Contrastive Self-supervised EEG Representation Learning for Emotion Classification

Keya Hu[1], Ren-Jie Dai[2], Wen-Tao Chen[1], Hao-Long Yin[1], Bao-Liang Lu[1, 3] *Fellow IEEE* and Wei-Long Zheng[1, *]
[1]*Department of Computer Science and Engineering, Shanghai Jiao Tong University*
[2]*School of Software Engineering, Tongji University*
[3]*RuiJin-Mihoyo Laboratory, Clinical Neuroscience Center, RuiJin Hospital, Shanghai Jiao Tong University School of Medicine*

***Abstract*—Self-supervised learning provides an effective approach to leverage a large amount of unlabeled data. Numerous previous studies have indicated that applying self-supervision to physiological signals can yield better representations of the signals. In the paper, we aim to apply this method to the crucial field of emotion recognition. We perform the experiment with several state-of-the-art contrastive self-supervised methods to explore their effectiveness in pre-training feature encoders on raw electroencephalography (EEG) signals and fine-tuning the pre-trained encoders on the downstream emotion classification tasks. We attempt to vary the proportion of labeled data used during fine-tuning and find that the improvement from self-supervised methods is more pronounced when the proportion of labeled data is small. Additionally, we explore the transferability of the feature encoders pre-trained on various datasets and observe that most self-supervised methods exhibit a certain degree of transferability. Methods that effectively utilize the temporal information in EEG signals show superior stability, accuracy, and transferability.**

***Index Terms*—Self-supervised learning, EEG emotion classification, Affective computing**

I. Introduction

Recognizing and comprehending emotions, which play a pivotal role in human daily life [1], are fundamental steps in human interaction. Affective Brain-Computer Interfaces (aBCIs) offer a technological means to directly detect human emotions from electroencephalogram (EEG) signals. Performing emotion classification tasks on EEG signals collected through aBCIs represents the initial stride toward grasping human emotions [2]. It also has practical value in the objective assessment of potential emotional disorders in mental health.

Nowadays, the volume of data used for training plays an increasingly crucial role in the performance of models. Based on vast amounts of unlabeled data, self-supervised learning (SSL) has shown remarkable advantages in fields such as speech recognition [3], [4] and natural language processing [5]–[7]. Previous research on SSL for temporal physiological signals has also yielded promising results. For instance, in datasets related to sleep stage detection, epilepsy detection, human activity recognition [8]–[10], and other temporal data sets, the SSL approaches that utilize a large amount of unlabeled data for pre-training and fine-tuning on a smaller labeled dataset have proven to achieve favorable outcomes.

In the task of emotion classification, previous studies with manually extracted various EEG features have achieved good results in multiple subject-dependent training scenarios [11]. Emotion recognition using SSL with extracted EEG features has also yielded excellent results [12]. To better utilize the vast amount of EEG data, especially unlabeled data, and to learn more robust representations of EEG data, it is ultimately necessary to perform SSL on raw EEG signals.

In our paper, we perform the experiment with several state-of-the-art contrastive-learning-based SSL methods to explore their effectiveness in pre-training on raw EEG signals. Throughout our experiments, we vary the quantity of labeled data used in the fine-tuning stage to investigate the impact of SSL on classification performance. The experimental results indicate that most SSL methods show significant improvement in accuracy when the amount of labeled data is limited. We also attempt cross-dataset pre-training, aiming to validate the potential of enlarging the pre-training dataset by incorporating a wider variety of EEG signals. The results show that most of the SSL methods are transferable.

II. Method

We use several SSL methods to pre-train our feature encoders to learn better representations and then evaluate them on downstream tasks by fine-tuning the feature encoders. The architecture of the pre-training methods and fine-tuning process are shown in Figure 1.

A. Self-supervised Learning Algorithm

We aim to learn better EEG representations by applying seven SSL methods, namely SimCLR, MoCo, ContraWR, BYOL, CPC, and TS-TCC, to pre-train EEG feature extractor in our experiments. The following is a brief introduction to these SSL methods.

1) SimCLR: A simple Framework for Contrastive Learning of Visual Representation [5] first applies two different data augmentation methods to the same data sample. Then it uses one encoder to generate anchor, positive, and negative samples

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 62376158), STI 2030-Major Projects+2022ZD0208500, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Shanghai Pujiang Program (Grant No. 22PJ1408600), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

*Corresponding author: weilong@sjtu.edu.cn

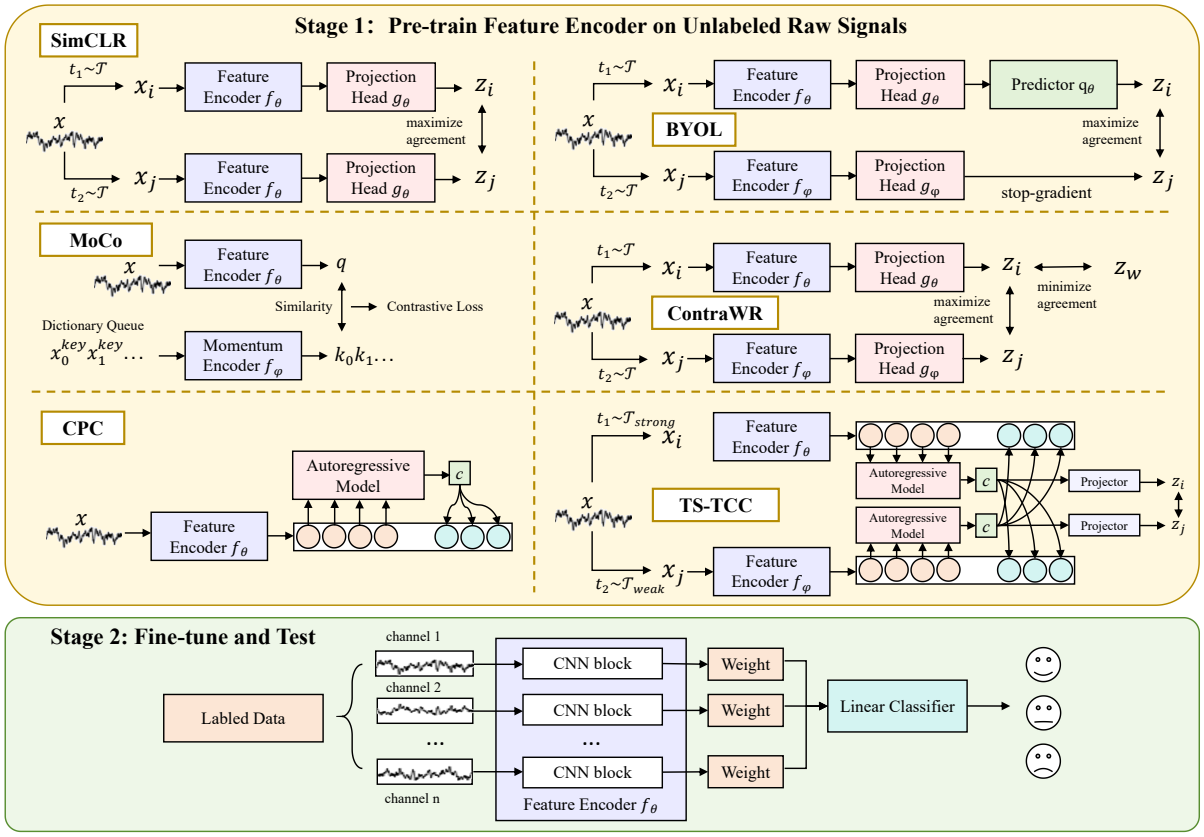


Fig. 1. The schematic diagram shows the architecture of our experiment. It illustrates the principles of various SSL methods in the pre-training stage on unlabeled raw signals and the emotion recognition model in the fine-tuning stage on limited labeled signals

from augmented samples. It maximizes agreement between differently augmented views of the same data samples while minimizing agreement between views of different samples within the mini-batch.

2) MoCo: Momentum Contrast [6] maintains a queue of encoded samples as a dictionary and updates it using a momentum encoder, the positive samples are obtained by augmenting the data within the current batch and then encoding it through an encoder. At the same time, negative samples are derived from the queue, and their encodings are obtained through another encoder.

3) ContraWR: Contrast with the World Representation [13] replaces the large number of negative samples with a single average representation over the dataset, called the world representation. The purpose of the loss function is to ensure that the similarity in representations between the anchor and the positive sample is stronger than the similarity between the anchor and the world representation.

4) BYOL: Bootstrap Your Own Latent [14] is a contrastive SSL method without negative pairs. It uses two neural networks, referred to as online and target networks, that interact and learn from each other. Starting from an augmented view of an image, BYOL trains its online network to predict the target network's representation of another augmented view of the same image.

5) CPC: Contrastive Predictive Coding [9] is a predictive contrastive SSL method that utilizes temporal information to learn robust representations of time series data. It extracts meaningful representations and then feeds them into an autoregressive model to predict future sequences. It aims to maximize the agreement between the correct and predicted future representations.

6) TS-TCC: Time-Series Representation Learning via Temporal and Contextual Contrasting [8] is a cross-view contrastive prediction SSL method based on CPC. The data are transformed using weak and strong augmentations, encoded by a feature extractor, and sent into an autoregressive model to generate context vectors. One augmented view's context vector predicts future sequences of the other one. The loss function aims to maximize the agreement between the correct and predicted future representations and the agreement between the context vectors of the two augmentations.

B. Raw EEG Emotion Classification Model

Our emotion classification feature encoder is constructed by 1D CNN models. Each CNN model contains three CNN blocks with 1D Convolution, BatchNorm, ReLU, and Max-Pooling layers. In the downstream task, we connect the learnable weights to the feature encoder and a single-layer linear classifier. Our goal is for the 1D CNN model to capture

valuable information from each channel of raw EEG signals, while the learned weights associated with each channel are intended to discern the significance of different channels in emotion classification. We utilize the pre-training methods mentioned earlier to obtain a well-trained feature encoder. Fine-tuning is then performed on the labeled data in the downstream emotion classification tasks.

III. Experiment

A. Datasets

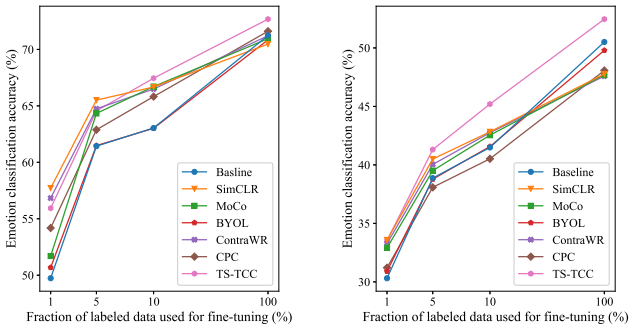
We use SEED and SEED-IV datasets for pre-training and evaluation. The SEED dataset contains EEG data of 15 subjects with different types of emotion, which are positive, negative, and neutral [15]. The SEED-IV dataset is an evolution of the original SEED dataset. It also contains EEG data of 15 subjects, the subjects in two datasets are different. The number of categories of emotions changes to four: happy, sad, fear, and neutral [16]. We downsample the continuous raw EEG signals from SEED and SEED-IV to 200 Hz. Each sample is taken as a 1-second window without overlap between consecutive sample points, forming the datasets.

B. Experimental Settings

We pre-train the feature encoder separately by the SSL methods of SimCLR, MoCo, ContraWR, BYOL, TS-TCC, and CPC for 100 epochs. We also get the randomly initialized feature encoder as the baseline of our experiment. Then, we append channel weights, followed by a linear classifier, to the encoder and fine-tune it for 30 epochs on SEED and SEED-IV datasets, obtaining the final accuracy. For the SEED dataset, we split it into a 3:1:1 ratio for training, validation, and testing sets, while for the SEED-IV dataset, we use a 4:1:1 ratio. We evaluate the models using three different random seeds to obtain three distinct training sets, testing sets, and validation sets while ensuring balanced label quantities within each fold. We use the validation dataset to find the best hyperparameters of various models, then finally test them on the test dataset to get results. The pre-training learning rate is 3e-4, the weight decay is 3e-4, and the batch size is 256. The fine-tuning learning rate is 1e-3, the weight decay is 1e-2, and the batch size is 16 when using no more than 5% dataset, 32 when using no more than 10% dataset, 256 otherwise.

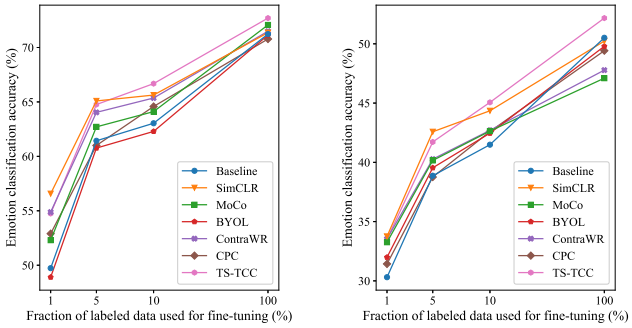
During the experiments, we attempt to fine-tune each subject from the SEED and SEED-IV datasets using the feature encoders, respectively, that are pre-trained on all the SEED and SEED-IV training datasets. We also explore cross-dataset transferability by fine-tuning each subject from SEED using the feature encoders pre-trained on the entire SEED-IV training dataset and vice versa.

During the fine-tuning stage, we randomly sample a specified percentage of data with three fixed seeds from the original dataset, train the feature encoders on this subset of data, and finally compute the average accuracies over all subjects. In our experiment, We divide the dataset into four gradients of data volume: 1%, 5%, 10%, and 100% to evaluate the performance.



(a) SEED (b) SEED-IV

Fig. 2. The average classification results of fine-tuning the feature encoders pre-trained on the unlabeled data, respectively, with different percentages of labeled data in the SEED and SEED-IV datasets.



(a) SEED-IV → SEED (b) SEED → SEED-IV

Fig. 3. The average classification results of cross-dataset pre-training and fine-tuning with the labeled SEED dataset with different percentages of the labeled data in the SEED and SEED-IV datasets, respectively.

IV. Results

A. Performance Comparison of Pre-trained Encoders

We test the feature encoders pre-trained by the above SSL methods, with the unlabeled data in the SEED and SEED-IV datasets serving as the pre-training datasets, respectively. The results are illustrated in Figure 2. It shows that with no more than 10% of the data volume, most self-supervised methods maintain a good lead in accuracy during fine-tuning. SimCLR, ContraWR, and TS-TCC methods consistently show great performance. They gain a 10% relative accuracy improvement with only 1% data volume in both datasets and achieve accuracy of 65.51%, 64.74%, and 64.55% on the SEED dataset with only 5% data, respectively.

We also explore the cross-dataset transfer performance after pre-training on different datasets. We fine-tune the feature encoders well pre-trained on the SEED-IV dataset with various SSL methods by each subject in the SEED dataset and vice versa. The results are shown in Figure 3. It can be observed that most SSL methods also show a significant advantage, especially in scenarios with limited data. SimCLR, ContraWR, and TS-TCC methods can also achieve approximately a 10%

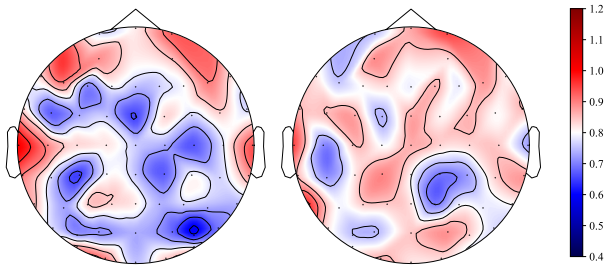


Fig. 4. The topography maps represent the average weights of each channel obtained from fine-tuning on the SEED dataset (the left one) and SEED-IV dataset (the right one) with fixed pre-trained encoders, respectively.

relative accuracy improvement with only 1% data volume in both datasets.

When all available labeled data is used, the volume of data is sufficient to fine-tune the model effectively. Thus, the improvement from self-supervised methods is not significant. However, when data volume is insufficient, the encoder obtained through self-supervised pretraining can better extract features, thereby improving classification accuracy to some extent.

To summarize, most SSL methods demonstrate good improvements and transferability when limited data is used, especially for SimCLR, ContraWR, and TS-TCC methods. Among them, the TS-TCC method based on temporal sequence SSL consistently outperforms others only based on augmentations in various data quantities, demonstrating excellent stability and remarkable transferability. We guess BYOL does not perform well since the signal-to-noise ratio of EEG signals is relatively low, and training only with augmented positive sample pairs is insufficient.

B. Key Brain Regions

After fine-tuning with the fixed feature encoders pre-trained by TS-TCC, we visualize the weights of different channels, represented as a topography map in Figure 4. Closer to red indicates higher values, while closer to blue indicates lower values. This topography map reflects the importance of different channels in emotional cognition. It can be observed that among these samples, the regions with high weights are relatively consistent with previous studies [15], [16], including the prefrontal lobe, temporal lobe, and occipital lobe. It can be inferred that these brain regions play a more significant role in tasks related to emotion.

V. Conclusion

In this paper, we attempt to apply some effective contrastive learning SSL methods to emotion classification tasks using raw EEG signals. In our experiments, it can be observed that pre-training on the raw EEG signals, followed by fine-tuning on a labeled dataset for emotion classification, leads to an improvement in classification accuracy, particularly in scenarios with limited labeled training data. Most SSL methods demonstrate the ability to learn more robust EEG features from raw signals. Furthermore, the pre-trained feature extractors exhibit good transferability, capturing EEG features that are not specific to the training dataset. This assures future experiments involving diverse EEG datasets, aiming to learn more robust representations of EEG signals.

Among these methods, TS-TCC, the method specifically designed for time series signals, demonstrates the highest stability, accuracy, and best transferability in most cases. This suggests that, for signals like EEG, it is essential to design self-supervised methods that prioritize temporal relationships.

References

[1] R. J. Dolan, "Emotion, cognition, and behavior," *Science*, vol. 298, no. 5596, pp. 1191–1194, 2002.

[2] D. Wu, B.-L. Lu, B. Hu, and Z. Zeng, "Affective brain–computer interfaces (aBCIs): A tutorial," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1314–1332, 2023.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[4] B. a simple, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning*, pp. 1597–1607, PMLR, 2020.

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Gira simpleck, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics*, 2019.

[8] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "Time-series representation learning via temporal and contextual contrasting," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 2352–2359, 2021.

[9] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv*, vol. abs/1807.03748, 2018.

[10] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Self-supervised learning for label- efficient sleep stage classification: A comprehensive evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1333–1342, 2023.

[11] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84, 2013.

[12] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," in *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, (New York, NY, USA), p. 6–14, Association for Computing Machinery, 2022.

[13] C. Yang, D. Xiao, M. B. Westover, and J. Sun, "Self-supervised EEG representation learning for automatic sleep staging," *Journal of Medical Internet Research AI*, 2023.

[14] J.-B. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[15] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[16] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[17] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Self-supervised learning for label-efficient sleep stage classification: A comprehensive evaluation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1333–1342, 2023.