Compensated Integrated Gradients to Reliably Interpret EEG Classification

Kazuki Tachikawa, Yuji Kawai, Jihoon Park, Minoru Asada
Graduate School of Engineering,
Osaka University
`kazuki.tachikawa@ams.eng.osaka-u.ac.jp`

Abstract

Integrated gradients are widely employed to evaluate the contribution of input features in classification models because it satisfies the axioms for attribution of prediction. This method, however, requires an appropriate baseline for reliable determination of the contributions. We propose a compensated integrated gradients method that does not require a baseline. In fact, the method compensates the attributions calculated by integrated gradients at an arbitrary baseline using Shapley sampling. We prove that the method retrieves reliable attributions if the processes of input features in a classifier are mutually independent, and they are identical like shared weights in convolutional neural networks. Using three electroencephalogram datasets, we experimentally demonstrate that the attributions of the proposed method are more reliable than those of the original integrated gradients, and its computational complexity is much lower than that of Shapley sampling.

1 Introduction

Deep learning has become a promising method for image recognition, natural language processing, speech recognition, and even classification of diseases [1, 2] and evaluation of brain and body physiology [3, 4] from raw electroencephalogram (EEG) signals. In EEG classification, besides the results it is important to determine the reasons of the classification for its proper interpretation by a specialist. For instance, in medical decision support systems, interpretation enables doctors to corroborate automatic classification based on machine learning against medical knowledge and possibly unveil previously unnoticed features of a disease.

Several methods for visualizing the separate contribution of input features to classification have been proposed [4–8]. Among them, integrated gradients (IG) and Shapley sampling (SS) have been shown to be theoretically superior to other methods, because they satisfy axioms for the fair attribution of contributions [5, 7, 9]. Although the IG method is computationally efficient, it requires an appropriate baseline (reference point) for determining reliable contributions. The baseline is an input assumed to not include any features and is empirically set (usually, the zero point), as no formal methods for finding the appropriate baseline have been devised. Setting an inappropriate baseline can undermine the reliability of the attributions [10]. In contrast, the SS method does not require setting a baseline [9], but its computational cost is extremely high [7].

In this study, we propose a method for compensating the contributions obtained from IG at an arbitrary baseline by using SS contributions. The proposed method satisfies the same axioms as the IG method for an appropriate baseline under specific classifier constraints. We experimentally evaluate the reliability and computational complexity of the proposed method on three EEG datasets.
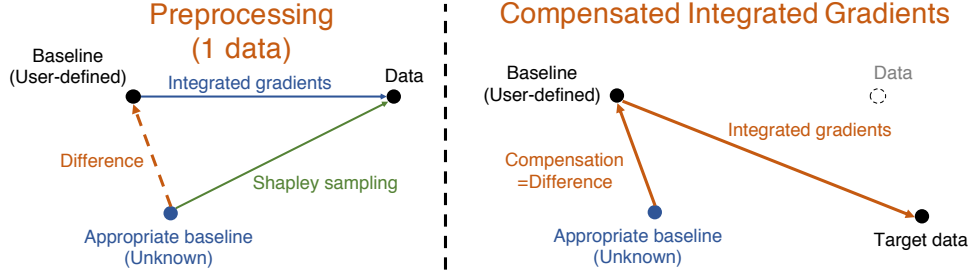
Figure 1: Diagram of the proposed method. Left panel: process to obtain the compensation amount. Right panel: path to compute a reliable contribution.

2 Compensated IG

The proposed compensated IG method is a type of path method [5, 11] that integrates gradients of the output with respect to the input of a classifier, usually before a softmax function in the output layer, along an arbitrary path from the baseline to the input data point. Given path function $\gamma(\alpha)$ for $\alpha \in [0, 1]$, classifier f, and input data x, the contribution of the i-th feature, $\mathrm{PathIG}_i^\gamma(x)$, is given by

$$\mathrm{PathIG}_i^\gamma(x) = \int_{\alpha=0}^{1} \frac{\partial f(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha.$$

This method satisfies the following four axioms [5, 11]:

1. Implementation invariance: If two classifiers are functionally equivalent, i.e., they retrieve equal outputs for every input, the attribution of contributions are also equivalent.

2. Dummy: If a classifier does not depend on some variables, the contribution of these variables is always zero.

3. Linearity: Let classifier f be represented by the weighted linear sum of two sub-classifiers f_1 and f_2, i.e., $f = af_1 + bf_2$ for scalars a and b. Then, the contribution of the classifier $\phi(f)$ is also obtained by the weighted linear sum of the contribution of each sub-classifier: $\phi(f) = a\phi(f_1) + b\phi(f_2)$.

4. Completeness: The sum of the contributions of all features equals the output value of the classifier.

If the path is a straight line, the path method further satisfies the symmetry axiom defined below (this method is called IG) [5].

5. Symmetry: Let the output of a classifier not always change, even if exchanging features x_i and x_j, i.e., $f(x_i, x_j) = f(x_j, x_i)$. Then, the contributions of these features are identical.

The SS method also satisfies all of these axioms, and therefore, this method corresponds to the IG method appropriately setting the reference point (green arrow in the left panel of Fig. 1).

In most cases, a zero input is used as a baseline instead of a better input (blue arrow in the left panel of Fig. 1). However, this zero input is often inappropriate, resulting in unreliable attribution of prediction [10]. To compensate the contribution from the user-defined zero input, we calculate the difference between the contribution obtained from this input and a reliable contribution obtained using the SS method for one data record (orange dashed arrow in the left panel of Fig. 1). Note that the computational cost of this step is not very high because the SS method is only applied to one data record. This difference corresponds to the integral of the gradients along an unknown path from the true baseline to the user-defined baseline. The value of the integral does not depend on the data, as it is determined only by the two baselines. The integral value is added to the contribution obtained from the user-defined baseline and into an arbitrary target data point to obtain the input/output gradients integrated along the orange path in the right panel of Fig. 1.

This operation satisfies axioms 1-4, besides the symmetry axiom (5) if the classifier processes of the input features are independent and identical, as proved in Appendix A. For example, the identity constraint is realized by a weight sharing technique in convolutional neural networks (CNNs).

Table 1: Spearman's correlation compared to contributions obtained from Shapley sampling. The temporal model corresponds to one-dimensional convolutional neural networks (CNNs), and the spatiotemporal model to two-dimensional CNNs.

		Temporal Model					Spatiotemporal Model		
Dataset	Class	C-IG	SS	IG	Dataset	Class	C-IG	SS	IG
PhysioNet	N1	**0.983**	0.970	0.180					
	N2	**0.970**	0.953	0.655					
	N3	**0.988**	**0.988**	0.925					
	R	0.963	**0.970**	0.665					
	W	**0.987**	0.972	0.326					
CHB-MIT	Szr.	**0.996**	0.994	0.817	CHB-MIT	Szr.	0.806	**0.983**	0.695
	No S.	**0.993**	0.990	0.293		No S.	0.917	**0.982**	-0.037
UCI EEG	Alc.	**0.994**	0.991	0.260	UCI EEG	Alc.	0.793	**0.989**	0.323
	Ctr.	**0.995**	0.992	0.258		Ctr.	0.739	**0.988**	0.331

C-IG, compensated integrated gradients (proposed method); SS, Shapley sampling; IG, integrated gradients; R, rapid eye movement; W, wakefulness; Szr., seizure; No S., no seizure; Alc., alcoholism; Ctr., control.

However, the contributions in spatial CNNs, which are widely used in image recognition, do not satisfy the symmetry axiom because each filter processes several input features, and therefore, they violate the independence constraint. In contrast, when using temporal CNNs, in which each filter convolutes an input time-series feature, the contributions satisfy all the axioms. Still, an inappropriate baseline in the original IG produces the violation of all axioms.

3 Experiments

We evaluated the reliability and computational cost of the proposed method on three publicly available EEG datasets, namely, the PhysioNet polysomnography dataset [12–14], UCI EEG dataset [15, 16], and CHB-MIT Scalp EEG dataset [17, 18]. For the PhysioNet polysomnography dataset, we trained six-layer temporal CNNs to classify the data into five sleep stages (N1, N2, N3, rapid eye movement, and wakefulness). These CNNs are one-dimensional and convolve input EEG signals separately from each other to acquire time-domain features that are integrated in the fully connected output layer. For both the UCI EEG and CHB-MIT Scalp EEG datasets, we trained five-layer temporal CNNs and four-layer spatiotemporal CNNs to classify the data into two classes (alcoholism/control and seizure/no seizure, respectively). The spatiotemporal CNNs are two-dimensional and prevent the proposed method from satisfying the symmetry axiom.

In each dataset, we randomly selected 200 data records of each class. For each classifier, we computed the contributions of input features using the proposed method, the IG method using the zero-input baseline, and the SS method. We used 10 additional data records on the proposed method for compensation to mitigate the sampling error in the SS method, although theoretically, one data record should be enough for compensation. Ideally, contribution comparison should be performed against true contributions, but they are unknown. Therefore, we considered the contributions obtained by the complete SS method as the true contributions and measured the similarity among contributions by the Spearman's correlation [19, 20]. Large correlation coefficients (close to 1) indicate high similarity between two compared contributions. Note that the comparison with the contributions obtained by the SS method reflects the sampling errors. Furthermore, we determined the contributions of EEG sensors (electrodes) to the classification of alcoholism in the UCI EEG dataset to qualitatively compare the methods. We used a temporal model and averaged the contributions from the 200 analyzed data records for this comparison.

4 Results

The Spearman's correlation coefficients of all datasets and models are listed in Table 1. The coefficient values of the proposed method, the compensated IG, were almost the same as those of SS on the temporal CNNs. In contrast, IG exhibited the lowest values, particularly in the UCI EEG dataset. Therefore, the proposed method can suitably compensate the unreliable contributions obtained from the IG method.
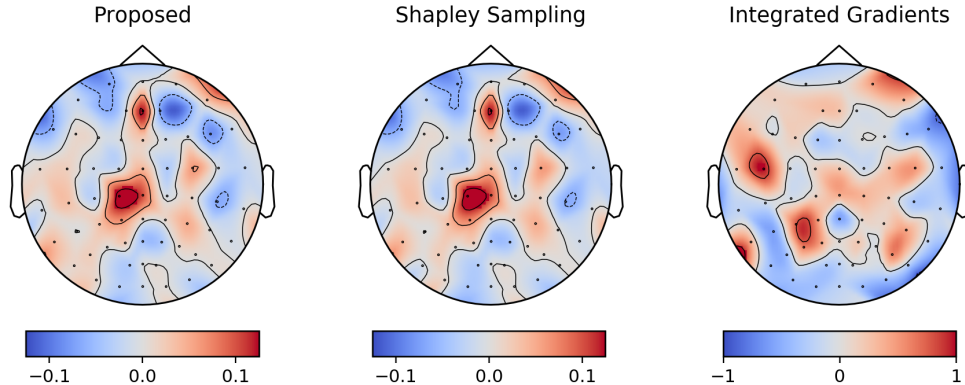
Figure 2: Contributions of EEG sensors (electrodes) over the scalp to classify alcoholism. The average contribution of the methods applied to 200 data records from the UCI EEG dataset is depicted. The red and blue areas represent positive and negative contributions to the classification of alcoholism, respectively.

Even when the compensated IG does not satisfy the symmetry axiom on the spatiotemporal CNNs, its coefficient values are larger than those of the original IG with its zero-input baseline. However, the values of the compensated IG were lower than those of SS in this case, indicating that the violation of the symmetry axiom undermines the reliability of the estimated contributions. Still, compensation improved the reliability of the original IG.

We defined the computational complexity by the number of backpropagations (for IG) × number of data records + number of forward propagations (for SS) × number of sensors (electrodes) × number of data records. Assuming equal computational costs for forward and backpropagation, we obtained computational costs of 40,000 for IG, 650,000 for the proposed method, and 12,200,000 for SS when applied to the UCI EEG dataset. If 1000 data points are targeted, the computational cost ratio is 20:81:6100.

Fig. 2 shows the contributions of the EEG sensors (electrodes) for alcoholism classification. The contributions obtained from the proposed method (left panel) were indistinguishable from the true contributions obtained from the SS method (middle panel), whereas those obtained from the original IG method (right panel) exhibit a different distribution from those obtained from the other methods.

5 Conclusion

We propose a compensated IG method using SS to improve the reliability to interpret classification outcomes. The proposed method satisfies four axioms, namely, implementation invariance, dummy, linearity, and completeness, besides an additional symmetry axiom under classifier constraints (see Appendix A). Using three EEG datasets, we demonstrate that the proposed method can compute more reliable contributions than the IG method with an inappropriate baseline and presents much lower computational cost than the SS method. The contributions obtained from the proposed method were very similar to those obtained from the SS method especially for temporal CNNs, which meet the constraints for the symmetry axiom. However, the classifier constraints decrease the classification accuracy (see Appendix B). In contrast, spatiotemporal CNNs exhibit higher classification accuracy but lower interpretation reliability than the temporal CNNs. Therefore, classifier selection should depend on whether reliability or classification accuracy are emphasized. Even when the symmetry axiom is violated given the convolution among input features, we demonstrated that compensation effectively improves the reliability of the original IG. In future developments, we will apply the proposed method to multichannel time-series data different from EEG, e.g., acceleration of human activities [21] and electrocardiography [22].

References

[1] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. Learning robust features using deep learning for automatic seizure detection. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 178–190, 2016.

[2] Kazuki Tachikawa, Yuji Kawai, Jihoon Park, and Minoru Asada. Effectively interpreting electroencephalogram classification using the Shapley sampling value to prune a feature tree. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 672–681, 2018.

[3] Orestis Tsinalis, Paul M. Matthews, Yike Guo, and Stefanos Zafeiriou. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. *arXiv:1610.01683*, 2016.

[4] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3319–3328, 2017.

[6] Yitong Li, michael Murias, samantha Major, geraldine Dawson, Kafui Dzirasa, Lawrence Carin, and David E. Carlson. Targeting EEG/LFP synchrony with neural nets. In *Advances in Neural Information Processing Systems 30*, pages 4620–4630. 2017.

[7] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4768–4777, 2017.

[8] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153, 2017.

[9] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[10] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. *NIPS workshop on Explaining and Visualizing Deep Learning*, 2017.

[11] Eric J. Friedman. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4):501–518, 2004.

[12] Ary L. Goldberger, Luis A.N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[13] Bob Kemp, Aeilko H. Zwinderman, Bert Tuk, Hilbert A.C. Kamphuisen, and Josefien J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[14] Bob Kemp. The Sleep-EDF Database. `https://www.physionet.org/physiobank/database/sleep-edfx/`.

[15] Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang, and Ann Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995.

[16] Henri Begleiter and Lester Ingber. UCI Machine Learning Repository: EEG Database Data Set. `https://archive.ics.uci.edu/ml/datasets/eeg+database`.

[17] Ali Hossam Shoeb. *Application of machine learning to epileptic seizure onset detection and treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[18] Ali Hossam Shoeb. CHB-MIT Scalp EEG Database. `https://www.physionet.org/pn6/chbmit/`.

[19] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *NIPS workshop on Machine Deception*, 2017.

[20] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local explanation methods for deep neural networks lack sensitivity to parameter values. *In Proceedings of the Workshop of the 6th International Conference on Learning Representations*, 2018.

[21] Ming Zeng, Le T. Nguyen, Bo Yu, Ole J. Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*, pages 197–205, 2014.

[22] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *Proceedings of the 15th International Conference on Web-Age Information Management*, pages 298–310, 2014.

Appendices

A Proof: all path methods satisfy the symmetry axiom when using temporal CNNs

Let an input time-series be n-dimensional: $x = [x_1, ..., x_i, ..., x_n]$, the length of each time-series z be L: $x_i = [z_{i1}, ..., z_{ik}..., z_{iL}]$, and a classifier $f(x)$ be given by

$$f(x) = \sum_i^n \sum_j^m W_{ij} g^m(x_i), \tag{1}$$

where, W_{ij} denotes a weight, $g^m(x_i)$ is a nonlinear function that converts time-series input x_i into outputs m. For the CNN considered in this study, $g^m(x_i)$ represents one-dimensional (temporal) convolutional layers or filters that share weights and W_{ij} represents the fully connected layer. The contribution of p-th feature x_p is represented as $\phi(f, x_p)$.

From Eq. (1),

$$\phi(f, x_p) = \phi\left(\sum_i^n \sum_j^m W_{ij} g^m(x_i), x_p\right).$$

Using linearity,

$$\phi\left(\sum_i^n \sum_j^m W_{ij} g^m(x_i), x_p\right) = \phi\left(\sum_j^m W_{pj} g^m(x_p), x_p\right).$$

Using the implementation invariance axiom and the symmetry assumption (let symmetric features be x_q),

$$W_{pj} = W_{qj}.$$

and

$$g^m(x_p) = g^m(x_q).$$

Therefore,

$$\phi\left(\sum_j^m W_{pj} g^m(x_p), x_p\right) = \phi\left(\sum_j^m W_{qj} g^m(x_q), x_q\right).$$

Then,

$$\phi(f, x_p) = \phi(f, x_q).$$

Consequently, all path methods satisfy the symmetry axiom if the processes of input features are independent and identical in a classifier.

B Classification Accuracy

Table 2: Classification accuracy of different models on three datasets.

Dataset	Model	Accuracy
PhysioNet	Temporal	81.6%
CHB-MIT	Temporal	83.4%
	Spatiotemporal	88.7%
UCI EEG	Temporal	69.5%
	Spatiotemporal	77.1%