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

深層学習の可視化による神経科学的知見の抽出

Extraction of Neuroscientific Findings by Visualization of Deep Neural Network

佐久間一輝 *1 Kazuki Sakuma　森田純哉 *1 Junya Morita　野村太輝 *2 Taiki Nomura　平山高嗣 *3 Takatsugu Hirayama　榎堀優 *3 Yu Enokibori　間瀬健二 *3 Kenji Mase

*1静岡大学情報学部 Faculty of Informatics, Shizuoka University　*2名古屋大学大学院情報科学研究科 Graduate School of Information Science, Nagoya University

*2名古屋大学大学院情報学研究科 Graduate School of Informatics, Nagoya University

Recently, research using deep learning has been conducted in various fields. Additionally, research on visualization methods learned by deep learning has also been actively conducted. Furthermore, the relationship between the human subjective state and electroencephalogram (EEG) has been clarified in the psychophysiological field. In this research, we apply the visualization method developed in the image field to the analysis of EEG. Using this method, we examine whether we can abstract physiologically reasonable structure of brain activity from the network visualizing EEG signals.The result of our experiment indicated the two important brain structures showing consistency with the previous neuroscience studies. We consider that our proposed method has some utilities as a tool to progress scientific understanding of human mind.

1. 研究背景

以前から，感情に関わる研究は生理心理学分野で脳波 (EEG: Electroencephalogram) などの生体信号を用いて行われてきた．EEG は脳内部活動の情報を保持する有益なものであると考えられてきたが，EEG は複数の異なる主観的状態の情報が集積されたものであり，ノイズも多く実利用が困難であるとされてきた．これに対し，生理心理学では EEG を周波数帯域成分に分離し，電極位置の情報と併用して用いることで，分析対象の主観的状態との相関が検討されてきた [Sarlo 05]．このような EEG を用いた感情分析においては，熟練した高度な技能によって，ノイズ除去や結果の解釈がなされてきた．

一方で，深層ニューラルネットワーク (DNN: Deep Neural Network) は，高精度のモデルを作成することができる機械学習の手法として盛んに研究が行われている．DNN は入力特徴にほとんど加工することなく，特徴抽出器をボトムアップに学習可能であるという利点がある．しかし，高精度のモデルを作成できたとしてもそのモデルが内部でどのように特徴抽出を行っているのかわからない．そこでモデルがどのように特徴抽出しているのかを理解するため，DNN の特徴抽出の様子や内部状態を可視化する研究が行われてきた [Ramprasaath 16, Schirrmeister 17]．しかしそのような研究は特定の入力特徴に対する特徴抽出を可視化するものであり，特徴抽出がモデル作成者の意図するものであるかを確認するためのものである．DNN には高度な特徴抽出機能があるにも関わらず，そこから対象に対する新たな知見を得ることは考慮されていない．

本研究の目的は，DNN を用いて学習したモデルから得られる知見と，生理心理学分野での解析から得られた感情と EEG の関係の知見を比較することで，DNN から妥当性の高い知見を得ることができるかを検討することである．この目的を達成することで，神経科学において，扱うデータの精度の範囲内で信頼性の高い知見を得るツールを導くことができると考える．

連絡先: 佐久間一輝，静岡大学情報学部情報科学科，静岡県浜松市中区城北 3 丁目 5 − 1，053-478-1452，sakuma.kazuki.15@shizuoka.ac.jp

2. 関連研究

ここでは本研究のターゲットとなる領域（EEG と感情）と，手法（DNN）に関する先行研究について述べる．

2.1 EEG と主観的状態との関係

EEG は生理心理学の分野で脳活動と人間の主観的状態との関係性を調査するために利用されてきた．その中で，被験者の主観的状態と脳の部位や δ 波，θ 波，α 波，β 波，γ 波などの周波数帯域別の活動に関連があることが報告されてきた．例えば，左前頭部の θ 波と α 波が快感情と，右前頭部の θ 波と α 波が不快感情と関連していることがわかっている [Davidson 03, Sarlo 05]．

神経科学の分野でも，感情と脳活動の関連を科学的に解明する試みが行われている．その中で，前頭葉の背外側前頭前野（DLPFC）は高次認知活動に関与しており，また，感情を抑制する機能を持つと考えられている．Beauregard らは，被験者に悲しくなるようなビデオや性的な写真を見せ，意識的に悲しみや性的な感情を抑制させるようにすると DLPFC の活動が上昇することを報告した [Beauregard 01]．

2.2 DNN を用いた研究

また，近年機械学習の発展に伴い，EEG の認識に DNN を用いる研究が行われている．Schirrmeister らは未加工の EEG から右手，左手，足，安静の 4 つの運動に関係した主観的状態の識別に CNN (Convolutional Neural Network) を適用し，CNN は未加工の EEG から周波数帯域成分を学習可能であることを示した [Schirrmeister 17]．

上記の研究は，EEG 解析に DNN を用いるアプローチの有用性を示唆しているが，学習したモデルがどのようにして脳活動のパターンを識別するのかを説明することが困難である．この問題に対して，学習した DNN の可視化を行う研究が行われている [Ramprasaath 16]．Ramprasaath らは CNN の可視化に Gradient-weighted Class Activation Mapping (Grad-CAM) を適用した [Ramprasaath 16]．彼らはさらに詳細な可視化を行うために，Grad-CAM と Guided Backpropagation(Guided BP) を組み合わせた Guided Grad-CAM(GGC) を示した．し

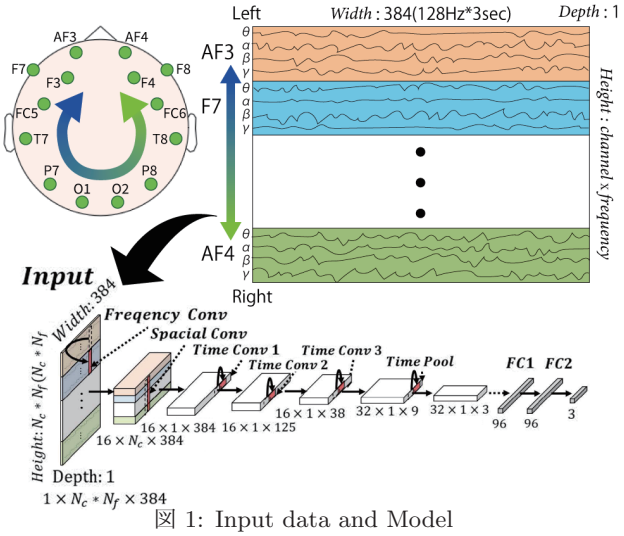

図 1: Input data and Model

表 1: Parameters of CNN

Type	Structure
Input	$Depth \times Height(N_f * N_c) \times Width(Times)$
Frequency Conv	Kernel:16 × N_f × 1, Stride:N_f × 1, pad=0, elu
Spacial Conv	Kernel:16 × N_c × 1, Stride:N_c × 1, pad=0, elu
Batch Norm 1	Dimensions:16
Dropout 1	Wight Decay:0.5
Time Conv 1	Kernel:16 × 1 × 12, Stride:1 × 3, pad=0, elu
Time Conv 2	Kernel:16 × 1 × 12, Stride:1 × 3, pad=0, elu
Time Conv 3	Kernel:32 × 1 × 12, Stride:1 × 3, pad=0, elu
Time Pool	Kernel:1 × 3,Stride:1 × 3, pad=0, Max
Batch Norm 2	Dimensions:32
Dropout 2	Wight Decay:0.5
FC 1	96, elu, Dropout:0.5
FC 2	Classes, Softmax

かしながら，EEG による感情認識にこれらの技術を適用する研究は行われていない．

3. 提案手法

本提案手法では，CNN と GGC を用いて脳活動の抽象化を行い，神経科学や生理心理学の知見の抽出を行う．

3.1 学習フェイズ

3.1.1 入力特徴

前処理では，EEG のデータセットに対して高周波成分を除去するために [1-50Hz] の FIR フィルタを適用する．次に，外れ値除去として [-500μV,500μV] の範囲を超える電位を含むデータを除外する．最後に，ウェーブレット変換により，θ 波 (4-7Hz)，α 波 (8-13Hz)，β 波 (14-30Hz) 及び γ 波 (31-50Hz) として各周波数帯域成分を算出する．

全電極位置とその全周波数帯域成分を 2 次元の画像に変換し，その各行は，各電極位置の周波数帯域成分の一連の信号入力を表す（図 1 の上部を参照）．

3.1.2 モデル

学習に用いた CNN のモデルの概要を図 1 の下部に示す．中間層では，周波数帯域軸，電極軸，時間軸の順で独立した畳み込みを行うような Kernel を設計し，出力層は快，中立，不快を示す 3 クラス分類となっている．その他のモデルの構造に関わるパラメータを表 1 に示す．

3.2 中間層の可視化と集約フェーズ

EEG と GGC を用いた中間層の可視化と集約の概要を図 2 に示す．本フェーズは三段階に分かれている．

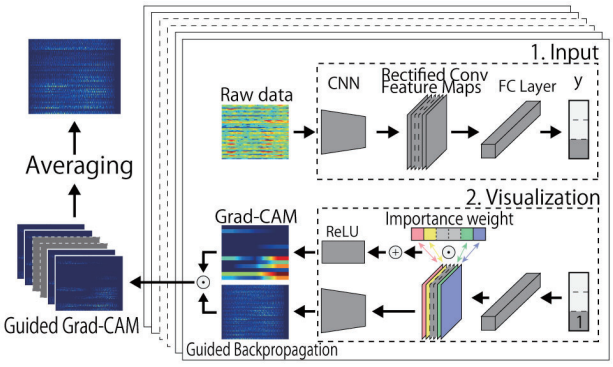

図 2: Guided Grad-CAM Procedure

3.2.1 入力

学習済みモデルに，正しく分類を行うことのできた EEG の信号（2 次元の画像）を入力する．

3.2.2 可視化

出力層における，可視化したいクラスの値を 1 としパラメータを更新せずに Backpropergation を行う．Guided BP は，以上の操作を行った後の入力層の勾配の値である．Grad-CAM は CNN 内の全結合層の直前の畳み込み層が出力した特徴マップを用いて出力する．各特徴マップの勾配の平均値と，各特徴マップの weight を掛け合わせたものに ReLU 関数を適用する．作成した Guided BP と Grad-CAM の積が GGC となる．

3.2.3 集約

クラス毎に，作成した GGC を平均した画像を作成する．これにより，被験者間の差や試行間の差を跨いだクラスの特徴が抽象化される．抽象化された平均画像を比較することで，神経科学や生理心理学における知見の抽出を試みる．

4. 評価実験

本実験は，提案手法を用いて感情に関連する特徴的な脳活動を抽出できるかどうかを検討することを目的としている．

4.1 データセット

EEG と感情に関するデータセットの作成にあたり，32 人の被験者に日常生活を記録した写真を提示した後に，感情（快不快と覚醒度）と記憶の主観的評定を行わせた．写真が被験者に提示されている間，簡易脳波計である EMOTIV EPOC を用いて 14 チャンネル（図 1 の左上部参照）サンプリング周波数 128 Hz で EEG の記録を行った．

写真は，画面中央に注視点が表示されてから 4 秒間表示され，その後被験者は SAM（Self Assessment Maskin）に従って主観的状態の評価を行った [Bradley 94]．この手順の結果，写真を閲覧する試行に対応する 4,536 個の EEG データが得られた [野村 17]．

4.2 実験条件

DNN の学習に用いるデータセットは，写真提示の各試行の始めから 3 秒間の EEG データを抽出し，前節で示した前処理を適用した．その結果、表 1 に示されるパラメータは Width=384 (128Hz × 3sec)，Height=56 (4 周波数帯域 × 14 チャンネル) となる．これらに対して平滑化，ノイズ付与，及び Time Cropping を Data Augumentation として採用し，元の学習データを 45 倍に水増しした [Schirrmeister 17]．これらのデータのうち，訓練用データとして 4/5（学習データとして 3/4,

検証データとして 1/4)，およびテストデータとして 1/5 を用いて 5 分割交差検証を行った．

5. 結果

5.1 識別性能

学習したモデルの識別精度は 5 つの fold の平均で 46.31%であった．全 3 クラス分類問題のチャンスレートである 33.33%を上回っていることから，EEG の時間周波数成分から快-不快に関わる主観的状態を学習可能であることが分かった．

5.2 GGC を用いた可視化

5 つの fold において最高の精度であった GGC によって得られたヒートマップを図 3 に示す．これらの画像は，快，平常，不快のそれぞれのラベルが付与された EEG について，GGC によって特徴を抽象化したものである．図 3 より，快の場合にはおよそ下半分（右半球に相当するチャンネルが配置），不快の場合にはおよそ上半分（左半球に相当するチャンネルが配置）が高い値になっている．

図 3 のヒートマップについて，時間軸の値を平均 することで，頭皮上での注目強度を示す topomap[Gramfort 14] を作成した（図 4）．図に示されるように，本研究において学習された DNN は快と不快とラベルづけされた脳波に対し， 左右で異なる部位に注目していることがわかる．

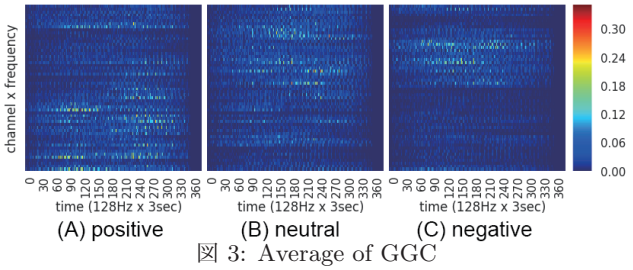

図 3: Average of GGC

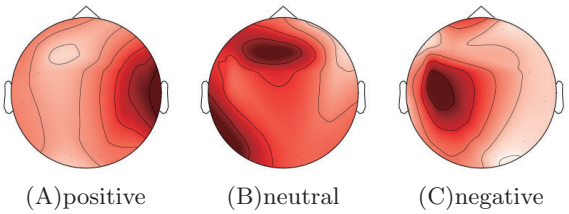

図 4: Topomap of GGC (α band)

5.3 Raw データの可視化画像

上記の結果は，DNN が快と不快を区別する脳波の特徴として,「左右半球の差分」を学習したことを示す．この結果が，DNN による学習の結果であることを確認するため，RAW データを図 3 と同様の形式で 平均化した画像（図 5），およびその topomap（図 6）を作成した．これらより，単純な平均画像においては，快と不快を区別する明確な特徴が示されないことが確認される．

5.4 神経科学的に合理的な構造の抽出 (局所度の比較)

神経科学の分野では，前頭葉の一部である DLPFC に不快感情を抑制する機能があることが知られている．それに対して快感情の生起に関わる局所的な脳部位は明らかでない．本実験においても画像を提示され不快感を抱いているときには，DLPFC などの局所的な部位が活性化している可能性がある．

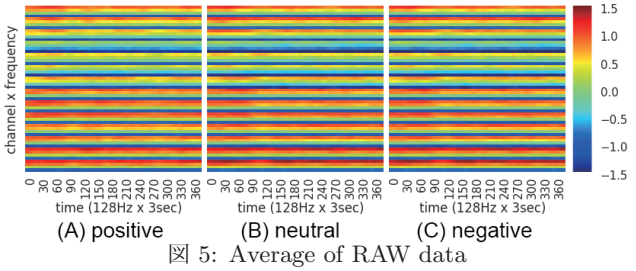

図 5: Average of RAW data

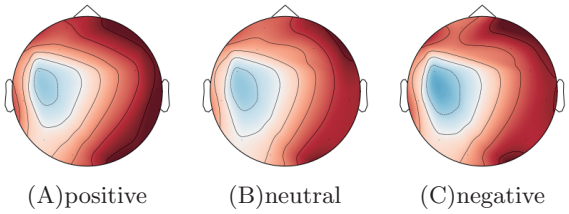

図 6: Average topomap of RAW data (α band)

そこで，各 fold における快不快の GGC 同士で局所度の比較を行った．この分析では，GGC の局所度を，GGC 画像の各セルを活性度によってソートした配列に対して得られる回帰直線の傾きとして定義する．配列の index に対して対数を取ることで片対数グラフとし，それに対して回帰分析を行った．実際に得られるグラフの例を図 7 に示す．図 7 において青線は不快時，緑線は快時の GGC の活性値と回帰直線に対応する．

図 8 は各 fold における全被験者から求められた快不快それぞれの局所度の平均である．fold02 と fold05 の間には快不快間で局所度の差に有意差がみられた (fold01: t (32) $= -2.03$, $p = 0.06$, fold02: t (32) $= -3.97$, $p < 0.01$, fold03: t (32) $= -1.55$, $p = 0.13$, fold04: t (32) $= -2.83$, $p = 0.01$, fold05: t (32) $= -4.57$, $p < 0.01$). .

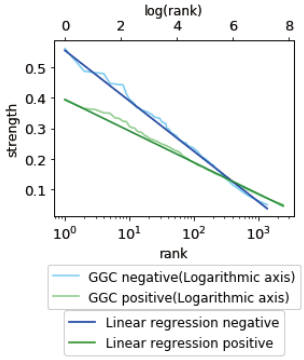

図 7: Rank regression obtained from a participant

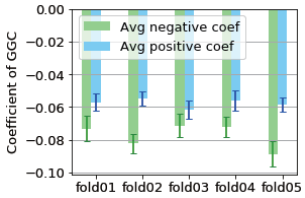

図 8: The average of the localities of each fold

5.5 信頼性の検証

前項までの結果は，提案手法により神経科学的な知見の抽出が可能であることを示唆する. しかし，今回の実験におけるモデルの識別精度は高いものではなく，得られた知見が本当に神経学的な構造を反映しているのか疑問が残る．ただし，脳波と感情という対象の特質から完全な分類は困難である．そのため，本研究では結果の一貫性を評価することで信頼性を検証する．

各 fold における快不快の GGC の類似性を比較検討した．こ

の検証を行うため，2 つの GGC によって生成された 2 つのベクトルを比較するコサイン類似度と，2 つのハッシュ行列間の距離を計算する *perceptual hash* の 2 種類の類似度を使用した [Zauner 10].

図 9 と 10 は、2 つの類似度の結果を示している． 各行列の行と列は，各 fold（*fold01 … fold05*）に対して取得された快と不快の GGC に対応し，行列のセルは類似性の値となっている．行列全体に対して，同じ感情としてラベル付けされた GGC の間に高い類似性がみられ，前項で得られた GGC の特徴に fold 間に高い一貫性があることが示されている．

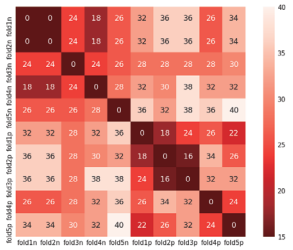

図 9: Cosine similarity between Guided Grad-CAM in each fold

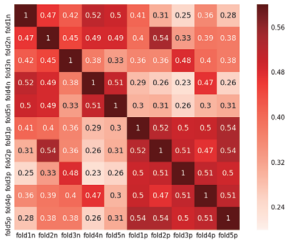

図 10: Difference in hash value of Guided Grad-CAM in each fold

6. 考察

快不快の主観的状態を識別する DNN を作成し，学習したモデルがどのような特徴抽出を行っているのか検討を行った．GGC の平均を用いることで，学習したモデルが EEG の左右差を用いて，快不快のクラス分類を行っていることがわかった．これは生理心理学分野で得られている知見とも一致する．このことから，提案手法を用いることで学習済み DNN から感情に関わる特徴的な脳活動を抽出し，妥当性の高い知見を得ることができたと考えられる．

しかし，生理心理学分野の知見では快感情のときに左前頭部，不快感情のときに右前頭部が活性化するとされているにも関わらず，GGC の平均においては逆の部分に注目している．これは DNN が EEG のどの部分が活性化していないのかに注目してクラス分類を行っているからであると考えられる (図 4c と図 6c 参照)．本実験で用いた DNN の識別精度は 46.31%であり高くない．これは，脳波と感情という扱いや分析が難しい対象を扱っていることと，DNN が注目する箇所が最も妥当とは言えないものであったことが原因であると考えられる．

GGC における注目箇所の局所性についての検証も行い，全ての fold で不快時の局所性が快時より大きく，特に fold02 と fold05 では有意差が見られた．このことから DNN は不快の識別においてより明確な構造を持っていると考えられる．この結果は，不快感情は局所的な部位が活性化し，快感情が前頭葉の広範な部位が活性化する構造を反映していると考えられる．しかし今回の検証において局所性は GGC 全体に対して求めたものであった．そのため，空間と時間で分離した局所性を求めることが今後の課題として挙げられる．

また，得られた知見には異なるデータ間で学習したモデル間で一貫性があった．しかし学習したモデルの識別精度は決して高い値ではなかった．一見矛盾しているようだが，精度と一貫性は異なる指標であると考える．識別精度はあらゆる状況に当てはまる普遍性，一貫性は得られた傾向の信頼性を表す．神経科学における研究では，あらゆる状況に対する普遍性を求めるのではなく，揺らぎのある特定の状況における傾向の信頼性を求める．そのため，本研究は神経科学において，扱うデータの精度の範囲内で信頼性の高い知見を得るツールのひとつになり得ると考える．

7. 結論

本研究の目的は，DNN を用いて神経科学的知見を得る方法を提案することであった．そのために，EEG によって快不快感情を認識する DNN を設計し，学習したモデルから感情に関連する特徴的な脳活動の抽出を試みた．結果，快不快感情間の脳活動の左右差と，不快感情における脳活動の局所性という，これまでの神経科学の研究と一貫性を示す二つの重要な脳構造を示した．これらの結果から，提案手法は人間の心の科学的理解を深めるためのツールとして有用性があると考えられる．

参考文献

[Beauregard 01] Beauregard, M., Lévesque, J., and Bourgouin, P.: Neural correlates of conscious self-regulation of emotion., *The Journal of neuroscience* (2001)

[Bradley 94] Bradley, M. M. and Lang, P. J.: Measuring emotion: the self-assessment manikin and the semantic differential, *Journal of behavior therapy and experimental psychiatry*, Vol. 25, No. 1, pp. 49–59 (1994)

[Davidson 03] Davidson, R. J.: Affective neuroscience and psychophysiology: Toward a synthesis, *Psychophysiology*, Vol. 40, No. 5, pp. 655–665 (2003)

[Gramfort 14] Gramfort, A., Luessi, M., Larson, E., D., E., Strohmeier, C., Brodbeck, D., and Parkkonen, M., L.and Hamalainen: MNE software for processing MEG and EEG data, *Neuroimage*, Vol. 86, No. 1, pp. 440–460 (2014)

[Ramprasaath 16] Ramprasaath, R., Abhishek, D., Ramakrishna, V., Michael, C., Devi, P., and Dhruv, B.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization, *CVPR 2016* (2016)

[Sarlo 05] Sarlo, M., Buodo, G., Poli, S., and Palomba, D.: Changes in EEG alpha power to different disgust elicitors: the specificity of mutilations, *Neuroscience letters*, Vol. 382, No. 3, pp. 291–296 (2005)

[Schirrmeister 17] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization, *Human brain mapping*, Vol. 38, No. 11, pp. 5391–5420 (2017)

[Zauner 10] Zauner, C.: Implementation and benchmarking of perceptual image hash functions (2010)

[野村 17] 野村太輝ら FCNN を用いた感情認識における生理心理的制約の効果, *JSAI* (2017)