Identifiable EEG Embeddings by Contrastive Learning from Differential Entropy Features

Zhen Zhang[1,2,3], Feng Liang[1,2(✉)], 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. Encoding EEG data into low-dimension latent embeddings greatly facilitates data analysis and interpretation in neuroscience studies, clinical diagnosis, and human-computer interaction. But generating informative and identifiable latent embeddings that are representative of the origin EEG is not an easy mission. Contrastive learning has the potential to utilize large amounts of unlabelled EEG data and extract informative and identifiable latent embeddings for a wide range of downstream tasks. We explore the feasibility of applying the contrastive learning method to train the EEG latent encoder from the feature of differential entropy of short-time window frequency domain signals. The encoder minimizes the noise-contrastive estimation loss by comparing the embeddings with positive and negative embedding samples, where the distinction of samples is guided by time nearness information or task-specific labels. We test encoders with different output dimensions and the outcome latent embeddings can be identifiable via visualization of a few dimensions. The decoding result also shows that the embeddings preserve information about the original EEG features and can be potentially used for a wide range of downstream tasks. The source code is available at: https://www.github.com/liangfengsid/deContrastiveLearning.

Keywords: EEG · Contrastive learning · Latent embedding

1 Introduction

Electroencephalogram (EEG) are electrical signals on the scalp collected by a set of electrodes and has been widely applied in neuroscience research [15], clinical diagnosis [16], and behavior and affection analysis [11,12]. To relate EEG

V. C. M. Leung et al. (Eds.): Qshine 2023, LNICST 573, pp. 227–235, 2024.
https://doi.org/10.1007/978-3-031-65126-7_21

with specific properties of interest, much work has been done on extracting various EEG features and using different statistical or machine-learning models to retrieve useful information about behavior or health status.

Using proper features or latent embeddings of EEG is critical for EEG analysis. EEG signals usually have large sizes and come with significant noises. Most existing work uses frequency domain signals as features [13,19]. For example, it has been proved that differential entropy (DE) of different bands [3,21] incorporates useful emotional information. The latent approach [4,5] extracts invariant and identifiable latent embeddings of EEG, which can significantly reduce the representation size and extract useful information out of noises. Most work [6,17] uses supervised learning models to get latent embeddings of EEG for tasks with specific outcome labels. But much EEG data such as clinical EEG are recorded without labels, where supervised learning cannot be applied. Methods to generate general EEG latent embeddings that are independent of specific tasks can benefit the application of EEG to a wide range of downstream tasks. Recently, some studies [8,14,22] have worked on self-supervised learning methods and yielded promising results. Cebra [18] indicates that contrastive learning [1,2,10,20] has a great potential to extract invariant and identifiable EEG latent embeddings, which motivates us to explore the feasibility of extracting the general EEG latent embeddings for downstream tasks.

We apply contrastive learning, a powerful self-supervised learning algorithm, to transform DE features of EEG into lower dimensional latent embeddings for downstream tasks. We retrieve the DE of frequency power spectrum density and train a deep neural network by contrastive learning which minimizes the noise-contrastive estimation (NCE) [8] loss between generated latent embeddings of samples in a batch of training data. We use either the (self-supervised) implicit time information or (supervised) specific labels to identify positive and negative samples in the NCE loss. The first case can train the model in scenarios without labels, while the learning in the second case is guided by labels and can generate latent embeddings tuned for the specific downstream task. We explore the visual representation of the latent embeddings of different output dimensions generated by encoders guided by different information and find that the embeddings can be identifiable intuitively. We also decode the embeddings in different tasks to investigate the potential to apply the embeddings to a wide range of EEG applications.

2 Method

2.1 Dataset

We use the SEED [21] dataset, which is 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 related to 3 emotional labels. The signals are collected by 62 electrodes, downsampled to 200 Hz, and filtered to bandpass frequency from 0 to 75 Hz. With data divided by movie clips, we use 90% of the data for training both the encoder and the decoder, and the remaining 10% for testing.

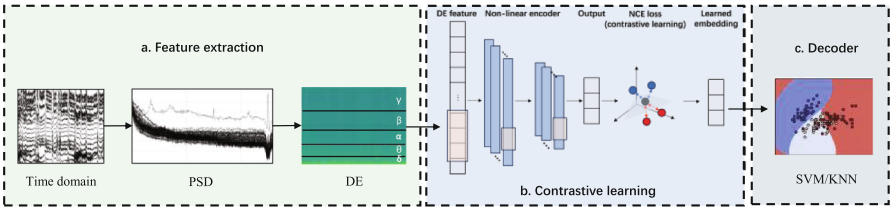

Fig. 1. The procedure of encoding EEG DE into embeddings by contrastive learning and decoding the learned embeddings

2.2 Model

The whole procedure of the model is depicted in Fig. 1. It composes three steps: the DE feature extraction, which generates DE of the frequency domain data from the origin time domain representation; the contrastive learning encoder, which encodes the DE features into latent embeddings by contrastive learning; and the decoder, which decodes the latent embedding to labels of interest.

DE Feature Extraction. DE [3] has the ability to discriminate signals between high and low frequency energy. We first transform the time domain signals to the frequency domain power in non-overlapped short-time Hanning windows and then follow a similar process to [21] to extract the DE of different frequency bands in each electrode channel (Fig. 1.a). The difference is, instead of using the magnitude spectrum as the input, we use the power spectrum density (PSD), which is recognized better than the magnitude spectrum for analyzing random vibration signals as its value is independent of frequency.

If we assume the PSD within a specific frequency band in the electrode channel i, represented by X_i, follows Gaussian distribution, i.e., $X_i \sim \mathcal{N}(\mu, \sigma^2)$, the DE is calculated as

$$\begin{aligned} h(X_i) &= -\int_{X_i} f(x) log(f(x)) dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} log \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} dx \\ &= \frac{1}{2} log 2\pi e\sigma^2, \end{aligned}$$

where $f(x)$ is the probability density of $x \in X_i$. For each electrode channel, we divide the frequency into five bands (delta $\in [1, 4)$ Hz, theta $\in [4, 8)$ Hz, alpha $\in [8, 14)$ Hz, beta $\in [14, 31)$ Hz, and gamma $\in [31, 50)$ Hz), and calculate the DE for each frequency band, respectively. Therefore, for each short-time window, we extract 62×5 DE features, which is represented as $h(X)$.

Contrastive Learned Embeddings. As shown in Fig. 1.b, DE features are fed into a learnable encoder and the output is the EEG latent embedding. The

encoder is non-linear and is usually a convolutional neural network (CNN) or a deep neural network (DNN) that applies contrastive learning, which follows a similar procedure as in [18].

For the DE features h and g, where g is a positive or negative sample of h, let $p(h)$ be the probability density function of h, $p(g|h)$ and $q(g|h)$ be the probability density function of the positive and negative samples conditioned on h, respectively. After encoding h and g, $c(h)$ and $c(g)$ are their normalized latent embeddings, respectively. The similarity function between $c(h)$ and $c(g)$ is denoted as $\psi(h, g)$. The objective is to minimize the NCE loss, which is:

$$\mathop{\mathbb{E}}_{\substack{h\sim p(h), g_+\sim p(g|h)\\ g_1,g_2,\ldots,g_n\sim q(g|h)}} [-\psi(h, g_+) + log \sum_{i=1}^{n} e^{\psi(h,g_i)}].$$

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. It can be based on time nearness between h and g if no label is provided, where samples close to h in time are considered positive and those far from h in time are considered negative. We can also provide labels to guide the training so that samples with the same label as that of h are considered positive and others are negative. The label-guided approach is supervised contrastive learning. With the SEED dataset with emotion labels from different subjects, we learn different encoder models based on time, emotion labels, and subject labels, respectively, and compare their embedding performance.

As to the similarity function, we use the dot product of the normalized latent embeddings adjusted with a temperature parameter τ, i.e., $\psi(h, g) = c(h)^T c(g)/\tau$.

Embedding Decoding. In application, EEG embeddings can be decoded for classification and regression tasks, as shown in Fig. 1.c. We use K-nearest neighbors (KNN) and non-linear support vector machine (SVM) models to classify the EEG embeddings generated by different encoders into emotion and subject labels, respectively. The embedding decoder is trained separately from the embedding encoder, and the training embeddings for the decoder are generated by the well-trained encoder from training DE features.

3 Results

3.1 Contrastive Learning Convergence

We explore the convergence performance of the encoder by contrastive learning with different criteria for identifying the positive and negative samples. Figure 2 shows the NCE loss of training a 4-layer neural network using GELU activation functions to encode EEG to 16-dimension embeddings guided by time nearness (when no label is provided), emotion labels, and subject labels, respectively. The encoder is trained for 10,000 iterations with a minibatch size of 1024 and

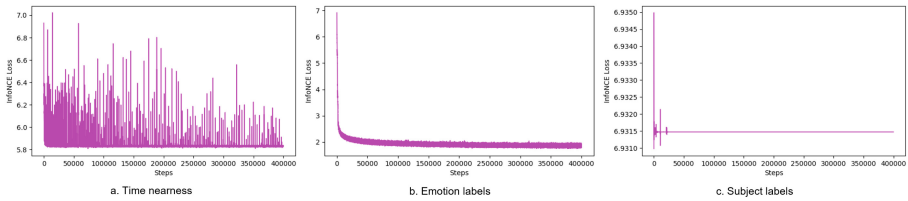

Fig. 2. NCE loss the encoder guided by different criteria

a learning rate of 0.001. The encoder does not converge in the time nearness and subject label cases. The NCE loss in the time nearness case jitters with a lower bound of about 5.8 limited by the time window length, while that in the subject label case is a straight line at the level of about 6.9315. The NCE loss in the emotion label case drops quickly in the first 1,000 iterations and gradually converges to about 1.9 after that.

3.2 Embedding Visualization

Visualizing the embeddings can help to interpret the encoding quality [9]. As shown in Fig. 3, we generate embeddings guided by different information (time or task related labels) with different output dimension sizes, 2, 8, and 16. The first two or three dimensions of the embeddings are drawn where values of the related information are indicated by colors, where embeddings related to the same information have the same color. When the embeddings are identifiable, the embedding points will cluster by color and have a clearer contour intuitively, where points of the same color are closer to each other and farther away from embedding points with different colors.

3.3 Decoding Accuracy

The results of top-1 accuracy of decoding embeddings of different dimensions from different encoder models to different labels are shown in Table 1. The highest classification accuracy for emotion labels is 0.507, which achieved by using the non-linear SVM method with 16-dimension embeddings generated by the emotion-guided encoder, and that for subject labels is 0.234, which achieved by using KNN with 8-dimension embeddings generated by the subject-guided encoder. The accuracy of decoding label-guided embeddings is higher than decoding embeddings that are guided by time nearness when no label is provided. The higher dimension of the embeddings tends to increase classification accuracy, except for decoding the subject-guided embeddings to subject labels. The embeddings generated by time nearness information can also be used for emotion classification, which indicates the potential application of contrastive learning to EEG to a wider range of downstream tasks. The classification accuracy for subject labels is much poorer than that for emotion labels. The possible

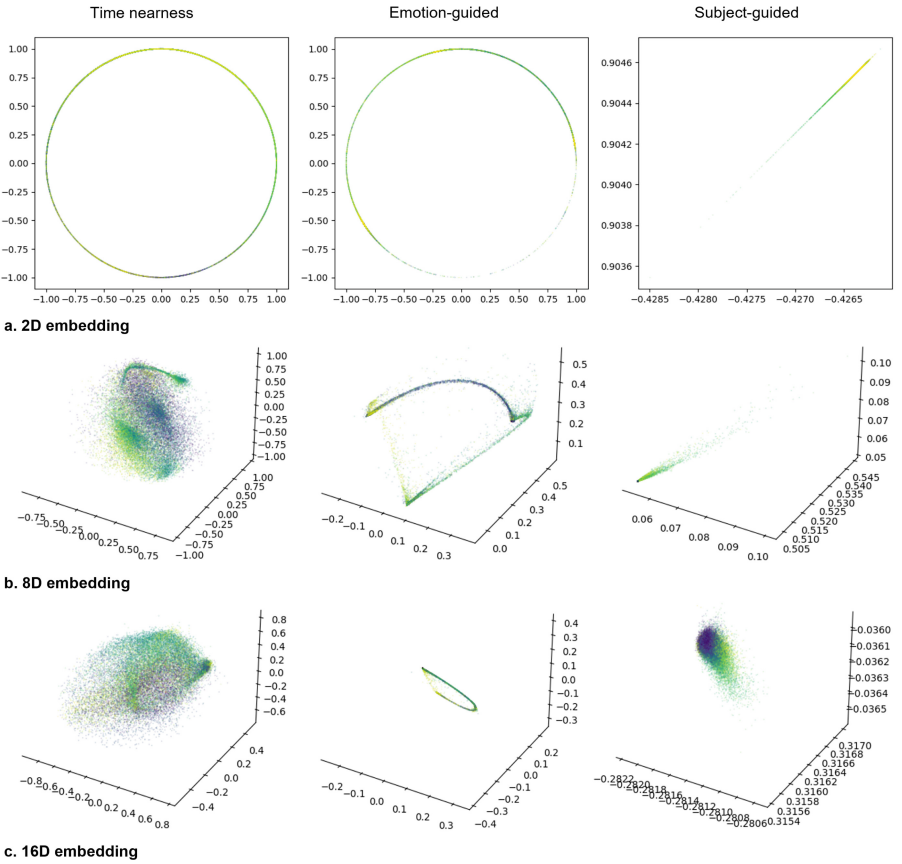

Fig. 3. The first two/three dimensions of learned embeddings of 2D, 8D, and 16D guided by time nearness, emotion labels, and subject labels, respectively.

reason is that the DE of different frequency bands varies more with people's emotional changes, but is more consistent across different people subjects. Decoding subject-guided embeddings to emotion labels also generates low accuracy, because the subject-guided embeddings have removed information about distinguishing emotions.

4 Discussion

About de Feature. The DE is proven a significant single feature which greatly reduces the feature dimension to 5 values for each channel in each short-time window, with some important frequency domain information remained and some noises filtered. But it also leaves out much useful information for encoding an invariant and identifiable embedding. Since the embedding encoder is supposed

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

Decoding Label	Emotion		Subject
	SVM	KNN	KNN
Time nearness embedding-2D	0.338	0.382	0.040
Time nearness embedding-8D	0.359	0.398	0.044
Time nearness embedding-16D	0.415	0.428	0.065
Emotion-guided embedding-2D	0.417	0.429	0.026
Emotion-guided embedding-8D	0.500	0.447	0.054
Emotion-guided embedding-16D	**0.507**	0.464	0.072
Subject-guided embedding-2D	0.352	0.341	0.097
Subject-guided embedding-8D	0.326	0.426	**0.234**
Subject-guided embedding-16D	0.369	0.383	0.078

to extract low-dimension latent features where EEGs with similar characteristics should have similar embeddings close in distance, the purpose of the DE extraction and the contrastive learning encoder is somewhat overlapped. Besides, as EEG signals tend to vibrate in a short time and exhibit more distinguishable characteristics in a longer observation, the features extracted from short-time windows only fluctuates and may not be representative of specific properties. More input features besides DE, including statistical features about frequency domain and time domain signals and asymmetric features between electrode channels [7,11], can hopefully improve the embedding performance.

About Encoder Model. The encoder we use in this paper is a four-layer DNN. We also tested with a similar complexity CNN, alternatively. Both the encoding convergence and the visualization of the outcome embeddings are similar and the later decoding accuracy is slightly lower. We also used deeper neural networks (up to 16 1-D convolutional layers). The encoding convergence and decoding accuracy do not improve either. The reason is that the dimension of the DE feature, 310, is not large and a very deep network is not necessary. When we add more features for the encoder input, a more complex neural network may improve the embedding quality, which will be left to our future work.

About Embedding Dimension. For time-nearness-guided and emotion-guided embedding, the higher the dimension, the higher the top-1 classification accuracy for emotion labels. It indicates that high-dimension embeddings have the ability to include more useful information than low-dimension ones. But it is not the same case with subject-guided embeddings. Lower-dimension embeddings may lack representation ability, while higher-dimension embeddings may involve more noise than useful information for subject classification.

5 Conclusion

In this paper, we explore encoding EEG into identifiable low-dimension latent embeddings from differential entropy powers by self-supervised contrastive learning. The latent embedding can be an informative representation uesd for downstream tasks. Using contrastive learning to extract latent embedding for EEG data is an interesting and promising topic and still needs a lot of studies. In the future, we will explore more traditional EEG features or even the raw signals for encoding EEG embeddings with contrastive learning and other self-supervised alternatives. We aim to find the algorithm to generate invariant and identifiable EEG embeddings for general tasks, and explore a wider application of EEG in the fields of neural studies, clinical screening and diagnosis, and human-computer interaction.

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. 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)
2. 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)
3. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for EEG-based emotion classification. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 81–84. IEEE (2013)
4. 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)
5. 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)
6. 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)
7. Hinrikus, H., et al.: Electroencephalographic spectral asymmetry index for detection of depression. Med. Biol. Eng. Comput. **47**, 1291–1299 (2009)
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. Khosla, P., et al.: Supervised contrastive learning. Adv. Neural. Inf. Process. Syst. **33**, 18661–18673 (2020)

11. Li, Y., et al.: A novel bi-hemispheric discrepancy model for EEG emotion recognition. IEEE Trans. Cogn. Dev. Syst. **13**(2), 354–367 (2020)
12. Lin, Y.P., et al.: EEG-based emotion recognition in music listening. IEEE Trans. Biomed. Eng. **57**(7), 1798–1806 (2010)
13. 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)
14. Pandarinath, C., et al.: Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods **15**(10), 805–815 (2018)
15. Pang, J.C., et al.: Geometric constraints on human brain function. Nature **618**, 566–574 (2023)
16. 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)
17. Sadtler, P.T., et al.: Neural constraints on learning. Nature **512**(7515), 423–426 (2014)
18. Schneider, S., Lee, J.H., Mathis, M.W.: Learnable latent embeddings for joint behavioural and neural analysis. Nature **617**, 360–368 (2023)
19. Wang, X.W., Nie, D., Lu, B.L.: Emotional state classification from EEG data using machine learning approach. Neurocomputing **129**, 94–106 (2014)
20. 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)
21. 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)
22. 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)