Discriminative and Robust Feature Learning for MIBCI-based Disability Rehabilitation

Xiuyu Huang (✉ xiuyu.huang@connect.polyu.hk)  
HK PolyU: The Hong Kong Polytechnic University  https://orcid.org/0000-0003-1600-9109

Nan Zhou  
The Hong Kong Polytechnic University

KupSze Choi  
The Hong Kong Polytechnic University

Research

Keywords: Convolutional neural networks, Electroencephalogram, Motor imagery, Correntropy based center loss

Posted Date: October 14th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-963653/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Discriminative and robust feature learning for MIBCI-based disability rehabilitation

Xiuyu Huang\textsuperscript{1*}, Nan Zhou\textsuperscript{1}, and Kup-Sze Choi\textsuperscript{1}

\textsuperscript{1}Centre for Smart Health, The Hong Kong Polytechnic University, Hong Kong, SAR China

*xiuyu.huang@connect.polyu.hk

Abstract.

**Background:** In the past few years, motor imagery brain-computer interface (MIBCI) has become a valuable assisting technology for the disabled. However, how to effectively improve the motor imagery (MI) classification performance by learning discriminative and robust features is still a challenging problem.

**Methods:** In this study, we propose a novel loss function, called correntropy-based center loss (CCL), as the supervision signal for the training of the convolutional neural network (CNN) model in the MI classification task. With joint supervision of the softmax loss and CCL, we can train a CNN model to acquire deep discriminative features with large inter-class dispersion and slight intra-class variation. Moreover, the CCL can also effectively decrease the negative effect of the noise during the training, which is essential to accurate MI classification.

**Results:** We perform extensive experiments on two well-known public MI datasets, called BCI competition IV-2a and IV-2b, to demonstrate the effectiveness of the proposed loss. The result shows that our CNNs (with such joint supervision) achieve 78.65\% and 86.10\% on IV-2a and IV-2b and outperform other baseline approaches.

**Conclusion:** The proposed CCL helps the learning process of the CNN model to obtain both discriminative and robust deeply learned features for the MI classification task in the BCI rehabilitation application.

**Keywords:** Convolutional neural networks; Electroencephalogram; Motor imagery; Correntropy based center loss

1 Background

Individuals who have paralysis have significant difficulties in engaging in social interactions due to functional limitations. Based on the statistics from the World Health Organization (WHO) \cite{1}, one billion, approximately 15\% of the world’s population, are experiencing different types of disability. Many of them have varying degrees of activity and participation restrictions. To interact with the world, they rely on assistive technologies, such as Brain Computer-Interfaces (BCI). The BCI is a computer-based technology that collects brain activities, analyzes them, and translates them into commands that drive an external machine to execute the desired actions \cite{2, 3}. Motor imagery brain-computer interfaces (MIBCI) is one of the most promising approaches in the medical research area \cite{4}. Its popularity is mainly because of the non-invasiveness of the electroencephalogram (EEG, a brain signal collection technique) and the easy implementation of the motor imagery paradigm. Such a convenient system has been performed in a variety of applications, such as orthoses \cite{5}, prostheses \cite{6}, robotic arms \cite{7}, and mobile robots \cite{8}, across different fields, including clinical practices \cite{9, 10} and military \cite{11}.

The translation of command (e.g., classification) from the raw MI EEG signals remains a significant challenge. The key to an accurate classification is to obtain discriminative features. The conventional strategies usually rely on handcrafted feature extractors, such as common spatial pattern (CSP) \cite{12}, filter bank common spatial pattern (FBCSP) \cite{12}, band power \cite{13}, and Riemannian covariances \cite{14}. The extracted features usually integrate with linear classifiers, including logistics regression (LR) and support vector machine (SVM) \cite{15, 16}, for the MI classification in the BCI rehabilitation devices.

The conventional approaches dominate classification performance for a relatively long time until the recent development of graphic processing units (GPU) \cite{17}. The modern GPU offers a powerful computational ability, enabling researchers to explore deep learning (DL) approaches that require an enormous computational burden to address the BCI classification task. The DL strategies make distinguished improvements in the classification accuracy compared to the traditional methods. It also
provides an extra benefit that the MI EEG data can be processed in an end-to-end manner without any preprocessing [17]. Two well-known and also the earliest DL approaches, called EEGNet [18] and shallow ConNet [19], deliver favorable classification outcomes for the MI recognition. Since then, many research groups have thoroughly investigated multiple innovative convolutional neural networks (CNNs, one of the deep learning structures). The most common framework (Fig. 1) of the CNN model is to perform two steps in succession, learning deep features from raw EEG data by CNN blocks, then making label predictions using the learned features. The training of the model on the classification task is usually guided by the softmax loss [20].

The deep features need to be separable and discriminative to achieve excellent performance in the general classification task. Such features can be acquired by the joint supervision of the softmax loss and center loss, which was proposed in [20] for face recognition in the computer vision field. Specifically, the softmax loss ensures the separability of the features (Fig. 2 (a)), and the center loss concurrently increases the discriminative power upon the separability by pulling deep features towards their class centers (intra-class distance minimization, Fig. 1(b)). This joint supervision signal is also widely adopted for the MI detection task due to its efficiency and easy setup [21, 22]. However, the center loss is sensitive to the non-Gaussian noise since it is based on the quadratic $L_2$ norm distance [23]. Few noise points/outliers far from the class centers may dominate the objective function and degrade the classification performance (Fig. 2 (c)). Coincidentally, raw EEG data have a low signal-to-noise ratio (SNR) and contain much noise [13], so it is not suitable to directly use the center loss to optimize the CNN model for MI classification.

For this concern, we propose a new loss, referred to corentropy-based center loss (CCL), based on the correntropy induced distance (CID). Like the original center loss, the CCL also simultaneously learns the class centers and penalizes the distance between deep features and their corresponding class centers to minimize the intra-class variation. However, there are two major differences between the CCL and center loss. First, the class centers learned by the CCL are based on maximum correntropy criterion mean (MCCM) [24]. The outliers have a low weight in the MCCM calculation. Second, unlike the $L_2$ norm distance that heavily penalizes the deep features far from class centers, the CID used in the CCL assigns minor penalizations to these far feature points. In the Fig. 2 (c) and (d), we can notice that class centers updated by the center loss have a significant deviation from the $C_{true}$, while the class MCCMs nearly

![Fig. 1. The common framework of the CNN model](image1)

![Fig. 2. The distribution of deeply learned features under different training schemes on MINST dataset (the same one for the toy example in the original paper [20] of center loss) for intuition, the noise represents the data of the digit '3', but mislabelled as '0' or '1': (a) softmax loss only and clean dataset; (b) joint loss 1 (softmax loss and center loss) and clean dataset; (c) joint loss 1 and noisy dataset; (d) joint loss 2 (softmax loss and CCL) and noisy dataset; $C_{true}$s are updated during the training, $C_{true}$s are computed by the conventional mean of feature points respectively in each class excluded from noise.](image2)
overlap with $C_{\text{true}}$. The noise feature points are also filtered out by the CCL as shown in Fig. 2 (d). Such characteristics can decrease the negative effect of the non-Gaussian outliers during the model training, which enables the CNN model to learn a robust feature pattern from the raw MI EEG data that usually contain much noise. The main contributions of our method are summarized as follows:

1. To our best knowledge, this is the first attempt to use CCL to guide the training of the CNN model. With joint supervision of the softmax loss and CCL, the discriminative features can be attained for the MI classification, which can be easily equipped for the BCI applications in either clinical or military fields.

2. The proposed loss is based on the CID and MCCM that can significantly reduce the adverse effect of the noise during the CNN model training.

3. We present extensive experiments on two MI public datasets, called BCI-competition IV-2a and IV-2b. The results show the effectiveness of the proposed method.

2 Method

In this section, we firstly elaborate the network architecture used in the study. Then, we intuitively and mathematically introduce the proposed loss (CCL).

2.1 Network architecture

In this study, we follow the model architecture of the EEGNet [18], but make a slight modification. We decrease the kernel size of the first CNN block to capture the feature pattern above 8 Hz, as previous studies show that the MI is mainly related to the EEG signals in the range of 8-30 Hz [13]. Our CNN model (see Fig. 3) consists of two temporal filters with average-pooling, one spatial filter with average-pooling, and two dense layers. One dense layer is applied with the CCL, and the other is involved with the softmax loss [25].

![Fig. 3. Illustration of network architecture.](image)

Fig. 3. Illustration of network architecture. $x_i$ and $\hat{y}_i$ are $i$th trial of the raw MI EEG data and $i$th predicted label, respectively. $n$ is the number of classes. $N_e$ is the number of EEG electrodes. $N_t$ is the time stamp. $k_1$ and $k_2$ are the sizes of temporal filters.

2.2 The proposed CCL

As illustrated in Fig. 2 (b), the center loss can increase the discriminative power by reducing the intra-class variation. However, it is based on the $L_2$ quadratic norm distance and is easily affected by the noise or non-Gaussian outliers. It is well-known that the MI-EEG series carry significant noise during the collection. We proposed the CCL, which combines maximum correntropy criterion (MCC) [24] and center loss. The CCL not only preserves the primary function of the center loss that increases the discriminative power of deep features by reducing intra-class variations but also decreases the anti-effect of the noise feature points by assigning them 'minor' significance. It is mathematically defined as follows.

$$L_{\text{CCL}} = \frac{1}{n} \sum_{i=1}^{m} \left[ 1 - \exp \left( -\frac{||x_i - c_{\hat{y}_i}||^2}{2\sigma_{\hat{y}_i}^2} \right) \right], \quad (1)$$

where $x_i \in \mathbb{R}^d$ represents the $i$th sample’s deeply learned features with a label of $y_i \in \{1,2,\ldots,n\}$. $d$ denotes the feature dimension. $c_{\hat{y}_i}$ is the $y_i$th class MCCM of deep features. $\sigma_{\hat{y}_i}$ is the size of the Gaussian kernel for data of $y_i$th class. The sample size and the number of classes are $m$ and $n$. Intuitively, the MCCM $c_{\hat{y}_i}$ (class center) is similar to the conventional mean (averaged value) of each class but assigns a very low weight to the outlier/noise in its calculation. More details and its mathematical definition can be found in Appendix A. The merit of this loss is that we can dynamically adjust kernel size $\sigma_{\hat{y}_i}$ to control the significance of the feature points that are far from the MCCMs. We implement a
two-step approach to update the deep feature $x_i$ and class MCCM $c_{y_i}$ in succession within one epoch. We first use the gradient descent method in each epoch to update $x_i$. Then, the self-adaptive half quadratic (HQ) approach [24] is subsequently performed to update class MCCM $c_{y_i}$. The two-step update approach is given in Appendix A. Similar to the center loss, we adopt the joint supervision of softmax loss ($L_\mathcal{S}$) [25] and the CCL to train the CNN for a discriminative and robust feature learning. The joint supervision signal ($L$) is formulated

$$L = L_\mathcal{S} + \lambda L_{CCL}$$  

where the $\lambda$ is the trade-off scalar for balancing these two losses. We summarize the learning details of the CNN model in Algorithm 1 (Appendix B).

3 Evaluation details

In this section, we firstly introduce the dataset used for the evaluation. Then, the details of the experimental settings are presented.

3.1 Data

Two well-known public datasets, called BCI competition IV-2a and IV-2b [26], provided by the Technical University of Graz, are used in this study.

The BCI competition IV-2a was collected from 9 individuals (A01-A09) with a 250 Hz sample rate using 22 electrodes. It is based on the cue-based BCI paradigm that consists of four MI classes, referred to as the imagination of movement of the left hand, right hand, both feet, and tongue. Each subject employed two separate sessions, with 288 trials (72 trials for a single MI type) each session. The same data division scheme of the competition is adopted where the first session is for training and the second for testing whose labels are to be predicted. We use the 4s temporal segment from the starting point of the MI-cue until the end of the MI. Given the 250 Hz sample rate, 1000 samples are used in one trial, and the raw MI data for the input of CNN is in a matrix form ($22 \times 1000$).

The BCI Competition IV-2b was also collected from 9 healthy people (B01-B09) but only with 3 EEG electrodes (C3, Cz, and C4) with a sampling frequency of 250 Hz. The recordings were also based on a cue-based screening paradigm that only comprises two classes, e.g., the MI of the left hand (class 1) and right hand (class 2). Five sessions of data were collected for each subject. We also used the same data division as the competition. The first three sessions are for training, and the rest two are for testing. Like the IV-2a, the 4-second interval, from the starting point of the cue until the end of the MI, is treated as an input trial to the model. Given the recording frequency of 250 Hz, the raw input data for one trial is in a $3 \times 1000$ matrix.

The examples of one trial input data of BCI competition IV-2a and IV-2b are shown in Fig. 4. The fluctuations seem different between subfigure (a) and (b), but they are similar if plotted on the same scale. Table 1 and Fig. 5 present the demographics of subjects and electrode positions, respectively.

![Fig. 4. Examples of the trials of MI EEG raw data. (a) One trial of A07; (b) One trial of B07.](image)

| Table 1. Demographics. |
|-------------------------|
| Gender | Mean age (SD) |
|-------|---------------|
| IV-2a | female: 4; male: 5 | 23.11 (2.57) |
| IV-2b | female: 4; male: 5 | 24.44 (3.50) |
3.2 Experimental settings

We conducted the experiments on an NVIDIA Tesla P100 16GB GPU running online on the Colab provided by Google. The CNN model and the objective function ($\mathcal{L}$) are defined using Pytorch. The learning rate and the mini-batch size for $x_i$ update are set as 0.001 and 64, respectively. The model is trained with 750 epochs. The trade-off scalar $\lambda$ is set as 0 at the first 75 epochs (10%) and empirically set as 0.6 after the 75th epoch until the end of training. This setting is because the softmax loss can dominate the objective function at the initial training stage, making the features separable first. Other hyperparameters for CNN architectures are displayed in Table 2. The sampling frequency is 250 Hz for both datasets, so we set $k_1$ as 32 to obtain the temporal information above 8 Hz. The codes for experiments are given in the Supplemental files.

To evaluate the effectiveness of the proposed method, we compare our approach to other baselines, including one vector-based method (e.g., the BCI competition winner algorithm FBCSP [12]), three matrix-based approaches (e.g., SMM [27], CCSP [28], and SSCSP [29]), and three DL-based methods (EEGNet [18], shallow ConNet [19], and DRDA [22]). The evaluation metric is the classification accuracy, which is consistent with other articles mentioned above.

### Table 2. Hyper-parameter settings.

| Hyper-parameter | CNN for IV-2a | CNN for IV-2b |
|-----------------|---------------|---------------|
| $N_e$           | 1000          | 1000          |
| $N_t$           | 22            | 3             |
| $k_1$           | 32            | 32            |
| $k_2$           | 16            | 16            |

4 Results

4.1 Comparison to other baselines

Our proposed method is first evaluated on the BCI competition IV-2a and IV-2b datasets. The classification accuracy of each subject, averaged accuracy, and standard deviation (SD) are shown in Tables 3 and 4. The model with the best performance for each subject and the average accuracy are highlighted in boldface for a clear illustration. We observe that the proposed method (CNN model jointly supervised by the softmax loss and CCL) achieves the best classification across most subjects on both datasets. It has the largest improvement of 11.26% accuracy better than the second-best model on subject A05 in dataset IV-2a and 14.63% than the second-best model on subject B03 in IV-2b. On the average level, our method also achieves the best classification accuracy with an improvement of at least 3.90% and 2.12% on IV-2a and IV-2b compared to other baselines.

In addition, we also notice that the standard deviation of our method is smaller than others, which shows the high stability of the proposed approach across subjects. More interestingly, the proposed method has a relatively more significant improvement on the subjects A02, A05, and A06, who have EEG signals with a low SNR [22].
Table 3. Classification accuracies (%) obtained with the dataset BCI competition IV-2a

| Methods | Subject | Average (SD) |
|---------|---------|--------------|
| FBCSP   | A01     | 76.00        |
|         | A02     | 56.50        |
|         | A03     | 81.25        |
|         | A04     | 61.00        |
|         | A05     | 55.00        |
|         | A06     | 45.25        |
|         | A07     | 82.75        |
|         | A08     | 81.25        |
|         | A09     | 70.75        |
| SMM     |         | 67.75 (12.94) |
| EEGNet  |         | 74.50 (14.36) |
| ConNet  |         | 72.53 (13.42) |
| CCSP    |         | 66.51 (15.12) |
| SSCSP   |         | 70.72 (12.35) |
| DRDA    |         | 74.75 (12.22) |
| Ours    |         | 80.01 (13.06) |

Table 4. Classification accuracies (%) obtained with the dataset BCI competition IV-2b

| Methods | Subject | Average (SD) |
|---------|---------|--------------|
| FBCSP   | B01     | 70.00        |
|         | B02     | 60.36        |
|         | B03     | 60.94        |
|         | B04     | 97.50        |
|         | B05     | 93.12        |
|         | B06     | 90.63        |
|         | B07     | 93.13        |
|         | B08     | 80.13        |
|         | B09     | 82.50        |
| SMM     |         | 72.68 (12.77) |
| EEGNet  |         | 72.70 (14.72) |
| ConNet  |         | 79.37 (16.27) |
| CCSP    |         | 75.88 (12.57) |
| SSCSP   |         | 72.62 (12.77) |
| DRDA    |         | 74.56 (11.94) |
| Ours    |         | 80.01 (13.06) |

4.2 Sensitivity analysis

The hyperparameter $\lambda$ dominates the intra-class variation and noise influence. It is critical to our model, so we conduct an experiment to investigate its sensitiveness. Fig. 6 illustrates the averaged classification accuracy across subjects on the dataset BCI Competition IV-2a as $\lambda \in \{0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1\}$. It is observed that simply using the softmax loss (when $\lambda = 0$) may not be a good option, and the CNN model has the lowest averaged classification accuracy. We can also observe a ‘Log’ curve when $\lambda$ varies between 0 to 1. The classification accuracy increases with the increase of $\lambda$ at the beginning but remains relatively stable as the $\lambda \geq 0.4$. The highest classification performance is achieved when the $\lambda = 0.6$. Therefore, we empirically set the $\lambda$ as 0.6 for the final implementation of our approach.

4.3 Feature visualization

To better understand the function of the proposed CCL, we visualize the deep learning features under four conditions. The models in first two conditions, Model A ($\lambda = 0$) and Model B ($\lambda = 0.6$), are trained using the original data from A08 (random selection). In addition, we also try to intuitively observe the effectiveness of CCL on reducing the negative effect of the noise/outliers. It is hard to know which trial can be regarded as the noise in the original dataset, so we manually add ten trials of electricity flow...
simulation (50 Hz \(sin\) wave), one of the most frequent noises in the EEG signal, to the data of A08. Five trials are labeled as ‘left hand’, and the other five as ‘right hand’. Therefore, the models in the last two conditions, Model C (center loss, trade-off value = 0.01) and Model D (CCL, \(\lambda = 0.6\)), are trained using the data from A08 with 10-trial noises. For models A and B, as the dimension of the second last layer (deeply learned feature) is eight, we use the principle component analysis (PCA) [30] technique to convert the high-dimensional features into 2-D vectors for the visualization in the plane. For Models C and D, as the popular dimension reduction techniques such as PCA [30] and t-SNE [31] are easily affected by outliers, we decrease the output dimension of the second last layer to 2 for direct visualization. Fig. 7 shows the feature distributions of these four different models. We have the following two key findings. First, by comparing the subfigures (a) and (b), we observe that the intra-class variation significantly decreases when using the CCL, which has the same function as the center loss. Second, it is also clearly noted that the CCL filters out the noise/outliers, while the center loss does not (see. Fig. 7 (c) and (d)).

Fig. 7. The distribution of deeply learned features under different training schemes. (a) Model A; (b) Model B; (c) Model C; (d) Model D.

5 Discussion

The current study investigated the feasibility of a new loss function, called CCL, as a supervision signal for the CNN model training to obtain a high classification performance in the MI recognition task. The results reported in Section 4 demonstrate the superiority of the joint supervision of the softmax loss and CCL. The CNN models trained by such combined loss outperform other baselines and show better stability across individuals. The reasons behind such performance increases may be diverse. The conventional/traditional machine learning methods have their limitations. The FBCSP only extracts 1-D feature vectors for the classification, which may discard the 2-D latent information of the MI EEG signal [12]. For the matrix-based approach, including SMM, CCSP, and SSCSP, although the 2-D structure information can be preserved, they can only take the linear transformation [17]. Alternatively, as a DL-based approach, the proposed method can process the EEG data in 2-D space without destroying its structural information and generate the non-linear feature patterns for the MI classification. In addition, our approach also outperforms previous DL models. The objective function may be one of the significant reasons for the increment of accuracy. The EEGNet and shallow ConNet are only supervised by softmax loss that only ensures the separable features but does not guarantee their discriminative power nor consider the negative effect of noise. In contrast, the joint supervision of the softmax loss and CCL can extract discriminative deeply learned features and eliminate the negative impact of noise. These encouraging results demonstrate that, apart from the model structure that has been heavily investigated by previous research groups, an efficient and well-designed objective function can also be another research focus to break through the bottleneck of the CNN models on MI classification.

The center loss has two outstanding characteristics. First, it is stable/insensitive to the trade-off hyper-parameter \(\lambda\). Second, it increases the discriminative power of the deeply learned feature by minimizing the intra-class variation [20]. The Result section shows that the CCL, initially inspired by the center loss, also carries on these two characteristics to a certain degree in the MI classification task. The verification performance does not show significant variance when \(\lambda \geq 0.4\) (section 4.2), and the model trained by the joint supervision of softmax loss and CCL has a smaller intra-class variation than the one only trained by softmax loss (Fig. 7). These phenomena show that CCL preserves the significant properties and functionality of the center loss by changing the quadratic L2 norm distance into CID (applying the Gaussian kernel). It still delivers stable performance across different \(\lambda\) (in a particular range) and increases the discriminative power of the deep features by reducing the intra-class variation.

The unique property of the CCL is its effectiveness in handling the negative effect of the noise in the MI classification task. We have observed how this property benefits the MI classification in two
dimensions in the Result section. First, for the subjects (A02, A05, and A06) with a low signal-to-noise ratio, the CNN trained by joint supervision of softmax loss and CCL shows a more significant improvement. In these subjects, more noise is mixed with the signals during the data collection. We assumed that deep features extracted by the CNN pipeline from the trials containing a large amount of noise are more likely far from the class feature centers. Our CCL only assigns very low significance on the further points during the training. The greater performance increments on these three subjects may come from such low significance assignments, allowing the CNN to learn more helpful information from the high SNR trials than the low ones. Second, by seeing Fig. 7 (d), it is evident that the features of noise (stimulated electricity signals) are filtered out from the feature distributions of the standard EEG data when using the joint supervision of the softmax loss and CCL for the CNN training. This phenomenon may provide direct evidence on the effectiveness of the CCL on reducing the negative effect of the ‘pure’ noise in the MI classification task. According to these two exciting findings, the CCL may provide benefits on reducing the adverse effects of both low SNR and pure noise trials.

The proposed CCL show its efficacy on the MI classification task in our experiments. It may offer similar benefits on other EEG-based BCI classification tasks such as P300 [32], steady-state visually evoked potentials (SSVEP) [33], and Steady-State Somatosensory Evoked Potential (SSSEP) [34]. Each type of classification has its traits. It is not appropriate to estimate the effectiveness of CCL on these tasks without actual experiments, although they share similar characteristics of the EEG signals. In the future, we consider extending this work into other EEG-based BCI classification tasks to have a deeper understanding of how the general effectiveness of CCL on the BCI classifications. Such insight enables us to gain more confidence in applying the proposed algorithm in the real-world BCI application in both clinical and military fields.

6 Conclusion

In this study, we proposed a new loss function called correntropy-based center loss (CCL). By combining softmax loss with CCL to jointly supervise the learning of CNNs, the discriminative power of the deeply learned features can be significantly increased for the MI classification. Unlike the center loss based on the quadratic $L_2$ norm distance, the CCL is based on the CID that assigns minor penalization to the outliers, therefore eliminating the noise’s adverse impact. The novel self-adaptive strategy was used to learn the class MCCMs, releasing the burden of manually setting the kernel size $\sigma$. Experiments on two well-known MI datasets demonstrate encouraging results and show the effectiveness of the proposed method. The proposed algorithm has the potential to be used in the practical MIBCI rehabilitation systems.

Funding

The work in this paper is supported in part by the Hong Kong Innovation and Technology Fund (MRP/015/18) and the Hong Kong Research Grants Council (PolyU 152006/19E).

Availability of data and materials

The datasets BCI Competition IV-2a and IV-2b for this study can be found on the BCI Competition IV website: http://www.bbci.de/competition/iv/.

Ethics approval and consent to participate

The BCI Competition IV-2a and IV-2b datasets are used to evaluate the proposed approach. These two datasets are publicly available. The people involved in the datasets have obtained ethic approval. Users can download the data for free for research and publish relevant articles, so the ethical review and approval were waived for the current study.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors’ contributions

HX conceptualized the study, performed the majority of the experiments and analyses, made the Fig.s, and wrote the first draft of the manuscript. ZN and KC performed some experiments, updated the Fig.s,
performed the statistics, and edited the manuscript. All authors approved the submitted version.

Supplemental Files
The file folder 'Source codes' contains all the necessary codes to reproduce the results presented in the manuscript.

Appendix A

The CCL is defined as (1). For updating deep feature \( \mathbf{x}_i \), the gradient descent method can be adopted. The gradients of \( \mathcal{L}_{CCL} \) with respect to \( \mathbf{x}_i \) is computed as:

\[
\frac{\partial \mathcal{L}_{CCL}}{\partial \mathbf{x}_i} = \frac{1}{2\sigma_{y_i}} \exp \left( -\frac{\| \mathbf{x}_i - \mathbf{c}_{y_i} \|^2}{2\sigma_{y_i}^2} \right) \left( \mathbf{x}_i - \mathbf{c}_{y_i} \right) \tag{A.1}
\]

For updating the class MCCM \( \mathbf{c}_{y_i} \), the half quadratic (HQ) strategy [24] is used, where the convergence can be achieved without concerning the sensitivity of the learning rate. Let \( \varphi(x) = x - x\ln(-x) \), it is easy to verify that

\[
\exp(-x) = \max_{z} [zx - \varphi(z)]. \tag{A.2}
\]

Let the \( q = -z \), (A.2) can be elaborated as

\[
1 - \exp(-x) = \min_{q} [1 + qx + \varphi(-q)]. \tag{A.3}
\]

When we update the class MCCM \( \mathbf{c}_{y_i} \), the \( \{\mathbf{x}_i\}_{i=1}^m \) is fixed. Therefore, according to (A.3), the problem (1) for \( \mathbf{c}_{y_i} \) update can be reformulated as

\[
\min_{\mathbf{c}_{y_i}, p} \frac{1}{2} \sum_{i=1}^{m} \left[ p_i \left( \frac{\| \mathbf{x}_i - \mathbf{c}_{y_i} \|^2}{2\sigma_{y_i}^2} \right) + \varphi(p_i) \right]. \tag{A.4}
\]

where \( p \) is the auxiliary variable. The problem (A.4) can be solved by alternative optimization. By setting \( \frac{\partial}{\partial p}(zx - \varphi(z)) = x + \ln(-z) = 0 \), one can find that the maximum value of (A.2) is reached at \( z = -\exp(-x) \). This is the HQ approach. By \( q = -z \), when \( \{\mathbf{c}_j\}_{j=1}^n \) is fixed, the variable \( p \) can be updated by

\[
p_i^{t+1} = \exp\left( -\frac{\| \mathbf{x}_i - \mathbf{c}_{y_i} \|^2}{2\sigma_{y_i}^2} \right), \quad i = 1, \ldots, m. \tag{A.5}
\]

where \( t \) is the iteration step. When \( p \) is fixed, the subproblem with respect to \( \{\mathbf{c}_j\}_{j=1}^n \) update is

\[
\min_{\mathbf{c}_{y_i}} \frac{1}{2} \sum_{i=1}^{m} \left[ p_i \left( \frac{\| \mathbf{x}_i - \mathbf{c}_{y_i} \|^2}{2\sigma_{y_i}^2} \right) \right]. \tag{A.6}
\]

which has the closed-form solution for \( \{\mathbf{c}_j\}_{j=1}^n \) as

\[
\mathbf{c}_j^{t+1} = \frac{\sum_{i=1}^{m} \delta(y_i = j) p_i^{t+1} \mathbf{x}_i}{\sum_{i=1}^{m} \delta(y_i = j) p_i^{t+1}}, \quad j = 1, \ldots, n, \tag{A.7}
\]

where \( \delta(\text{condition}) \) is the indicator function, and it is equal to 1 if the condition is satisfied and equal to 0 otherwise. Many studies show that the adaptive kernel width \( \sigma \) is a practical solution for fast convergence and better performance [35]. According to [36], the hyperparameter \( \{\sigma_j^2\}_{j=1}^n \) is updated in
each iteration by:
\[(\sigma^t_j)^2 = \frac{\sum_{i=1}^{m} \delta(y_i = j) \| x_i^t - c_{y_i} \|^2}{2 \sum_{i=1}^{m} \delta(y_i = j)},\quad j = 1, \ldots, n,\]  

(A.8)

**Appendix B**

Based on [25], the softmax loss is defined as
\[\mathcal{L}_S = -\sum_{i=1}^{m} \log \frac{e^{W_j^T x_i + b_j}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}\]  

(B.1)

where \(W_j \in \mathbb{R}^d\) denotes the \(j\)th column of the weights \(W \in \mathbb{R}^{d \times n}\) in the last dense layer and \(b \in \mathbb{R}^d\) is the bias term. By combining the softmax loss (\(\mathcal{L}_S\)) and CCL (1) the objective function is given as follows,
\[
\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_{CCL} = -\sum_{i=1}^{m} \log \frac{e^{W_j^T x_i + b_j}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2} \sum_{i=1}^{m} \left[ 1 - \exp \left( -\frac{\| x_i - c_{y_i} \|^2}{2\sigma^2_{y_i}} \right) \right],\]  

(B.2)

The discriminative and robust feature learning process is summarized in **Algorithm 1**, and we did not include the bias term in the implementation.

**Algorithm 1** The discriminative and robust feature learning algorithm
1: **Input**: Training data \(\{a_i | i = 1, \ldots, m\}\). Initialized parameters \(\theta\) in convolution layers. Parameters \(W\) and \(\{c_j | j = 1, \ldots, n\}\) in loss layers. Hyperparameter \(\lambda\) and learning rate \(\mu^t\). The number of iteration \(t = 0\).
2: **Output**: The parameter \(\theta\).
3: **while** Not convergent \(t \leftarrow t + 1\).
4: \(t \leftarrow t + 1\).
5: Compute \(x_i^t = a_i^t \theta^t\).
6: Compute \(L_S^t, L_{CCL}^t, \text{ and } L^t\).
7: Compute \(\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L_S^t}{\partial x_i^t} + \lambda \frac{\partial L_{CCL}^t}{\partial x_i^t}\).
8: Update \(W^{t+1} = W^t - \mu^t \frac{\partial E}{\partial W^t}\).
9: Update \(\theta^{t+1} = \theta^t - \mu^t \sum_{i=1}^{m} \frac{\partial L^t}{\partial x_i^t} \frac{\partial x_i^t}{\partial \theta^t}\).
10: Update \(\sigma^t_j \leftarrow (A.8), j = 1, \ldots, n\).
11: Update \(p_i^{t+1} \leftarrow (A.5), i = 1, \ldots, m\).
12: Update \(c_j^{t+1} \leftarrow (A.7), j = 1, \ldots, n\).
13: **end while**
References

1. Who W: _World report on disability_. Geneva: WHO 2011.
2. Shih JJ, Krusienski DJ, Wolpaw JR: Brain-Computer Interfaces in Medicine. *Mayo Clinic Proceedings* 2012, 87(3):268-279.
3. Jiang Y, Zhang Y, Lin C, Wu D, Lin CT: EEG-Based Driver Drowsiness Estimation Using an Online Multi-View and Transfer TSK Fuzzy System. *IEEE Transactions on Intelligent Transportation Systems* 2021, 22(3):1752-1764.
4. Aih M, Jun SC: Performance variation in motor imagery brain-computer interface: a brief review. *Journal of neuroscience methods* 2015, 243:103-110.
5. Pfurtscheller G, Guger C, Müller G, Krausz G, Neuper C: Brain oscillations control hand orthosis in a Tetraplegic. *Neuroscience Letters* 2000, 289(2):211-214.
6. Elstob D, Secco E: A Low Cost Eeg Based Bci Prosthetic Using Motor Imagery. *International Journal of Information Technology Convergence and Services* 2016, 6:23-36.
7. Sunny TD, Aparna T, Neethu P, Venkateswaran J, Vishnupriya V, Vyas PS: Robotic Arm with Brain – Computer Interfacing. *Procedia Technology* 2016, 24:1089-1096.
8. Bi L, Fan X, Liu Y: EEG-Based Brain-Controlled Mobile Robots: A Survey. *IEEE Transactions on Human-Machine Systems* 2013, 43(2):161-176.
9. Birbaumer N: Breaking the silence: brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology* 2006, 43(6):517-532.
10. Zhang Y, Jiang Y, Qi L, Bhiuian MZA, Qian P: Epilepsy Diagnosis Using Multi-view & Multi-medoid Entropy-based Clustering with Privacy Protection. 2021, 21(2) %J ACM Trans. Information Technol. ): Article 48.
11. Kotchekov IS, Hwang BY, Appelboom G, Kellner CP, Connolly ES: Brain–computer interfaces: military, neurosurgical, and ethical perspective. *Neurosurgical Focus* 2010, 28(5):E25.
12. Kai Keng A, Zheng Yang C, Haibong Z, Cuntai G: Filter Bank Common Spatial Pattern (FB CSP) in Brain–Computer Interface. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence): 1-8 June 2008 2008; 2008: 2390-2397.
13. Wierczala P, Zapala D, Wojciek GM, Masiak J: Most Popular Signal Processing Methods in Motor-Imagery BCI: A Review and Meta-Analysis. *Frontiers in neuroinformatics* 2018, 12:78.
14. Barachant A, Bonnet S, Congedo M, Jutten C: Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing* 2013, 112:172-178.
15. Xu Q, Zhou H, Wang Y, Huang J: Fuzzy support vector machine for classification of EEG signals using wavelet-based features. *Medical Engineering & Physics* 2009, 31(7):858-865.
16. Zhang Y, Wang G, Cheng F-I, Wang S: Support vector machines with the known feature-evolution priors. *IEEE Transactions on Image Processing* 2011, 20(9):2731-2742.
17. Craik A, He Y, Contreras-Vidal JL: Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of Neural Engineering* 2019, 16(3):031001.
18. Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ: EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* 2018, 15(5):056013.
19. Schirrmieister RT, Springenberg JT, Fiederer LDJ, Glasheteter M, Eggensperger K, Tangermann M, Hutter F, Burgard W, Ball T: Deep learning with convolutional neural networks for EEG decoding and visualization. 2017, 38(11):5391-5420.
20. Wen Y, Zhang K, Li Z, Qiao Y: A Discriminative Feature Learning Approach for Deep Face Recognition. In: 2016: *Cham*: Springer International Publishing; 2016: 499-515.
21. Hang W, Feng W, Du R, Liang S, Chen Y, Wang Q, Liu X: Cross-Subject EEG Signal Recognition Using Deep Domain Adaptation. *IEEE Access* 2019, 7:128273-128282.
22. Zhao H, Zheng Q, Ma K, Li H, Zheng Y: Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification. *IEEE Transactions on Neural Networks and Learning Systems* 2021, 32(2):535-545.
23. Zhou N, Cheng H, Qin J, Du Y, Chen B: Robust High-Order Manifold Constrained Sparse Principal Component Analysis for Image Representation. *IEEE Transactions on Circuits and Systems for Video Technology* 2019, 29(7):1946-1961.
24. He R, Hu B, Zheng W, Kong X: Robust Principal Component Analysis Based on Maximum Coentropy Criterion. *IEEE Transactions on Image Processing* 2011, 20(6):1485-1494.
25. Bishop CM: Pattern recognition. *Machine learning* 2006, 128(9).
26. Tangermann M, Müller K-R, Aertsen A, Birbaumer N, Braun C, Brunner C, Leeb R, Mehring C, Miller K, Mueller-Putz G et al: Review of the BCI Competition IV. *Frontiers in Neuroscience* 2012, 6:55.
27. Luo L, Xie Y, Zhang Z, Li W-J: Support Matrix Machines. In: *Proceedings of the 32nd International Conference on Machine Learning*. Edited by Francis B, David B, vol. 37. Proceedings of Machine Learning Research: PMLR; 2015: 938–947.
28. Kang H, Nam Y, Choi S: Composite Common Spatial Pattern for Subject-to-Subject Transfer. *IEEE Signal Processing Letters* 2009, 16(8):683-686.
29. Samek W, Meinecke FC, Müller K: Transferring Subspaces Between Subjects in Brain–Computer Interfacing. *IEEE Transactions on Biomedical Engineering* 2013, 60(8):2289-2298.
30. Jolliffe IT, Cadima J: Principal component analysis: a review and recent developments. 2016, 374(2065):20150202.
31. Van der Maaten L, Hinton G: Visualizing data using t-SNE. *Journal of machine learning research* 2008, 9(11).
32. Fazel-Rezai R, Allison BZ, Guger C, Sellers EW, Kleih SC, Kübler A: P300 brain computer interface: current
challenges and emerging trends. *Frontiers in neuroengineering* 2012, 5:14.

33. Guger C, Allison B, Grosswindhager B, Prückl R, Hintermüller C, Kapeller C, Bruckner M, Krausz G, Edlinger G: How Many People Could Use an SSVEP BCI? 2012, 6(169).

34. Ahn S, Kim K, Jun SC: Steady-State Somatosensory Evoked Potential for Brain-Computer Interface-Present and Future. *Frontiers in human neuroscience* 2015, 9:716.

35. Huang F, Zhang J, Zhang S: Adaptive Filtering Under a Variable Kernel Width Maximum Correntropy Criterion. *IEEE Transactions on Circuits and Systems II: Express Briefs* 2017, 64(10):1247-1251.

36. He R, Zheng W, Hu B: Maximum Correntropy Criterion for Robust Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2011, 33(8):1561-1576.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Sourcecodes.rar