MoGE: Mixture of Graph Experts for Cross-subject Emotion Recognition via Decomposing EEG

Xuan-Hao Liu, Wei-Bang Jiang, Wei-Long Zheng✉, and Bao-Liang Lu✉ *Fellow, IEEE*
Department of Computer Science and Engineering, Shanghai Jiao Tong University
800 Dongchuan Rd., Shanghai 200240, People's Republic of China.
{haogram_sjtu, 935963004, weilong, bllu}@sjtu.edu.cn

***Abstract*—Decoding emotions of previously unseen subjects from electroencephalography (EEG) signals is challenging due to the inter-subject variability. Domain Generalization (DG) methods aim to mitigate the domain shift among different subjects. Once trained, a DG model can be directly deployed on new subjects without any calibration phase. While existing DG studies on cross-subject emotion recognition mainly focus on the design of loss function for domain alignment or regularization, we introduce Sparse Mixture of Graph Experts (MoGE) model to explore DG issues from a new perspective, i.e. the design of the neural architecture. In the MoGE model, routers allocate each EEG channel to a specialized expert, thereby facilitating the decomposition of the intricate brain into distinct functional areas. Extensive experiments on three public datasets demonstrate that compared to other DG methods, our MoGE model trained with empirical risk minimization (ERM) achieves the state-of-the-art (SOTA) accuracies, 88.0%, 74.3%, and 81.8% on SEED, SEED-IV, and SEED-V datasets, respectively. Our code is available at https://github.com/XuanhaoLiu/MoGE.**

***Index Terms*—EEG, emotion recognition , mixture-of-experts, graph neural networks, domain generalization**

I. Introduction

Recently, EEG-based emotion recognition (ER) has attracted great interest. Stable neural patterns of different basic emotions over time have been observed [1], which guaranteed the potential of EEG-based emotion recognition. However, the well-known individual differences in EEG data across subjects greatly limit the generalization of BCIs in real-world applications. The reasons for explaining the inter-subject variability range from external factors like electrode impedance and head shapes to internal factors like mental states, life experiences, and personalities.

To address the challenge mentioned above, transfer learning methods were introduced in the past few years. Domain Adaptation (DA) and Domain Generalization (DG) are two basic approaches in transfer learning. In contrast to DA methods, DG methods do not require any data from the target domain, which matches the most common real-world medical scenarios that training a model with multiple subjects' data (multi-source domains) to diagnose a new subject (target domain). Thus, DG methods make more sense than DA methods for practical applications.

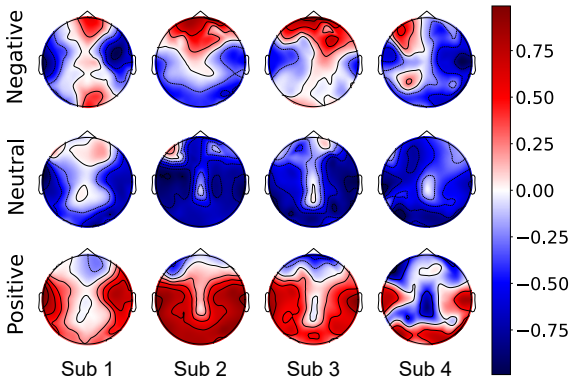


Fig. 1: The average topography of the EEG features in the gamma band of different subjects in the SEED dataset, calculated by averaging EEG data of each subject from 3 sessions.

There are several DG methods for EEG-based ER task, including domain alignment [2], ensemble learning [3], and contrastive learning [4], [5]. These previous works are mostly focused on designing specific loss functions. For example, the adversarial losses were applied for domain distribution alignment, and the contrastive losses were adopted for learning the meta attributes of EEG signals. Different from the special design of loss functions, the generalization ability of the neural networks' architectures has been largely unexplored.

Figure 1 depicts the average topography of differential entropy (DE) features of different subjects. DE features are related to the energy levels of brain activities. Although neural patterns exhibit variability among individuals, certain energy activations in specific brain regions can be discerned, e.g., for the positive emotions, the frontal areas are low activated while the temporal areas have high energy.

Based on the observation of the various atrributes of different brain areas, several studies found learning region-wise representations of different brain areas [5] helps generalization tasks. However, these methods partitioned EEG in a constant approach, lacking flexibility and adaptability to specific tasks.

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), Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University (No. 21TQ1400203) and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

Wei-Long Zheng and Bao-Liang Lu are co-corresponding authors.

Hence, we propose a novel Sparse Mixture of Graph Experts (MoGE), which learns to distribute each EEG channel to particular experts through sparse mixture-of-experts [6] on a Graph-Transformer model. Extensive experiments demonstrate that MoGE trained with ERM outperforms previous DG-based transfer learning methods.

In summary, the contributions of this paper are as follows:

- **New perspective:** Different from the previous methods for EEG-based ER which focused on aligning domains with well-designed loss function, we investigate the importance of decomposing brain signals and the architecture of networks from theoretical perspectives.
- **New model:** According to the theoretical derivation, we propose a novel Sparse Mixture of Graph Experts (MoGE) model which learns transferable features by decomposing the brain areas into several simple parts.
- **Excellent performance:** Extensive experiments on three public datasets demonstrate the outstanding performance of our MoGE model. Ablation studies are conducted to verify the effectiveness of the adjacency matrix and the EEG channels router in the MoGE model.

II. Related Work

1) Domain Generalization: The inter-subject variability of EEG signals has hindered the development and promotion of BCIs for a long period of time. Previous DG methods predominantly concentrated on designing specific loss functions. The domain adversarial losses are adopted to promote networks align with the invariant correlations [2]. Contrastive learning enables networks to learn the difference among features, which is applied in both a graph-based multi-task self-supervised learning (GMSS) [5] and a prototype contrastive domain generalization (PCDG) [4].

2) Transformer: Transformers have achieved remarkable success across various domains, including natural language processing [7] and computer vision [8]. Learning EEG features through Transformers facilitates emotion recognition [9].

3) Mixture-of-Experts: The Mixture-of-Experts (MoE) concept originated from a straightforward insight: breaking down a complex task into smaller queries and coordinating multiple experts to address them [10]. By a sparsely-gated which dispatch each token to a limited number of experts, Sparse MoE was firstly introduced to neural networks in 2017 [6]. However, MoE was largely unexplored in the domain of EEG. Although previous works using MoE to detect seizure [11] and emotions [12], the experts in their model are still Feed-forward Networks (FFN), which may not be well-suited for capturing spatial information crucial for EEG processing.

4) Graph Neural Network: Spatial convolution aggregates data of the nodes and their neighbors of the graph, for example, GMSS adopts a fixed graph of EEG channels to extract spatial information of EEG signals [5]. Meanwhile, many studies argue that learnable graphs are more suitable for EEG data. Jia *et.al.* proposed multi-view spatial-temporal graph convolutional networks (MSTGCN) for sleep stage classification [13]. An elastic graph transformer (EmoGT) was designed by replacing the FFN in Transformer with GCN for EEG-based ER tasks [14]. Excellent performances were acquired by these two works, which both learn non-fixed graphs from EEG data, suggesting that adaptive GNNs benefit modeling the spatial graph of EEG.

III. Neural Architecture for Cross-subject ER

A. Distribution Shift

The domain distribution shift, i.e. the inter-subject variability of EEG signals is caused by not only external factors but mostly internal factors about the physiological activities that vary among subjects [15]. However, the invariant neural attributes do exist among subjects [1]. From Figure 1, some stable neural patterns can be observed, e.g. for positive emotion, the temporal areas exhibit more activation than other emotions while the energy of the prefrontal area is significantly low, for negative emotion, the prefrontal area is highly activated.

According to the distribution shift analysis [16], we assume the EEG data of emotions from different domains are generated by a latent variable model. A joint distribution p of EEG data $\boldsymbol{x}$ can be factorised into corresponding attributes $a^1, a^2, \ldots, a^K$ (denoted as $a^{1:K}$) with $a^k \in \mathbb{A}^k$, where $\mathbb{A}^k$ is a finite set. One of these K attributes is correlated to emotions, denoted as $a^{emo} \in \mathbb{A}^{emo}$, which is significant for ER tasks. For example, EEG from SEED datasets has the $\mathbb{A}^{emo} = \{\text{Positive, Neutral, Negative}\}$. Some of the attributes are correlated to different brain areas, e.g. assuming that a^1 and a^2 are correlated to the energy of the temporal and prefrontal areas. Then for $\boldsymbol{x}_{pos}$ with a^{emo} = Positive, the attributes a^1 has high energy while a^2 is low activated.

Minimizing the risk $R(f) = \mathbb{E}_{(\boldsymbol{x},a^{emo})\sim p}[\mathcal{L}(a^{emo}, f(\boldsymbol{x}))]$ is the objective of classifier f, where $\mathcal{L}$ is the loss function. However, the exact finite sets $\mathbb{A}^{1:K}$ except $\mathbb{A}^{emo}$ are unknown in the real world. As a result, all classification models are trained by minimizing empirical risk $\hat{R}$ based on a finite set of EEG data $\boldsymbol{x}$ of size n_t:

$$\hat{R}(f;p) = \frac{1}{n_t} \sum_{(a_i^{emo},\boldsymbol{x}_i)\sim p} \mathcal{L}(a_i^{emo}, f(\boldsymbol{x}_i)). \tag{1}$$

The generation process of EEG data $\boldsymbol{x}$ is:

$$\boldsymbol{z} \sim p(\boldsymbol{z}), \quad a^i \sim p(a^i|\boldsymbol{z}), \quad \boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{z}), \tag{2}$$

where $\boldsymbol{z}$ denotes latent factor. EEG data $\boldsymbol{x}$ can be refactorized and by using the Bayes rule, we can derive:

$$\begin{aligned} p(a^{1:K}, \boldsymbol{x}) &= p(a^{1:K}) \int p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{x}|a^{1:K}) d\boldsymbol{z} \\ &= p(a^{1:K}) p(\boldsymbol{x}|a^{1:K}). \end{aligned} \tag{3}$$

The domain distribution shift occurs when these K attributes $a^{1:K}$ have different marginal distributions among different domains. Thus, we have $p_{sub_i}(a^{1:K}) \neq p_{sub_j}(a^{1:K})$, where p_{sub_i} and p_{sub_j} are the probability density functions of two subjects and share the same conditional generative process described in Formula 3.

B. Importance of Neural Architecture

Let $\varepsilon_s, D_s, \varepsilon_t$, and D_t denote the source dataset, source distribution, target dataset, and target distribution, respectively. By the algorithm alignment analysis [17], let $\mathcal{N} = \{\mathcal{N}_i\}_{i=1}^n$ denotes a neural network with n modules, which is trained to learn the target function $F = g(\boldsymbol{x})$. However, it is relatively hard to align $\mathcal{N}$ with function g. Assuming that the target function can be decomposed into n subfunctions $g_1, g_2, \ldots, g_n$. Replacing the alignment of $\mathcal{N}$ and g with the alignment of $\{\mathcal{N}_i\}_{i=1}^n$ and $\{g_i\}_{i=1}^n$ respectively is a feasible method:

$$\begin{aligned} &Alignment(\mathcal{N}, g, \epsilon, \delta) = n \cdot \max_i \mathcal{M}(\mathcal{N}_i, g_i, \epsilon, \delta) \\ &= n \cdot \max_i \{\mathbb{P}_{\boldsymbol{x}\sim D}[\|\mathcal{N}_i(\boldsymbol{x}) - g_i(\boldsymbol{x})\| < \epsilon] > 1 - \delta\}. \end{aligned} \tag{4}$$

Based on Equation 4, if we could find a neural architecture that can be decomposed and trained separately to align the subfunctions of g, the challenging task of learning g can be simplified. Hence, separating the multi-channel EEG into different brain areas, e.g. temporal lobe area and parietal lobe area, helps models learn each lobe's attributes respectively.

Among these K attributes $a^{1:K}$ of EEG data $\boldsymbol{x}$, there are attributes only correlated to emotions but not correlated to subjects (domains), which is called invariant correlation. Assuming there exists a function g_{inv} that for $\boldsymbol{x}_1 \in \varepsilon_s$, we have $g_{inv}(\boldsymbol{x}_1) = y$. The invariant correlation has:

$$\forall\, \boldsymbol{x}_2 \in \varepsilon_t, \quad \mathbb{P}_{\boldsymbol{x}_2\sim D_t}[\|g_{inv}(\boldsymbol{x}_2) - y\| < \epsilon] > 1 - \delta, \tag{5}$$

where $\mathbb{P}$, ϵ and δ is the probability, precision, and failure probability. The spurious correlation is the attributes correlated to subjects, i.e. inter-subject variability. Let ω be asymptotic notations, similar to Formula 5, we have:

$$\forall\, \boldsymbol{x}_2 \in \varepsilon_t, \quad \mathbb{P}_{\boldsymbol{x}_2\sim D_t}[\|g_{spu}(\boldsymbol{x}_2) - y\| > \omega(\epsilon)] > 1 - \delta. \tag{6}$$

According to the analysis above, a network $\mathcal{N}$ trained on the source dataset with ERM satisfies:

$$\begin{aligned} If \quad & Alignment(\mathcal{N}, g_{inv}, \epsilon, \delta) \le |\varepsilon_s|, \\ & \mathbb{P}_{\boldsymbol{x}\sim D_t}[\|\mathcal{N}(\boldsymbol{x}) - y\| \le O(\epsilon)] > 1 - O(\delta); \end{aligned} \tag{7}$$

$$\begin{aligned} If \quad & Alignment(\mathcal{N}, g_{spu}, \epsilon, \delta) \le |\varepsilon_s|, \\ & \mathbb{P}_{\boldsymbol{x}\sim D_t}[\|\mathcal{N}(\boldsymbol{x}) - y\| \ge w(\epsilon)] > 1 - O(\delta). \end{aligned} \tag{8}$$

Formulas 7 and 8 stand for the alignments of network $\mathcal{N}$ with invariant correlation and spurious correlation. When ϵ is sufficiently small, only one of these two formulas holds. If the network $\mathcal{N}$ only satisfies Formula 7, i.e. $\mathcal{N}$ aligns with the invariant correlation, $\mathcal{N}$ has outstanding generalization performance on target domains. In conclusion, we theoretically illustrate that neural networks is sufficient to achieve good generalization performance even solely by ERM training, which has been proven by previous works [18]–[21].

IV. Methods

A. Preliminaries

The overall architecture of MoGE is shown in Figure 2. The differential entropy (DE) features extracted from the raw EEG signals are fed into L MoGE blocks. The MoGE block is composed of multi-head attention (MHA) and mixture-of-experts (MoE), where each expert is a graph network.

The DE samples of each subject are divided by an overlapping window to generate $X'_{in} = \{X_1, X_2, \ldots, X_T\} \in \mathbb{R}^{T\times C\times D}$, where C, T, and D denote the number of EEG channels, the size of the overlapping window, and the dimension of extracted features, respectively. Before feeding the X'_{in} into the MoGE model, the D-dimension tensors of each channel are embedded into d dimension, a class token is attached at the beginning of X'_{in}, and learnable position embeddings with random initiation are added to each token. Finally, the $X_{in} \in \mathbb{R}^{(T+1)\times C\times d}$ is fed into the MoGE model. Only the class token is used for classification by a linear layer.

B. Mixture of Graph Experts Block

The Mixture of Graph Experts (MoGE) Block is the most essential part of the proposed model. We replace the FFN in the Transformer layer by MoE and each expert is a Spatial-Temporal Graph Convolution Network (GCN) [6]. The output of the MoGE block is:

$$f_{\textbf{MoGE}}(X_{in}) = f_{\textbf{MoE}}(f_{\textbf{MHA}}(X_{in})), \tag{9}$$

where $f_{\textbf{MHA}}$ is the MHA layer and $f_{\textbf{MoE}}$ is the MoE layer. Denoting the $f_{\textbf{MHA}}(X_{in})$ as $\boldsymbol{x}$, we have:

$$f_{\textbf{MoGE}}(\boldsymbol{x}) = \sum_{i=1}^{N} G(\boldsymbol{x})_i \cdot E_i(\boldsymbol{x}), \tag{10}$$

where $G()$ is the gate function, $E_i()$ is the i-th experts, and N is the number of experts. The gate function $G()$ decides which EEG channel is processed by which expert, and is composed of a linear layer with softmax to calculate the probability of each expert:

$$G(\boldsymbol{x}) = \mathrm{TOP}_k(\ \mathrm{Softmax}(W_g\boldsymbol{x})\ ), \tag{11}$$

where $W_g \in \mathbb{R}^{d\times N}$ is the learnable parameter for the gate. Each EEG channel is allocated to the k experts who have the largest k probability and k is a hyperparameter. Specifically, the input tensor X_{Ei} of each expert has the same shape of X_{in}, but only the EEG channels sent to Experts i have values while the other channels are set to zero.

As shown in the bottom right of Figure 2, if we divide the brain area into several critical parts, e.g. frontal lobe area, occipital lobe area, and some special links that are not discovered by neuroscientists but can be learned by the networks, the complex EEG signals could be decomposed into several relatively simple parts and solved separately.

C. Spatial-Temporal Graph Convolution Network

The effectiveness of GCN has been validated by many existing approaches in EEG-based classification tasks since GCN can build a graph of the electrodes on the scalp and capture their spatial relationships by aggregating the information based on a graph. For the MoGE block, instead of using simple FFNs as experts, we propose a spatial-temporal graph convolution network as a graph expert with a learnable graph

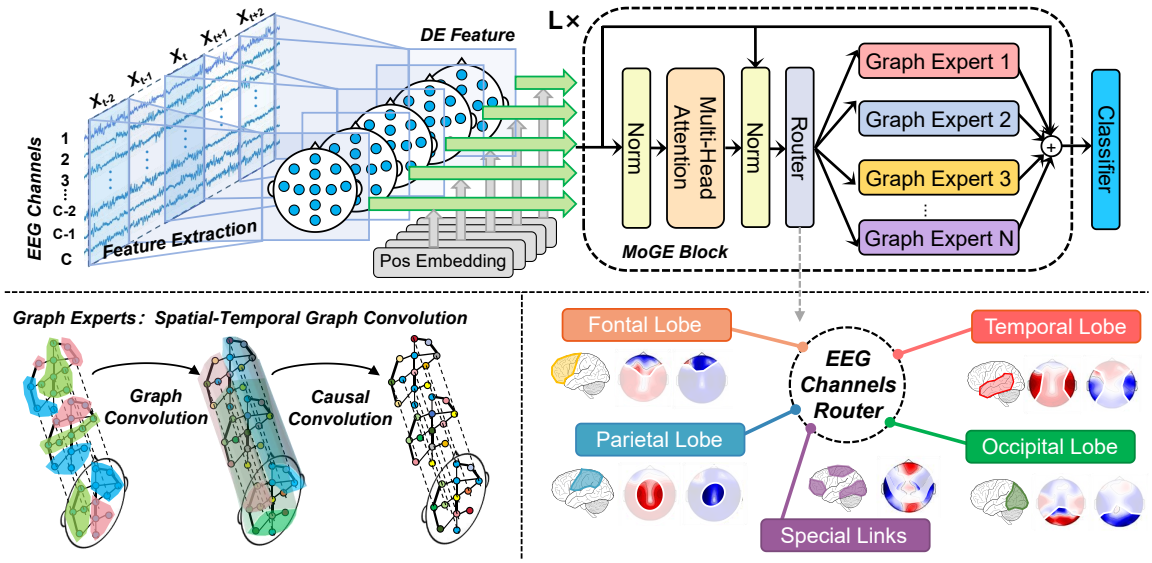


Fig. 2: The overall architecture of our MoGE model. DE features extracted from the raw EEG signals with position embedding are fed into L MoGE blocks. Each graph expert has graph convolution and causal convolution. The EEG channels router distributes each EEG channel to the graph experts, which focus on particular brain areas.

$G = \{V, E, \boldsymbol{A}\}$, where $\boldsymbol{A} \in \mathbb{R}^{C\times C}$ is the symmetric matrix adjacency matrix of graph G, V and E stand for the set of vertexes and edges, respectively.

The graph convolutional layer is written as follows [22]:

$$H_g^l = \hat{D}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{D}^{-\frac{1}{2}}H^l W^l, \tag{12}$$

where $\hat{\boldsymbol{A}} = \boldsymbol{A} + I$, I is identity matrix, and diagonal degree matrix $\hat{D}$ satisfies $\hat{D}_{ii} = \sum_j \hat{\boldsymbol{A}}_{ij}$. $H^l \in \mathbb{R}^{(T+1)\times C\times d}$ stands for the input tensor of the l-th layer, $W^l \in \mathbb{R}^{d\times d}$ is the weight matrix, and $H_g^l \in \mathbb{R}^{(T+1)\times C\times d}$ represents the output tensor. Hence, $H^0 = X_{in}$. For all MoGE blocks, we use the same shared adjacency matrix $\boldsymbol{A}$. The causal convolutional layer is adopted for extracting temporal information from each EEG channel, whose outputs of time t only depend on the inputs from time t or earlier.

V. Experiments

A. Datasets and Implementation Details

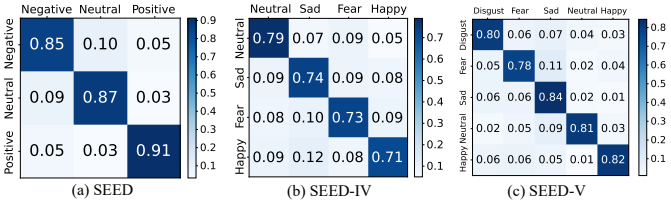
We validate the generalization ability of our MoGE model compared with multiple transfer learning methods on three publicly available datasets, the SEED [26], SEED-IV [27], and SEED-V [28] datasets by leave-one-subject-out (LOSO) cross-validation experiments. The size of overlapping window $T = 5$. The batch size is 64 and the dropout rate is 0.3. The number of MoGE blocks is $L = 4$. The hyperparameters of learning rate are tuned from {1e-4, 3e-3, 1e-3}, the number of the heads of MHA is ranging from 2 to 4, the embedding dimension d is tuned from {16, 32, 64}, and the number of experts is ranging from 3 to 6. We use the default Adam optimizer and cross-entropy loss function for training our MoGE model.

B. Cross-subject Results

To thoroughly validate the generation ability of the proposed MoGE model, we compare the MoGE trained with ERM with baseline methods and recent SOTA models or DG algorithms on three public datasets. The LOSO experimental results are presented in Table I, which demonstrate that the MoGE model trained with ERM achieves the best performance among DG-based models on the SEED and SEED-V datasets. The methods in the last five rows of Table I are trained with ERM, it can be observed that the Transformer-based [7], [25] and

	Methods	Results (Avg. ± Std.)		
		SEED	SEED-IV	SEED-V
	SVM	0.567 ± 0.163	0.515 ± 0.087	0.318 ± 0.068
DA	DAN [23]	0.838 ± 0.086	0.589 ± 0.081	0.679 ± 0.072
DA	wMADA-β [24]	0.893 ± 0.040	-	-
DA	PPDA [3]	0.867 ± 0.071	0.712 ± 0.062	-
DG-based	DG-DANN [2]	0.843 ± 0.083	0.698 ± 0.097	0.702 ± 0.117
DG-based	DResNet [2]	0.853 ± 0.080	0.721 ± 0.127	0.717 ± 0.128
DG-based	PPDA_NC [3]	0.854 ± 0.071	0.706 ± 0.064	-
DG-based	GMSS [5]	0.865 ± 0.062	0.735 ± 0.074	-
DG-based	PCDG [4]	0.873 ± 0.021	0.736 ± 0.051	-
DG-based	MAET [9]	0.868 ± 0.076	0.730 ± 0.110	0.783 ± 0.088
DG-based	EmoGT [14]	0.853 ± 0.081	0.722 ± 0.113	0.757 ± 0.122
DG-based	Transformer [7]	0.837 ± 0.121	0.684 ± 0.096	0.698 ± 0.083
DG-based	GMoE [25]	0.846 ± 0.093	0.712 ± 0.085	0.763 ± 0.092
DG-based	**MoGE (Ours)**	$\mathbf{0.880 \pm 0.045}$	$\mathbf{0.743 \pm 0.061}$	$\mathbf{0.818 \pm 0.100}$

TABLE I: Overall classification accuracies (Avg. ± Std.) of different transfer learning methods on three datasets.



(a) SEED (b) SEED-IV (c) SEED-V

Fig. 3: The confusion matrices of our MoGE model on three datasets.

	Variants	SEED	SEED-IV	SEED-V
Adjacency	Fully Connected $\boldsymbol{A}$	0.817	0.671	0.686
	Zero Connected $\boldsymbol{A}$	0.849	0.721	0.747
	Random $\boldsymbol{A}$	0.804	0.663	0.635
	GMSS's $\boldsymbol{A}$	0.852	0.726	0.801
Router	No Router	0.854	0.726	0.761
	Binary Router	0.864	0.732	0.769
	Cerebral Lobes Router	0.871	0.739	0.792
	GMSS's Router	0.827	0.699	0.725
	MoGE 2 experts	**0.880**	0.741	**0.818**

TABLE II: The average DG performance of different ablation variants on three datasets.

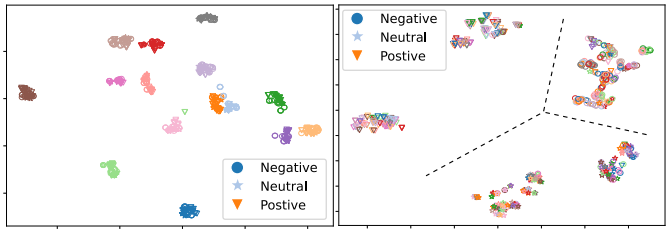
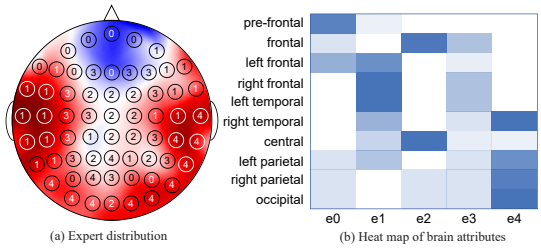


Fig. 4: Feature visualization of the SEED dataset. (Left) The color represents different subjects, and the shapes represent the emotional categories. (Right) The output of the last MoGE block.



(a) Expert distribution (b) Heat map of brain attributes

Fig. 5: (a) The expert selection of the MoGE's router decision of the first block with positive emotion. (b) The heat map of the expert selection. The y-axis represents 10 brain attributes according to GMSS. The x-axis corresponds to the selected expert id.

GNN-based models [29] already have decent generalization ability, and better results are acquired by combining GNN and Transformer [14]. Notably, the EmoGT has better generalization ability than DResNet [2] on the SEED and SEED-V datasets, while the latter method learns the invariant attributes by adversarial losses. However, these studies treat EEG signals as a unified entity, without considering the differences in the attributes across distinct brain regions. Thus, we leverage the mixture-of-experts to decompose complex EEG signals into simpler components and acquire 88.0%, 74.3%, and 81.8% on the three datasets by only ERM training. Moreover, the confusion matrices of our MoGE model are depicted in Figure 3, which represents that the MoGE has a very balanced ability to classify various emotions on three datasets. The balanced ability confirms the generalization ability of our MoGE model from another perspective, indicating our MoGE model aligns well with the invariant correlation of each emotion's EEG features even with no DG methods.

C. Ablation Study

Table II exhibits the DG performances of different ablation variants. We fully investigate two modules in MoGE: adjacency matrix and router.

1) Learnable Adjacency Matrix: We evaluate the effectiveness of the learnable adjacency matrix by replacing the adjacency matrix $\boldsymbol{A}$ with other unlearnable adjacency matrices, which are defined as follows:

- Fully Connected Adjacency Matrix: A matrix with all elements are 1.
- Zero Connected Adjacency Matrix: A matrix with all elements are 0.
- Random Adjacency Matrix: A randomly generated matrix whose elements are ranging from 0 and 1.
- GMSS's Adjacency Matrix: A matrix with the same structure in previous work GMSS [5].

We can see that the MoGE with learnable adjacency matrix is better than with the constant adjacency matrix, which demonstrates the generalization capability of elastic GNN.

2) EEG Channels Router: The most essential part of the MoGE model is the EEG channels router that distributes the normalized EEG features from different brain areas by their neural attributes to corresponding experts. The EEG channels router in our MoGE model is implemented by a linear learnable gate, as defined in equation 11. We verified the effectiveness of this learnable router by replacing it with some specially designed routers according to neuroscience. These specially designed routers are defined as follows:

- No Router: All EEG channels are fed into a single graph expert, i.e., normal Graph-Transformer.
- Binary Router: All EEG channels are divided into the left and right hemispheres of the brain based on the structural symmetry of the brain.
- Cerebral Lobes Router: The router distributes the EEG channels to experts corresponding to each cerebral lobe.
- GMSS's Router: The brain areas are divided in the same way as the previous work GMSS [5].
- MoGE 2 experts: The router of MoGE send EEG each channel to 2 experts, i.e. $k = 2$.

It can be seen that even without routering, the MoGE, which reduces to a GNN-Transformer, already has great generalization ability. Roughly distributing EEG channels to 2 experts enhances DG performance. By cerebral lobe-based router, the MoGE achieves a great boost in classification accuracies. However, a finer-grained GMSS's router results in a decrement, indicating that too many experts are harmful to the generalization ability of models. Moreover, we observe that sending each channels to more than one experts increases the computational complexity but does not improve the performance.

D. Visualization

Figure 4 demonstrates the distribution of the EEG signals before processing and the features extracted by MoGE. It can be seen the extracted features are clearly distinguishable.

To further understand the expert selections of the EEG signals, we collect the router's strategy of the first MoGE block and visualization a record of positive emotion. The top-1 selections of each EEG channel are shown in Figure 5 (a), and a heat map of the brain attributes each expert concentrates on is displayed in Figure 5 (b). The brain attributes are the same as the brain areas described in GMSS [5]. It can be observed that experts 0, 1, 2, and 4 are focusing most on the pre-frontal, temporal, central, and occipital areas, respectively. Expert 3 processes the scalp areas around the central areas, paying attention to a large range of brain attributes, which helps build the connection of each attribute. The visualization represents that different experts specialize in different brain regions, which decompose the complex EEG signals into several simple tasks and solve them separately.

VI. Conclusion

In this paper, we propose a Sparse Mixture of Graph Experts (MoGE) model which leverages the mixture-of-experts architecture for the decomposition of EEG signals. We analyze the domain distribution shift and the importance of neural architecture from the theoretical perspective. Experimental results demonstrate the excellent DG performance of our MoGE model trained by only ERM method, which are 88.0%, 74.3%, and 81.8% on SEED, SEED-IV, and SEED-V datasets, respectively. Even trained only by ERM, our MoGE model still outperforms other SOTA DG methods or other SOTA deep models, indicating the effectiveness of the neural architecture and the decomposition of multi-channel emotion EEG. that has been long ignored in the previous research on EEG tasks. We visualize the experts' selections and discover that the experts are able to specialize in specific brain regions.

References

[1] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2017.

[2] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of EEG signals with adversarial domain generalization," in *26th International Conference Neural Information Processing, ICONIP*. Springer, 2019, pp. 30–42.

[3] L.-M. Zhao, X. Yan, and B.-L. Lu, "Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[4] H. Cai and J. Pan, "Two-phase prototypical contrastive domain generalization for cross-subject EEG-based emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Y. Li, J. Chen, F. Li, B. Fu, H. Wu, Y. Ji, Y. Zhou, Y. Niu, G. Shi, and W. Zheng, "Gmss: Graph-based multi-task self-supervised learning for EEG emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.

[6] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2016.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[9] W.-B. Jiang, X.-H. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5975–5984.

[10] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.

[11] Z. Du, R. Peng, W. Liu, W. Li, and D. Wu, "Mixture of experts for EEG-based seizure subtype classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 4781–4789, 2023.

[12] L. Yang, D. Liu, Q. Zhang, S. Chao, P. Ni, Q. Wang, and H. Sun, "EEG emotion recognition via identity based multi-gate mixture-of-experts network," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2022, pp. 2498–2505.

[13] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, Y. Zhou, and H. L. Li-wei, "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[14] W.-B. Jiang, X. Yan, W.-L. Zheng, and B.-L. Lu, "Elastic graph transformer networks for EEG-based emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] W. Samek, F. C. Meinecke, and K.-R. Müller, "Transferring subspaces between subjects in brain–computer interfacing," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2289–2298, 2013.

[16] O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, K. Dvijotham, and T. Cemgil, "A fine-grained analysis on distribution shift," *arXiv preprint arXiv:2110.11328*, 2021.

[17] K. Xu, J. Li, M. Zhang, S. S. Du, K.-i. Kawarabayashi, and S. Jegelka, "What can neural networks reason about?" in *International Conference on Learning Representations*, 2019.

[18] Y. Chen, W. Huang, K. Zhou, Y. Bian, B. Han, and J. Cheng, "Understanding and improving feature learning for out-of-distribution generalization," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson, "On feature learning in the presence of spurious correlations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 516–38 532, 2022.

[20] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *The Eleventh International Conference on Learning Representations*, 2022.

[21] E. Rosenfeld, P. Ravikumar, and A. Risteski, "Domain-adjusted regression or: ERM may already learn features sufficient for out-of-distribution generalization," *arXiv preprint arXiv:2202.06856*, 2022.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[23] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu, "Cross-subject emotion recognition using deep adaptation networks," in *25th International Conference Neural Information Processing, ICONIP*. Springer, 2018, pp. 403–413.

[24] Y. Luo and B.-L. Lu, "Wasserstein-distance-based multi-source adversarial domain adaptation for emotion recognition and vigilance estimation," in *2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM*. IEEE, 2021, pp. 1424–1428.

[25] B. Li, Y. Shen, J. Yang, Y. Wang, J. Ren, T. Che, J. Zhang, and Z. Liu, "Sparse mixture-of-experts are domain generalizable learners," in *The Eleventh International Conference on Learning Representations*, 2022.

[26] 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.

[27] 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, 2018.

[28] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.

[29] R. Li, Y. Wang, and B.-L. Lu, "A multi-domain adaptive graph convolutional network for EEG-based emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5565–5573.