1G4-OS-11a-6in1

行列因子分解を用いた動画刺激による脳活動データからの言語表象推定への取り組み

Toward Estimation of the Language Representation of the Brain Activity evoked by Visual Stimulation with Matrix Factorization

川瀬千晶 *1 Chiaki Kawase　小林一郎 *1 Ichiro Kobayashi　西本伸志 *2 Shinji Nishimoto　西田知史 *2 Satoshi Nishida　麻生英樹 *3 Hideki Asoh

*1お茶の水女子大学 Ochanomizu University　*2情報通信研究機構 脳情報通信融合研究センター National Institute of Information and Communications Technology

*3産業技術総合研究所 人工知能研究センター National Institute of Advanced Industrial and Technology

It is known that primary visual cortex uses a sparse code to efficiently represent natural scenes. Based on the fact, we build up a hypothesis that the same phenomenon happens at the higher cognitive function, here we focus on language representation, in the cerebral cortex. To proof the hypothesis, in the experiments, we have adopted two methods using sparse coding in which language representation of the brain activity evoked by visual stimulation, expressed with distributed semantics, is estimated through the reconstruction of the original data by means of sparse coding, and confirmed that both methods outperform the method without sparse coding. By this fact, we have shown an evidence that the cerebral cortex uses a sparse code to represent the meaning of language.

1. はじめに

ヒトの大脳皮質の初期視覚野での処理において，情報はスパースに処理をされていることが知られている [1]．初期視覚野には様々な刺激に反応する細胞が多数あり，視覚情報を受け取ると，その中から少数の細胞のみが反応し，入力信号を表現する．これにより，複雑な入力信号から本質的な情報を抽出し，効率よく視覚情報を処理している．また，近年，脳神経科学分野において，脳神経活動を定量的に理解する研究が盛んに行われており [2][3]，脳活動と言語表象の関係について様々な知見が得られている．本研究では，特に，初期視覚野におけるスパース表象と同様に高次表象である言語表象でも相同のスパースコーディングが行われているという仮説を立て，その仮説を立証することを目的とする．具体的な方法として，動画視聴時のヒトの脳活動の観測データと，その動画の説明文との対応関係がスパースコーディングを介在させることにより予測精度が向上するかを二つの手法を用いて確認する．

2. 先行研究

近年，動画像などを視聴した際の脳の活動パターンから人がどのような意味カテゴリを想起しているかを調査する研究が盛んになってきており，多くの新しい知見が得られている [3][4][5][6][7][8][9]．Huth ら [7] は，動画像中に現れる物体や動作を類義語体系である WordNet の語彙で表現し，動画像の刺激（WordNet 語彙 [10]）と脳神経活動との関係について調査し，脳の皮質における意味のマップを作成した．Stansbury ら [3] は，潜在的意味解析手法 LDA[11] を用いて，静止画に対して付与された語彙からシーンに対するラベル付けを教師なし学習で行い，その結果と静止画に対する脳神経活動の関係を結びつけ，カテゴリに対する脳の意味解釈の活動領域を明確にするとともにモデルを構築した．Cukur ら [2] は，動画像中の物体に注意を払い認識する際に，どのように認識の意味形態が変化しているかを脳活動データから推定している．このように統計的な言語モデルは脳活動における感覚や文脈の情報に基づく表象表現を説明するのに適したモデルであることが指摘されてきたが，さらに近年，本研究申請の連携研究者である西本，西田らは，Mikolov ら [13] によって提唱された word2vec を構築する際に採用された skip-gram が潜在意味解析手法等のこれまでの言語モデルに較べて，より適していることを同じ実験設定の下で確認し，日本語 Wikipedia をコーパスとし，skip-gram と呼ばれる言語モデルを利用することで得られる日本語の語彙の分散意味表現と血中酸素飽和度で計測される脳神経活動の間に相関関係が存在することを示している [9]．

連絡先: 川瀬千晶，お茶の水女子大学，g1220516@is.ocha.ac.jp

3. スパースコーディング

スパースコーディングとは，信号を少数の基底の線形和で表現する方法である [1]．入力信号を辞書の中から少数の基底を選び復元することで，特徴抽出や次元削減の方法として用いられる．

$$X^* \approx \arg\min_X \frac{1}{2}\|Y - AX\|_2^2 \ \text{ s.t. } \ \|X\|_0 \leq \epsilon \qquad (1)$$

$$X^* \approx \arg\min_X \frac{1}{2}\|Y - AX\|_2^2 + \lambda\|X\|_1 \qquad (2)$$

X を係数，Y を入力信号，A を辞書とするとき，式 (1) を最小化する X を求めることにより，スパースな係数 X が得られる．スパースコーディングの制約として，辞書の基底数はデータの次元よりも多くなくてはならない．式 (1) の $argmin$ 内の第一項が入力信号 Y と復元信号 AX との二乗和誤差，第二項が係数のスパース性を表している．これを解くアルゴリズムには OMP(Orthogonal Matching Pursuit) や LARS(Least Angle Regression) がある．また，係数 X のスパース性を L1 ノルムを用いて書き換えることにより式 (2) に書き換えることができる．これを解くアルゴリズムには Lasso-LARS や Lasso-CD(Coordinate Decent) がある．また，辞書と係数を同時に求めたい場合には式 (3) を用いる．この式の最適化は，まず A

を固定して，スパースコーディングによって X を求める．次にその X を固定してこの式を最小化する A を求める，という処理を繰り返しながら収束させる．前者のスパースコーディングのアルゴリズムは先と同様で，後者の辞書の最適化のアルゴリズムには Lasso-LARS，Lasso-CD がある．

$$(A, X)^* = \arg\min_{A,X} \frac{1}{2}\|Y - AX\|_2^2 + \lambda\|X\|_1 \qquad (3)$$

4. 脳活動情報からの言語表象推定

大脳皮質の言語表象においてもスパースコーディングが機能しているかを検証する方法として以下に示す 2 つの手法を提案する．

4.1 手法１：脳活動行列の言語係数行列への直接回帰による言語表象推定

本手法は学習フェーズと実行フェーズに分けられる．学習フェーズでは，まず，fMRI（functional magnetic resonance imaging）を用いて計測した脳活動データをサンプルごとに計測した各ボクセルの観測値を入れて行列化し，これを脳活動行列とする（図 1）．また，説明文もサンプルごとに出現する

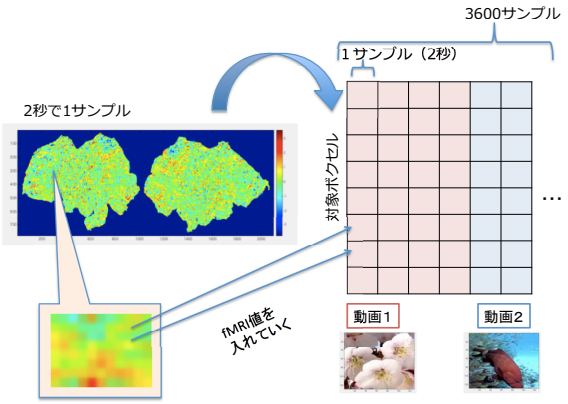

図 1: 脳活動行列の作成

単語の分散意味表現の和のベクトルからなる行列を作り，これを言語行列とする．言語行列はスパースコーディングを用いた辞書学習により言語辞書行列と言語係数行列に分解する．次に Ridge 回帰を用いて，脳活動行列を言語係数行列に写す写像 Φ を求める (式 (4))．

$$\Phi^* = \arg\min_{\Phi} \frac{1}{2}\|Y - \Phi X\|_2^2 + \lambda\|\Phi\|_2^2 \qquad (4)$$

実行フェーズでは，新たな脳活動データを入力として与え，脳活動行列を作成する．この脳活動行列を写像 Φ により写し，言語係数行列を求める．この言語係数行列と学習で作成した言語辞書行列によって復元された分散意味表表現のベクトルを，脳活動に対応する言語表象とみなす．これをパラレルコーパスである動画の説明文の分散意味表現ベクトルと比較することにより，言語表象にスパースコーディングが機能しているかを検証する（図 2）．

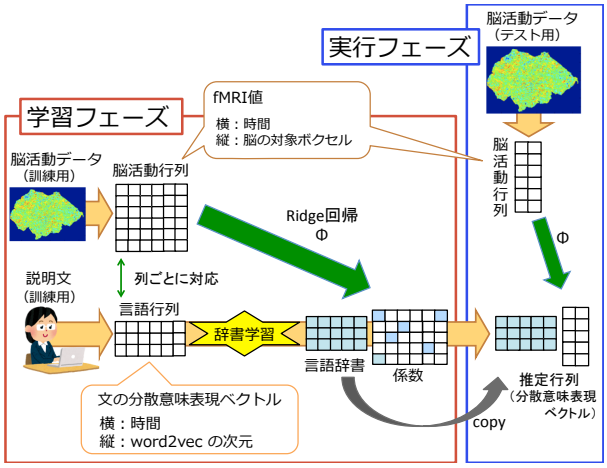

図 2: 手法 1 の概要

4.2 手法２：脳活動係数行列から言語係数行列への回帰に基づく言語表象推定

次に，脳活動行列に対してもスパースコーディングをすることにより，脳活動データの本質的な情報を抽出し，それに基づく言語表象を推定することを目指す．手法 1 と同様に，脳活動データから脳活動行列を作成し，説明文から言語行列を作成する．それぞれ脳活動行列と言語行列に対し辞書学習を行い辞書と係数に分解する．次に Ridge 回帰を用いて，脳活動行列を言語係数行列に写す写像 Φ を求める．実行フェーズでは，新たな脳活動データを入力として与え，脳活動行列を作成する．この行列を学習フェーズで作った脳活動辞書を用いてスパースコーディングをし，脳活動係数行列を求める．この脳活動係数行列を写像 Φ により，言語係数行列を求める．この係数行列と学習で作成した言語辞書行列によって復元された分散意味表現ベクトルを脳活動に対応する言語表象とみなす．これをパラレルコーパスである動画の説明文の分散意味表現ベクトルと比較する（図 3）．

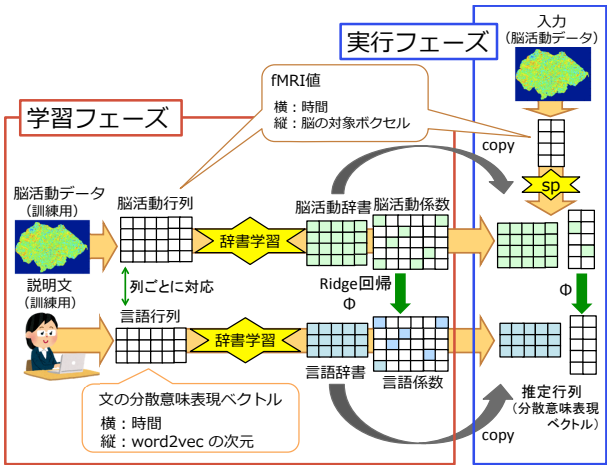

図 3: 手法 2 の概要

5. 実験

5.1 実験設定

使用するデータは，動画視聴時の脳活動データと動画説明文である [12]．このデータセットが訓練用に 3600 サンプル，テス

表 1: 推定行列と正解行列との cos 類似度

		脳活動 (学習)			言語 (学習)			テスト			時間のずれ			
手法	ボクセル	基底数	sp	辞書学習	基底	sp	辞書学習	sp	基底数	cos 類似度	2 秒	4 秒	6 秒	8 秒
Ridge 回帰	30662	-	-	-	-	-	-	-	-	-	0.153	0.261	0.262	0.200
手法 1	30662	-	-	-	2000	lasso-lars	lasso-lars	-	-	-	0.159	0.270	0.268	0.205
手法 1	30662	-	-	-	2000	lars	lasso-lars	-	-	-	0.156	0.266	0.265	0.202
手法 1	30662	-	-	-	1200	lasso-lars	lasso-lars	-	-	-	0.164	**0.273**	**0.272**	0.208
手法 1	30662	-	-	-	1200	lars	lasso-lars	-	-	-	0.160	0.268	0.267	0.205
Ridge 回帰	1404	-	-	-	-	-	-	-	-	-	0.066	0.176	0.190	0.120
手法 1	1404	-	-	-	1200	lasso-lars	lasso-lars	-	-	-	0.073	**0.191**	**0.202**	0.129
手法 1	1404	-	-	-	1200	lars	lasso-lars	-	-	-	0.070	0.186	0.198	0.126
手法 2	1404	1500	lasso-lars	lasso-lars	1200	lasso-lars	lasso-lars	lasso-lars	390	0.95	0.157	0.256	0.259	0.194
手法 2	1404	1500	lasso-lars	lasso-lars	1200	lasso-lars	lasso-lars	lars	31	0.79	0.144	0.197	0.198	0.151
手法 2	1404	1500	lasso-lars	lasso-lars	1200	lars	lasso-lars	lasso-lars	390	0.95	0.126	0.205	0.207	0.161
手法 2	1404	1500	lasso-lars	lasso-lars	1200	lars	lasso-lars	lars	31	0.79	0.096	0.146	0.142	0.115
手法 2	1404	1500	lars	lasso-lars	1200	lasso-lars	lasso-lars	lasso-lars	392	0.95	0.173	**0.265**	**0.269**	0.198
手法 2	1404	1500	lars	lasso-lars	1200	lasso-lars	lasso-lars	lars	29	0.78	0.158	0.230	0.218	0.171
手法 2	1404	1500	lars	lasso-lars	1200	lars	lasso-lars	lasso-lars	392	0.95	0.170	0.261	0.266	0.196
手法 2	1404	1500	lars	lasso-lars	1200	lars	lasso-lars	lars	29	0.78	0.156	0.228	0.214	0.169

ト用に 270 サンプルある．脳活動データは，一人の被験者に動画像を見せ，fMRI を用いてその時の脳神経活動を 2 秒で 1 サンプル記録したものである．脳活動の観測領域は 100×100×32 ボクセルであり，そのうち大脳皮質部分が 30662 ボクセルあり，これを処理対象とした．手法 2 の方で脳活動データの辞書学習をする際に，データ数 3600 サンプルよりもデータの次元を少なくしなくてはいけないため，30662 ボクセルのうち，先行研究 [9] で予測精度が 0.36 以上の 1404 ボクセルを抽出し，対象ボクセルとした．各ボクセルの大きさは，2.24×2.24×4.1mm^3 である．動画説明文は被験者に見せた動画像から 1 秒ごとに抽出した静止画に対し，アノテータ 60 人のうちランダムに抽出された 5 人が静止画を見て書いた説明文を使用した．説明文は静止画の説明や，画像を見て感じたこと，思ったことなど，静止画を見て想起したことを書いてもらったものである（図 4）．脳活動データには動画視聴から約 4〜6 秒の観測のずれが

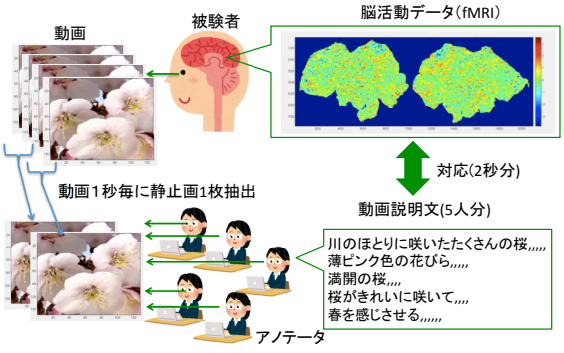

図 4: 使用データ

あるため，言語データは，脳活動データの 2,4,6,8, 秒前のデータと対応するように学習した．また，本研究ではスパースコーディングのアルゴリズムには Lasso-LARS または LARS，辞書学習のアルゴリズムには LARS と Lasso-LARS を組み合わせて用い，基底数は言語辞書行列では 1200 または 2000，脳活動行列では 1500 に設定した．Lasso-lars の場合，スパース性の重みのパラメータ λ は 1.0 に設定した．言語行列を作る際，時間ごとに説明文に出てくる単語（名詞，動詞，形容詞）を，Mikolov ら [13] によって提案された word2vec において，日本語 Wikipedia のコーパスを対象に skip-gram を利用して構築した 1000 次元の分散意味表現を用いて表し，その和を文全体の分散意味とし，これを言語表象とした．

5.2 評価方法

テストデータの動画説明文に対しても学習データと同様に言語行列を作成し，これを正解行列とする．これを文書の分散意味表現ベクトルとし，実験で推定される分散意味表現ベクトルに対して評価を行った．1 サンプルごとに正解ベクトルと推測されたベクトルとの cos 類似度を求め，マクロ平均をとり，[-1,1] の値で評価した．

5.3 実験結果

脳活動データと言語データのずれを 2，4，6，8 秒と変えた場合について，それぞれ推測された分散意味表現ベクトルと正解データの分散意味表現ベクトルとの cos 類似度のマクロ平均を示す（表 1）．また，提案手法との比較としてスパースコーディングを用いず脳活動行列から言語行列に直接 Ridge 回帰をして実験した．また，この Ridge 回帰のみによる推定と手法 1 では脳の対象部位を大脳皮質全体（30662 ボクセル）と予測精度の高いボクセル（1404 ボクセル）にした場合について求め，手法 2 では予測精度の高いボクセルにした場合について求めた．表には，用いた手法，対象座標のボクセル数，学習フェーズでの脳活動行列，言語行列のそれぞれの辞書学習に用いたアルゴリズムと辞書の基底数，テスト用の脳活動行列の復元に用いたスパースコーディングのアルゴリズム，サンプルごとの平均使用基底数，元の行列と復元行列との cos 類似度，時間のずれごとの推定の精度を示す．

5.4 考察

どの設定においても，脳活動と視聴動画に対する言語の分散意味ベクトルとのずれを 4 秒または 6 秒にした場合の精度がはるかに高く，脳活動と動画視聴には 4〜6 秒のずれがあることが確認できた．対象座標を大脳皮質全体（30662 ボクセル）の場合で比較すると，言語の処理にスパースコーディングを用いた手法 1 の方がリッジ回帰のみで推測した場合より精度が高かった．このことから，大脳皮質では初期視覚野と同様に言語表象でもスパースコーディングの処理が行われているのではないか，と考えられる．ボクセル数を 1404 ボクセルに絞ると，大脳皮質の対象になっていない部位の情報が落ちてしまう

ため，精度が低くなるが，この場合もリッジ回帰のみで推定するよりも手法 1 の方が精度が高く，さらに手法 2 の方が高くなった．このことは，入力された脳活動データを脳活動行列の中から少数の基底のみを用い復元することにより，ノイズを除き本質的な情報を抽出しているため，より明確に言語表象を推定できたと考えられる．また，マクロ平均をとる前のサンプルごとの精度を見てみると，2 秒間をまとめて 1 サンプルとしているため，その 2 秒間の間に動画の切り替えがあったサンプルでは精度が低く，0 に近い値になっている．また，表 1 の中で一番精度の高かった上から 4 番目の条件で時間のずれを 4 秒にしたときの中でも最も精度の高いサンプルは 0.628 になっている．動画の切り替えがあるサンプルの精度が他のサンプルと比べはるかに低い値になっているため，全体の平均が低くなってしまっていることが考えられる．このため，動画の切り替えがあるサンプルをノイズとして取り除くことが必要だと考える．高い精度が続いているところも見受けられ，このことから．動画によって言語表象が推定されやすいものとされにくいものがあることが考えられ，今後はこれについても分析していきたい．また，言語表象の推定に大脳皮質のどの部分がよく関係しているのかにも興味があり，これは手法 2 で作成された脳活動辞書や脳活動係数を分析することにより，どのような基底が作成され，どのような基底が選ばれているのかを見ることによって分かる可能性がある．手法 2 において表 1 のテストの cos 類似度は入力した脳活動を学習の基底を用いてどれだけ正確に表現できているかを表しており，表現に用いた基底数が多い方が高くなっていることが確認できる．さらに，この cos 類似度が高いところでは予測精度が高くなっている．手法 1 においても手法 2 においても，辞書学習やスパースコーディングに用いるアルゴリズムやパラメータによって精度が変わるため，良い組み合わせを見つけることも重要と考える．

6. まとめと今後の課題

本稿では，スパースコーディングを用い，脳活動から推測される言語表象を分散意味表現で表現した．手法 1 により，大脳皮質での言語表象の処理にスパースコーディングが用いられている可能性があることを示した．そして，手法 2 により，スパースコーディングの特徴抽出手法としての有効性を示した．今後は言語表象として，word2vec より精度が高いと言われている Glove[14] や文の単位で分散意味表現を可能にする paragrah2vec[15] などを用いて実験したいと考えている．

参考文献

[1] Olshausen BA,Field DJ,“Sparse coding of sensory inputs”,Current Opinion in Neurobiology,14:481-487,2004.

[2] Cukur T, Nishimoto S, Huth AG and Gallant JL,“Attention during natural vision warps semantic representation across the human brain”,Nature Neuroscience,16:240-252, 2013.

[3] Stansbury DE, Naselaris T, Gallant JL ,“Natural scene statistics account for the representation of scene categories in human visual cortex”,Neuron,79(5):1025-1034,2013.

[4] Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA,“Predicting Human Brain Activity Associated with the Meanings of Nouns”,Science,320(5880):1191-1195,2008.

[5] Pereira F, Detre G, and Botvinick M,“Generating text from functional brain images”, Frontiers in Human Neuroscience,5(72),2011.

[6] Pereira F, Botvinicka M, Detre G,“Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments”,Artificial Intelligence,194:240-252,2013.

[7] Huth AG, Nishimoto S, Vu AT, Gallant JL,“A continuous semantic space describes the representation of thousands of object and action categories across the human brain”,Neuron,76(6):1210-1224,2012.

[8] Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y,“Neural Decoding of Visual Imagery During Sleep”,Science,340(6132):639-642,2013.

[9] Nishida S, Huth AG, Gallant JL, Nishimoto S,“Word statistics in large-scale texts explain the human cortical semantic representation of objects, actions, and impressions”,Annual Meeting Society for Neuroscience,333(13),2015.

[10] Miller GA,“WordNet: A Lexical Database for English”,Communications of the ACM,38(11):39-41,1995.

[11] Blei DM, Ng AY, and Jordan MI,“Latent Dirichlet Allocation”,Journal of Machine Learning Research,3: 993-1022, 2003.

[12] Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL,“Reconstructing visual experiences from brain activity evoked by natural movies”,Current Biology,21(19):1641-1646,2011.

[13] Mikolov T, Sutskever I, Chen K,Corrado G and Dean J, “Distributed Representations of Words and Phrases and their Compositionality ”,Advances in Neural Information Processing Systems,26:3111-3119,2013.

[14] Pennington J, Socher R and Manning CD,“Glove: Global Vectors for Word Representation”, Conference on Empirical Methods in Natural Language Processing,1532-1543,2014.

[15] Le Q, Mikolov T, “Distributed Representations of Sentences and Documents”,Proceedings of the 31st International Conference on Machine Learning,Beijing,China,JMLR:W & CP,32,2014.

[16] Vinje WE and Gallant JL,“Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision”,Science,287(5456):1273-1276,2000.