MULTI-LEVEL SPATIAL-TEMPORAL ADAPTATION NETWORK FOR MOTOR IMAGERY CLASSIFICATION

Wei Xu[⋆†‡], *Jing Wang*[⋆†‡*], *Ziyu Jia*[⋆†‡], *Zhiqing Hong*[⋆], *Yunze Li*[⋆†‡], *Youfang Lin*[⋆†‡]

[⋆] School of Computer and Information Technology, Beijing Jiaotong University, China
[†] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, China
[‡] CAAC Key Laboratory of Intelligent Passenger Service of Civil Aviation, Beijing, China

ABSTRACT

Electroencephalogram (EEG) signals for motor imagery (MI) are easily influenced by the environment and the state of the subject, which exhibit temporal and spatial variance. And this variance is more significant across subjects and sessions, which imposes limitations on the cross-domain MI tasks. To address this problem, we propose a Multi-level Spatial-Temporal Adaptation Network (MSTAN), extracting domain-invariant multi-level spatial-temporal features to overcome domain differences. First, stacked spatial-temporal graph convolution (STGCN) layers and an attention-based readout module are designed to extract spatial-temporal patterns of EEGs at multiple levels. An adaptation scheme is then introduced to narrow domain differences: 1) Individual graph parameters for the source and target domains are designed at each STGCN layer to capture the domain-specific brain region dynamic relationships; 2) The differences of spatial-temporal features between the source and target domain are reduced by minimizing the distribution distance. Experiments are conducted to evaluate the proposed method on a public dataset and the results show that our method achieves state-of-the-art performance in cross-domain motor imagery classification.

Index Terms— Motor Imagery, STGCN, Domain Adaptation

1. INTRODUCTION

Electroencephalogram (EEG) signals are widely used in the field of Brain-Computer Interface (BCI) research, which helps to understand human intentions and bridge the gap between machine and the human brain. Motor imagery (MI) based on EEG signals, as a prominent BCI system, has much promising application in rehabilitation for patients [1], control of peripheral devices [2], etc.

The core of this area, also the challenge, is to identify human intentions by decoding the neural activity triggered by imagery body movements [3] [4], often representing as dynamic spatial-temporal patterns in EEG signals. Some researchers designed CNNs or RNNs [5] [6] [7] to extract the spatial-temporal features of EEG signals based on the characterization capability of deep learning to improve MI classification performance. However, the features obtained by simple CNNs or RNNs do not apply well to complex EEGs. Therefore, Zhao et al. try to extract multiple temporal patterns via a multi-branch temporal CNN to improve the classification performance [8]. In a word, the extraction of discriminative features is essential for MI classification.

EEG signals are easily disturbed by the environment and influenced by the subject's state, hence the signals are varying across different subjects and sessions. It is challenging to build a motor imagery machine learning model which is optimal for different subjects and sessions [9]. The simple CNN or RNN based approach does not consider reducing the variation between domains, resulting in a limited performance. So, researchers focused on reducing the differences brought by different domains. Fine-tuning the CNN-based model [10] can transfer the model on the source domain to the target domain and achieve better classification performance. However, this approach requires enough data and labels of the target domain to calibrate the model on the source domain, which is challenging to implement in real application scenarios. For this reason, Zhao et al. propose to match the distribution offsets of source and target domain features by domain adaptation [1], which can reduce the feature differences across domains with the help of data from the target domain only. For EEG decoding, how to extract domain-invariant spatial-temporal features is the key issue to improve cross-domain MI classification performance.

To address the problems above, we propose a Multi-level Spatial-Temporal Adaptation Network (MSTAN) to capture more discriminative and domain-invariant spatial-temporal features to improve performance in cross-domain MI classification. We design a spatial-temporal graph convolution (STGCN) layer to extract the spatial-temporal features of EEG signals. The STGCN layer includes a graph learning module and a stgcn module. The graph learning module learns multiple connectivity relationships between different

*Corresponding author.

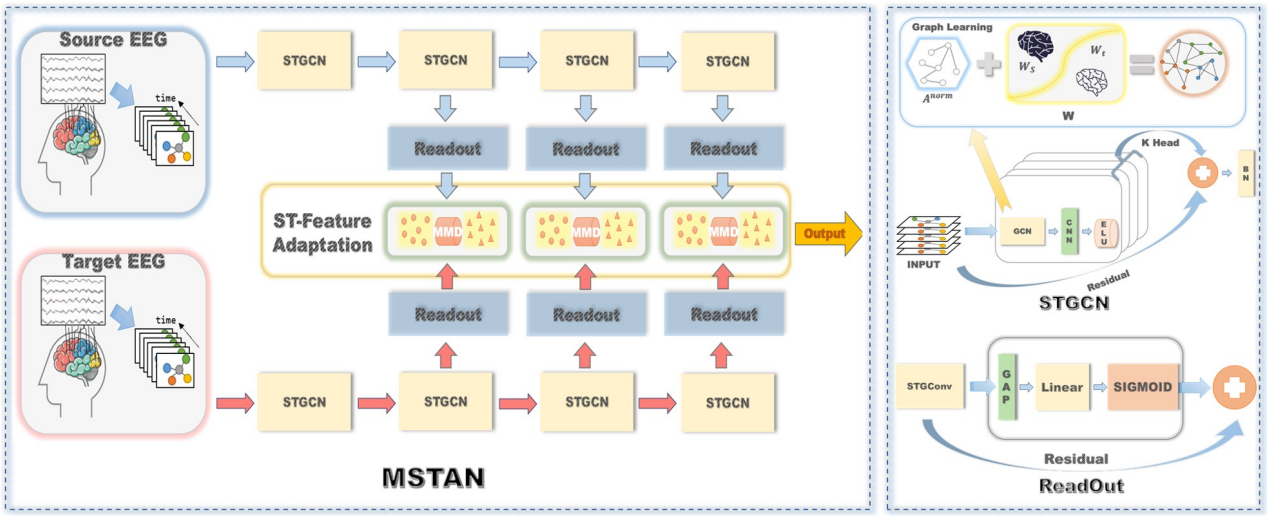

Fig. 1. The framework of the proposed MSTAN. The complete network consists of STGCN layers, Readout modules and ST-Feature Adaptation scheme. The specific details of STGCN and Readout are shown in the figure on the right. The ST-Feature Adaptation scheme is designed to narrow the domain difference at multi-level spatial-temporal features. The ELU indicates Exponential Linear Unit and the BN indicates BatchNorm.

brain regions, and the stgcn module extracts the spatial and temporal patterns. For the last several STGCN layers, we design a readout module that can extract more discriminative spatial-temporal features at multiple levels. Furthermore, to narrow the differences between the source and target domains, we propose an adaptation scheme to extract multi-level domain-invariant features. First, we design different graph parameters for the source and target domains respectively in the graph learning module to capture the specific brain region dynamic relationships. Second, we achieve inter-domain feature space adaptation by constraining the distribution distance between spatial-temporal features of different domains. In experiments, we evaluate the proposed method on the BCI-2a dataset [11] and the results show that our method achieves state-of-the-art performance.

2. METHODOLOGY

2.1. Overview

The framework of the proposed Multi-level Spatial-Temporal Adaptation Network (MSTAN) is shown in Figure 1. It includes the STGCN layer for extracting EEG spatial-temporal patterns, the Readout module for extracting multi-level features and the Spatial-Temporal Feature Adaptation scheme for capturing domain-invariant features.

2.2. Spatial-Temporal Graph Convolution Layer

As shown in Figure 1, we design a spatial-temporal graph convolution (STGCN) layer which combines GCN [12] and CNN [13] for extracting spatial-temporal features of EEG. For the l-th STGCN Layer, the input of EEG is represented as $\mathbf{H}_{l-1} \in \mathbb{R}^{C\times N\times T}$, where C is the channels of feature map (1 for first STGCN Layer), N is the number of EEG nodes and T represents the total number of sampling points.

2.2.1. *Graph learning module*

To model the complex functional regional relationships of the brain, we propose a graph learning module combined with the spatial location of EEG electrodes. We predefine an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ based on the spatial relationship of the electrodes. If node v_i and node v_j are not adjacent, then $A(i,j) = 0$, otherwise $A(i,j) = 1$. The normalized form of the adjacency matrix is $\mathbf{A}^{\text{norm}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, in which $\mathbf{D}$ is the degree matrix corresponding to the adjacency matrix.

Because of the complexity of the connectivity between functional brain regions, we use the idea of multi-head mechanism in order to ensure that the model learns a sufficient set of connectivity relationships. We combine K graph parameters with a normalized adjacency matrix to capture the complete dynamic brain functional area connectivity, which can be expressed as:

$$\mathbf{A}_i^{dynamic} = \mathbf{A}^{\text{norm}} + \mathbf{W}_i \odot \mathbf{A} \tag{1}$$

in which $\mathbf{W}_i \in \mathbb{R}^{N\times N}$ is the i-th graph parameter and $\mathbf{A}_i^{dynamic}$ represents the i-th dynamic connectivity relationships.

2.2.2. STGCN module

In STGCN module, we implement the extraction of spatial-temporal patterns by GCN and CNN. For the feature map $\mathbf{H}_{l-1}^t \in \mathbb{R}^{C\times N}$ at moment t, the k times graph convolution operations that can be performed are as follows:

$$\mathbf{Z}_l^t = \left[\mathbf{H}_{l-1}^t \mathbf{A}_0^{dynamic}, \ldots, \mathbf{H}_{l-1}^t \mathbf{A}_{k-1}^{dynamic}\right] \tag{2}$$

in which $\mathbf{Z}_l^t \in \mathbb{R}^{kC\times N}$, and $\mathbf{Z}_l \in \mathbb{R}^{kC\times N\times T}$ represents the output of the graph convolution.

After completing the graph convolution operation, to further extract the temporal pattern of EEG, we use f_l to denote the 1D temporal convolution of the l-th STGCN layer, and the final output can be expressed as:

$$\mathbf{H}_l = f_l(\mathbf{Z}_l) \tag{3}$$

in which, $\mathbf{H}_l \in \mathbb{R}^{C'\times N\times T}$ and C' is the channel of the output feature map.

2.3. Readout Module

The readout module we have designed is shown in Figure 1. Assuming that the output of the l-th STGCN is $\mathbf{H}_l \in \mathbb{R}^{C\times N\times T}$, node-wise Global Average Pooling (GAP) is employed for reducing computational costs, which is defined as:

$$\mathbf{O}_l = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_l^n \tag{4}$$

where $\mathbf{O}_l \in \mathbb{R}^{C\times T}$ is the output of the GAP, N is the number of EEG nodes, $\mathbf{H}_l^n \in \mathbb{R}^{C\times T}$ is the feature of the n-th node.

The importance of each time slice can be obtained through a linear layer and a Sigmoid activation function, which is defined as:

$$\alpha_l^t = \frac{1}{1+\mathrm{e}^{-(\mathbf{W}_l\mathbf{O}_l^t + b_l)}} \tag{5}$$

where $\alpha_l^t \in \mathbb{R}^1$ is the attention score of t-th time slice, $\mathbf{W}_l \in \mathbb{R}^{1\times C}$ and $b_l \in \mathbb{R}^1$ is the parameter of the linear layer.

After completing the above steps of generating the attention score, we use this score to complete the weighted summation over all time slices. Also, we use residual connection to ensure effective learning of the module. The last step of the attention mechanism can be expressed as :

$$\overline{\mathbf{H}}_l = \sum_{t=1}^{T}(\alpha_l^t + 1)\mathbf{H}_l^t \tag{6}$$

where $\overline{\mathbf{H}}_l \in \mathbb{R}^{C\times N}$ is the spatial-temporal feature extracted by the readout module.

2.4. Spatial-Temporal Feature Adaptation

We assume there are two different domains: the source domain $\mathcal{D}^s$ and the target domain $\mathcal{D}^t$. In particular, we denote the N_s labeled data from the source domain as $\{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, and the N_t samples from the target domain as $\left\{\left(x_j^t, y_j^t\right)\right\}_{j=1}^{N_t}$, where $x_i^s \sim \mathcal{D}^s, x_j^t \sim \mathcal{D}^t \in \mathbb{R}^{1\times N\times T}$.

2.4.1. Domain-specific Graph Learning

The graph structure is important for the STGCN layers. If the graph structure learned in the source domain is used directly on the target domain, it will introduce unnecessary spatial-temporal feature bias which limits the classification performance. We design specific graph learning parameters $\mathbf{W}_s$ and $\mathbf{W}_t$ for the source and target domains in the graph learning module of the spatial-temporal graph convolution layer. In the training phase, both the source domain data and the unlabelled target data are passed into the above model. To help the graph parameters of the target domain to be learned, we fuse the graph parameters of the two domains for modelling the source data by a factor γ, which can be expressed as:$\mathbf{W}_{s-t} = \gamma\mathbf{W}_s + (1-\gamma)\mathbf{W}_t$. In both training and testing phase, we model the target data with $\mathbf{W}_t$ only.

2.4.2. Feature Adaptation

In the training phase, the features of source and target domains are obtained after the l-th readout module which can be expressed as: $\overline{\mathbf{H}}_l^s, \overline{\mathbf{H}}_l^t$. To further reduce the spatial-temporal feature differences between the source and target domains, we introduce the maximum mean difference (MMD) [14] to measure the distribution distance of spatial-temporal features in the source and target domains, while optimizing by back propagation, which can be expressed as

$$\mathcal{L}_{\mathcal{M}}^l = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s}\phi\left(\overline{\mathbf{H}}_l^s\right) - \frac{1}{n_t}\sum_{j=1}^{n_t}\phi\left(\overline{\mathbf{H}}_l^t\right)\right\|_{\mathcal{H}} \tag{7}$$

where ϕ indicates a map from the original space to Hilbert space, n_s is the number of samples in the source domain and n_t is the number of samples in the target domain.

2.5. Loss function

With the above network, we can obtain multi-level spatial-temporal features $\overline{\mathbf{H}}_1^s, \overline{\mathbf{H}}_2^s, \overline{\mathbf{H}}_3^s$ (here are 3 levels). We apply a flatten operation to the multi-level spatial-temporal features, which are finally used for classification and the classification loss defined on the source domain can be expressed as:

$$\mathcal{L}_{\mathcal{C}} = -\sum_i y_i^s \log \hat{y}_i^s \tag{8}$$

where y_i^s is the i-th samples's true label of source domain and $\hat{y}_i^s$ is the probability of model prediction.

The complete loss function can be expressed as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{C}} + \sum_{l=1}^{L} \alpha_l \mathcal{L}_{\mathcal{M}}^{l} + \beta \mathcal{L}_{\theta} \tag{9}$$

where $\mathcal{L}_{\theta}$ is the l2 regularization loss, α and β is the coefficient of $\mathcal{L}_{\mathcal{M}}^{l}$ and $\mathcal{L}_{\theta}$ respectively and L is the number of levels.

Table 1. The mean accuracies (ACC) and kappa scores of different methods running on Bci-2a dataset. #ATD and #ATL is the abbreviation for the amount of target data and label used for model training.

Method	#ATD	#ATL	Bci-2a	
			acc(%)	kappa(%)
FBCSP	None	None	67.7	57.0
ShallowNet	None	None	72.9	63.9
ShallowNet-TL	Most	Most	76.8	69.2
TS-SEFFNET	None	None	74.7	66.3
MCNN	None	None	75.1	64.4
DARA	ALL	None	74.7	66.3
MSTAN w/o multi-level	ALL	None	77.8	70.4
MSTAN w/o DA	None	None	77.0	69.6
MSTAN	ALL	None	79.2	72.0

*'w/o' denotes without and 'DA' denotes 'Domain Adaptation.

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset

The BCICIV 2a [11] dataset consists of EEG data from 9 subjects. The BCI paradigm consists of four different Motor Imagery tasks. Each subject contains data from two sessions, and each session contains 72 trails for each task. In this dataset, we used a cross-session experimental protocol: data from two sessions of each subject are treated as one source domain and one target domain. We train the model for each of the nine subjects and end up with average test results on the target domain.

3.2. Experimental Results

To validate the proposed MSTAN model, we compare our model with state-of-the-art methods on BCI-2a, respectively, including FBCSP [15], ShallowNet [16], ShallowNet-TL [17], TS-SEFFNet [3], DRDA [1], MCNN [8]. To demonstrate the validity of each module of the model, we remove the multi-level readout module and the domain adaption process respectively.

The results of the motor imagery classification are shown in Table 1. The proposed MSTAN achieves state-of-the-art performance in cross-session experiments with the BCI-2a dataset. Traditional methods like SVM perform poorly due to the low signal-to-noise ratio and the variability of EEG. Deep learning models, such as ShallowNet and MCNN, outperform the traditional method by achieving 72.9% and 75.1% accuracy, respectively. However, these two methods can't overcome the differences across domains. The ShallowNet-TL uses some data from the target domain for fine-tuning to achieve transfer of the model from the source domain to the target domain with a classification accuracy of 76.8%. DARA uses a domain adaptive approach to reduce the domain variance and achieve higher performance. Ablation experiments on MSTAN demonstrate the effectiveness of extracting multi-level spatial-temporal features and narrowing the difference between domains. The ability of the proposed MSTAN to extract domain-invariant multi-level spatial-temporal features allows it to achieve a higher accuracy of 79.2%.

(a) MSTAN w/o DA (b) MSTAN with DA

Fig. 2. Feature visualization of MSTAN. (a) Feature visualisation of the MSTAN model without domain adaptation. (b) Feature visualisation of the complete MSTAN.

To demonstrate that our proposed method is effective in reducing inter-domain differences, we downscale the source and target domain features of the last layer of STGCN with t-SNE [18] and visualize them. Figure 2(a) shows the difference in features between the different domains without transfer: the source domain features are concentrated in the upper left corner while the target domain features are concentrated in the lower right corner. On the contrary, we can see from Figure 2(b) that the features of the source and target domains are mostly mixed together, proving that the transferred model can reduce the feature differences between the different domains.

4. CONCLUSION

In this paper, we propose a Multi-level Spatial-Temporal Adaptation Network (MSTAN) for cross-domain motor imagery classification. MSTAN based on multi-level STGCN layers and readout modules can extract more discriminative spatial-temporal features. In addition, the spatial-temporal feature adaptation scheme helps the model extract domain-invariant features. The experiments demonstrate that our method can reduce the domain differences of spatial-temporal features to a certain extent and achieve the improvement of classification accuracy.

5. REFERENCES

[1] He Zhao, Qingqing Zheng, Kai Ma, Huiqi Li, and Yefeng Zheng, "Deep representation-based domain adaptation for nonstationary eeg classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 535–545, 2020.

[2] Kai Ang and Cuntai Guan, "Brain–computer interface for neurorehabilitation of upper limb after stroke," *Proceedings of the IEEE*, vol. 103, pp. 944–953, 06 2015.

[3] Yang Li, Lianghui Guo, Yu Liu, Jingyu Liu, and Fangang Meng, "A temporal-spectral-based squeeze-and-excitation feature fusion network for motor imagery eeg decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1534–1545, 2021.

[4] Yu Zhang, Chang S Nam, Guoxu Zhou, Jing Jin, Xingyu Wang, and Andrzej Cichocki, "Temporally constrained sparse group spatial patterns for motor imagery bci," *IEEE transactions on cybernetics*, vol. 49, no. 9, pp. 3322–3332, 2018.

[5] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, pp. 056013, 2018.

[6] Ping Wang, Aimin Jiang, Xiaofeng Liu, Jing Shang, and Li Zhang, "Lstm-based eeg classification in motor imagery tasks," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 26, no. 11, pp. 2086–2095, 2018.

[7] Ziyu Jia, Youfang Lin, Jing Wang, Kaixin Yang, Tianhang Liu, and Xinwang Zhang, "Mmcnn: A multi-branch multi-scale convolutional neural network for motor imagery classification," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2020, pp. 736–751.

[8] Xinqiao Zhao, Hongmiao Zhang, Guilin Zhu, Fengxiang You, Shaolong Kuang, and Lining Sun, "A multi-branch 3d convolutional neural network for eeg-based motor imagery classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 10, pp. 2164–2177, 2019.

[9] Dongrui Wu, Yifan Xu, and Bao-Liang Lu, "Transfer learning for eeg-based brain-computer interfaces: A review of progress made since 2016," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[10] Xiaying Wang, Michael Hersche, Batuhan Tömekce, Burak Kaya, Michele Magno, and Luca Benini, "An accurate eegnet-based motor-imagery brain–computer interface for low-power edge computing," in *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2020, pp. 1–6.

[11] Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller, "Bci competition 2008–graz data set a," *Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces), Graz University of Technology*, vol. 16, pp. 1–6, 2008.

[12] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[13] Krizhevsky Alex, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional networks," in *volume-1; pages-1097–1105; NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*.

[14] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan, "Learning transferable features with deep adaptation networks," in *International conference on machine learning*. PMLR, 2015, pp. 97–105.

[15] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 2390–2397.

[16] Robin Tibor Schirrmeister, Lukas Gemein, Katharina Eggensperger, Frank Hutter, and Tonio Ball, "Deep learning with convolutional neural networks for decoding and visualization of eeg pathology," *arXiv e-prints*, pp. arXiv–1708, 2017.

[17] Hauke Dose, Jakob S Møller, Helle K Iversen, and Sadasivan Puthusserypady, "An end-to-end deep learning approach to mi-eeg signal classification for bcis," *Expert Systems with Applications*, vol. 114, pp. 532–542, 2018.

[18] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.