ATTENTION-BASED TRANSFER LEARNING FOR BRAIN-COMPUTER INTERFACE

Chuanqi Tan *Fuchun Sun* *Tao Kong* *Bin Fang* *Wenchang Zhang*

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University
{tcq15@mails, fcsun@mail, kt14@mails, fangbin@mail, zhangwc14@mails}.tsinghua.edu.cn

ABSTRACT

Different functional areas of the human brain play different roles in brain activity, which has not been paid sufficient research attention in the brain-computer interface (BCI) field. This paper presents a new approach for electroencephalography (EEG) classification that applies attention-based transfer learning. Our approach considers the importance of different brain functional areas to improve the accuracy of EEG classification, and provides an additional way to automatically identify brain functional areas associated with new activities without the involvement of a medical professional. We demonstrate empirically that our approach out-performs state-of-the-art approaches in the task of EEG classification, and the results of visualization indicate that our approach can detect brain functional areas related to a certain task.

Index Terms— Attention Mechanism, Brain-computer Interface, Transfer Learning, Adversarial Network

1. INTRODUCTION

Brain-computer interface (BCI) based systems can read the brain information of the subject and decode it into instructions for controlling an external device, thereby interacting more naturally with the user. The key issue in BCI-based systems is the accuracy of Electroencephalography (EEG) classification.

One of the most important problems in EEG classification methods is that *the relationship between the functional areas of the human brain and the specific activities is not effectively utilized*. This problem makes it difficult to find key electrodes with higher signal-to-noise ratios, therefore, it is difficult to obtain an effective EEG classifier. Medical research has shown that the functional areas of the human brain have strong regional correlation with specific activities. In the past, when we designed BCI-based systems, we needed medical experts to specify key electrodes involved in special activity. For a new activity, it is difficult to construct a usability system without the help of medical experts, which severely limits the applicability of BCI-based systems. It would be meaningful to have a non-medical approach that automatically discovers activity-related functional areas from brain signals. In addition, another key issue is the lack of training data. Because the cost of biosignal acquisition and labeling is extremely high, it is almost impossible to construct a large, high-quality EEG signal dataset. It is difficult to train advanced classifiers without sufficient training samples.

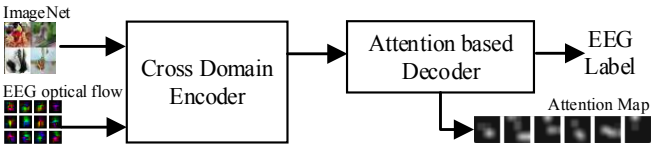

Fig. 1. Overview of our approach. We applied ImageNet as the source domain and EEG optical flow as the target domain to an attention-based transfer learning framework. In addition to obtain EEG label, it gets an extra **attention map** to reflect the activity of the human brain.

To solve these problems, we propose an *attention-based transfer learning framework* that includes two main components: a cross domain encoder and an attention-based decoder with recurrent neural network (RNN). An overview of our approach is shown in Figure 1. A cross domain encoder has the ability to transfer knowledge from natural images domain by representing the original EEG signal in a new form - EEG optical flow. It uses a large amount of training data in the source domain (image classification) to help train the complex feature extractor in the target domain (EEG classification), which solves the problem of a lack of training data by using adversarial transfer learning. The feature extractor will be transferred as knowledge to the target domain. An attention-based decoder uses the attention mechanism to automatically discover the weights of the brain functional areas, which effectively improves the accuracy of EEG classification. This mechanism can reflect the brain functional areas related to a specific activity and overcomes the reliance on medical experts when dealing with new activity.

This work is jointly supported by National Natural Science Foundation of China under with Grant No.91848206, 61621136008, U1613212.

The main **contributions** of this paper are as follows: (1) We introduce attention-based transfer learning to the EEG classification task. (2) Our approach provides a novel way to automatically discover brain functional areas associated with new activities, reducing reliance on medical experts. (3) Experiments show that our approach out-performs the state-of-the-art approaches in an EEG classification task and verify the usability of our approach.

2. RELATED WORK

Many works have been conducted to improve EEG classification accuracy and a great variety of hand-designed features have been proposed. With the rapid development of deep learning in recent years, many excellent networks have been presented by researchers. In recent years, many public works have discussed deep learning applications in bioinformatics research [1].

Transfer learning [2] and deep transfer learning [3] enable the use of different domains, tasks, and distributions for training and testing. [4] reviewed the current state-of-the-art transfer learning approaches in BCI. [5] proposed a novel EEG representation that reduces the EEG classification problem to an image classification problem that implicates the ability of transfer learning. [6] transferred general features via a convolutional network across subjects and experiments. [7] evaluated the transferability between subjects by calculating distance and transferred knowledge in comparable feature spaces to improve accuracy. [8] designed a deep transfer learning framework which is suitable for transferring knowledge by joint training.

[9] and [10] discussed whether the human visual system has attention mechanism. [11] reviewed the recent works on attention-based RNN and its application in computer vision, and categorized the approaches into four classes: item-wise soft attention, item-wise hard attention, location-wise hard attention, and location-wise soft attention. [12] applied the visual attention mechanism in an RNN network to obtain the ability to extract information from images or video by adaptively selecting a sequence of regions or locations. In [13], an attention-based model is applied to identify multiple objects in an image by using reinforcement learning to identify the most relevant regions of the input image. [12] demonstrated that attention not only works on object detection tasks but many other computer vision tasks like image classification. [14] introduced an attention-based model to an image caption task. [15] proposed to extract the feature vector by using the intermediate layer of VGG, and the feature can be associated with a specific region in the image through the network map. In natural language processing tasks, [16] applied a soft attention mechanism to machine translation. [17] showed the latest attention model Google use in machine translation, which uses only attention without a convolutional neural network (CNN) or an RNN in a traditional encoder-decoder model.

To the best of our knowledge, no researchers have attempted to automatically discover brain functional areas associated with new activities.

3. METHOD

Our approach has a traditional encoder-decoder structure that consists of **two main components**: a cross domain encoder and an attention-based decoder.

3.1. Cross domain encoder

To obtain the ability of transfer learning, the raw EEG signal was converted to a new representation - EEG optical flow, which was proposed in our previous work [5]. Many benefits can be gained from using the EEG optical flow. In particular, EEG optical flow can *enhance the ability of transfer learning from natural images*.

Many studies have demonstrated that the front layers in a convolutional neural network (CNN) can extract the general features of images, such as edges and corners. Therefore, we were able to transfer the front layers of a CNN network trained on ImageNet to extract the general features of the EEG optical flow. However, the general feature extractor trained by natural images does not fully match the EEG optical flow.

Inspired by generative adversarial nets (GAN), we apply an adversarial network to train a better general feature extractor which is described in our previous work [8]. The pipeline of adversarial transfer learning is shown in Figure 2.

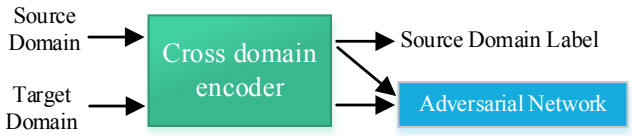

Fig. 2. Pipeline of adversarial transfer learning.Adversarial network used to identify the origins of input features.

We use features extracted from natural images and the EEG optical flow as the inputs for the adversarial network and train it to identify their origins. If the adversarial network achieves inferior performance, it indicates a small difference between the two types of feature and better transferability, and vice versa. It can be achieved by optimizing this loss function:

$$\mathcal{L} = -\sum_{k} \mathbb{I}[y = k] \log p_k + \alpha \mathcal{L}_{adver} + \beta \Re(v), \quad (1)$$

where k is the number of categories, p_k is the softmax value of the classifier activations, $\mathcal{L}_{adver}$ is the cross entropy of the adversarial network, $\Re(v)$ is the regularization of manifold constraints, and α and β are hyperparameters.

Manifold constraints force the learning algorithm to transfer useful knowledge from the source domain and ignore the knowledge which may destroy the manifold structure of the

target domain. [18] demonstrated that keeping the geometric structure can be reduced to the regularization of:

$$\Re(v) = \frac{1}{2}\sum_{i,j=1}^{n} \zeta(v_{i*}, v_{j*})(W)_{ij} \tag{2}$$

where v_i is the embedded representation of sample x_i, $\zeta(v_{i*}, v_{j*})$ is the loss function to measure the euclidean distance of v_{i*} and v_{j*}, $(W)_{ij}$ is the cosine similarity measure of $p - nearest$ neighbor in the adjacency matrix.

To train this adversarial network, we applied an *iteratively optimizing algorithm* with two steps, which has been described in our previous work [8]. In this section, we have introduced a cross domain encoder that extracts features suitable for both the source and target domains and obtains the high-quality features of an EEG signal with help from natural images.

3.2. Attention-based decoder

In this section, we use the features extracted by the cross domain encoder to obtain the final EEG label and attention map of the brain through the attention-based decoder. The attention-based decoder is an RNN network, and we feed the features obtained from each EEG optical flow frame into the RNN network, and treat the output of the last timestamp as the final EEG label.

In a traditional encoder-decoder network, the input of the decoder is the output of the last fully connected layer of the encoder, which raises a crucial problem. The features extracted from the last fully connected layer *lose the location information* of the brain functional areas, so these features do not reflect the importance of different brain functional areas for specific activities.

Encouraged by recent works in computer vision and neural language process, and inspired by recent success in employing attention in these research works, we applied an attention-based decoder that can attend to salient parts of an EEG optical flow while carrying out EEG classification. The attention mechanism provided a powerful tool to overcome the important issue mentioned above. The *location-wise attention mechanism* allows us to consider the weight of different parts in the EEG optical flow, which reflect the different functional areas of the human brain. The pipeline of our attention-based decoder is shown in Figure 3.

The encoder can obtain feature vectors of each EEG optical flow. In order to link the items in the feature vector to the parts of the EEG optical flow one by one, we use the feature map of the *convolutional layer* instead of the output of the *fully connected layer*. Since a low-level feature retains more information, it will be lost in the fully-connected layer. In this way, we can extract L vector features of D dimension as feature vectors, each dimension of feature vector corresponding to a part of the EEG optical flow, as shown in the following equation:

$$a = \{\mathbf{a}_1, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D, \tag{3}$$

where L is the number of frames and D is the number of areas on the EEG optical flow. The items of feature vector are linked to the spatial location of the EEG optical flow by convolutional operation, which is demonstrated in Figure 4:

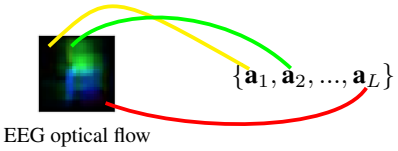

Fig. 4. Link between items in the feature vector and spatial location of the EEG optical flow.

In the attention mechanism, we need to obtain the context vector as the input of the RNN at each time t. The following equations are applied to calculate the context vector: $\hat{\mathbf{z}}_t$:

$$e_{ti} = f_{att}(\mathbf{a}_i, h_{t-1}) \tag{4}$$

$$\alpha_{ti} = \frac{exp(e_{ti})}{\sum_{k=1}^{L} exp(e_{tk})} \tag{5}$$

$$\hat{\mathbf{z}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_i\}), \tag{6}$$

where h_{t-1} is the hidden state of the previous step, f_{att} is the map of a multilayer perceptron (MLP), e_{ti} is the output of the MLP, α_{ti} is the attention weights and ϕ is the function combining feature vectors and attention weights.

There are two types of attention mechanism, the soft attention mechanism and the hard attention mechanism. The main difference is the definition of the ϕ function. In the soft attention mechanism, $\phi(\{\mathbf{a}_i\}, \{\alpha_i\}) = \sum_i^L \alpha_i \mathbf{a}_i$ that means all parts of the EEG optical flow will be considered in the context vector $\hat{\mathbf{z}}_t$. In the hard attention mechanism, ϕ is a function that returns a sampled $\mathbf{a}_i$ at every point in time according to the multinouilli distribution parameterized by α .

The soft attention mechanism is a smooth function; it can be solved directly by the back propagation algorithm, which is equivalent to optimizing the following loss function:

$$\mathcal{L} = -log(P(y|x)) + \alpha \sum_{i}^{L}(1 - \sum_{t}^{C} \alpha_{ti})^2. \tag{7}$$

The hard attention mechanism is a non-smooth function that can be approximated by the Monte Carlo algorithm.

4. EXPERIMENTS

We applied our approach to a dataset called Open Music Imagery Information Retrieval (OpenMIIR) [19]. OpenMIIR is compiled during music perception and imagination, which involves 10 subjects listening to and imagining 12 short music

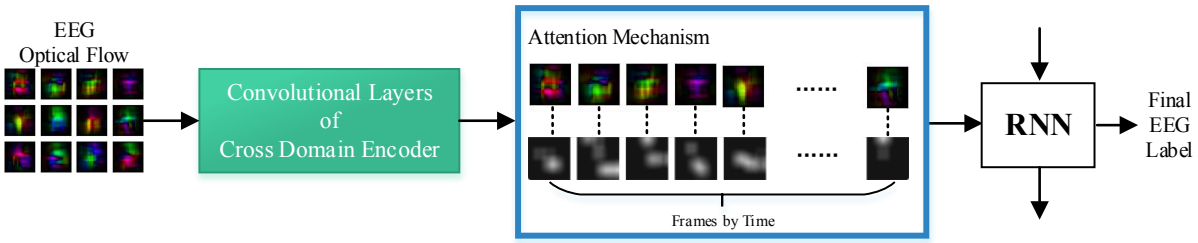

Fig. 3. Pipeline of attention-based decoder. First, the feature vectors that can maintain spatial information are produced by the convolutional layers of the cross domain encoder. Then, they are combined with the location-wise attention mechanism on each frame. Finally, these feature vectors are sent to the RNN network one by one to obtain the final EEG label.

fragments taken from well-known pieces. These signals were recorded using 64 EEG electrodes at 512 Hz, and 240 trials were recorded per subject. The following parameters were used in our approach. We converted raw EEG signals into EEG videos with thirteen frames and a resolution of 32*32. These frames were converted to EEG optical flow with twelve frames. We employed VGG16 and VGG19 [20] as the targets of the cross domain encoder .

The OpenMIIR dataset does not distinguish between training and test sets, so we randomly selected 10% of the dataset to use as the test dataset. As the baseline, we tested some recently proposed approaches: the deep neural network (DNN) described in [19] and the CNN described in [21]. In addition, we made comparisons to our previous work [8], that without an attention mechanism. Experiments on the OpenMIIR dataset were conducted to compare the performance of our approach and that of the baseline approaches, and the results are shown in Table 1.

Table 1. Classification accuracy (%) on the OpenMIIR dataset and comparisons to the baseline approaches. For example, the corner mark in $Our_{(Soft+VGG16)}$ refers to use of the soft attention mechanism and application of the VGG16 network as the encoder.

[19]	27.22	[21]	27.80
$[8]_{VGG16}$	32.08	$[8]_{VGG19}$	35.00
$Our_{(Soft+VGG16)}$	37.92	$Our_{(Soft+VGG19)}$	36.67
$Our_{(Hard+VGG16)}$	37.08	$Our_{(Hard+VGG19)}$	35.84

As the results show in Table 1, the soft attention mechanism achieves better classification results than the hard attention mechanism. One possible reason is that the soft attention mechanism considers the interaction between multiple functional areas, while the hard mechanism only considers one functional area, as shown in Figure 5. Medical knowledge tells us that the reflection of an activity in the brain is the result of a combination of multiple functional areas, which is more in line with the soft attention mechanism.

We visualized an attention map while a subject was listening to an intense piece of music, as shown in Figure 6. It was found that the learned weights of attention are somewhat similar to the result from medical experts [22].

(a) $Soft$ (b) $Hard$

Fig. 5. Visualization of soft attention mechanism and hard attention mechanism on the same frame.

Fig. 6. Visualization of soft attention mechanism when listing to an intense music fragment.

We can draw the following conclusions from the experimental results presented in this section: (1) The experimental results shown in Table 1 demonstrate that our proposed approach performs better than traditional approaches; (2) VGG16 is a better choice for encoder than VGG19 in our attention-based transfer learning for EEG classification task; (3) The performance of the soft attention mechanism is better than that of the hard attention mechanism; (4) Attention mechanisms can be used to automatically discover brain functional areas associated with new activities and reduce the dependence on medical experts.

5. CONCLUSIONS

We propose a novel approach to improve the accuracy of EEG classification in BCI. This approach takes advantage of the medical fact that different brain functional areas play different roles in activities. It applies an attention mechanism to automatically assess the importance of functional areas of the brain during activity. It can be concluded that our approach is superior to other state-of-the-art approaches. In addition, our approach can be used to automatically discover brain functional areas associated with activities, which is very useful when dealing with EEG data related to a new activity.

6. REFERENCES

[1] P Mamoshina, A Vieira, E Putin, and A Zhavoronkov, "Applications of deep learning in biomedicine.," *Molecular Pharmaceutics*, vol. 13, no. 5, pp. 1445, 2016.

[2] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu, "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 270–279.

[4] Vinay Jayaram, Morteza Alamgir, Yasemin Altun, Bernhard Scholkopf, and Moritz Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 20–31, 2016.

[5] Chuanqi Tan, Fuchun Sun, Wenchang Zhang, Jianhua Chen, and Chunfang Liu, "Multimodal classification with deep convolutional-recurrent neural networks for electroencephalography," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 767–776.

[6] Mehdi Hajinoroozi, Zijing Mao, Yuan-Pin Lin, and Yufei Huang, "Deep transfer learning for cross-subject and cross-experiment prediction of image rapid serial visual presentation events from eeg data," in *International Conference on Augmented Cognition*. Springer, 2017, pp. 45–55.

[7] Yuan-Pin Lin and Tzyy-Ping Jung, "Improving eeg-based emotion classification using conditional transfer learning," *Frontiers in Human Neuroscience*, vol. 11, 2017.

[8] Chuanqi Tan, Fuchun Sun, and Wenchang Zhang, "Deep transfer learning for eeg-based brain computer interface," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 916–920.

[9] Ronald A Rensink, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[10] Maurizio Corbetta and Gordon L Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, pp. 201, 2002.

[11] Feng Wang and David MJ Tax, "Survey on the attention based rnn model and its applications in computer vision," *arXiv preprint arXiv:1601.06823*, 2016.

[12] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al., "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[13] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[14] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, 2014, pp. 595–603.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[18] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[19] Sebastian Stober, "Learning discriminative features from electroencephalography recordings by encoding similarity constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6175–6179.

[20] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[21] Sebastian Stober, Avital Sternin, Adrian M Owen, and Jessica A Grahn, "Deep feature learning for eeg recordings," *arXiv preprint arXiv:1511.04306*, 2015.

[22] Vernon L Towle, José Bolaños, Diane Suarez, Kim Tan, Robert Grzeszczuk, David N Levin, Raif Cakmur, Samuel A Frank, and Jean-Paul Spire, "The spatial location of eeg electrodes: locating the best-fitting sphere relative to cortical anatomy," *Electroencephalography and clinical neurophysiology*, vol. 86, no. 1, pp. 1–6, 1993.