、被験者一人あたり 8 個のモデルを作成した。

また、脳活動の予測精度向上のために、fMRI 信号が時間的に緩やかに変化する特性を考慮して、自己回帰成分（過去の脳活動系列）を説明変数に含む脳活動予測モデルも構築した。このモデルは、知覚入力の特徴表現系列に加え、過去の脳活動系列（主成分分析により 1000 次元に圧縮）も回帰項とするような線形回帰モデルとなる。このモデルも同様に、VGG の 8 つの中間層に対応する形で、被験者一人あたり 8 個のモデルを作成した。

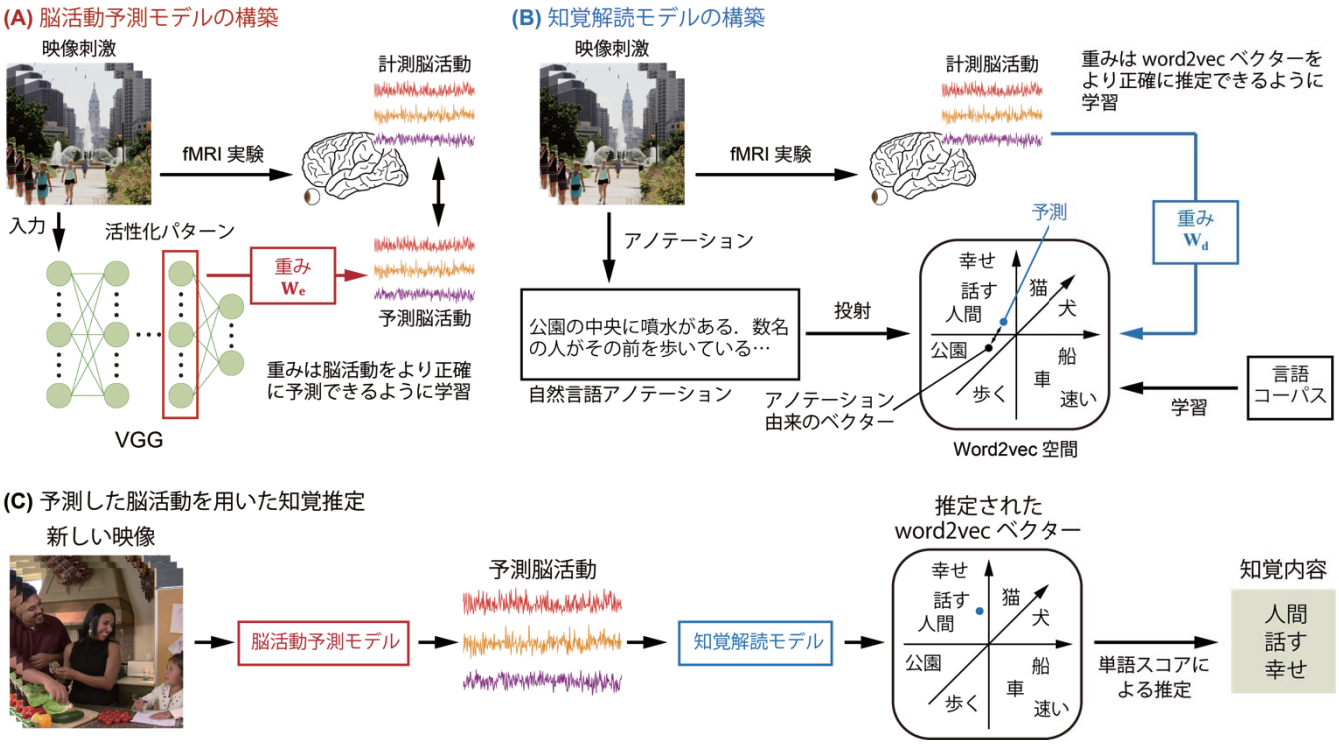

図 1：脳活動予測モデルと知覚解読モデルによる知覚推定の概要

2.2 知覚解読モデル

知覚解読モデルの構築（図 1B）は、復号化モデリングに相当する。このモデリングは、符号化モデリングとは逆に、脳活動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。

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

一旦 $\mathbf{W}_d$ が学習された後は、新たな $\mathbf{R}$ が得られれば、知覚内容を反映する特徴表現 $g(\boldsymbol{S})$ を推定可能なモデルとして機能する。

特に、本研究の知覚内容解読モデルでは、特徴空間に word2vec 空間 [Mikolov 13] を用いた。Word2vec 空間では、単語と単語の意味関係が、対応するベクター表現の空間内の位置関係として適切に表現されることが知られている。また、脳情報デコーディングに特徴空間として用いると、映像に誘発される知覚内容を高い精度で推定可能なことが報告されている [Nishida 17]。本研究では、100 次元の word2vec 空間を、日本語 Wikipedia コーパスを用いてあらかじめ学習しておいた。

知覚入力の word2vec 空間上の表現を得るために、知覚刺激として用いた映像の自然言語アノテーションを取得した。アノテーションは映像の 1 秒ごとのキャプチャ画像に対して、50 文字以上のシーン記述の形で収集した。また、1 枚の画像に対し 5 人以上からのアノテーションを得た。各アノテーションは形態素解析を行った後、word2vec の単語ベクター表現に変換し、1 枚の画像に対するアノテーションに含まれる全単語ベクターの平均として、映像の各シーンに対する特徴表現を算出し、モデル学習に用いた。

知覚内容の推定を単語の形で行う手順は次の通りとなる。新たな脳活動データが得られると、学習済みモデルの重みを利用して word2vec ベクターを算出した。そのベクターと、コーパスに頻出の 1 万単語のそれぞれに対応する word2vec ベクターの相関係数（単語スコア）を計算した。より高い単語スコアを持つ単語ほど知覚内容を反映する単語とみなし、単語スコアの上位単語を名詞・動詞・形容詞に分けてリストアップした。このようにして、単語による知覚推定を行った。

2.3 予測した脳活動を用いた知覚推定

新たな視覚入力に対する脳活動予測を行うために、入力となる映像をフレーム画像に分割し、VGG に入力して活性化パターンを算出した。その活性化パターンの値に、予め学習しておいた脳活動予測モデルの重みを適用し、脳活動を予測した。各ボクセルにおいて、脳活動予測値は 8 つの中間層のモデルに対応して 8 つの値が得られるが、学習データにおける交差検定により算出した予測精度（脳活動の予測系列と計測系列の相関）に基づき、そのボクセルにおいて最も精度の高いモデルの予測値を使用した。

続いて、その予測脳活動と活性化パターンを、自己回帰成分を持つ学習済み脳活動予測モデルに入力し、再度脳活動を予測した。同様に、学習データにおける交差検定により算出した予測精度に基づき、8 つの中間層に対応するモデルのうち、ボクセルごとに最も精度の高いモデルの予測値を使用した。

その後、全ボクセルについての脳活動予測値に対して知覚解読モデルを適用し、単語による知覚推定を行った（図 1C）。

3. MRI 実験

3.1 被験者

fMRI による脳計測実験には 8 名の被験者（男性 5 名、女性 3 名、23～40 歳）が参加した。全ての被験者から実験前に書面による同意を得た。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ている。

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（64ch receiver coil、multiband gradient

従来手法の結果（計測脳活動からの解読内容）

	名詞	動詞	形容詞
1	金髪	着る	優しい
2	髪	憧れる	可愛い
3	髪型	かぶる	若い
4	服	嫌う	幼い
5	黒髪	悩む	親しい
6	着用	慕う	悪い
7	女性	喋る	怖い

提案手法の結果（予測脳活動からの推定内容）

	名詞	動詞	形容詞
1	金髪	着る	優しい
2	女性	話す	親しい
3	男性	悩む	可愛い
4	髪型	憧れる	幼い
5	言動	嫌う	怖い
6	独身	喋る	若い
7	母親	殴る	悪い

図 2：1 つの映像シーンにおける知覚推定結果の比較

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）。また同じ装置を用いて、脳構造画像の取得も行った。

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、16 スキャンに分けて呈示した。映像視聴中、スクリーン中央に呈示される小さな点を固視するよう被験者に教示した。

映像は動画共有サイトの Vimeo（https://vimeo.com）から、投稿者が明示的にダウンロードを許可しているもののみ入手した。映像には主に、映画の予告編、自主制作映画、プロモーションビデオ、風景映像などが含まれる。それらを 10 秒から 20 秒のランダムな長さで切り抜き、ランダムな順番で再びつなげることにより、10 分 10 秒×16 本の映像刺激を作成した。

16 スキャンで取得した脳活動データのうち、12 スキャン分はモデル学習用に用いるデータ（学習データ）であり、それらのスキャンにおいては、全て異なる映像を呈示した。残りの 4 スキャン分はモデル評価用に用いるデータ（評価データ）であり、データの SN 比を上げるため、各スキャンで 4 回同じ映像を繰り返し呈示し、それに対する活動平均を評価データとして利用した。

4. 結果

4.1 知覚推定における提案手法と従来手法の比較

評価データ取得時の 10 分間の映像に対して、提案手法を用いて、脳活動予測および知覚推定を行った。また一方で、同じく評価データの脳活動を使用して、知覚解読モデルのみを用いた場合の、従来の手法における知覚解読も行った。図 2 に、1 つの映像シーンにおいて、2 つの手法のそれぞれを用いて行った推定結果の例を示す。図中左の画像は、被験者が閲覧した、または脳活動予測モデルに入力した映像シーンを表す。図中右の単語リストは従来手法（左）および提案手法（右）による知覚推定結果を名詞・動詞・形容詞に分けて表示したものである。1 万単語の中から、それぞれ最も単語スコアが高かった（最も知覚内容を反映しているであろう）上位 7 単語を上から順に並べている。提案手法は、従来手法と比較しても遜色ない程度に、シーンを適切に記述する単語を推定結果として出力していることが見て取れる。

続いて、映像の全シーンの推定結果を利用して精度を計算した。精度は解読モデルが算出した word2vec ベクターと、シーンに対するアノテーションから算出した word2vec ベクターの間の相関係数で定量化した。そして、その精度について従来手法と提案手法で比較を行った（図 3）。その結果、8 名の被験者全員において、提案手法はむしろ従来手法に比べて高い推定精度を示すことが分かった（Wilcoxon test、$p < 10^{-14}$）。

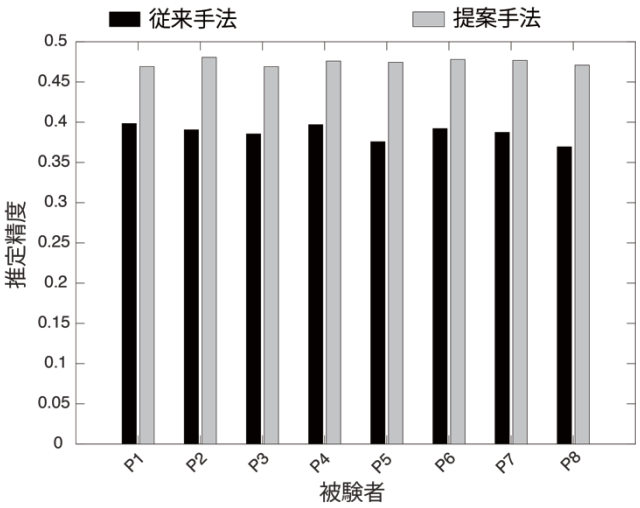

図 3：推定精度の比較

4.2 知覚の個人差との関係

過去の研究において、word2vec による知覚解読モデルを用いた際の、各シーンにおける知覚推定結果の被験者間のバラつきが、同じシーンにおけるアノテーションのアノテーター間のバラつきと相関することが報告されている [Nishida 17]。これは、映像によって誘起される知覚の個人間のバラつきが、知覚解読結果に反映されることを示唆している。まずその再検証のため、本研究の評価データにおいて、アノテーションのバラつきをアノテーション由来の word2vec ベクターにおけるアノテーター間距離の総和として評価し、知覚推定のバラつきを脳活動から解読した word2vec ベクターにおける被験者距離の総和として評価した。そして、それらの相関を計算したところ、確かに有意な相関が認められた（$r = 0.20$、$p < 10^{-6}$）。次に、提案手法の推定結果においても、同様の相関関係が成り立っているか検証するため、脳活動から解読した word2vec ベクターの代わりに、提案手法により推定された word2vec ベクターの被験者間距離の総和を用いて、相関を計算した。その結果、有意な相関が認められ（$r = 0.20$、$p < 10^{-5}$）、提案手法による知覚推定の結果も、映像によって誘起される知覚の個人間のバラつきを反映することが示唆された。

5. 考察

本研究では、符号化・復号化モデリングの枠組みを応用して、任意の映像に対する知覚内容の推定を、新たな fMRI 実験を要さずに行う手法について提案を行い、その性能を検証した。現段階では、性能を厳密に評価するため、少量のデータに対して手法を適用し評価を行ったが、従来手法に劣らない性能を示し、またその推定結果は知覚の個人差を反映しうることが示唆された。今後、さらに大規模なデータに対しても提案手法を適用し、有効性を確認することで、提案手法が fMRI 計測コストを劇的に削減する、画期的な知覚推定手法として脳情報デコーディング技術の応用可能性を大幅に広げることが期待される。

知覚推定精度の比較では、提案手法は従来手法より高い精度を示した（図 3）。脳活動予測では、計測脳活動を完全に予測できるわけではなく、この予測のための誤差が知覚推定の精度も下げると直観されるが、なぜこの直観と相反する結果となったのだろうか？一つの理由として考えられるのは、脳活動予測によって、計測脳活動に内在する計測ノイズや課題と無関係な変

動が除去された可能性である。脳活動予測モデルは 2 時間分の学習データに基づいて構築されるため、ノイズや変動を無視した純粋な脳応答信号をモデル化したと考えられる。本提案手法で採用した符号化モデリングは、過去の研究においても、時間周波数視覚特徴の脳内受容野マップ推定 [Nishimoto 11] や、意味情報の脳内表現マップの可視化 [Huth 12]などで有用であったように、脳内情報表現の定量化とそれに基づく脳活動予測を行ううえで効果的に機能するといえる。

提案手法を、脳表象を取り入れた知覚推定システムとして、既存の人工知能技術の発展とみなしたとき、どのような利点があるだろうか？一つの利点は、現状の人工知能技術では性能を上げにくい、人間らしい主観にもとづく識別などの課題への適用である。実際、過去の研究では、既存の識別モデルや自然言語処理モデルに脳表象を組み込み、人間の主観が強く影響するような課題に対して、モデルの精度向上を確認した例がある [Fong 17、Fyshe 14、Ruan 16]。本提案手法でも、今後様々な課題において、既存の人工知能技術との性能比較を行っていく予定である。もう一つの利点は個人差の反映である。提案手法では、個々人の脳を基にシステムが構築できるため、本研究でも示したように、知覚の個人差をうまく取り込める可能性がある。今後さらなる検証が必要ではあるが、知覚の個人差つまり個性を人工知能に実装することができ、特定の人物の代役をつとめるコピーロボットの開発や、個性のデジタル・アーカイブ化につながる技術になることが期待される。

以上のように、提案手法は、脳活動計測のコストを劇的に削減しつつも高い精度で知覚推定を行うデコーディング技術として利用可能であると同時に、個人の脳表象を組み入れることで主観や個人差を反映可能な新しい人工知能システムとしても利用可能となる。また、今回は単語で知覚推定を行う解読モデルを利用したが、提案手法の枠組みでは、任意の解読モデルに置き換えることができる。したがって、例えば感性や意思決定、経済的行動などを解読するモデルを適用することで、個性を考慮した知覚・認知推定システムとして、幅広い分野で提案手法が重要な役割を果たす可能性を秘めている。

参考文献

[Fong 17] Fong R, Scheirer W, Cox D (2017) Using Human Brain Activity to Guide Machine Learning. arXiv:1703.05463.

[Fyshe 14] Fyshe A, Talukdar PP, Murphy B, Mitchell TM (2014) Interpretable semantic vectors from a joint model of brain- and text-based meaning. In: ACL, pp 489–499.

[Huth 12] Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76:1210–1224.

[Huth 16] Huth AG, Lee T, Nishimoto S, Bilenko NY, Vu AT, Gallant JL (2016) Decoding the semantic content of natural movies from human brain activity. Front Syst Neurosci 10:1–16.

[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.

[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 17] Nishida S, Nishimoto S (2017) Decoding naturalistic experiences from human brain activity via distributed representations of words. Neuroimage, in press.

[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.

[Ruan 16] Ruan Y-P, Ling Z-H, Hu Y (2016) Exploring semantic representation in brain activity using word embeddings. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp 669–679.

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