Toward Subject Invariant and Class Disentangled Representation in BCI via Cross-Domain Mutual Information Estimator

Eunjin Jeon¹, Wonjun Ko¹, Jee Seok Yoon¹, and Heung-Il Suk¹²
¹Department of Brain Cognitive Engineering, Korea University
²Department of Artificial Intelligence, Korea University
Anam-Ro 145, Seoul 02841, Republic of Korea
{eunjinjeon, wjko, wltjr1007, hisuk}@korea.ac.kr

Abstract
In recent, deep learning-based feature representation methods have shown a promising impact in electroencephalography (EEG)-based brain-computer interface (BCI). Nonetheless, due to high intra- and inter-subject variabilities, many studies on decoding EEG were designed in a subject-specific manner by using calibration samples, with no much concern on its less practical, time-consuming, and data-hungry process. To tackle this problem, recent studies took advantage of transfer learning, especially using domain adaptation techniques. However, there still remain two challenging limitations: i) most domain adaptation methods are designed for labeled source and unlabeled target domain whereas BCI tasks generally have multiple annotated domains. ii) most of the methods do not consider negatively transferable to disrupt generalization ability. In this paper, we propose a novel network architecture to tackle those limitations by estimating mutual information in high-level representation and low-level representation, separately. Specifically, our proposed method extracts domain-invariant and class-relevant features, thereby enhancing generalizability in classification across. It is also noteworthy that our method can be applicable to a new subject with a small amount of data via a fine-tuning, step only, reducing calibration time for practical uses. We validated our proposed method on a big motor imagery EEG dataset by showing promising results, compared to competing methods considered in our experiments.

Introduction
Brain-computer interface (BCI) allows users to directly communicate or control external devices by thoughts, typically measured with electroencephalography (EEG) (Graimann, Allison, and Pfurtscheller 2009). EEG signals that measure the electrical activity of brain are usually categorized into two types, evoked and spontaneous, depending on their inducing manner in non-invasive BCIs. Evoked EEGs, e.g., steady-state visually evoked potentials, steady-state somatosensory evoked potentials, and event-related potentials, are derived from immediate automatic responses to an external stimulus regardless of a user’s own will, whereas spontaneous EEGs induce activation of event-related (de)synchronization (ERD/ERS) when carrying out mental tasks at a user’s own will. Of various EEG signal types, we focus on motor imagery signals characterized by ERD/ERS induced by imagining body movements without any physical movements. Taking advantage of controlling system without explicit commands, there have been many studies on decoding motor imagery using machine learning-based methods.

For motor imagery EEG, because of high variability among subjects (inter-subject) and sessions for the same subject (intra-subject) on account of inherent background neural activities, fatigue, concentration levels, etc. (Jayaram et al. 2016), it is challenging to train a generic model, that can be applicable to different datasets or subjects. Therefore, training a model for each subject is a typical approach to decode brain signals despite time-consuming and amount of data-quirifiable process (Lin and Jung 2017). In order to address the limitation, previous studies involved multiple subjects and/or sessions data simultaneously in training a single model in the way of transfer learning (Jayaram et al. 2016; Lin and Jung 2017) and presented its potential.

We focus on, in this work, boosting model generalization among subjects in a transfer learning manner (Jayaram et al. 2016; Lin and Jung 2017), by considering a subject as a domain (Pan and Yang 2009). In the meantime, due to the property of motor imagery EEG, i.e., inter-subject high variability (Jayaram et al. 2016), it may be observed that a large distributional discrepancy on the feature representation space between different domains, i.e., different subjects, which referred to as a domain shift (Pan and Yang 2009; Chai et al. 2016; Jeon, Ko, and Suk 2019). Therefore mitigating the domain shift is one of the important issues in the transfer learning for BCI tasks (Chai et al. 2016; Jeon, Ko, and Suk 2019).

Regarding the domain shift problem, there have been numerous studies in machine learning, which referred to domain adaptation (Ganin et al. 2016; Wang and Deng 2018). However, a direct application of domain adaptation techniques to BCI tasks is challenging because BCI tasks generally exploit multiple annotated domains, i.e., subjects, whereas the previous domain adaptation methods have been mostly studied between two domains, labeled source and unlabeled target domain (Ganin et al. 2016; Long et al. 2015;
Meanwhile, recent studies have reported that the domain adaptation methods may not only improve the feature representation of target domain, but corrupt it. This corruptive phenomena is called negative transfer (Wang et al. 2019; Peng et al. 2019; Xu et al. 2018). To address this issue, (Wang et al. 2019) indicated that a source domain includes not only transferable data, but also untransferable factors with respect to domain adaptation. Further, (Peng et al. 2019) sorted the source domain as domain-invariant, domain-specific, and class-irrelevant. To this end, it is of great importance to differentiate positive and negative transferable factors from data.

In respect of applying domain adaptation methods to BCI tasks, we consider domain shifts from multiple domains in a single procedure jointly by estimating mutual information from multiple subjects, which is different from the previous studies that mostly considered two domains (Ganin et al. 2016; Tzeng et al. 2017; Long et al. 2015) or calculated one-to-one discrepancies even for multiple domains (one source vs. target) (Zhao et al. 2018; Xu et al. 2018). In other words, the proposed method aims at finding domain-invariant, i.e., domain-unspecific, feature representations, applicable to multiple domains, allowing to quick fine-tuning with a minimal number of samples for a target domain, if necessary.

In terms of transferability, we decompose the domain-invariant features into class-irrelevant and class-relevant using another mutual information. Technically the proposed method minimizes mutual information between class-relevant and irrelevant features, thereby disentangling feature representations. Further, in the viewpoint of meta-learning (Vanschoren 2018), we build a generic model that can classify EEG signals of different subjects and can quickly adapt to a new subject by fine-tuning with a small amount of data of the target subject.

We evaluated our proposed network on GIST-Motor Imagery dataset (Cho et al. 2017) for a classification task. Our experimental results demonstrated that (i) the class disentanglement enhances the performance and (ii) maximizing mutual information between subjects diminishes the distributional differences. Our results show promising performance compared to the comparative methods trained individually, even with fewer samples. The main contributions of this work can be three-fold:

- First, we propose a deep neural network that disentangles domain-invariant, class-irrelevant, and class-relevant feature representations by means of mutual information in an end-to-end manner.
- Second, the proposed network required smaller calibration data to generalize on a new subject, thereby achieving an impactful utility in practice.
- Finally, although the proposed network is trained with multiple subjects’ data, it shows plausible results on classification performance in comparison with baseline methods trained individually.

Related Work

Transfer learning in BCI

In general, many methods of decoding EEGs are devised for an individual use with independent training per subject due to high inter- and intra-subject variability. However, it is required plenty of data and time-consuming (Lin and Jung 2017). To cope with the limitation, various studies have been focused on training a model with multiple subjects and/or sessions with the goal of learning a subject-invariant representation. They are applied in two ways: (i) zero-shot or few-shot learning for decoding unseen subjects’ data in training (¨Ozdenizci et al. 2019; Kindermans et al. 2014; Kang and Choi 2014; Azab et al. 2019) and (ii) improving performance by incorporating other subjects’ data (Lin and Jung 2017; Wei et al. 2018; Jeon, Ko, and Suk 2019). Especially, deep learning-based methods (Chai et al. 2016; Jeon, Ko, and Suk 2019; ¨Ozdenizci et al. 2019) approach this issue from a domain adaptation standpoint. They all constrain the distributional discrepancy between subjects by minimizing the maximum mean discrepancy in the latent space (Chai et al. 2016) or making the encoder to confuse the domain (Jeon, Ko, and Suk 2019; ¨Ozdenizci et al. 2019). Among them, two studies fundamentally devised for nothing but two subjects (Chai et al. 2016; Jeon, Ko, and Suk 2019).

In (¨Ozdenizci et al. 2019), a classifier is trained from the pre-trained subject-invariant representation, hence, confusing the domain and identifying the class are not trained in an end-to-end manner. Our proposed network achieves the domain adaptation among multiple domains, i.e., subjects, and classification simultaneously in an end-to-end manner.

Domain Adaptation

There are various studies to mitigate differences between source and target domains. (Wang and Deng 2018) categorized into one-step and multi-step approaches by the presence of intermediate domain which decreases a gap between source and target domain. In one-step domain adaptation, most methods achieve it by means of a domain discriminator, which guides its features to be indistinguishable between domains by reversing its gradient during training via adversarial learning (Ganin et al. 2016; Tzeng et al. 2017; Zhang et al. 2018). However, they consider a single target domain by using also one source domain. In the case of multiple source domains, there are multiple domain discriminators which identify the domain between a target domain and each source domain, separately (Zhao et al. 2018; Xu et al. 2018). Since we should denote a specific target domain to take advantage of existing methods, they are inapplicable for our task dealing with several source domains. Therefore, we propose a novel paradigm based on mutual information for minimizing the distributional discrepancy. In order to do this, we assume that a latent space maximized the mutual information between multiple domains can be viewed as a common shared space among them.

Representation Disentanglement

Recent researchers have concentrated on disentangled representation learning in various fields such as image trans-
Subject 1
Subject 2

Subject N

Figure 1: Overview of the proposed network. We randomly select trials regardless of subjects for a mini-batch. After a local encoder $E_l$ maps an input $x$ to a local feature $f_l$, a global encoder $E_g$ receives the local feature, then, extracts a global feature $f_g$. To embed subject-invariant feature, we maximize mutual information between two features by using a local discriminator $T_l$ and global discriminator $T_g$. We split the global feature $f_g$ into class-relevant feature $f_{re}$ and class-irrelevant feature $f_{ir}$. Then, the classifier $C$ takes $f_{re}$ and discriminates the corresponding class. Through a gradient reversal layer (GRL), we reduce mutual information between $f_{re}$ and $f_{ir}$ estimated from a mutual information neural estimator $M$. Finally, the decoder $D$ reconstructs the original input from the concatenated feature $f_g$.

Proposed Methods

In this work, we regard each subject as one domain. Thus, we assume that it is given for $S$ subjects $D^s = \{ (x^{(i)}, y^{(i)}) \}_{i=1}^{n_s}$ with $n_s$ labeled samples, where $s \in \{ 1 \ldots, S \}$. For uncluttered, we omit the superscript (i) without loss of generality. Let $x \in \mathbb{R}^{n_c \times n_f}$ denotes a power spectral density transformed from a raw EEG trial including spatio-spectral information, where $n_c$ and $n_f$ are the number of electrode channels and frequency bins, respectively. We also define $y$ as the corresponding class label. The goal of this work is to build a deep neural network robustly applicable for multiple subjects. In other words, we want to develop an intention identification system that can be generalized for all subjects, i.e., minimize the classification errors. Further, the generalized network can be applied to quickly decode a new subject’s intention through fine-tuning.

For the classification, we propose a method to maximize the mutual information among subjects and to disentangle a latent representation to the class-relevant and class-irrelevant features.

An overall framework of our proposed method is shown in Figure 1. Basically, our proposed network has an autoencoder structure. In the encoding-related block (blue box in Figure 1), an input $x$ is embedded to the subject-invariant latent space through an encoder $E$, i.e., $E_l + E_g$. In the decoding-related block (yellow box in Figure 1), there are three components; (i) a classifier $C$ to identify the class of an input $x$; (ii) a mutual information estimator $M$ to diminish dependency among two features, i.e., class-relevant feature $f_{re}$ and class-irrelevant feature $f_{ir}$; (iii) the decoder $D$ to reconstruct an input $x$ with a concatenated feature $f_g$, i.e., $[f_{re}, f_{ir}]$. Later, the trained network is fine-tuned for a new subject by using his/her data only.

Subject-Invariant Feature Embedding

First, we introduce a method to embed an input to the subject-invariant latent space by maximizing mutual information among all subjects. We exploit a deep neural network, i.e., deep infomax (DIM), maximizing the mutual information between input and encoder’s output (Hjelm et al. 2019). By setting a mini-batch in stochastic gradient descent based learning with samples of multiple subjects randomly, mutual information across subjects can be estimated by the DIM framework. Further, maximizing the mutual information leads to minimize the discrepancy among subjects.

We utilize an encoder $E$ and a discriminator $T$ to maximize mutual information analogous with DIM (Hjelm et al. 2019). Here, both encoder $E$ and discriminator $T$ have two types (local and global characteristic). The local encoder $E_l$
maps an input $x$ to a local feature $f_l = E_l(x) \in \mathbb{R}^{h \times w \times d_l}$, where $h$, $w$, and $d_l$ represent the height, width, and depth of the feature. The global encoder $E_g$ computes a global feature $f_g = E_g(f_l) \in \mathbb{R}^{d_g}$ from the corresponding local feature, where $d_g$ is the depth of the global feature. In other words, the encoder $E_l$ is structured such that it produces an intermediate local feature and a global feature in order.

Then, we estimate and maximize the mutual information between the local feature and the global feature using the local discriminator $T_l$ and the global discriminator $T_g$. While both discriminators take the same local and global features as inputs, they maximize the mutual information in similar but different ways. First, two encoders $E_l, E_g$, and the global discriminator $T_g$ are trained by maximizing

$$\max_{E_l, E_g, T_g} \sum_{i=1}^{N} \mathcal{L}_g(f_g^{(i)}, f_l^{(i)})$$

where $N$ is the number of samples in a mini-batch and $\mathcal{L}_g$ is derived from Jensh Shannon divergence (JSD)-based mutual information between the joint distribution $P_{j}$ and the product of marginal distributions $P_{m}$ (Nowozin, Cseke, and Tomioka 2016). Eq. (1) is rewritten as follows:

$$\max_{E_l, E_g, T_g} \sum_{i=1}^{N} \mathbb{E}_{p_j}[-\text{sf}(-T_g(f_g^{(i)}, f_l^{(i)}))]$$

$$- \mathbb{E}_{p_m}[\text{sf}(T_g(f_g^{(i)}, f_l^{(i)}))]$$

where $f_g^{(i)}$ and $f_l^{(i)}$ are different but sampled from the same distribution by shuffling the samples from the joint distribution along the batch axis (Belghazi et al. 2018) and $\text{sf}(z) = \log(1 + e^z)$.

In the meantime, two encoders $E_l, E_g$, and the local discriminator $T_l$ are also trained by maximizing the mutual information as follows:

$$\max_{E_l, E_g, T_l} \frac{1}{h \times w} \sum_{i=1}^{N} \sum_{j=1}^{h \times w} \mathcal{L}_l(f_g^{(i)}, f_l^{(i),j})$$

where $\mathcal{L}_l$ is also formulated as JSD-based mutual information and $f_l^{(i),j}$ denotes a $j$-th local patch of the local feature $f_l^{(i)}$. After calculating the mutual information between the global feature and a local patch of the local feature for all patches, they are averaged and used to update the parameters of the local encoder, the global encoder, and the local discriminator, respectively. In the end, the calculated global feature is subject-invariant and represents the input in both global and local viewpoints.

### Class Disentanglement

Although we find subject-invariant features representation by maximizing local/global mutual information as described above, as those are learned in an unsupervised manner, the output global feature is not necessarily useful per se for classification. (Peng et al. 2019; Zheng and Sun 2019) addressed that feature learned in an unsupervised generally includes both task-relevant and task-irrelevant information. Hence, we assume that the subject-invariant feature needs disentangling to class-relevant and class-irrelevant factors. Lastly, the class-relevant factors are enough for the classification task at the end.

In order for feature decomposition, we split the global feature $f_g \in \mathbb{R}^{d_g}$ into two parts such that one part is related to the class-relevant feature $f_{re}$ and the other part is related to the class-irrelevant feature $f_{ir}$, where $f_{re}, f_{ir} \in \mathbb{R}^{d_g/2}$. Then, the classifier $C$ is trained by minimizing the softmax cross-entropy loss with only the class-relevant feature as input as follows:

$$L_{cls} = - \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}^{(i)} \log(C(f_{re}^{(i)}))$$

where $\mathbf{y}^{(i)}$ denotes an one-hot label vector for an input $x^{(i)}$.

For the decomposition of the global feature $f_g$ into two factors of $f_{re}$ and $f_{ir}$ to support the classifier $C$, we also exploit mutual information neural estimation (MINE) (Belghazi et al. 2018), for which another networks $M$ is introduced. In (Belghazi et al. 2018), the parameters of MINE need to be trained by gradient ascent due to the Donsker-Varadhan representation (Donsker and Varadhan 1983). However, in our case, we want to minimize the mutual information between two factors $f_{re}$ and $f_{ir}$, we apply a gradient reversal layer (GRL) (Ganin et al. 2016). A GRL passes features innate to MINE $M$ during forward propagation, but reverses the gradients during back propagation, i.e., reversed gradients.

$$\max_{M} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_g(f_{re}^{(i)}, f_{ir}^{(i)})$$

### Reconstruction

Since the disentangled features are expected to keep spatio-spectral information in the input, we lastly build a decoder network for reconstruction of the global feature $f_g$. The split features, i.e., class-relevant feature $f_{re}$ and class-irrelevant feature $f_{ir}$, are concatenated to the global feature along the depth axis $f_g \in \mathbb{R}^{d_g}$. Then, the concatenated global feature $f_g$ is regarded as an input to the decoder $D$. In our case, we denote a reconstruction loss as follows:

$$L_{recon} = \frac{1}{N} \sum_{i=1}^{N} ||x^{(i)} - D(f_{g}^{(i)})||_2^2.$$
Algorithm 1: Learning subject-invariant and class-disentangled representation

**Input**: local encoder $E_l$, global encoder $E_g$, local discriminator $T_l$, global discriminator $T_g$, decoder $D$, mutual information estimator $M$, classifier $C$

**Output**: well-generalized local encoder $\hat{E}_l$, global encoder $E_g$, classifier $C$

1. **while not converged**
   1. // Subject-Invariant Feature Embedding
      1. Update $E_l$, $E_g$, $T_l$, $T_g$ by $I_g$ (1) and $I_l$ (3);
      2. // Class Disentanglement
         2. Split global feature $f_g$ to $f_{re}$ and $f_{ir}$;
      3. Update $E_l$, $E_g$, $C$ by $L_{cs}$ (4);
   4. // Mutual Information Estimation
       4. Calculate mutual information between the disentangled feature pair ($f_{re}$, $f_{ir}$) with $M$;
   5. Update $M$ by $L_e$ (5);
   6. Update $E_l$, $\hat{E}_g$ by $-\mathcal{I}_g$;
   7. // Reconstruction
      7. Reconstruct $\hat{g}$ by the concatenated feature $\hat{f}_g$ with $f_{re}$ and $f_{ir}$;
   8. Update $E_l$, $E_g$, $D$ by $L_{recon}$ (6);
10. **end**

\[ \hat{E}_l = E_l, \hat{E}_g = E_g, \hat{C} = C \]

Data & Preprocessing

GIST-motor imagery dataset\(^1\) (Cho et al. 2017) is a big-data set of 52 subjects (19 females, 24.8 ± 3.86 years old). It consists of EEG signals of two different motor imagery tasks of left-hand and right-hand. All EEG signals were recorded from 64 Ag/AgCl electrodes according to a 10-20 system and sampled with 512 Hz. In a subject, each class has 100 or 120 trials, acquired from four sessions. All subjects were asked to take a rest for 2 seconds, then to imagine the hand movement for 3 seconds by following the given instruction in the monitor. Since two subjects (subject 29 and 34) had a high correlation with electromyography (EMG), they were known as bad subjects and not used for analysis in (Cho et al. 2017). Therefore, we also conducted experiments using only the other 50 subjects’ data.

The signals were preprocessed by a large Laplacian filtering to reduce noise and a bandpass filtering in the range of 4 to 40 Hz related to sensorimotor rhythms. After segmenting the signals to a baseline (2 seconds) and task (3 seconds), we subtracted a mean value of the baseline to each trial for baseline correction. In an attempt to exploit frequency properties of the data, we computed the power spectral density (PSD) by applying Welch's method. Consequently, we obtained $[64 \times 37]$-sized datum per sample and use it as an input to our method.

Experimental Settings

As mentioned above, a subject performed motor imagery during four sessions in total. Therefore, to perform 5-fold cross-validation, we randomized each subject’s trials regardless of sessions and divided the dataset into training, validation, and test set to the ratio 7 : 1 : 2. For each training and validation, we utilized all subjects’ training and validation data. We applied Gaussian normalization to all inputs for each channel to preserve spatial information. To verify the generalization of the network, we conduct experiments in two different scenarios.

- **Scenario I**: Since our network is regarded as learning subject-invariant and class-disentangled representations, we train our network using all subjects’ data within a single session, then, test it for each subject.
- **Scenario II**: After training our proposed network, the network is able to make a common feature among subjects. Thus, we take advantage of it to reduce calibration time for a new subject. Here, by assuming that we have a few trials (samples) for a new subject, i.e., a target subject. Without the target subject, we train our network using the remaining subjects (49 subjects in our dataset), then, randomly select 30%, 50%, and 70% of the unseen subject’s training trials. Subsequently, we perform fine-tuning by minimizing all losses, with the target subject’s selected data in a subject-wise manner.

Model Implementation

**Architecture** To capture spatial and spectral information respectively, the local encoder is a spectral convolutional layer with $1 \times 37$, in the meantime, the global encoder is composed of a spatial convolutional layer with $64 \times 1$. The dimension of depth at each convolutional layer was set to 50 and 100, respectively. Consequently, we acquired the local feature $f_l \in \mathbb{R}^{64 \times 1 \times 50}$ and the global feature $f_g \in \mathbb{R}^{100}$. In the global discriminator, after embedding the local feature through another spatial convolutional layer with 100 filters, it was concatenated with the global feature along the depth axis. Then, the concatenated feature was taken as an input to two fully-connected layers, with 50 and 1 units per layer. We used the encode-and-dot-product architecture for the local discriminator in accordance with (Hjelm et al. 2019). In the table.

Table 1: Average of 5-fold classification accuracy [%] on all subjects in GIST-motor imagery dataset (Cho et al. 2017) with common spatial pattern (CSP) (Ramoser, Muller-Gerking, and Pfurtscheller 2000), filter bank common spatial pattern (FBCSP) (Ang et al. 2008), Shallow ConvNet (Schirrmeister et al. 2017), and our proposed method without and with the class disentanglement ($w/o$ and $w/$ CD).

| Method          | Accuracy [%] |
|-----------------|--------------|
| CSP             | 66.1 ± 13.8  |
| FBCSP           | 67.9 ± 14.5  |
| Shallow ConvNet | 63.3 ± 11.0  |
| Ours ($w/o$ CD) | 65.9 ± 10.6  |
| Ours ($w/$ CD)  | 71.5 ± 9.7   |

\(^1\)Available at http://gigadb.org/dataset/100295
local discriminator, each local and global feature passed to a $1 \times 1$ convolutional layer and a fully-connected layer, respectively, where the dimension of both layers was 120. Since the local feature still had a channel-axis, we considered a channel of the local feature as a local patch. After taking the dot-product between the global feature and each channel of the local feature, we took an average of them to perform Eq. (3). A classifier $C$ was composed of 2 fully-connected layers, where each layer had $\{100, 2\}$ units. Regarding the mutual information estimator, there were two fully-connected layers to receive class-relevant and class-irrelevant feature respectively, where the number of hidden nodes was 20. Then, we figured out the sum of two outputs from the two layers and passed it on 2 fully-connected layers with $\{20, 1\}$. Finally, our decoder was composed of spatial deconvolutional layer and spectral convolutional layer with 50, 1 units.

Training Settings
Exponential linear units (ELU) was used for a nonlinear function in our network. Except for two discriminators and the mutual information estimator, we applied the batch normalization (Ioffe and Szegedy 2015). In addition, we applied an $l_2$-regularization with a coefficient of 0.05 to the classification loss Eq. (4) and dropout (Srivastava et al. 2014) with a rate of 0.5 to prevent over-fitting. We trained models using momentum optimizer (momentum = 0.9) with a learning rate of $10^{-4}$, where the dimension of a mini-batch size was 253. In the scenario II, we need to do fine-tuning for unseen subjects. Thus, we trained the same network under the different settings by ablation studies. Our network was implemented based on Tensorflow\(^3\).

Baselines
For evaluation, we compared our proposed method to other methods of decoding motor imagery. We briefly introduce two machine-learning based models and a deep-learning based model.

- **Common Spatial Pattern (CSP)** (Ramoser, Muller-Gerking, and Pfurtscheller 2000): Thanks to its popularity and simplicity, this is a de facto method in motor-imagery BCI. CSP learned spatial filters to disassemble multi-channel EEG signals into distinguishable patterns for class by maximizing the differences of the variance of classes. We used 2 spatial filters for each class, i.e., left-hand and right-hand.

- **Filter Bank Common Spatial Pattern (FBCSP)** (Ang et al. 2008): This method achieved good performance in various BCI tasks. First, raw EEG signals were bandpass-filtered into $4 \sim 40$ Hz bands at an interval of 4 Hz. In each of bands, we applied the CSP (filter=2 for each class) first for feature extraction and then selected the CSP features through mutual information-based feature selection (Ang et al. 2008) with the 10 highest scores.

- **Shallow ConvNet** (Schirrmeister et al. 2017): This deep network showed its power of learning feature representations in a fully data-driven manner and end-to-end learning. Shallow ConvNet was composed of two convolutional layers capturing spatio-temporal information of raw EEG signal and a fully-connected layer. However, we modified the spatial filter size as $64 \times 1$ because the number of electrode channels is 64 in GIST-motor imagery dataset (Cho et al. 2017). For fair comparison, we did not use ‘cropped-training’ strategy, which considered in the original work (Schirrmeister et al. 2017).

---

\(^3\)https://www.tensorflow.org/
Since the baseline methods were designed with raw EEG trials, we also utilized the raw EEGs including spatio-temporal information different from our proposed method. After conducting CSP and FBCSP, we employed a linear discriminant analysis for classification (Vidaurre et al. 2010). It is worth noting that we trained the baseline methods individually.

**Results**

**Scenario I** We trained two models with all subjects’ data to validate the effectiveness of class disentanglement under the same experimental settings. On one hand, ours (w/o CD) was composed of an encoder, two discriminators, and classifier, then, it was trained by minimizing Eq. (1), (3), and (4). On the other hand, ours (w/ CD) is our full proposed network. While ours (w/o CD) achieved 65.9% ± 10.6% accuracy on average over 50 subjects in 5-fold cross-validation, ours (w/ CD) showed 71.5% ± 9.7% accuracy, referred to Table 1. In other words, we could enhance the performance by discarding class-irrelevant information from subject-invariant feature.

Although we trained our network using all subjects within one training session, our network showed reasonably high results compared to other baseline methods trained individually. Especially, from Figure 2, a few subjects who empirically considered as BCI illiterate, not making discriminative motor imagery patterns, showed promising performance improvements thanks to training with other subjects including discriminative patterns.

**Scenario II** In the scenario II, we conducted the experiments for CSP (Ramoser, Muller-Gerking, and Pfurtscheller 2000), FBCSP (Ang et al. 2008), and our proposed method. The results with baseline methods are shown in Table 2. In the results, our proposed method presented superiority to others in most cases. Additionally, we estimated p-value to validate statistical significances between the proposed method and baselines. We used two-tailed, Wilcoxon’s signed rank test. This supports our hypothesis that the feature representation learning in our framework is effective to transfer and to be updated with a minimal number of samples for a new subject.

**Conclusions**

In this paper, we proposed a deep neural network that learns subject-invariant and class-disentangled representation via mutual information estimation among features in different levels for BCI tasks in an end-to-end manner. The subject-invariant latent space can be deployed in decoding a subject even with smaller training data and improving performance by adding other subjects’ data, i.e., transfer learning. We evaluated our proposed method with a big motor imagery EEG dataset. Further, we expect our proposed method can be applied to other types of EEG signals.

**Acknowledgments**

This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451; Development of BCI based Brain and Cognitive Computing Technology for Recognizing User’s Intentions using Deep Learning).

**References**

Ang, K. K.; Chin, Z. Y.; Zhang, H.; and Guan, C. 2008. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In IEEE International Joint Conference on Neural Networks, 2390–2397.

Azab, A. M.; Mihaylova, L.; Ang, K. K.; and Arvaneh, M. 2019. Weighted Transfer Learning for Improving Motor Imagery-Based Brain–Computer Interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27(7):1352–1359.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Hjelm, D.; and Courville, A. 2018. Mutual Information Neural Estimation. In International Conference on Machine Learning, 530–539.

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2017. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3722–3731.

Chai, X.; Wang, Q.; Zhao, Y.; Liu, X.; Bai, O.; and Li, Y. 2016. Unsupervised Domain Adaptation Techniques based on Auto-Encoder for Non-Stationary EEG-based Emotion Recognition. *Computers in Biology and Medicine* 79:205–214.

Chen, T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating sources of disentanglement in variational autoencoders. In Advances in Neural Information Processing Systems, 2610–2620.

Cho, H.; Ahn, M.; Ahn, S.; Kwon, M.; and Jun, S. C. 2017. EEG Datasets for Motor Imagery Brain–Computer Interface. *GigaScience* 6(7):gix034.

Donsker, M. D., and Varadhan, S. S. 1983. Asymptotic Evaluation of Certain Markov Process Expectations for Large Time. *Communications on Pure and Applied Mathematics* 36(2):183–212.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research* 17(1):2096–2030.
Gräffmann, B.; Allison, B.; and Pfurtscheller, G. 2009. Brain–Computer Interfaces: A Gentle Introduction. In *Brain-Computer Interfaces*. Springer. 1–27.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning Deep Representations by Mutual Information Estimation and Maximization. In *International Conference on Learning Representations*.

Ioffe, S., and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*.

Jayaram, V.; Alamgir, M.; Altun, Y.; Scholkopf, B.; and Grosse-Wentrup, M. 2016. Transfer Learning in Brain-Computer Interfaces. *IEEE Computational Intelligence Magazine* 11(1):20–31.

Jeon, E.; Ko, W.; and Suk, H.-I. 2019. Domain Adaptation with Source Selection for Motor-Imagery based BCI. In *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 1–4. IEEE.

Kang, H., and Choi, S. 2014. Bayesian Common Spatial Patterns for Multi-Subject EEG Classification. *Neural Networks* 57:39–50.

Kindermans, P.-J.; Tangermann, M.; Müller, K.-R.; and Schrauwen, B. 2014. Integrating Dynamic Stopping, Transfer Learning and Language Models in an Adaptive Zero-Training ERP Speller. *Journal of Neural Engineering* 11(3):035005.

Lin, Y.-P., and Jung, T.-P. 2017. Improving EEG-based Emotion Classification using Conditional Transfer Learning. *Frontiers in human neuroscience* 11:334.

Liu, Y.-C.; Yeh, Y.-Y.; Fu, T.-C.; Wang, S.-D.; Chiu, W.-C.; and Frank Wang, Y.-C. 2018. Detach and Adapt: Learning Cross-Domain Disentangled Deep Representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8867–8876.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. I. 2015. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 97–105. JMLR. org.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems*, 271–279.

Özdenizci, O.; Wang, Y.; Koike-Akino, T.; and Erdogmus, D. 2019. Transfer Learning in Brain-Computer Interfaces with Adversarial Variational Autoencoders. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, 207–210. IEEE.

Pan, S. J., and Yang, Q. 2009. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Peng, X.; Huang, Z.; Sun, X.; and Saenko, K. 2019. Domain Agnostic Learning with Disentangled Representations. In *International Conference on Machine Learning*, 5102–5112.