Contrastive Learning Consistent and Identifiable Latent Embeddings for EEG

Feng Liang[1,2(✉)], Zhen Zhang[1,2,3], Jiawei Mo[4], and Wenxin Hu[1,2]

[1] Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen, China
{fliang,huwenxin}@smbu.edu.cn
[2] Guangdong-Hong Kong-Macao Joint Laboratory for Emotional Intelligence and Pervasive Computing, Shenzhen MSU-BIT University, Shenzhen, China
[3] School of Information Science and Engineering, Lanzhou University, Lanzhou, China
zhangzhen19@lzu.edu.cn
[4] School of Computer Science and Engineering, Central South University, Changsha, China
mojiawei@csu.edu.cn

Abstract. Extracting informative EEG data into low-dimension latent embeddings is important for storing and analyzing these neuron signals and applying them to various applications, such as modern human-computer interaction (HCI) techniques. We use the contrastive learning algorithm on time-domain features of EEG in both discovery-driven (self-supervised) and hypothesis (supervised) manners to encode the EEG data into latent embeddings that are proven consistent and identifiable. The self-supervised embeddings have the potential to be used for a range of downstream tasks, while the supervised embeddings have very high decoding accuracy for specific tasks. With embeddings encoded from EEG features collected within every 0.5-s window, the accuracy of recognizing the identities of persons by decoding the self-supervised and supervised embeddings is as high as 96.2% and 99.6%, respectively. Our method and results can promote new HCI techniques, e.g., automatically connecting users to their roles in AR games once they wear EEG-capable devices. The source code is available at: https://www.github.com/liangfengsid/timeEegContrastive.

Keywords: EEG · Contrastive learning · Latent embedding · Neural representation · Identity recognition

1 Introduction

Electroencephalography (EEG), the technique that intensively collects time-series electronic signals from the scalp, is widely used in clinical screening [13] and has a great potential application in modern human-computer interaction [12]

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 236–244, 2024.
https://doi.org/10.1007/978-3-031-65126-7_22

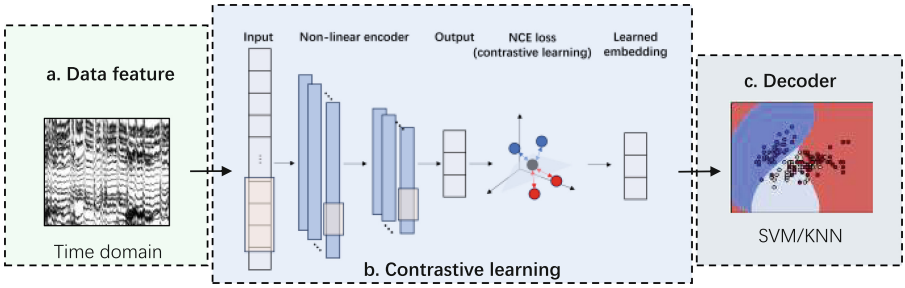

Fig. 1. The procedure of encoding time-domain EEG into embeddings by contrastive learning and decoding the learned embeddings

(HCI) and affective computing [10,16]. Encoding EEG data into consistent and identifiable latent embeddings [5,6] can greatly extend the application of EEG in various downstream tasks. The generated embeddings facilitate EEG applications by incorporating valuable information on EEG characteristics and filtering some irrelevant noise in the reduced-dimension representation. Most existing work on encoding EEG [7,14] uses supervised learning methods that depend on specific tasks, which limits its representation ability and application to other tasks. Recently, unsupervised learning [2,4,17] and self-supervised learning [3,8,11,19] methods have shown their capability of learning discriminative latent embeddings which can be generally used for downstream tasks. For example, [15] shows that contrastive learning can generate consistent and identifiable neural latent embeddings from instant spike-stimulated signals in a discovery-driven or hypothesis manner. However, neural signals are usually affected by continuous long-lasting stimuli, e.g., disease and environmental influence. The ability of contrastive learning to generate embeddings from continuous long-lasting stimulated EEG data is unknown.

In this paper, we investigate the ability of contrastive learning to generate consistent and identifiable EEG latent embeddings We use features of the time-domain representation of EEG as the input of a learnable encoder, which optimized by the noise-contrastive estimation [8] (NCE) in both the discovery-driven (self-supervised) and hypothesis (supervised) manners. We explore properties of consistency and interpretability by testing the convergence performance of the model and the visualization effect of the latent embeddings, respectively. We also verify the identifiablility of the embeddings by exploring the decoding performance of different downstream tasks. We find that encoding the time-domain features by contrastive learning can generate general EEG latent embeddings for some downstream applications. Excitingly, using the EEG latent embeddings that are encoded by 0.5-s time window features, the accuracy of recognizing the identities different persons is 96.2% in the self-supervised case and as high as 99.6% in the supervised case. The close-to-perfect decoding performance proves the potential of applying our method to emerging HCI and other EEG-related scenarios.

2 Method

Model. The whole procedure of the model is depicted in Fig. 1. We use an EEG dataset of emotion detection with neural activity with continuous long-lasting stimuli (Fig. 1.a). The encoder learns from time-domain features via a neural network with contrastive learning either in a discovery-driven manner or guided by task-specific labels and outputs EEG latent embeddings (Fig. 1.b). The decoder classifies the latent embeddings into different task-specific labels (Fig. 1.c).

Dataset. We use the SEED [18] dataset, which is a widely used open dataset designed for exploring the relationship between EEG and emotions. The dataset comprises EEG data from 15 people subjects joining a 3-session testing, with each testing session stimulated by watching 15 movie clips of a total of about 3600 s, which are continuous lasting stimuli. The movie stimuli are related to 3 emotions, i.e., positive, neutral, and negative. The EEG signals are collected by 62 electrode channels, down-sampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data grouped by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

Most work extracts EEG features from the frequency-domain representation [1,10,16], but some recent work [5,6] shows that interpretable latent embeddings can be learned from time-domain representation. We further down-sample the time-domain signals to 2 Hz so that within every 0.5 s, we can extract 5 voltage features for each of the 62 channels, namely the maximum, minimum, mean, median, and standard deviation. Finally, each encoder input is a 310-dimension vector with a 0.5-s time window, the volume of the training set is $\mathbb{R}^{270855\times310}$, and that of the testing set is $\mathbb{R}^{35280\times310}$.

Contrastive Learned Embeddings. The encoder is a non-linear convolutional neural network (CNN) or deep neural network (DNN) that applies contrastive learning optimizing the NCE loss, which follows a similar procedure as in [15].

For the input features x and y, where y is a positive or negative contrastive sample of x, let $p(x)$ be the probability density function of x, $p(y|x)$ and $q(y|x)$ be the probability density function of the positive and negative samples conditioned on x, respectively. Encoding x and y can be represented by a function f with normalized outputs, and $f(x)$ and $f(y)$ are the normalized latent embeddings, respectively. We use the dot product of $f(x)$ and $f(y)$ adjusted with a temperature parameter τ as the similarity function between these two latent embeddings, which is denoted as $\psi(x, y) = f(x)^T f(y)/\tau$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{x\sim p(x), y_+\sim p(y|x)\\ y_1,y_2,\ldots,y_n\sim q(y|x)}} [-\psi(x, y_+) + log \sum_{i=1}^{n} e^{\psi(x,y_i)}].$$

Positive and negative samples are taken from a minibatch of the training input. The identification of positive and negative samples depends on the scientific problem we are solving. In the discovery-driven manner when no label is provided, samples near x along the timeline are positive and those far away from x along the timeline are negative. While in the hypothesis manner, specific labels are provided and the learning process is similar to supervised contrastive learning in concept. Samples with the same label as that of x are positive, while those with different labels from x are negative. We train different encoder models without any label, with emotion labels, and with subject labels, respectively, and compare their convergence, latent visualization, and decoding performance, respectively.

We test the encoder with two different neural network structures, where one is a five-layer 1D convolutional network with skipping connections (CNN) and the other is a four-layer fully connected network (DNN). Perceptrons in both networks are activated by GELU functions. The mini-batch size of the input is 1,024, the learning rate is 0.001. The encoder output dimension can be 8, 16, and 32, and the number of mini-batch training iterations is 10,000 times the encoder output dimension.

Embedding Decoding. The decoder is a classification or regression model that fits the EEG latent embedding to the task-specific labels. We use the K-nearest neighbors (KNN) method (where $k = 5$) as the model and emotions and subjects as the labels, respectively. Both types of labels are discrete, where emotion labels have 3 values and subjects have 15. The embedding decoder is trained separately from the embedding encoder, where the encoder output embeddings are generated from the training input features by the well-learned encoder and the corresponding labels are the training inputs of the decoder.

3 Results and Discussions

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder using contrastive learning in discovery-driven and hypothesis manners. Figure 2 shows the NCE loss of the DNN and CNN encoders to encode EEG to 32-dimension latent embeddings provided without labels, with emotion labels, and with subject labels, respectively. The encoder tends to converge in all cases, which indicates that the encoder will generate consistent latent embeddings for EEG features that are considered similar. The NCE losses of encoding 8-dimension and 16 dimension latent embeddings also converge but jitter in a smaller magnitude, which is not depicted here to prevent redundancy.

3.2 Embedding Visualization

Low-dimension representations easily can be visualized to help the interpretation [9]. We visualize the first three dimensions of the testing embeddings generated by encoders of different neural network structures, contrastively learned

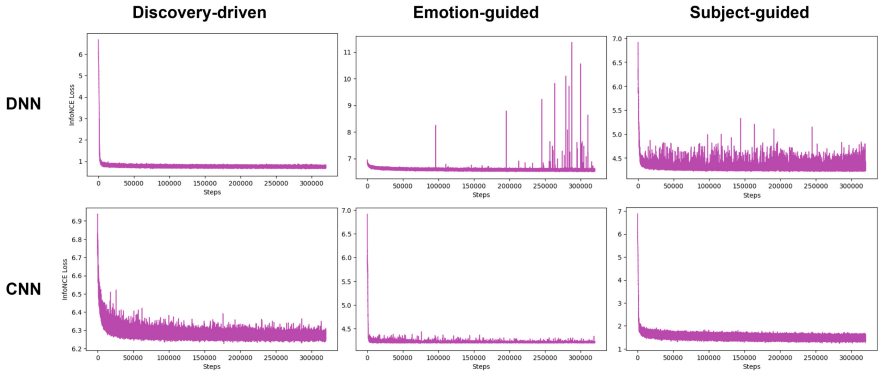

Fig. 2. The convergence performance of the DNN and CNN encoders using contrastive learning in discovery-driven, emotion-label-guided, and subject-label-guided manners, respectively

with different labels, and with different output dimensions (Fig. 3). The related labels of the data points can be distinguished by colors. From the figures, we can see that in the CNN encoder cases, where points of the same color or similar colors tend to cluster into color clouds according to some patterns, generally generate more identifiable embeddings than DNN ones, where color point clouds are more chaotic. The reason is that the CNN model considers the sequential context of the input data.

Specifically, in the CNN subject-guided cases even when the output dimension is as low as 8, we can clearly recognize 15 color clouds, which corresponds exactly to 15 subjects in the dataset. It shows that we can use a low-dimension EEG latent embedding to distinguish the identities of people, which is an exciting result and can greatly benefit new applications of HCI.

For CNN encoders trained in a discovery-driven manner, color clouds are formed in a more complex pattern and are not completely separated from each other. It indicates that the embeddings are generally informative and have great potential to be applied to various tasks. The higher the embedding dimension, the clearer the patterns and the more identifiable the color clouds.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings from different encoders to different labels are shown in Table 1. The accuracy of distinguishing different subjects has reached as high as 99.6% using 32-dimension latent embedding generated by the CNN encoder via supervised contrastive learning. Even for these latent embeddings of dimensions 8 and 16, the accuracy is as high as 99.5%, which is almost the same as the 32-dimension ones. It shows that we can compress EEG data to float vectors as small as 8 dimensions to represent

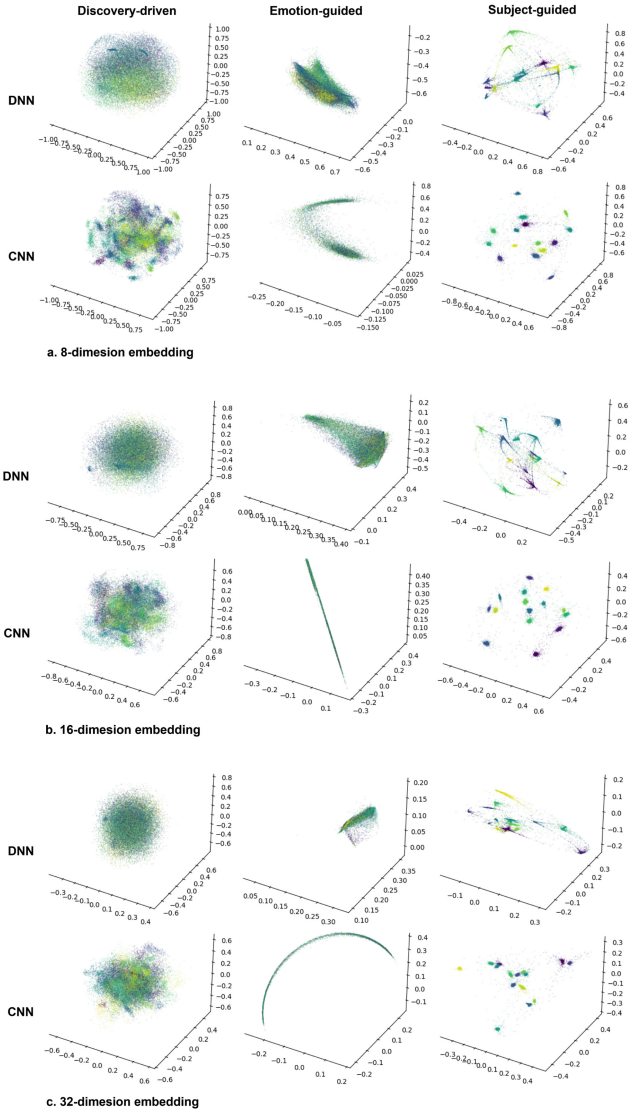

Fig. 3. The first three dimensions of DNN or CNN-learned embeddings of 8D, 16D, and 32D by discovery-driven, emotion-label-guided, and subject-label-guided contrastive learning, respectively.

Table 1. Decoding accuracy from different embeddings to different labels with different decoders

Decoding Label	Emotion		Subject	
	DNN	CNN	DNN	CNN
Discovery-driven embedding-8D	0.327	0.357	0.722	0.958
Discovery-driven embedding-16D	0.355	0.355	0.746	0.962
Discovery-driven embedding-32D	0.346	0.340	0.785	**0.962**
Emotion-guided embedding-8D	0.417	0.413	0.427	0.108
Emotion-guided embedding-16D	0.420	0.410	0.357	0.091
Emotion-guided embedding-32D	0.394	0.409	0.304	0.091
Subject-guided embedding-8D	0.339	0.341	0.979	**0.995**
Subject-guided embedding-16D	0.327	0.340	0.980	**0.995**
Subject-guided embedding-32D	0.343	0.331	0.978	**0.996**

the identity of a person. Moreover, in the self-supervised learning case where no labels are provided, the accuracy of identifying a subject is still 96.2%. Recall that the input feature to generate the embedding is only EEG collected within 0.5 s, which is quite a short time. The feasibility of input features and the close-to-perfect classification accuracy indicate the promising future of our method be to applied to identify persons in various HCI scenarios.

The accuracy of detecting emotions is low, probably because the emotional changes are rarely reflected in temporal voltage features. Including more EEG features in the inputs may improve emotion detection performance.

4 Conclusion and Future Work

In this paper, we propose to use contrastive learning to generate low-dimension EEG latent embeddings that are consistent and identifiable. The contrastive learning encoder can be trained in either the supervised or self-supervised manner and the encoder trained in both manners can have very high decoding accuracy. This indicates the potential of using the EEG latent embeddings for various downstream tasks. Excitingly, we can use the EEG latent embeddings to identify different persons at close-to-perfect accuracy of 99.6% with EEG input data with only a 0.5-s window. Our method can be promisingly used in emerging modern HCI devices and applications, e.g., automatically connecting people to their roles in video games via EEG-capable AR devices.

In the future, we will try to integrate our EEG latent embedding method into industrial HCI solutions for entertainment and health management. To achieve this, we need to improve our method for a wider range of downstream applications, which may require exploring more informative EEG features as inputs. We will also verify it with various datasets and in EEG devices of different specifications.

Acknowledgment. The work was supported in part by the National Natural Science Foundation of China (under grant 12102267) and the Shenzhen Sustainable Development Special Project (under grant KCXFZ20201221173411032).

References

1. Alarcao, S.M., Fonseca, M.J.: Emotions recognition using EEG signals: a survey. IEEE Trans. Affect. Comput. **10**(3), 374–393 (2017)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Adv. Neural. Inf. Process. Syst. **33**, 9912–9924 (2020)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning, pp. 1597–1607. PMLR (2020)
4. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process. Syst. **27**, 1–9 (2014)
5. Duncker, L., Bohner, G., Boussard, J., Sahani, M.: Learning interpretable continuous-time models of latent stochastic dynamical systems. In: International Conference on Machine Learning, pp. 1726–1734. PMLR (2019)
6. Duncker, L., Sahani, M.: Temporal alignment and latent gaussian process factor inference in population spike trains. Adv. Neural Inf. Process. Syst. **31**, 1–11 (2018)
7. Gao, Y., Archer, E.W., Paninski, L., Cunningham, J.P.: Linear dynamical neural population models through nonlinear embeddings. Adv. Neural Inf. Process. Syst. **29** (2016)
8. Hyvarinen, A., Sasaki, H., Turner, R.: Nonlinear ICA using auxiliary variables and generalized contrastive learning. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868. PMLR (2019)
9. Jazayeri, M., Ostojic, S.: Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. Curr. Opin. Neurobiol. **70**, 113–120 (2021)
10. Lin, Y.P., Yang, Y.H., Jung, T.P.: Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening. Front. Neurosci. **8**, 94 (2014)
11. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
12. Quitadamo, L.R., et al.: Support vector machines to detect physiological patterns for EEG and EMG-based human-computer interaction: a review. J. Neural Eng. **14**(1), 011001 (2017)
13. Rossini, P.M., et al.: Early diagnosis of Alzheimer's disease: the role of biomarkers including advanced EEG signal analysis: report from the IFCN-sponsored panel of experts. Clin. Neurophysiol. **131**(6), 1287–1310 (2020)
14. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
15. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 1–9 (2023)
16. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)

17. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)
18. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Trans. Auton. Ment. Dev. **7**(3), 162–175 (2015)
19. Zhou, D., Wei, X.X.: Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. Adv. Neural. Inf. Process. Syst. **33**, 7234–7247 (2020)