Parameter-Free Attention in fMRI Decoding

YONG QI, HUAWEI LIN, YANPING LI, AND JIASHU CHEN

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China
Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China

Corresponding author: Yong Qi (qiyong@sust.edu.cn)

This work was supported in part by the National Training Program of Innovation and Entrepreneurship for Undergraduates of China under Grant 201910708024, in part by the Research Plan of Shaanxi Provincial Department of Education under Grant 15JK1086, and in part by the Shaanxi University of Science and Technology Ph.D. Project under Grant BJ14-07.

ABSTRACT

An fMRI decoder aims to infer the corresponding type of task stimulus from the given fMRI data. Recently, deep learning techniques have attracted fMRI decoding attention. Yet, it has not demonstrated an outstanding decoding performance because of ultra-high dimensional data, extremely complex calculations, and subtle differences between different tasks. In this work, we propose a parameter-free attention module called Skip Attention Module (SAM) consisted of weight branch and skip branch, which can pay attention to areas with more information to enhance data features. SAM does not contain any parameters that need to be trained, and does not increase any burden of training. Thus, it can stack on any Convolutional Neural Networks (CNN) architecture or even on a pre-trained model. Our experiments on seven tasks of the large-scale Human Connectome Project (HCP) S1200 data set containing about 1200 subjects show that the architecture with SAM achieves a significant performance improvement compared with the non-attention architecture. We have conducted many experiments, and the average decoding accuracy is up to 88.7%. Besides, the average decoding error of the architecture using SAM is 1.2% ~ 3.1% lower than the architecture without SAM. For a single task, the architecture decoding accuracy using SAM has the highest increase of 11.1%. In addition, the proposed method also shows excellent performance on the ADHD-200 dataset, indicating the universality of the method. These results establish that the proposed SAM can be superimposed on any architecture and can effectively improve fMRI decoding accuracy.

INDEX TERMS

fMRI decoding, functional magnetic resonance imaging, parameter-free attention mechanism.

I. INTRODUCTION

A brain decoding model aims to predict the relevant information and corresponding task stimulus-based on the activity of the voxel area via functional magnetic resonance imaging (fMRI). The brain decoding technique provides many novel understandings about task stimulus’s evoked neural activity in recent years.

Most of the conventional methods are to extract features and build mathematical models manually. For instance, Multi-voxel Pattern Analysis (MVPA) has delivered a remarkable performance in the brain decoding field. Still, MVPA struggles to achieve excellent performance on high-dimensional data and requires experts to extract features [1] manually. Thus, researchers urgently need to seek an automated approach to extract features from high-dimensional fMRI brain data.

In recent years, the successful application of deep learning in computer vision has led researchers to seek approaches for the fusion of neuroscience and deep learning. At present, plenty of studies demonstrated that deep learning has superior brain decoding ability. For example, Riaz et al. presented the first end-to-end model for decoding neurological disorder [2]. Vu et al. proposed an end-to-end convolutional neural network to label tasks and yielded a significantly low error rate (3.3 ± 0.7%) in a 12 young subjects dataset [3]. Wang et al. proposed a deep neural network for decoding brain task states and reached an average accuracy of 93.7% in a state-of-the-art dataset [4]. Gao et al. utilized transfer learning to design a cross-subject decoding model to decipher the behavior tasks and achieved significant performance improvements [1].

Currently, two major obstacles are hinder the application of deep learning in the field of brain decoding: (1) Almost all neuroimaging data are high-dimensional, and the amount of data is particularly small [5]. Most data sets contain only a few hundred subjects, and some even a few dozen, but each data sample contains more than a hundred
thousand dimensions. In such a setting, deep learning methods will suffer from overfitting. Besides, these high-dimensional data also bring many challenges, such as huge parameters, complex calculations, etc. (2) The deep learning method is regarded as a black-box model in neuroscience because the mapping between input and decoding output is unknown.

To address these real problems, we propose a parameter-free attention module which can combine with the state-of-the-art architecture of Convolutional Neural Networks (CNN). This parameters-free attention module has been shown to effectively improve the decoding accuracy without increasing the amount of calculation and parameters. The attention mechanism was originally used in the field of Computer Vision (CV) to increase attention to specific areas [6]–[8]. In fact, humans also have similar attention mechanisms when watching photos or movies. They will subconsciously pay attention to meaningful areas [9]–[11]. In this study, we attempt to design an attention module applies to the field of fMRI decoding so that while improving the decoding ability, the number of network parameters and model size will not grow with the increase of the scale of the dataset. It means that the rise of the calculation depends only on the number of attention module occurrences.

This paper proposes a parameter-free attention module for processing 3D fMRI data to address excessive parameters and overfitting. To explore whether the parameter-free attention module improves decoding ability in fMRI data we utilize 3D-CNN architecture extended from 2D-CNN, which decodes the behavior tasks from fMRI data, by combining inception and skip connection.

This paper’s main contribution is to propose a parameter-free attention module, which can be superimposed on any architecture without additional concern about the size of input and output. Our experiment proved that the parameter-free attention module is effective. All the experiments results establish that the proposed SAM can be superimposed on any architecture and effectively improve fMRI decoding accuracy. The corresponding experimental data and code are available at https://github.com/huawei-lin/fMRI_parameter-free_attention.

### II. METHOD

#### A. DATASET

1) HCP S1200

The Human Connectome Project (HCP) is a large-scale public dataset in the brain research field, whose target is mapping the human brain and looking at the connection between function and behavior. In this study, we used HCP S1200 tfMRI dataset, which was collected from approximately 1200 young, healthy adult subjects (age 22-35) with 7 tasks, including emotion, gambling, language, motor, relational, social, and working memory (WM). Detailed information is shown in Tab.2, and the acquisition parameters are shown in Tab.1. Furthermore, details of the dataset can be found in the HCP Data Release Reference Manual [12], and the design of the experiment is summarized in previous papers [13], [14].

2) ADHD-200

Attention deficit hyperactivity disorder (ADHD) is a common brain disorders among children. Because the cause of ADHD is not yet known, it is particularly important to use algorithms to help distinguish health conditions from those with ADHD [15]. ADHD-200 is a grassroots project that promotes ADHD research by sharing data. The dataset includes 776 rest-state fMRI data and anatomical data collected on 8 independent imaging sites, of which 491 are from typical developmental individuals and 285 children and adolescents with ADHD (age: 7-21 years old). The collection parameters of different sites are shown in Tab. 1. In this study, we used datasets from NeuroImage (NI), New York University Medical Center (NYU) and Peking University (PKU) to evaluate the proposed module.

#### B. DATA PREPROCESSING

1) HCP S1200

The HCP S1200 dataset was used in this study had been preprocessed with HCP functional pipeline, including slice time correction, spatial artifact removal, spatial distortion removal, skull removal, surface generation, cross-modal registration, and alignment to standard space [13], [14].

2) ADHD-200

These data have been preprocessed as part of the connectome project, including time correction, motion correction, and mapping to MNI standard space [16]. The ADHD-200 gives fMRI data in a time series. In this study, these fMRI data will be divided into a unit time fMRI image according to time to convert a fMRI time series into whole brain images.

#### C. 3D CONVOLUTIONAL NEURAL NETWORKS

CNN is a hierarchical structure that is quite effective for extracting high-level features, especially high-dimensional data [17]. However, many prior studies mainly focused on 2 dimensions, but these do not make available for 3D volumetric images such as fMRI data and potentially overlook meaningful information of 3D structure. In this work, we utilize 3D-CNN architecture extended from state-of-the-art 2D-CNN architecture, such as PlainNet, ResNet, and InceptionNet, which cooperate with the parameter-free attention mechanism to decode the stimulus of tasks from fMRI data.

### TABLE 1. Detailed acquisition parameters of the fMRI data.

| Parameter            | HCP  | NI    | NYU  | PKU  |
|----------------------|------|-------|------|------|
| Slices               | 72   | 37    | 33   | 33   |
| TR (ms)              | 720  | 1960  | 2000 | 2000 |
| TE (ms)              | 33.1 | 40    | 15   | 30   |
| Flip Angle (deg)     | 52   | 80    | 90   | 90   |
| Thickness (mm)       | 2.0  | 3.0   | 4.0  | 3.5  |
| Bandwidth (Hz/Px)    | 2290 | 1628  | 3906 | 2520 |
FIGURE 1. The PlainNet, ResNet and InceptionNet architectures. (a) A PlainNet with 8 Convolutional layers and 4 SAMs. (b) A ResNet with 4 residual blocks and 3 SAMs. (c) An InceptionNet with 4 inceptions and 3 SAMs.
To verify the parameter-free attention mechanism’s effectiveness, we used PlainNet, ResNet, and InceptionNet for controlled trials. In these experiments, the input data size is $91 \times 109 \times 91$ and is aligned with the MNI152 standard-space template. Due to memory restriction on the GPU, all CNN architecture’s batch size was set to 16, and the learning rate was set to $10^{-4}$. Particularly, we also utilize the early-stopping algorithm, which could stop training in time when overfitting occurs in training. Before training, the 1113 available subjects in the HCP S1200 Dataset will be assigned to the training set (400, 35.9%), validation set (25, 2.3%), and test set (688, 61.8%). Besides, the datasets of all sites of ADHD-200 are divided into training set (70%), validation set (15%) and test set (15%). Fig.1 shows the PlainNet, ResNet, and InceptionNet architectures used in the experiment in this study.

**D. ATTENTION MECHANISM**

Because of the attention mechanism, humans and animals mainly focus on the most information when receiving vast amounts of information. These biological mechanisms have been revealed by many researchers [18]–[20]. Thus, researchers attempt to achieve some inspiration from this mechanism.

Recently, the attention mechanism has been successfully applied in many deep learning fields. Such as machine translation [21], speech recognition [22], image captioning [23], fine-grained image recognition [24], visual question answering [25], person re-identification [26] and panoptic segmentation [27].

Woo et al. proposed a attention module named Convolutional Block Attention Module (CBAM) [8], which is simple to implement and effective. CBAM introduces channel attention and spatial attention, and can infer attention maps in the channel and space respectively, so that the feature map can be enhanced in both the channel and the space.

In this study, we will extend CBAM from 2D to 3D, and conduct a control experiment with the SAM proposed in this article.

**E. PARAMETER-FREE ATTENTION MODULE**

An attention model’s key idea is to utilize the hidden unit to weigh the specific parts based on the importance of input,
which equips the model to pay attention to the particular regions involving the most information.

Inspired by the previous studies about the attention mechanism in the field of computer vision [28]–[30], we propose a parameter-free attention module. As shown in Fig.2, our parameter-free attention module is divided into weight branch and skip branch, and it can be able to adapt and stack to any architecture.

1) SKIP BRANCH

The shortcut connection has been studied for a long time [31]–[33], which is a beneficial approach to
TABLE 3. In the HCP dataset, a comparison between the previous method and the proposed method.

| Authors             | Method       | Subjects | Architecture               | Mean Accuracy |
|---------------------|--------------|----------|----------------------------|---------------|
| S. Koyamada et al.  | Logistic Regression | 499      | Logistic Regression         | 20.81%        |
|                     | SVM          | 499      | Linear Kernel              | 20.87%        |
|                     | SVM          | 499      | RBF Kernel                 | 47.97%        |
|                     | DNN          | 499      | 116-500-7                  | 48.94%        |
|                     | DNN          | 499      | 116-500-500-7              | 50.74%        |
|                     | DNN          | 499      | 116-500-500-500-7          | 50.57%        |
| Y. Gao et al.       | LSTM-8       | 965      | LSTM-8                     | 62.3%         |
|                     | LSTM-16      | 965      | LSTM-16                    | 65.1%         |
|                     | LSTM-32      | 965      | LSTM-32                    | 67.5%         |
|                     | AlexNet      | 965      | 8-Layer AlexNet            | 63.9%         |
|                     | InceptionNet | 965      | 22-Layer Inception V3      | 69.7%         |
|                     | ResNet       | 965      | 34-Layer ResNet 3          | 69.2%         |
| A. W. Thomas et al. | DeepLight    | 400      | CNN + bi-directional LSTM + FC | 81.91%   |
| Ours                | PlainNet     | 1113     | 8-Layer PlainNet           | 87.9%         |
|                     | ResNet       | 1113     | 12-Layer ResNet            | 83.35%        |
|                     | InceptionNet | 1113     | 32-Layer InceptionNet      | 85.08%        |
|                     | PlainNet-SAM | 1113     | 8-Layer PlainNet with 4 SAMs | 88.69% |
|                     | ResNet-SAM   | 1113     | 12-Layer ResNet with 3 SAMs | 86.41%  |
|                     | InceptionNet-SAM | 1113 | 32-Layer InceptionNet with 3 SAMs | 86.07% |

TABLE 4. In the ADHD-200 dataset, a comparison between the previous method and the proposed method.

| Authors             | Method               | NI       | NYU     | PKU     |
|---------------------|----------------------|----------|---------|---------|
| A. Riaz et al.      | AP Clustering [44] [45] | 32.00%   | 56.10%  | 58.82%  |
|                     | SVM [45]             | 44%      | 24.3%   | 58.8%   |
|                     | FCNet [46]           | 60.0%    | 58.5%   | 62.7%   |
|                     | DeepFMRI [2]         | 67.9%    | 73.1%   | 62.7%   |
| L. Shao et al. [47] | Revised gcForest     | 72.00%   | 73.13%  | 64.87%  |
| Ours                | PlainNet             | 71.30%   | 69.49%  | 71.63%  |
|                     | PlainNet-CBAM [8]    | 79.41%   | 70.56%  | 70.90%  |
|                     | PlainNet-SAM         | 80.05%   | 72.66%  | 75.28%  |
|                     | ResNet               | 74.15%   | 71.98%  | 70.05%  |
|                     | ResNet-CBAM [8]      | 71.08%   | 72.74%  | 70.93%  |

solve the problems of gradient vanishing and gradient exploding [34]–[39].

In this work, Skip Branch exploits a shortcut connection to connect the input of the attention module and the result of the weight branch, an auxiliary route for addressing the problems of exploding/vanishing gradients not have any parameter. The skip branch can also retain the decoding features map and the gradient information to ensure that is effectively propagated forward.

2) WEIGHT BRANCH

Our weight branch serves as a decoding feature selector in SAM during forwarding updates, which contains downsample feed-forward sweep and upsample feedback. The former operation collects valuable information on the whole feature map; the latter operation regains the original shape for combining with the original feature map.

From input, the feature map will pass through several 3D max-pooling layers to reach a low resolution, which
will increase the receptive field and retain only valuable information. Then, as many upsample layers with trilinear interpolation, the number of max-pooling layers will change the low-resolution feature map to the same size as the input feature map. Finally, the sigmoid function layer will become a feature map in the range $[0, 1]$.

The weight branch can enhance features through re-weight areas with important information and increase the network’s non-linear expression ability.

3) SKIP ATTENTION MODULE

Formally, in this work, our parameter-free attention module is defined as:

$$x_{i+1} = x_i \cdot (W(x_i) + 1)$$

Here the $W(x)$ is the output signals from the weight branch, and it ranges from $[0, 1]$. The $x$ is the input signal of the attention module, and it is called the skip branch.

In conclusion, the proposed Skip Attention Module (SAM) has the following properties:

1. End-to-end: the entire network using SAM can still be trained end-to-end by stochastic gradient descent (SGD) with backpropagation.
2. Feature enhancement: SAM can work through re-weight to focus on areas with important information on the entire feature map.
3. Ability improvement: the architecture using SAM will be more robust and more precise.
4. Parameter-free: the proposed SAM does not have any parameter that needs to be trained to increase the training burden.
5. Stack arbitrary: it can stack on any CNN architecture or even on a pre-trained model.
6. Faster training: the model with the proposed SAM was trained faster and easier to converge.
7. Preserve low-level detail: the skip branch’s low-level feature was retained and enhanced by the weight branch.
8. Easy and straightforward: the proposed SAM can be implemented in an easy code.

F. VISUALIZATION WITH GRAD-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a technique for producing visual explanations for decisions from CNN models, which can give an intuitive visual explanation without changing the structure of the network [40], [41]. In this study, Grad-CAM was used to explain the mechanism of SAM. As shown in Fig.4, this is an fMRI image of a patient with ADHD through a network architecture with and without SAM. (a) shows that the fMRI image of this patient passes through a network architecture without SAM, and only a few features are paid attention to, while (b) shows that this patient has more features that were noticed by SAM. Because SAM re-weights the feature map based on the previous weights, it makes the network more likely to focus on other potential features. More importantly, SAM does not contain any parameters, it will not bring any burden to the model.

III. RESULT AND ANALYSIS

The proposed SAM is evaluated on the HCP S1200 dataset. Also, we assessed the performance of our SAM with cross-validation. The experiments were conducted on a work-station with three NVIDIA TITAN X GPU. All the
experiments are trained by the early-stopping strategy to reduce the impact of overfitting, with a batch size of 16 and a learning rate of 10\(^{-4}\).

The comparison between our decoding task’s average accuracy and the existing methods is shown in Tab.3. Furthermore, Fig.5 shows the average decoding error rate using PlainNet, ResNet, and InceptionNet for each task. Overall, the average error rate of architectures stacked SAM compared to the non-attention architectures have decreased. The average decoding error rates of PlainNet, ResNet, and InceptionNet using SAM on the HCP S1200 dataset are reduced by 2.9%, 3.1%, and 1.2%, respectively. The error rate of Working Memory (WM) was significantly dropped from the ResNet with SAM and InceptionNet with SAM than the corresponding non-attention architectures. In ResNet results, compared with ResNet without SAM, the error rate of Working Memory (WM) of ResNet using SAM dropped by 11.1%, while in InceptionNet, it dropped by 8.5%.

We also visualize the decoding performance of the SAM using the confusion matrix depicted in Fig.3. And the part of the matrix less than 1% is ignored. The decoding performance of architectures with the SAM obviously outperformed non-attention architectures. The architectures with SAM have reduced most of the error decoding compared to the architectures without SAM. Besides, for the architectures without SAM, working memory tasks seem difficult to distinguish, but in architectures with SAM, decoding WM is more relaxed, and the accuracy rate is significantly improved.

**IV. CONCLUSION**

In this study, we designed a parameter-free attention module called Skip Attention Module (SAM) consisted of weight branch and skip branch, which can pay attention to areas with more information to enhance data features. It can stack on any architecture of convolutional neural network without any parameters that need to train and do not increase training burden. In fact, this article’s SAM layout method is the best in the experiment, not the only one.

Our experimental results show that compared with the architecture without SAM, the decoding error rate of the architecture with SAM generally decreases. In HCP experiments, the average decoding error rate of ResNet-SAM compared to ResNet was dropped by 3.1%. Some difficult tasks to decode in the architectures without SAM are easier to decode in the architectures using SAM. For example, in the Working Memory (WM) task, compared with ResNet without SAM, ResNet’s error rate using SAM dropped by 11.1%, and in InceptionNet, it dropped by 8.5%. In the ADHD-200 experiments, the error rate of using SAM in the PlainNet of dataset of NI site has been significantly improved, which is about 8.7% lower, and the error rate of other sites has decreased slightly. For CBAM, although some experiments performed slightly better than SAM, the overall performance was worse than SAM. The detailed experimental results of ADHD-200 are shown in Fig.5, and the comparison with previous studies is shown in Tab.4.

In summary, SAM can be superimposed on any architecture without additional concern about the size of input and output. On this basis, the accuracy and stability of fMRI decoding are also improved. In future work, we will continue to study the application and practice of attention mechanism in fMRI decoding and conduct interpretability research on attention mechanism. These results establish that the proposed SAM can be superimposed on any architecture and can effectively improve the fMRI decoding accuracy.

**ACKNOWLEDGMENT**

The authors thank the Amazon Sagemaker machine learning platform for providing experimental conditions and opportunities for this study.

**REFERENCES**

[1] Y. Gao, Y. Zhang, H. Wang, X. Guo, and J. Zhang, “Decoding behavior tasks from brain activity using deep transfer learning,” IEEE Access, vol. 7, pp. 43222–43232, 2019.

[2] A. Riaz, M. Asad, S. M. M. R. A. Arif, E. Alonso, D. Dima, P. Corr, and G. Slabaugh, “Deep fMRI: AN end-to-end deep network for classification of fMRI data,” in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Apr. 2018, pp. 1419–1422.

[3] H. Yu, H.-C. Kim, and H.-L. Lee, “3D convolutional neural network for feature extraction and classification of fMRI volumes,” in Proc. Int. Workshop Pattern Recognit. Neuroimag. (PRNI), 2018, pp. 1–4.

[4] X. Wang, X. Liang, Z. Jiang, B. A. Nguchu, Y. Zhou, Y. Wang, H. Wang, Y. Li, Y. Zhu, F. Wu, J. Gao, and B. Qiu, “Decoding and mapping task states of the human brain via deep learning,” Hum. Brain Mapping, vol. 41, no. 6, pp. 1505–1519, Apr. 2020.

[5] A. W. Thomas, H. R. Heekeren, K.-R. Müller, and W. Samek, “Interpretable LSTMs for whole-brain neuroimaging analyses,” 2018, arXiv:1810.09945. [Online]. Available: https://arxiv.org/abs/1810.09945

[6] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7794–7803.

[7] Z. Gao, J. Xie, Q. Wang, and P. Li, “Global second-order pooling convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 3024–3033.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[9] C. Bundesen, “A theory of visual attention,” Psychol. Rev., vol. 97, no. 4, p. 523, 1990.

[10] C. Bundesen, T. Habekost, and S. Kyllingsbæk, “A neural theory of visual attention: Bridging cognition and neurophysiology,” Psychol. Rev., vol. 112, no. 2, p. 291, 2005.

[11] L. Zhang and W. Lin, Background of Visual Attention—Theory and Experiments. Piscataway, NJ, USA: IEEE Press, 2013, pp. 25–71.

[12] H. Wu-Minn. (2017). 1200 Subjects Data Release Reference Manual. [Online]. Available: https://www.humanconnectome.org

[13] J. Lv, X. Jiang, X. Li, D. Zhu, S. Zhang, S. Zhao, H. Chen, T. Zhang, X. Hu, J. Han, J. Ye, L. Guo, and T. Liu, “Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function,” IEEE Trans. Biomed. Eng., vol. 62, no. 4, pp. 1120–1131, Apr. 2015.

[14] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, D. Nolan, E. Bryant, T. Hartley, O. Footer, J. M. Bjork, R. Poldrack, S. Smith, H. Johansen-Berg, A. Z. Snyder, and D. C. Van Essen, “Function in the human connectome: Task-fMRI and individual differences in behavior,” Neuroimage, vol. 80, pp. 169–189, Oct. 2013.

[15] T. Eslami and F. Saeed, “Similarity based classification of ADHD using singular value decomposition,” in Proc. 15th ACM Int. Conf. Frontiers, 2018, pp. 19–25.

[16] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, “The neuro bureau ADHD-200 preprocessed repository,” Neuroimage, vol. 144, pp. 275–286, Jan. 2017.
R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and T. Raiko, H. Valpola, and Y. LeCun, “Deep learning made easier by N. Schraudolph, “Accelerated gradient descent by factor-centering decomposition,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2204–2212.

D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.” 2014, arXiv:1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473

J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 577–585.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 2048–2056.

H. Zheng, J. Fu, T. Mei, and J. Luo, “Learning multi-attention convolutional neural network for fine-grained image recognition.” 2017, Proc. IEEE Int. Conf. Comput. Vis., pp. 5209–5217.

C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma, “Structured attentions for visual question answering,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1291–1300.

L. Zhao, X. Li, Y. Zhang, and J. Wang, “Deeply-learned part-aligned representations for person re-identification,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 3219–3228.

H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, “Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation,” 2020, arXiv:2003.07853. [Online]. Available: http://arxiv.org/abs/2003.07853

F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Conf., Jul. 2017, pp. 3156–3164.

Y. Zhu, C. Zhao, H. Guo, J. Wang, X. Zhao, and H. Lu, “Attention CoupleNet: Fully convolutional attention coupling network for object detection,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 113–126, Jan. 2019.

H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, “Parameter-free spatial attention network for person re-identification,” 2018, arXiv:1811.12150. [Online]. Available: http://arxiv.org/abs/1811.12150

C. M. Bishop, Neural Networks for Pattern Recognition. London, U.K.: Oxford Univ. Press, 1995.

B. D. Ripley, Pattern Recognition and Neural Networks. Cambridge, U.K.: Cambridge Univ. Press, 2007.

W. N. Venables and B. D. Ripley, Modern Applied Statistics With S-PLUS. Berlin, Germany: Springer, 2013.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Conf., Jun. 2015, pp. 1–9.

C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in Proc. Artif. Intell. Statist., 2015, pp. 562–570.

N. N. Schraudolph, “Centering neural network gradient factors,” in Neural Networks: Tricks of the Trade. Berlin, Germany: Springer, 1998, pp. 207–226.

N. Schraudolph, “Accelerated gradient descent by factor-centering decomposition,” IIDSIA, Lugano, Switzerland, Tech. Rep. IIDSIA-33-98, 1998, vol. 98.

T. Raiko, H. Valpola, and Y. LeCun, “Deep learning made easier by linear transformations in perceptrons,” in Proc. Artif. Intell. Statist., 2012, pp. 924–932.

T. Vatanen, T. Raiko, H. Valpola, and Y. LeCun, “Pushing stochastic gradient towards second-order methods–backpropagation learning with transformations in nonlineairities,” in Proc. Int. Conf. Neural Inf. Process. Berlin, Germany: Springer, 2013, pp. 442–449.

R. R. Selvaraju, M. Cogswell, À. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.

R. R Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-CAM: Why did you say that?” 2016, arXiv:1611.07450. [Online]. Available: http://arxiv.org/abs/1611.07450

S. Koyamada, Y. Shikauchi, K. Nakae, M. Koyama, and S. Ishii, “Deep learning of fMRI big data: A novel approach to subject-transfer decoding.” 2015, arXiv:1502.00093. [Online]. Available: http://arxiv.org/abs/1502.00093

Y. Qi was born in Tongchuan, Shaanxi, in 1982. He received the bachelor’s degree in computer science and technology and the master’s and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, in 2004 and 2011, respectively.

Since 2012, he has been a Lecturer with the Department of Computer Science, Shaanxi University of Science and Technology. He has one authorized invention patent. His research interests include the process of hemoglobin change in fMRI images and the topological structure division of brain function, OCT 3D fingerprint biomarker modeling, digital twins and computer simulation, machine learning, and deep learning algorithm design. He is currently a member of the China Computer Society and the Shaanxi Big Data and Cloud Computing Industry Technology Innovation Strategic Alliance. As a coach, he has won the Bronze Medal in the ACM-ICPC East Asian Region, three times.

HUAWEI LIN was born in Guangdong, in 1999. He is currently pursuing the bachelor’s degree in computer science with the Shaanxi University of Science and Technology. He was the Leader of a project funded by the National Ministry of Education of China for Undergraduates in 2019. His research interests include machine learning, deep learning, computer vision, and medical image processing. In his undergraduate period, he has received the ACM-ICPC Asia Bronze Medal in the ACM-ICPC International Collegiate Programming Contest and the Silver Medal in the ACM-ICPC Shaanxi Province Contest of China.

YANPING LI was born in Linfen, Shanxi, China, in 2001. She is currently pursuing the bachelor’s degree in computer science and technology with the Shaanxi University of Technology. Her research interests include medical image processing, multivariate statistical analysis, and other fields.

JIASHU CHEN is currently pursuing the bachelor’s degree in computer science and technology with the Shaanxi University of Science and Technology. His main research interests include computer vision and biometrics. 

[43] A. W. Thomas, K.-R. Müller, and W. Samek, “Deep transfer learning for whole-brain fMRI analyses,” in OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging. Cham, Switzerland: Springer, 2019, pp. 59–67.

[54] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” Science, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[44] B. Riaz, E. Alonso, and G. Slabaugh, “Phenotypic integrated framework for classification of ADHD using fMRI,” in Proc. Int. Conf. Image Anal. Recognit., 2016, pp. 217–225.

[45] B. Riaz, M. Asad, S. M. R. Al-Arif, E. Alonso, D. Dima, P. Corr, and G. Slabaugh, “FCNet: A convolutional neural network for calculating functional connectivity from functional MRI,” in Proc. Int. Workshop Connectomics Neuroimag., 2017, pp. 70–78.

[46] B. Miaou, L. L. Zhang, J. L. Quan, F. Q. Meng, and Y. L. Zhang, “Classification of ADHD individuals and neurotypicals using reliable RELIEF: A resting-state study,” IEEE Access, vol. 7, pp. 62163–62171, 2019.

[47] YONG QI was born in Tongchuan, Shaanxi, in 1982. He received the bachelor’s degree in computer science and technology and the master’s and Ph.D. degrees in computer science and technology from Northwestern Polytechnical University, in 2004 and 2011, respectively.

Since 2012, he has been a Lecturer with the Department of Computer Science, Shaanxi University of Science and Technology. He has one authorized invention patent. His research interests include the process of hemoglobin change in fMRI images and the topological structure division of brain function, OCT 3D fingerprint biomarker modeling, digital twins and computer simulation, machine learning, and deep learning algorithm design. He is currently a member of the China Computer Society and the Shaanxi Big Data and Cloud Computing Industry Technology Innovation Strategic Alliance. As a coach, he has won the Bronze Medal in the ACM-ICPC East Asian Region, three times.

[48] HUAWEI LIN was born in Guangdong, in 1999. He is currently pursuing the bachelor’s degree in computer science with the Shaanxi University of Science and Technology. He was the Leader of a project funded by the National Ministry of Education of China for Undergraduates in 2019. His research interests include machine learning, deep learning, computer vision, and medical image processing. In his undergraduate period, he has received the ACM-ICPC Asia Bronze Medal in the ACM-ICPC International Collegiate Programming Contest and the Silver Medal in the ACM-ICPC Shaanxi Province Contest of China.

[49] YANPING LI was born in Linfen, Shanxi, China, in 2001. She is currently pursuing the bachelor’s degree in computer science and technology with the Shaanxi University of Technology. Her research interests include medical image processing, multivariate statistical analysis, and other fields.

[50] JIASHU CHEN is currently pursuing the bachelor’s degree in computer science and technology with the Shaanxi University of Science and Technology. His main research interests include computer vision and biometrics.