勾配積分法に基づく脳波識別に貢献した周波数領域特徴の説明

Explanation of frequency domain features contributing to EEG classification using integrated gradients

立川 和樹 *1　河合 祐司 *2　朴 志勲 *2　浅田 稔 *2
Kazuki Tachikawa　Yuji Kawai　Jihoon Park　Minoru Asada

*1大阪大学大学院工学研究科
Graduate School of Engineering, Osaka University

*2大阪大学先導的学際研究機構
Institute for Open and Transdisciplinary Research Initiatives, Osaka University

Interpreting prediction of deep learning models is important in many applications, especially in medical diagnosis systems. Several methods have been proposed to interpret black-box predictions, but most of these studies are intended to calculate the contributions of input features. In this paper, we propose a method to compute the contributions in another feature space. The method applies differentiable transformations to input features and compute the contributions of the mapped features using integrated gradients. This approach enables us to calculate the contributions of amplitude and phase for each frequency in EEG classifications because the fast Fourier transform is differentiable. The proposed method is verified using three EEG datasets and the results show that the contributions of the proposed method are more reliable and has less computational cost than those of a conventional method. Our method will thus enhance the reliability of data-driven approaches in EEG analysis.

1. はじめに

深層学習は物体認識だけでなく，脳波を入力とした識別でも高い識別性能を有する [Schirrmeister 17, Hartmann 18]．特に，畳み込みニューラルネットは，脳波を周波数帯ごとに分けそれぞれの周波数帯で特徴量を設計する従来の手法とは異なり，簡単な前処理のみを施した脳波を入力として高い精度で識別できる [Schirrmeister 17]．この識別器を医療診断支援に応用する場合，どのような脳領域や周波数帯を根拠に識別したかを示すことで，医師が従来の知見との整合性を確認できる．

そのため，各入力特徴が識別に貢献した度合いを求める手法が提案されている [Sundararajan 17, Lundberg 17] が，これらの手法の多くは計算コストが非常に高く，一例の説明に数十分を要することもある [Zintgraf 17]．また，一部の手法は識別器とは関係のない説明をしている可能性があり [Adebayo 18]，計算コストを低く抑えつつ真に識別に貢献した特徴を求めることが重要である．これらの説明法の中で，勾配積分 (integrated gradients) 法は，特徴の貢献を求めるために望ましい性質を満たすという説明の正確さが理論的に裏付けられているだけでなく，計算コストも低いため，入出力勾配が計算できる場合には，優れた説明法であるといわれている [Sundararajan 17]．しかし，この手法は画像や言語のデータのように入力空間での貢献のハイライトを人が理解しやすいことを前提にしている．時系列の脳波データにおいては入力空間のハイライトを求める勾配積分法による説明を人が理解することは難しく，医学・神経科学での知見の蓄積のある周波数領域での貢献の提示が望ましい．そのため，周波数ごとの振幅成分と位相成分の貢献を求める Input Perturbation Network Prediction Correlation Map (IPNPCM) 法が提案されている [Schirrmeister 17]．しかし，この方法は勾配積分法のように説明の正確さが理論的に裏付けられておらず，また，計算コストは，ハイパーパラメータであるサンプル回数次第であるが，少ないとはいえない．

連絡先: 河合祐司，大阪大学先導的学際研究機構，〒 565-0871，
大阪府吹田市山田丘 2-1，kawai@ams.eng.osaka-u.ac.jp

そこで，本研究では勾配積分法に基づいて，周波数ごとの振幅成分と位相成分の貢献を，少ない計算量で正確に計算することを目指す．識別器入力の空間を人が理解しやすい空間へと変換することで，その変換が微分可能であれば，勾配積分法により変換後の空間での貢献の計算が可能になる．脳波の場合，高速フーリエ変換 (fast Fourier transform: FFT) が微分可能であることを利用して，周波数領域での貢献を計算する．今回，三つのデータセットを用いて IPNPCM 法と比べたときの提案法の有効性を実験的に示す．

2. 提案手法

図 1 に示すように，従来は識別器の入力 $\boldsymbol{x}$ に対する出力 $f(\boldsymbol{x})$ の勾配 $\partial f(\boldsymbol{x})/\partial \boldsymbol{x}$ を計算していたが，本研究では識別器の入力を，人が理解可能な空間である振幅 $\boldsymbol{a}$ と位相 $\boldsymbol{p}$ に変換したものを入力とした勾配 $\partial f(\boldsymbol{x})/\partial \boldsymbol{a}$ と $\partial f(\boldsymbol{x})/\partial \boldsymbol{p}$ を計算する．脳波識別の場合，脳波を複素数の周波数空間に変換する FFT，振幅成分に変換する複素数の絶対値の計算，および，位相成分に変換する複素数の偏角の計算はどれも微分可能である．そのため，周波数空間を $\boldsymbol{c}$ とすると、$\partial f(\boldsymbol{x})/\partial \boldsymbol{a}$ は連鎖律を用いて，

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{c}}{\partial \boldsymbol{a}} \times \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{c}} \times \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$$

のように，変換した入力に対する勾配が計算できる．

この勾配を用いて，周波数空間の i 番目の特徴の貢献 IG_i は，変数 α と基準点 $\tilde{a}$ を用いて以下の式から計算できる [Sundararajan 17]．

$$\mathrm{IG}_i = (a_i - \tilde{a}_i) \int_0^1 \frac{\partial f(\tilde{a} + \alpha(a - \tilde{a}))}{\partial a_i} d\alpha \qquad (1)$$

なお，位相の場合も同様に計算できる．

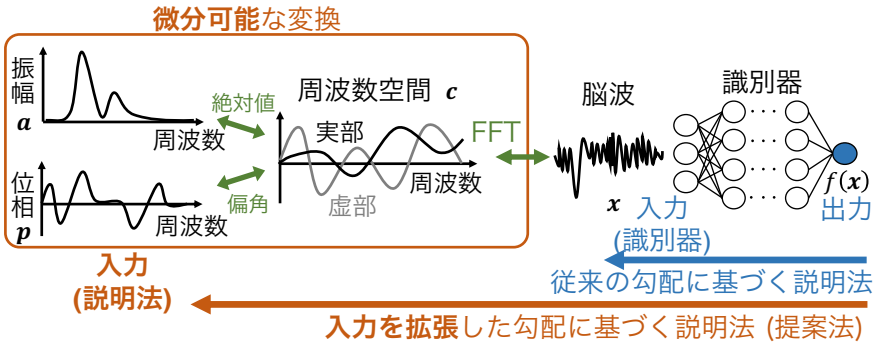

図 1: 提案法の概要．青線は識別器の入力に対する勾配を用いる従来法を示す．橙色部分は，微分可能な変換を入力特徴に施し，変換先の空間での特徴を入力として勾配を用いる提案法を示す．

3. 実験

3.1 データセットと識別器

一つの人工データセットと，二つの脳波データセットを用いた．脳波データの多くは，被験者の情報に基づいてラベル付けされ，貢献の真値は不明あるが，人工データセットの貢献の真値は既知である．

人工データセットは 5 チャンネル，60Hz で計測され，50 秒で 1 エポックとした．すなわち，1 データは 5×3000 のデータ点から成る．それぞれの時系列データは，2, 5, 10, 20, 30 Hz の正弦波が重ね合わせられたものからサンプリングされたものであり，それぞれの正弦波の振幅値は 0.3 ～ 0.7 でランダムで決定された．教師データはそれぞれの振幅値にそれぞれのチャンネルに設定された重み (-1.5, -0.5, 0.1, 1.0, 2.0) を掛けて合計した値とした．このとき、識別器が教師データを微小な誤差で回帰できるならば，それぞれの特徴の真の貢献は振幅値にチャンネルの重みをかけた値になる．2 層の畳み込みニューラルネットを用いて教師データを回帰し，誤差は 1.0×10^{-6} 以下になった．

一つ目の脳波データセットは，PhysioNet[*1] [Goldberger 00] で公開されている睡眠ポリグラフデータセット (PSG データセット) [*2] である．このデータセットは，男性 10 名，女性 10 名の計 20 名 [*3] の二晩分の睡眠データを 100Hz のサンプリング周波数で計測し， 30 秒を 1 エポックとしている．このデータセットには， Fpz-Cz と Pz-Oz に配置された電極から得られた脳波データと眼電データ（EOG），および，それらのデータに対する睡眠段階を Wake (W), REM (R), N1, N2, N3 に分類したラベルデータが付与されている．6 層の畳み込みニューラルネットを用いて 5 つの睡眠段階を識別した．

もう一つの脳波のデータセットは UCI Machine Learning Repository で公開されている EEG データセット [*4] である．このデータセットは 256Hz-64 チャンネル脳波計で，被験者に視覚刺激を提示した際の 1 秒分の脳波を計測したものである．被験者は 122 人であり，その内 45 人は健常であり， 77 人はアルコール依存である．4 層の畳み込みニューラルネットを用いてアルコール依存の有無を識別した．

*1 https://www.physionet.org/
*2 https://www.physionet.org/physiobank/database/sleep-edfx/
*3 version 1 を使用した
*4 https://archive.ics.uci.edu/ml/datasets/eeg+database

3.2 評価方法

まず，提案法と IPNPCM 法のそれぞれで，人工データセットを用いて周波数ごとの振幅成分と位相成分の識別への貢献を計算し，貢献の真値と比較した．人工データセットは貢献の真値が既知であるため，貢献の真値との誤差を top-k intersection [Ghorbani 17] と L2 ノルムの二つの指標で評価した．top-k intersection は貢献した特徴のインデックスを正しい順位で評価する能力を定量的に評価し，L2 ノルムは順位だけでなく，値の正確さも評価できる．このとき，貢献の正確さと計算量に影響するパラメータである，勾配積分法の積分区間の分割数と IPNPCM 法のサンプル回数を変化させ，貢献の正確さと計算量の両面から評価した．次に，二つの脳波データセットを用いてピクセルフリッピング法 [Bach 15] により貢献の正確さを評価した．これは識別の説明と実際に識別モデルから特徴が取り除かれたときの出力の値の変化量が一貫していることを定量的に示す方法である．全ての実験で，各クラスから 200 データずつランダムに選び，提案法と IPNPCM 法で識別に貢献した特徴を求めた．提案法の基準点 (baseline) は [Sundararajan 17] と同じく全ての値を 0 に設定した．また，提案法の積分区間の分割数は大きくするほど貢献が正確に求められるため，推奨されている値 [Sundararajan 17] の 2 倍である 500 に，IPNPCM 法のサンプル回数も同様の理由で既に推奨される回数 [Schirrmeister 17] の 2 倍の 1000 とした．

3.3 実験結果

図 2 に人工データセットを用いた場合の貢献の真値との誤差を示す．左のグラフは top-k intersection を箱ひげ図で示したものであり，縦軸の値が大きいほど貢献の順位が真値に近い説明であることを示す．また，横軸のステップ数は提案法と IPNPCM 法でそれぞれ必要な逆伝播計算の回数と順伝播計算の回数である．提案法はわずか 2 回の逆伝播計算で貢献の順位を正確に計算できることがわかる．右のグラフは縦軸に貢献の真値との L2 ノルムの平均値を示したものである．提案法はわずか 2 回の逆伝播計算で順伝播計算を 5000 回した IPNPCM 法よりもより真値に近い貢献の値を求めることができた．

図 3 に脳波データセットを用いた場合のピクセルフリッピング法の結果の平均と標準偏差を示す．縦軸は識別器の出力値で，横軸は除去された特徴の数である．同一データセットで比較したときに，曲線がより下方にある説明法は，説明法による貢献とその特徴が実際に除去された場合の出力値への影響が一貫しており，優れているといえる．この図から，提案法は IPNPCM 法よりも貢献を正確に求められたことがわかる．

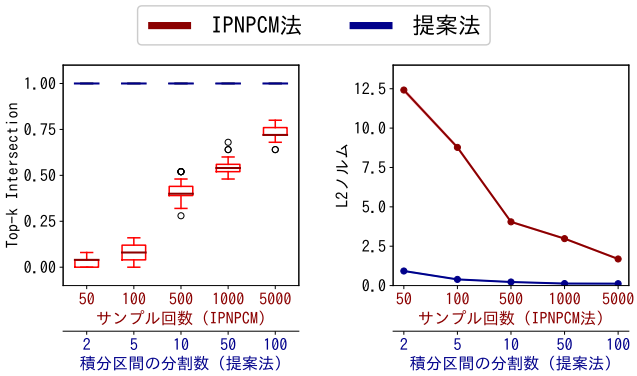

図 2: 人工データセットでの貢献の真値との誤差．200 データを用いた結果．左：top-k intersection．右：L2 ノルム．

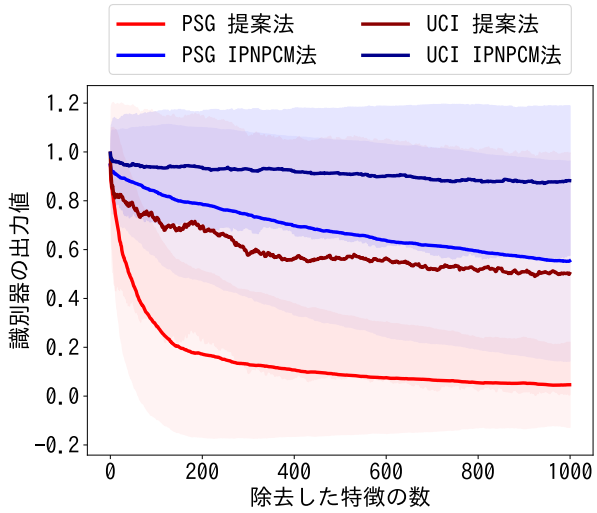

図 3: ピクセルフリッピングの結果．それぞれの曲線は 200 データでの平均値を示し，塗りつぶし範囲は標準偏差を示す．縦軸：識別器の出力値．横軸：除去された特徴の数．

図 4 に PhysioNet PSG データセットの二つの脳波センサの N2 クラス識別への貢献を可視化した．横軸は周波数であり，上段が提案法による説明で下段が IPNPCM 法による説明である．提案法により求まる貢献がより正確であることは既に図 3 に示しており，IPNPCM 法による説明はノイズが大きいことがわかる．また，デルタ帯 (0 ～ 2Hz) の活動の貢献は正であることは N3 クラスの定義と一致するため，この識別器は妥当であるといえる．

図 5, 6 に UCI EEG データセットのアルコール依存識別への貢献を可視化した．提案法と IPNPCM 法で説明の見た目が大きく異なるが，提案法による説明が正確であることは既に図 3 に示したため，二つの図の違いは従来法による誤説明の危険性を示す．

4. おわりに

脳波を入力とした識別器の周波数ごとの振幅成分と位相成分の貢献を少ない計算量で正確に計算するために，FFT の微分可能性を利用して，勾配積分法を用いて周波数領域である振幅成分と位相成分の貢献を求める方法を提案した．貢献の真値が既知である人工データセットを用いて，提案法は少ない計算量で貢献の真値を計算できることを示した．また，二つの脳波データセットを用いて提案法は従来法である IPNPCM 法よりも正確な説明が可能であることを示した．本研究は，微分可能な変換で人が理解しやすい空間に特徴を変換し，その空間を入力とみなして勾配を計算することを提案しており，振幅成分と位相成分への変換だけではなく，ほかの微分可能な変換にも適用可能である．例えば，脳波の信号源に変換できればより詳細に脳波識別の根拠と識別に使用した脳領域の関係を示すことができる．そのため，様々な種類のデータで人がデータの意味を理解しやすい特徴空間を調べることが今後の課題である．

謝辞

本研究は国立研究開発法人科学技術振興機構 (JST) の研究成果展開事業「センター・オブ・イノベーション (COI) プログラム」及び，JST，CREST，JPMJCR17A4 の支援によって行われた．

参考文献

[Adebayo 18] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., and Kim, B.: Sanity Checks for Saliency Maps, in *Advances in Neural Information Processing Systems 31*, pp. 9525–9536 (2018)

[Bach 15] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, *PLOS ONE*, Vol. 10, No. 7, pp. 1–46 (2015)

[Ghorbani 17] Ghorbani, A., Abid, A., and Zou, J.: Interpretation of Neural Networks is Fragile, in *NIPS workshop on Machine Deception* (2017)

[Goldberger 00] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals, *Circulation*, Vol. 101, No. 23, pp. e215–e220 (2000)

[Hartmann 18] Hartmann, K. G., Schirrmeister, R. T., and Ball, T.: Hierarchical Internal Representation of Spectral Features in Deep Convolutional Networks Trained for EEG Decoding, in *Proceedings of the 6th International Conference on Brain-Computer Interface*, pp. 1–6IEEE (2018)

[Lundberg 17] Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in *Advances in Neural Information Processing Systems 30*, pp. 4765–4774 (2017)

[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)

[Sundararajan 17] Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic Attribution for Deep Networks, in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, pp. 3319–3328 (2017)

[Zintgraf 17] Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M.: Visualizing Deep Neural Network Decisions: Prediction Difference Analysis, in *Proceedings of the 5th International Conference on Learning Representations* (2017)

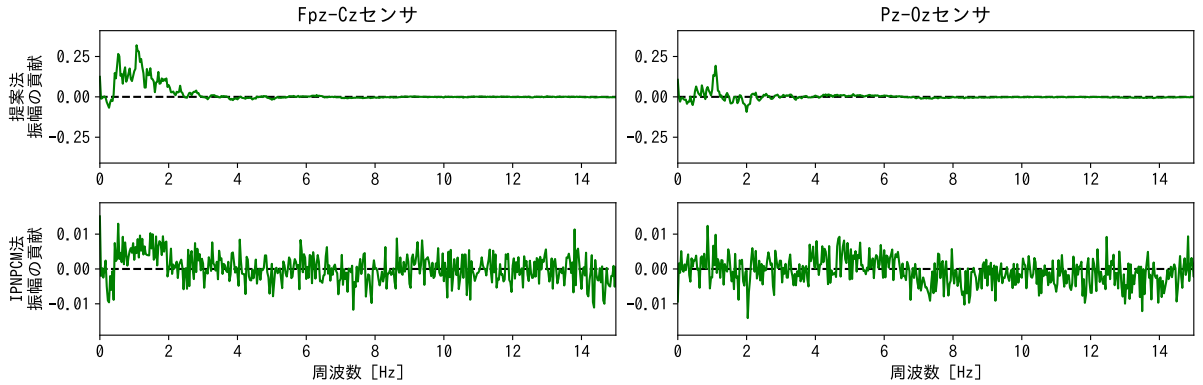

図 4: PhysioNet PSG データセットで N3 クラスの出力に貢献した振幅成分の 200 データ平均．位相成分は貢献が小さいため省略．上段：提案法での貢献．下段：IPNPCM 法での貢献．左：Fpz-Cz センサの貢献．右：Pz-Oz センサの貢献．

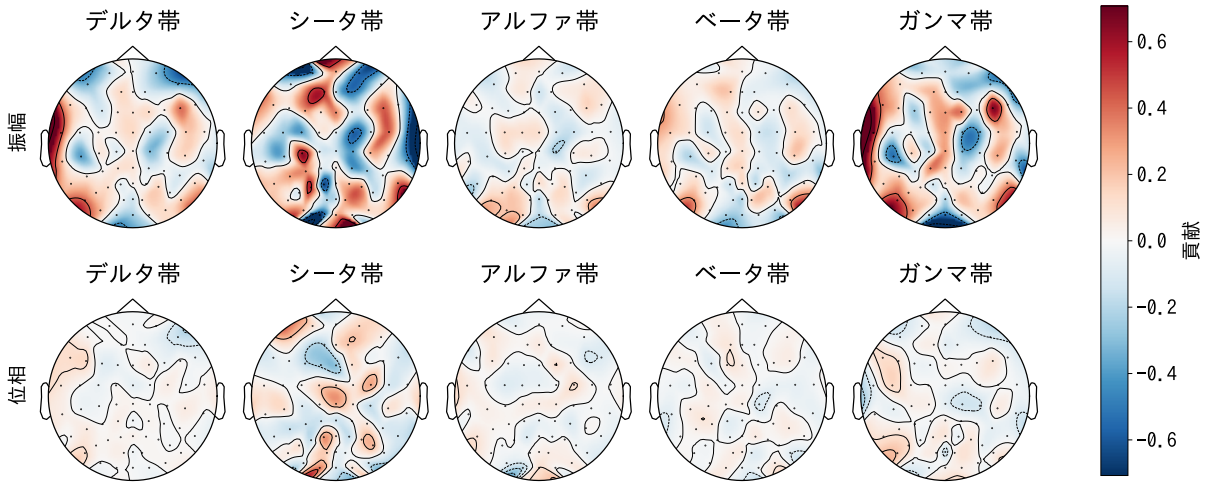

図 5: UCI EEG データセットでアルコール依存クラスの出力に貢献した特徴の 200 データ平均．提案法により貢献を計算．

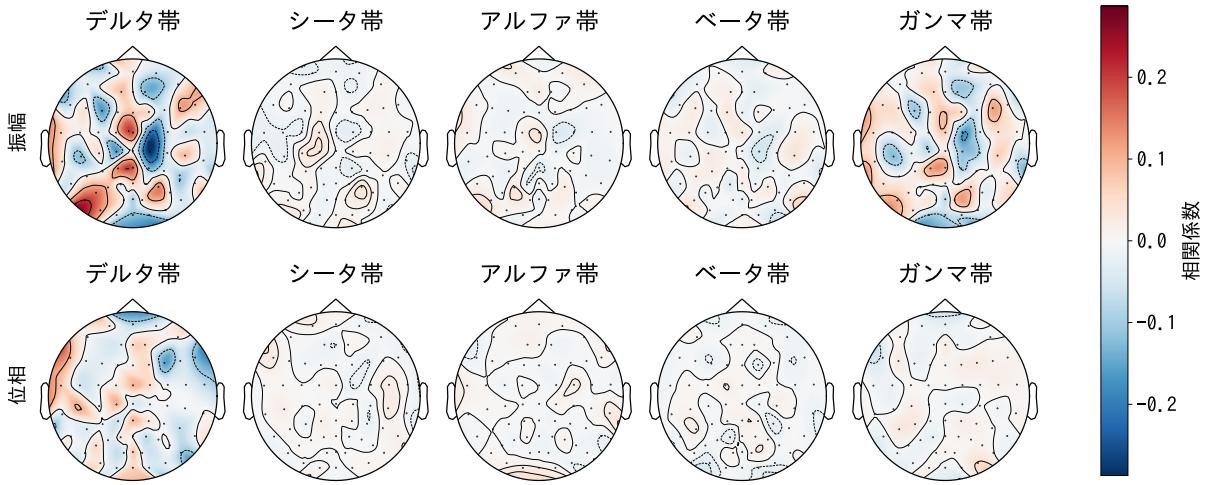

図 6: UCI EEG データセットでアルコール依存クラスの出力に貢献した特徴の 200 データ平均．IPNPCM 法により貢献を計算．