EEG Artifact Removal by Bayesian Deep Learning & ICA

Sangmin S. Lee, Kiwon Lee and Guiyeom Kang

Abstract— **Artifact removal is important for EEG signal processing because artifacts adversely affect analysis results. To preserve normal EEG signal, several methods based on independent component analysis (ICA) have been studied and artifacts are removed by discarding independent components (ICs) classified as artifacts. In this study, a method using Bayesian deep learning and attention module is presented to improve performance of the classifier for ICs. Probability value is computed through the method to predict if a component is artifact and to treat ambiguous inputs. The attention module achieves increasing classification accuracy and shows the map of the region where the classifier concentrates on.**

I. INTRODUCTION

Electroencephalogram (EEG) signals are electrical signals caused by brain activity. Analyzing EEG signal is regarded as useful method for monitoring brain function because of its high temporal resolution. However, EEG signal is disrupted by several types of artifact caused by eye blinking and muscle activity. Recent studies have developed the techniques to remove artifacts [1], [2].

To preserve EEG signal from artifact removal, the ICA is widely used [3], [4]. Artifact removal method with ICA based on the joint approximate diagonalization of eigen-matrices (JADE) algorithm was suggested [5]. They identified the artifactual components using features obtained by JADE components. However, it has a disadvantage that human should intervene for selection of artifactual components. To automate the artifact removal process, [6], [7] developed the classification method using the local binary pattern (LBP) feature which is widely used in image classification tasks. They computed LBP features from topographic maps, and used the linear discriminant analysis (LDA) to distinguish artifacts from topographic maps. However, this method was not reliable of its accuracy due to poor capability of classifier.

Classification methods based on the neural network have been studied in recent years [8], [9]. These can consider huge variation of features resulting high performance. Artificial neural network (ANN) was used as a classifier trained on several features from the topographic images obtained by ICA and produced high accuracy above 90% [10]. However, there is a disadvantage with the method as data with unseen distribution during the training process may not be predicted accurately. It is because that modern neural networks are not calibrated well and hence predictions do not reflect their confidence [11], [12]. It may be a problem because the neural network can be trained to classify the normal EEG signal as an artifact, resulting different analysis.

To overcome this problem novel classification method for artifact removal was developed by this study. Suggested method is based on Bayesian deep learning, which is used for measuring uncertainty of prediction [13]. Monte-carlo dropout (MC-dropout) method which can approximate deep gaussian process for Bayesian inference was used to estimate prediction results as a Beta distribution [14], [15]. After estimating the distribution of prediction, probability value (p-value) was computed for selecting artifacts with high confidence. In addition, the attention module was adapted to the neural network to improve performance by making the neural network concentrate on the region of interest according to input topographic map. The attention map from the attention module also can be used for interpreting suggested neural network.

II. METHOD

A. Independent Component Analysis

Let input EEG signal from N sensors be $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T$ where each EEG signal with length t from channel c is column vector $\mathbf{x}_c = [x(1), x(2), ..., x(t)]$. The purpose of the ICA algorithm is to find a unmixing matrix $\mathbf{W}$ which satisfies

$$\mathbf{S} = \mathbf{W}\mathbf{X} \tag{1}$$

where $\mathbf{S}$ is estimated source. To remove artifact of EEG, artifact-free sources $\mathbf{S}'$ is obtained by discarding artifact source vectors from $\mathbf{S}$. In addition, the position on scalp where each source came from is obtained by computing row vectors of $\mathbf{W}$. Therefore, the topographic map of source component can be drawn according to $\mathbf{W}$ and proper montage.

B. Convolutional Neural Network with Attention Module

The classifier based on the convolutional neural network (CNN) was devised to classify topographic maps on training set, $\mathbf{D} = \{\mathbf{I}, \mathbf{Y}\}$ where $\mathbf{I}$ is set of topographic maps from ICA and $\mathbf{Y}$ is set of labels. The whole process of classifier can be represented as $\mathbf{F}_\phi(\cdot)$ that is parameterized by ϕ. Proposed classifier was trained to distinguish if input topographic map is from normal EEG or one of three types of artifacts (e.g. EMG, VEOG, HEOG).

Figure 1 shows the architecture of suggested classifier. It consists of six convolutional layers and three fully connected layers. The attention module was added on the classifier to refine the spatial features of the classifier. Extracted features

This research was supported by the Ministry of Trade Industry & Energy (MOTIE, Korea), Ministry of Science & ICT (MSIT, Korea), and Ministry of Health & Welfare (MOHW, Korea) under Technology Development Program for AI-Bio-Robot-Medicine Convergence (20001650).

Authors are with the Ybrain, Seongnam-si, Republic of Korea (corresponding author to provide e-mail: sangmin.lee@ybrain.com)

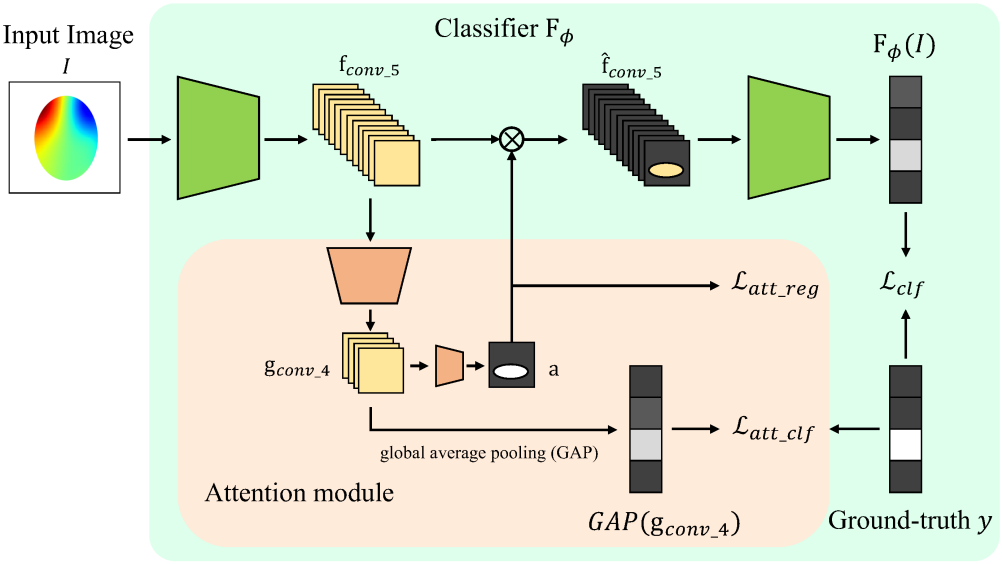

Fig. 1. Overall structure of the proposed classifier.

from fifth convolution layer, $\mathbf{f}_{\text{conv\_5}}$, passed the attention module, which consists of three convolutional layers and sigmoid activation function to make attention maps $\mathbf{a} = \mathbf{G}(\mathbf{f}_{\text{conv\_5}})$ ranged from zero to one. These attention maps were element-wise multiplied to original features for making refined features $\hat{\mathbf{f}} = \mathbf{a} * \mathbf{f}_{\text{conv\_5}}$. For more informative attention maps according to their labels, the output feature maps of the second layer of the attention module $\mathbf{g}_{\text{conv\_2}}$ passed the global average pooling (GAP) layer and were used to predict corresponding labels. These processes made the remaining layers concentrate on important spatial parts for classification [16], [17]. For multi-class classification, the cross-entropy loss was used as below

$$\mathcal{L}_{clf} = \sum_{i\in\{eeg,eog,veog,heog\}} y_i \log(\mathbf{F}_\phi(\mathrm{I}))_i \tag{2}$$

where $\mathrm{I} \in \mathbf{I}$ is an input topographic map and $y \in \mathbf{Y}$ is corresponding class label.

To make well-refined feature, two types of loss were devised. First, the attention classification loss $\mathcal{L}_{att\_clf}$ was to make the attention map have more accurate information about its corresponding label. The other loss, the attention regularization loss, made the attention map be sparse enough to emphasize important part of feature map. Attention loss is shown in below

$$\mathcal{L}_{att\_clf} = \sum_{i=\{eeg,eog,veog,heog\}} y_i \log(GAP(\mathbf{g}_{\text{conv\_2}})_i \tag{3}$$

$$\mathcal{L}_{att\_reg} = \| \mathbf{a} \|_2^2 \tag{4}$$

$$\mathcal{L}_{att} = \lambda_{att\_clf} \cdot \mathcal{L}_{att\_clf} + \lambda_{att\_reg} \cdot \mathcal{L}_{att\_reg} \tag{5}$$

where $\| \cdot \|_2$ means L2-norm and $\lambda_{att\_clf}, \lambda_{att\_reg}$ are weight terms of classification, regularization loss for attention, respectively. Finally, the overall loss can be expressed as below

$$\mathcal{L} = \lambda_{clf} \cdot \mathcal{L}_{clf} + \mathcal{L}_{att} + \lambda_{reg} \cdot \mathcal{L}_{reg} \tag{6}$$

where $\mathcal{L}_{reg}$ is regularization loss term for avoiding overfitting and $\lambda_{clf}, \lambda_{reg}$ are weight terms of multi-class classification loss, regularization loss, respectively.

C. Bayesian Deep Learning Inference

After the training process is over, Bayesian deep learning inference based on the MC-dropout was conducted to estimate output distribution. The soft-max output of the classifier $\mathbf{F}_\phi(\mathbf{i})$ can be assumed to follow Dirichlet distribution as it is the vector whose elements are all positive while sum of them is exactly one. For simplicity, probability density function (PDF) of the soft-max output was computed only for EEG class as the Beta distribution which is marginal distribution of the Dirichlet distribution. To approximate these parameters, MC-dropout was conducted and sample mean $\bar{\mu}_{eeg} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{F}_{\phi_n}(\mathbf{i})_{eeg}$, sample variance $\bar{\sigma^2}_{eeg} = \frac{1}{N-1}\sum_{n=1}^{N} (\mathbf{F}_{\phi_n}(\mathbf{i})_{eeg} - \bar{\mu}_{eeg})^2$ were obtained where N is total number of inference. As a result, the output distribution indicating probability, that the input image $\mathbf{i}$ was from the class EEG, can be obtained as below:

$$p(\mathbf{F}_\phi(\mathbf{i})_{eeg}) \sim \mathrm{Beta}(\hat{\alpha}, \hat{\beta}) \tag{10}$$

$$\hat{\alpha} = \bar{\mu}_{eeg}\left(\frac{\bar{\mu}_{eeg}(1-\bar{\mu}_{eeg})}{\bar{\sigma^2}_{eeg}} - 1\right) \tag{11}$$

$$\hat{\beta} = (1-\bar{\mu}_{eeg})\left(\frac{\bar{\mu}_{eeg}(1-\bar{\mu}_{eeg})}{\bar{\sigma^2}_{eeg}} - 1\right) \tag{12}$$

After estimating the distribution, p-value in case of the input $\mathbf{i}$ is the normal EEG can be computed with the cumulative distribution function (CDF) $F(x; \hat{\alpha}, \hat{\beta})$

$$Pr(\mathbf{F}_\phi(\mathbf{i})_{eeg}) \geq \frac{1}{2}) = 1 - F(\frac{1}{2}; \hat{\alpha}, \hat{\beta}) = 1 - I_{\frac{1}{2}}(\hat{\alpha}, \hat{\beta}) \tag{13}$$

where $I_x(\alpha, \beta)$ is the regularized incomplete beta function. To accomplish the artifact removal with handling uncertainty, only the independent source whose p-value is smaller than 0.05 is discarded.

III. Experiments & Results

A. Datasets & Environments

For the experiment, a set of 19-channel EEG signals from 10-20 system was used. Each signal was filtered to remove line noise and to preserve information across different bands of EEG data. After the filtering process, the artifact subspace reconstruction (ASR) method [18] was adapted to eliminate an artifact which is distributed throughout the whole of the scalp with large variance. The infomax-ICA method was then conducted to obtain a set of ICs constituting EEG and artifacts [19]. Finally, topographic maps were gained and labeled into 4 classes by EEG experts.

The image resolution is 300×300 and the total number of images is 12,020 (2,759, 2,425, 3,114, 3,722 for EEG, EMG, VEOG, HEOG, respectively). To avoid the situation where classifier is biased with imbalance dataset, total sample numbers of each class were equally set by data augmentation during the training. 5-fold cross validation was used for validating the method. $\lambda_{clf}, \lambda_{att\_clf}, \lambda_{att\_reg}, \lambda_{reg}$ were set to $1.0, 0.5, 0.01, 0.01$, respectively.

B. Qualitative Results

Figure 2 shows that the two samples of topographic map and their MC-dropout results. The first sample is a topographic map of HEOG. As the signals in frontal electrode show opposite polarity, the topographic map can be easily classified as an artifact. The p-value was 0.001 and it means that the classifier predicted the input as artifact with high confidence. However, in the case of the second topographic map, opposite polarity was not clearly shown at the frontal electrode and it is hard to classify it as HEOG. It resulted in the output distribution to be slightly right-biased and the p-value was 0.217 which would not be eliminated in our criteria.

Figure 3 shows the result of artifact removal using our method. First, second, third graphs show the signal before the artifact removal, after the artifact removal, and the signal reconstructed with artifact components only, respectively. The signal in the last graph was reconstructed with components whose p-values were between 0.05 and 0.95. It was shown that the EOG artifacts at channels Fp1, Fp2 were clearly removed at 2.5s and 4.5s and the EMG artifacts at the channels T3, T4 were removed during 5 seconds. It was notable that the ambiguous signals shown in Figure 3 (d) is not seen as an artifact.

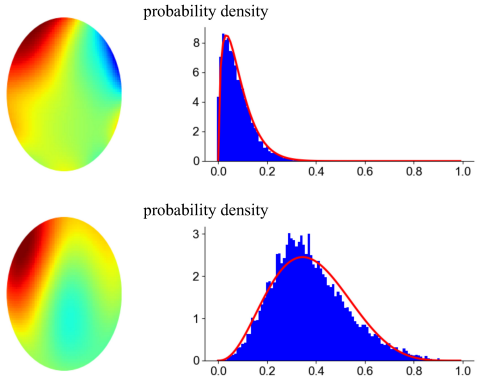

Fig. 2. Two samples of topographic map and their MC-dropout results. The first column contains similar two topographic maps. In the second column, there are two results from corresponding topographic maps. The x-axis is a soft-max output value for EEG. The blue graphs are histograms which show the normalized number of output for the case of EEG from MC-dropout. The red graphs are PDFs of estimated Beta distribution from MC-dropout results.

C. Classification Results

Figure 4 shows the visualization of attention maps from four samples of topographic map. As shown, attention maps highlighted the regions where the power was the most

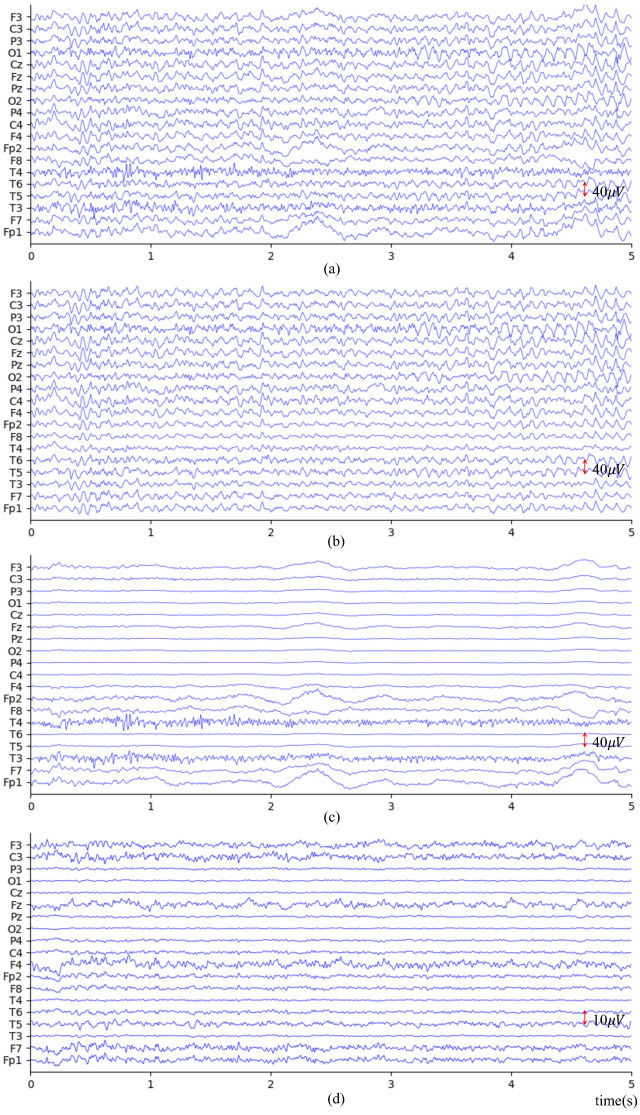

Fig. 3. The result of artifact removal from suggested method. (a) shows the signal before artifact removal. (b) shows the signal after artifact removal. The signal reconstructed with components that are classified as an artifact is shown in (c). The signal whose p-value is within [0.05, 0.95] which means that it is not able to declare clearly whether the signal is an artifact or an EEG is shown in (d).

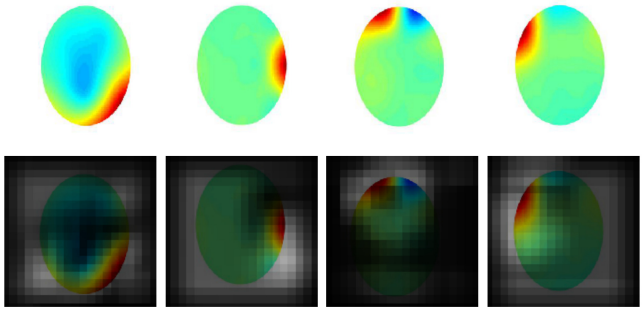

Fig. 4. Attention module visualization results. The first row consists of four samples of topographic map. The second row shows corresponding topographic maps multiplied with attention maps.

concentrated on. A baseline classifier was implemented to validate if refined features with attention maps improve performance. For a fair comparison, the baseline classifier consisted of same structure with suggested classifier except that it had no attention module. The accuracy of baseline classifier and suggested classifier were 94.48% and 95.86%, respectively. This result shows that the implemented attention module was helpful for classification task.

IV. Discussion

In this paper, a new artifact removal method based on Bayesian deep learning and ICA was developed. The method enables the confidence level of the classifier for its prediction results. The Bayesian inference with MC-dropout resulted classification results to be approximated with the Beta distribution. Probability values, which the topographic maps are obtained from EEG can be computed by using CDF of the Beta distribution. Indeed, the result showed that components whose p-value were ranged from 0.05 to 0.95 are not clear to be classified with artifacts. Therefore, unnecessary removal of ambiguous signal can be avoided. The artifact classification is just one example of applications and suggested method can be adapted to any other studies for classification where the confidence level of prediction is important.

Furthermore, devised attention module was helpful for interpreting the classifier. By visualizing the attention map, the region of interest where the classifier concentrates on can be obtained. If one wants to know the reason why the classifier predicts the label of corresponding input because of low confidence, it is possible to analysis with each result and an attention map from individual inference. The accuracy showed that the refined features with attention module are helpful for improving the performance of the classifier by comparing with the baseline classifier.

However, there are disadvantages in the method developed in this study. The threshold of p-value should be decided experimentally and it requires human intuition. In addition, the Bayesian inference based on MC-dropout costs huge computational resources because of iteration process for distribution estimation. It may cause too much overhead for pre-processing of EEG. As a result, suggested method cannot be regarded as real-time solution. Nevertheless, the method facilitates certainty in classification for artifacts in EEG and it could be helpful to acquire EEG signals from the sources badly contaminated by artifacts.

References

[1] I. Winkler, S. Debener, K. Müller, and M. Tangermann, "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-erp," in *Proc. 37th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 4101–4105.

[2] M. Kiamini, S. Alirezaee, B. Perseh, and M. Ahmadi, "Elimination of ocular artifacts from EEG signals using the wavelet transform and empirical mode decomposition," in *Proc. Telecommunications and Information Technology 2009 6th Int. Conf. Electrical Engineering/Electronics, Computer*, vol. 02, May 2009, pp. 1094–1097.

[3] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[4] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in neural information processing systems*, 1996, pp. 145–151.

[5] J. Iriarte, E. Urrestarazu, M. Valencia, M. Alegre, A. Malanda, C. Viteri, and J. Artieda, "Independent component analysis as a tool to eliminate artifacts in eeg: a quantitative study," *Journal of clinical neurophysiology*, vol. 20, no. 4, pp. 249–257, 2003.

[6] T. Radüntz, J. Scouten, O. Hochmuth, and B. Meffert, "Eeg artifact elimination by extraction of ica-component features using image processing algorithms," *Journal of neuroscience methods*, vol. 243, pp. 84–93, 2015.

[7] R. C. M. P. Gilberet, R. S. Roy, N. Sairamya, D. N. Ponraj, and S. T. George, "Automated artifact rejection using ica and image processing algorithms," in *2017 International Conference on Signal Processing and Communication (ICSPC)*. IEEE, 2017, pp. 354–358.

[8] P. Croce, F. Zappasodi, L. Marzetti, A. Merla, V. Pizzella, and A. M. Chiarelli, "Deep convolutional neural networks for feature-less automatic classification of independent components in multi-channel electrophysiological brain recordings," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 8, pp. 2372–2380, Aug. 2019.

[9] P. Nejedly, J. Cimbalnik, P. Klimes, F. Plesinger, J. Halamek, V. Kremen, I. Viscor, B. H. Brinkmann, M. Pail, M. Brazdil, *et al.*, "Intracerebral eeg artifact identification using convolutional neural networks," *Neuroinformatics*, vol. 17, no. 2, pp. 225–234, 2019.

[10] T. Radüntz, J. Scouten, O. Hochmuth, and B. Meffert, "Automated eeg artifact elimination by applying machine learning algorithms to ica-based features," *Journal of neural engineering*, vol. 14, no. 4, p. 046004, 2017.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1321–1330.

[12] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, 2018, pp. 3179–3189.

[13] Y. Gal, "Uncertainty in deep learning," *University of Cambridge*, vol. 1, p. 3, 2016.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[15] A. Damianou and N. Lawrence, "Deep gaussian processes," in *Artificial Intelligence and Statistics*, 2013, pp. 207–215.

[16] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[17] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[18] C. A. E. Kothe and T.-P. Jung, "Artifact removal techniques with signal reconstruction," Apr. 28 2016, uS Patent App. 14/895,440.

[19] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural computation*, vol. 7, no. 6, pp. 1129–1159, 1995.