Semi-Supervised Contrastive Learning for Generalizable Motor Imagery EEG Classification

Jinpei Han, Xiao Gu, and Benny Lo
Hamlyn Centre, Imperial College London
{j.han20, xiao.gu17, benny.lo}@imperial.ac.uk

***Abstract*—Electroencephalography (EEG) is one of the most widely used brain-activity recording methods in non-invasive brain-machine interfaces (BCIs). However, EEG data is highly nonlinear, and its datasets often suffer from issues such as data heterogeneity, label uncertainty and data/label scarcity. To address these, we propose a domain independent, end-to-end semi-supervised learning framework with contrastive learning and adversarial training strategies. Our method was evaluated in experiments with different amounts of labels and an ablation study in a motor imagery EEG dataset. The experiments demonstrate that the proposed framework with two different backbone deep neural networks show improved performance over their supervised counterparts under the same condition.**

***Index Terms*—motor imagery, EEG, generalization, semi-supervised learning, contrastive learning**

I. Introduction

Motor imagery (MI) based brain-computer interface (BCI) systems allow users to control external devices by mental execution. It plays an important role in rehabilitation engineering, such as interpreting the movement intentions of patients for assistance or therapeutic training. A successful MI BCI system requires a mechanism to record brain signals for interpretation. Among existing recording tools, Electroencephalography (EEG), which captures the brain electrical activities, is the most commonly used method due to its low cost, convenience, non-invasiveness and high temporal resolution. However, as the evoked potential of the brain activities is very weak and can easily be affected by artifact, EEG signals are often inherently noisy, and it is very difficult to relate the noisy signal with mental tasks. Research efforts have been devoted to the development of machine learning algorithms to enable automatic MI classification from EEG signals.

Among existing algorithms, conventional ones usually involve two steps, feature extraction and classification. Commonly feature extraction methods, such as Fourier/Wavelet transforms, Common Spatial Patterns [1] are followed by classification methods, such as Linear Discriminant Analysis [2], Support Vector Machine (SVM) [3]. These methods typically are deterministic and relatively less complex than deep learning methods and are less prone to overfitting [4].

In recent years, deep learning (DL) models have shown reasonable results in subject dependent classifications of EEG signals. Compared to conventional methods, deep learning models are well suited for end-to-end learning, performing inference from the raw data without prior feature selection [5]. Moreover, DL methods can scale well to large datasets and can simultaneously learn intricate high dimensional features from raw signals. The most commonly used DL models in MI-EEG classification are CNN based models, such as EEGNet [6] and DeepConvNet [5]. They have demonstrated superior performance on many tasks compared to conventional machine learning methods [5].

Despite the success of aforementioned DL methods, there remain several issues with regards to establishing a robust and accurate MI-based BCI system. It remains challenging to get access to large volumes of annotated high-quality data for MI classification training [7], [8]. In fact, knowing what the subjects are actually thinking or doing in cognitive neuroscience experiments could be challenging and which lead to difficulties in obtaining accurate, high-quality annotations and labels for motor-imagery EEG data [7].

Self-supervised learning (SSL) has opened the possibility of making use of self-generated pseudo labels for training on unlabelled data, with limited access to ground truth labels [9]. It performs training on a pretext task that tries to learn effective representations using the unlabeled data and pseudo labels and which is then followed by a downstream discrimination task. Example pretexts including relative positioning, temporal shuffling, contrastive predictive coding for EEG based sleep-staging [7] contrastive muti-segment coding and contrastive multi-lead coding for ECG based arrhythmia detection [10]. However, these pretexts are not suitable for motor imagery datasets since: (1) MI trials are recorded in discrete, short windows instead of continuous recordings like sleep monitoring or ECG. (2) Researchers usually ask participants to perform different motor imagery tasks in purely random order. (3) Different MI tasks involve activations in different parts of the brain; thus, recordings from different electrodes at one time might not share the same context. Therefore, the conventional SSL assumptions used in biosignals of both temporal and spatial invariance do not hold for MI EEG datasets.

Therefore, in our work, a semi-supervised learning structure is proposed, which makes use of a large quantity of unlabeled data and a small number of labels in an end-to-end manner. Inspired by the success of SimCLR for SSL [11], we apply the contrastive learning method to learn representations on unlabelled data. It involves applying different augmentations to the same unlabelled data and contrast against all different sets of data. This approach promotes the model to learn feature

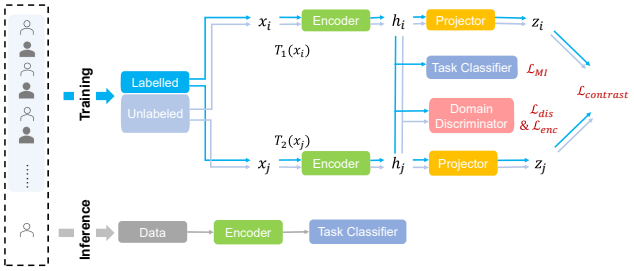


Fig. 1. Overview of our proposed semi-supervised learning framework.

representations that are invariant to transformations.

Another issue is raised from the complication of measuring biosignals. The performance of MI-based BCI systems depends heavily on the subject's neurophysiological and psychological conditions [8]. Inter-subject and inter-session variations caused by different mental conditions, psychological states, and electrode positions would bias the underlying distribution of recorded signals, thus degrading the performance of BCI systems. In fact, most existing computational models for motor imagery classification are domain dependent, yielding inferior results within cross-subject/session validation [12]. Recent research has successfully applied transfer learning or domain adaptation to address this issue [13]. However, this line of research assumes the visibility of partial data from the targeted domain during training, which limits its generalisation to a totally new domain. Inspired by recent work of applying adversarial training to solve domain generalization problems [14], we thus adopt an adversarial training module to disentangle the subject/session-specific information from the desired MI information in the latent representation.

Overall, to address the above issues of MI EEG classification, we propose a semi-supervised framework with a combination of self-supervised contrastive learning and adversarial training. The former contrastive learning is applied on all the training data without the involvement of any labels. Simultaneously, the latter adversarial domain discriminator is applied to diminish the feature distribution of different subjects/sessions in an adversarial way. Additionally, supervised cross entropy loss is applied on the partial labelled data to force the model to learn task-relevant features. It is demonstrated that the whole framework can learn subject/session-independent task-relevant representations effectively with minimal labels.

II. Methodologies

A. Methods Overview

Our framework consists of five main components: a set of data augmentation, an encoder, a task classifier, a domain discriminator and a projector. The overall pipeline of the proposed framework is summarised in Fig. 1. To tackle the real-world challenges, we assume only a small portion of each subject's data was recorded with user's label feedback for calibration purpose, denoted as L and the majority of data was collected unlabelled without the subject's attention, denoted as U. We aim to apply this semi-supervised learning framework to use the mixture of calibration data and the massive amount of unlabelled multi-subject/-session data to enable the prediction of the MI label from the new domain.

A batch of N EEG signals x containing both labelled L and unlabelled U data are first processed with a set of data augmentation T to facilitate contrastive learning as well as increasing the number of trainable data. The encoder E produces the latent vector h from the augmented input data. The latent vector is passed to the task classifier C for task label classification. At the same time, the domain discriminator D performs adversarial training with the latent vector to promote the encoder to learn domain-invariant features. The projector P projects the latent vector onto a lower-dimensional space to calculate the contrastive loss, which promotes the encoder to learn transformation-invariant features.

B. Data Augmentations for Contrastive learning

Chen *et al.* [11] have suggested that a composition of strong data augmentations is beneficial for contrastive learning. While augmentations for visual representations are visible and straightforward, it is non-intuitive to design data augmentations for EEG signals that can preserve the semantic information corresponds to different motor imaginations. We apply two different sets of augmentations to N sampled EEG data x in each iteration, yielding 2N of $T_1(x_i)$ and $T_2(x_j)$. Two random combinations of augmentations were used for each training sample. The designed augmentations include:

- Noise addition: addition of Gaussian noise.
- DC shift: the signal amplitude is shifted by a constant.
- Temporal roll: a section of the time-series data at the end of the window is rolled to the front and vice versa.
- Amplitude scale: the signal is scaled with a constant.
- Temporal cutout: A segment is masked by zeros.
- Crop and upsample: A segment of the data is cropped, and the data is replaced with the upsampled segment.

C. Self-Supervised Contrastive Learning

We tested two different CNN backbones as the encoder, namely: EEGNet and DeepConvNet. The two augmented batches $T_1(x_i)$ and $T_2(x_j)$ are passed through the encoder E to form latent representations $h_i = E(T_1(x_i))$ and $h_j = E(T_2(x_j))$. Following the inspiration of SimCLR, which applies a nonlinear projector head for contrastive learning for better representation quality [11], projector P is used to project the latent representation h_i and h_j to a lower dimension $z_i = P(h_i)$ and $z_j = P(h_j)$. z_i and z_j are then normalised with the l_2 norm and the contrastive loss is applied to measure the encoded mutual information in the same pair of windows with different augmentations, which is formulated as below,

$$\mathcal{L}_{contrast}(z_i, z_j) = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} exp(sim(z_i, z_k)/\tau)} \quad (1)$$

where $sim(z_i, z_j)$ represents the cosine similarity between vector z_i and z_j. τ is a temperature scaling parameter and $\mathbb{1}_{[k\neq i]}$ denotes a masking function which return 1 when $k \neq i$. The denominator sums up the exponential of cosine similarity of all the other 2(N-1) negative examples with respect to each anchor example within a batch. Ultimately, the contrastive loss function maximises the similarity between a pair of the same data with different augmentations and minimises the similarity between different data.

D. Supervised Learning with Cross Entropy Loss

In a semi-supervised setting, only a small portion of data in each batch contains labels. Thus only the $i, j \in L$ of the latent representation h are used to train the supervised classifier C. Where L represent the set of indices of the labelled data. Convolutional classification layers are used instead of the linear classification layer to reduce the number of trainable parameters and prevent overfitting from the small amount of labelled data. A cross-entropy loss $\mathcal{L}_{MI}$ is calculated from the outputs of the classification layer and the supervised labels:

$$\mathcal{L}_{MI} = -\sum_{k \in L} y_k \ log(C(h_k))) \tag{2}$$

E. Domain-Independent Adversarial Training

To force the encoder to learn domain-independent information, an adversarial training strategy is applied. A domain discriminator D is trained using the domain label s to predict the identities of the domains based on the concatenation of latent vectors h, whilst at the same time, the encoder E is encouraged to confuse the discriminator D so that the predicted label by D is random. In practice, we applied $\mathcal{L}_{dis}$ and $\mathcal{L}_{enc}$ to update D and E alternatively, which are formulated as below,

$$\begin{aligned} \mathcal{L}_{dis} &= -\sum_{k \in L \cup U} s_k \ log(D(h_k)) \\ \mathcal{L}_{enc} &= -\sum_{k \in L \cup U} s_{rand} \ log(D(h_k)) \end{aligned} \tag{3}$$

where s_{rand} denotes the randomly generated identity labels.

Overall, the encoder, projector, and task classifier are updated based on $w_1\mathcal{L}_{contrast} + w_2\mathcal{L}_{MI} + w_3\mathcal{L}_{enc}$, and the discriminator is updated based on $\mathcal{L}_{dis}$, where w_1, w_2, w_3 are the weights for each part of the loss. In the inference stage, the raw EEG sequence goes through encoder and classifier to predict the imagined actions.

III. Experiment and Results

A. Dataset and Experimental Setup

1) Dataset Description: The experiments were conducted using the BCIC IV 2a MI-EEG dataset from the MOABB library. EEG signals were recorded from 2 separate sessions of 9 different subjects where the subjects were asked to imagine the different types of movements. In each trial, a cue was shown to indicate the subjects to perform the desired MI task for 4 seconds, followed by a relaxed state between trials. During the sessions, 22-channels of EEG data were recorded at 250 Hz and bandpass-filtered between 0.5 Hz and 100 Hz.

2) Experiment Setup and Evaluation Protocols: The system was implemented using Python (v3.8.6), Pytorch (v1.7.0) and the Braindecode (v0.5) library. We filtered between 4Hz and 40Hz and converted it into microvolt. We used all 22 channels of the EEG recording and the entire 4 seconds of the trial windows. The EEG windows were then resampled from 250Hz to 128Hz resulting in a length of 512 sample points for each window and processed through channel-wise z-score normalisation. We tested two different convolution models, EEGNet-8,2 [6] and DeepConvNet [5], as the encoder backbone. The task classifier C consists of a single Conv2D layer with a kernel size equal to the latent vector's width and bias. The domain discriminator D consists of a linear layer that reduces the flattened latent vector to half of its length, followed by a leaky ReLU layer, a dropout layer and another linear layer with an output size of the domain size.

Experiments were conducted to evaluate the framework's robustness against inter-subject/session variations in binary classification (left hand and right hand) tasks. We used **Leave-One-Session-Out** to evaluate the framework's robustness against inter-session variations by treating different sessions of all the subjects as separate independent domains resulting in 18 domains in total. **Leave-One-Subject-Out** was used to evaluate the robustness against inter-subject variations by treating different subjects as separate domains, resulting in 9 domains in total. 10% of the train data was stratify sampled as a separate dataset for 10-fold cross-validation and early stopping. Since contrastive learning benefits by large batch sizes [11], we used a batch size of 1024 containing different labelled and unlabelled data sampler proportions. As Chen *et al.* [11] reported, longer training facilitates contrastive learning; thus, we used a maximum epoch of 1000 and early stopping with 200 patience. An Adam optimiser with a 0.0005 learning rate and a 1×10^{-4} weight decay was used for E,C,P. Another Adam optimiser with a 0.0002 learning rate was used for the domain discriminator D. The specific parameters applied for data augmentations are shown in Table I.

TABLE I
AUGMENTATION PARAMETER RANGES.

Augmentation	min	max	Augmentation	min	max
Gaussian Noise(σ)	0	0.2	Amplitude scale	0.8	1.2
DC shift	-0.8	0.8	Temporal cutout(s)	0	1
Temporal roll(s)	-1	1	Crop and upsample(s)	3	3

B. Classification Results

1) Compared Methods:

- Filter-Bank Common Spatial Pattern (FBCSP) [1] applies band-pass filtering followed by spatial filtering using Common Spatial Pattern on each frequency bands. SVM is used to perform classification based on the most discriminative CSP features from the filter bank.
- EEGNet [6] consists of a block with a 2D convolution filter and a Depthwise Convolution spatial filter, and a second block with a Separable Convolution operation.
- DeepConvNet [5] consists of a block with a temporal convolutional layer following by a spatial convolutional layer and max pooling, and three blocks that each made up of a convolutional layer and a max-pooling layer.

2) Quantitative Results: The inter-session classification results are presented in Table II. As shown, the fully supervised methods suffer significantly from the limited number of training data, whereas our proposed semi-supervised framework using either EEGNet or DeepConvNet encoder is less affected by the limited supervised labels than their counterparts. It demonstrates the feasibility of our methods when labels are scarce or of low-quality, which is well suited for MI BCI applications. Especially, the classification result of each session data when

TABLE II
INTER-SESSION CLASSIFICATION ACCURACY OF DIFFERENT METHODS WITH DIFFERENT RATIOS OF LABELS IN THE TRAINING.

Models	10%	20%	50%	100%
FBCSP	54.7%	58.8%	62.2%	64.8%
EEGNet	60.7%	68.0%	71.4%	75.8%
DeepConvNet	56.2%	65.4%	76.5%	**80.9%**
Semi-EEGNet	66.6%	71.5%	75.3%	75.6%
Semi-DeepConvNet	**67.6%**	**74.3%**	**77.4%**	79.4%

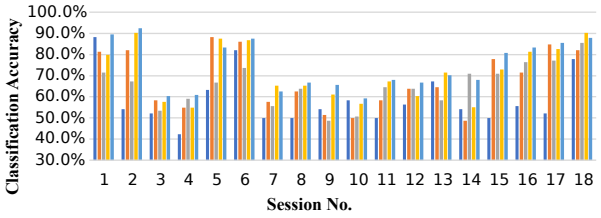


Fig. 2. Inter-session classification results with 20% labels.

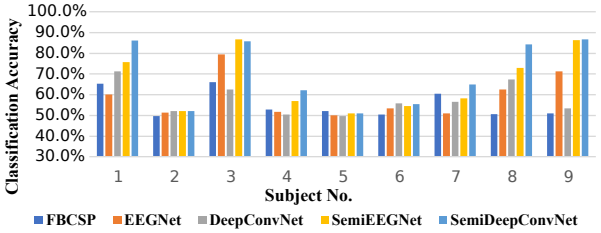


Fig. 3. Inter-subject classification results with 20% labels.

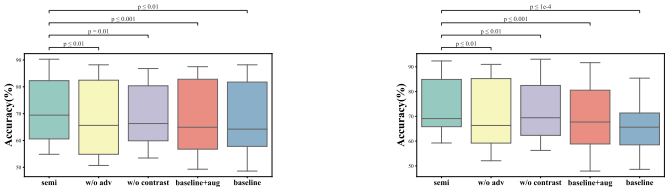
only 20% labels were used is displayed in Fig. 2. Furthermore, both semi-supervised methods perform much better than the baselines in inter-subject classification tasks on average shown in Fig. 3. Overall, our proposed method outperforms compared methods in both inter-session/subject settings when label is limited. However, when the training label is sufficient, e.g. when 100% labels were used, the proposed method might result in lower performance than the fully supervised ones. This might be caused by those negative pairs belonging to the same class caused confusion to the encoder, without providing any extra useful semantic information. This problem could be potentially alleviated if incorporating unlabelled data from additional sources for contrastive learning.

C. Ablation Study

We also conducted experiments to evaluate the effectiveness of both contrastive learning and adversarial training when only 20% of the labels were used. The results shown in Fig. 4b suggest that the DeepConvNet is data-intensive; with the help of data augmentation, the performance can be significantly improved in limited data condition. In contrast, EEGNet performs similarly with or without data augmentation in Fig. 4a. Nonetheless, both base networks are affected by the session-based information. The proposed semi-supervised networks achieve the highest accuracy by combining data augmentation, contrastive learning, and adversarial training.

IV. CONCLUSION

This work presented an end-to-end semi-supervised learning method for classifying MI EEG signals with limited labels. We compared the binary classification accuracy of our semi-supervised methods with three different baseline methods. The



(a) EEGNet (b) DeepConvNet

Fig. 4. Impact of different framework components (adv: adversarial loss, contrast: contrastive loss, aug: augmentation) removed, with 20% labels.

results showed that our proposed network significantly outperforms the baseline methods when there are limited labels. Furthermore, we conducted an ablation study to demonstrate how each component of our framework contributed to the improved performance. The study indicates that deep neural networks could learn subject/session invariant features from a small dataset with a small number of labels with the proposed framework. Overall, our work opened new possibilities for using deep neural networks for real-world applications without the need for tedious calibration processes.

REFERENCES

[1] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 Int Jt Conf Neural Netw. (IEEE WCCI)*. IEEE, 2008, pp. 2390–2397.

[2] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8659–8666, 2010.

[3] W. Xu, C. Guan, C. E. Siong, S. Ranganatha, M. Thulasidas, and J. Wu, "High accuracy classification of eeg signal," in *Proc. ICPR 2004.*, vol. 2. IEEE, 2004, pp. 391–394.

[4] M. Iftikhar, S. A. Khan, and A. Hassan, "A survey of deep learning and traditional approaches for eeg signal processing and classification," in *IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. 2018*. IEEE, 2018, pp. 395–400.

[5] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Hum. Brain Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

[6] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, p. 056013, 2018.

[7] H. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, "Uncovering the structure of clinical eeg signals with self-supervised learning," *J. Neural Eng.*, vol. 18, no. 4, p. 046020, 2021.

[8] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "Eeg-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, p. 1423, 2019.

[9] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal Mach Intell.*, 2020.

[10] D. Kiyasseh, T. Zhu, and D. A. Clifton, "Clocs: Contrastive learning of cardiac signals," *arXiv preprint arXiv:2005.13249*, 2020.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf Mach Learn.* Proc Mach Learn Res., 2020, pp. 1597–1607.

[12] H.-I. Suk and S.-W. Lee, "A novel bayesian framework for discriminative feature extraction in brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 286–299, 2012.

[13] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets," *IEEE Trans. Cogn. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, 2018.

[14] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of eeg signals with adversarial domain generalization," in *ICONIP*. Springer, 2019, pp. 30–42.