Received: 14 October 2021 | Revised: 29 January 2022 | Accepted: 9 February 2022
DOI: 10.1002/hbm.25813

RESEARCH ARTICLE

Attention module improves both performance and interpretability of four-dimensional functional magnetic resonance imaging decoding neural network

Zhoufan Jiang[1] | Yanming Wang[1] | ChenWei Shi[1] | Yueyang Wu[1] | Rongjie Hu[1] | Shishuo Chen[1] | Sheng Hu[1] | Xiaoxiao Wang[1,2] | Bensheng Qiu[1,2]

[1]Center for Biomedical Imaging, University of Science and Technology of China, Hefei, Anhui, China

[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, Anhui, China

Correspondence
Xiaoxiao Wang and Bensheng Qiu, Center for Biomedical Imaging, University of Science and Technology of China, Hefei, Anhui 230026, China.
Email: wang506@ustc.edu.cn (X. W.) and bqiu@ustc.edu.cn (B. Q.)

Funding information
Fundamental Research Funds for the Central Universities, Grant/Award Number: WK5290000002; National Natural Science Foundation of China, Grant/Award Numbers: 21876041, 81701665; University Synergy Innovation Program of Anhui Province, Grant/Award Number: GXXT-2021-003

Abstract

Decoding brain cognitive states from neuroimaging signals is an important topic in neuroscience. In recent years, deep neural networks (DNNs) have been recruited for multiple brain state decoding and achieved good performance. However, the open question of how to interpret the DNN black box remains unanswered. Capitalizing on advances in machine learning, we integrated attention modules into brain decoders to facilitate an in-depth interpretation of DNN channels. A four-dimensional (4D) convolution operation was also included to extract temporo-spatial interaction within the fMRI signal. The experiments showed that the proposed model obtains a very high accuracy (97.4%) and outperforms previous researches on the seven different task benchmarks from the Human Connectome Project (HCP) dataset. The visualization analysis further illustrated the hierarchical emergence of task-specific masks with depth. Finally, the model was retrained to regress individual traits within the HCP and to classify viewing images from the BOLD5000 dataset, respectively. Transfer learning also achieves good performance. Further visualization analysis shows that, after transfer learning, low-level attention masks remained similar to the source domain, whereas high-level attention masks changed adaptively. In conclusion, the proposed 4D model with attention module performed well and facilitated interpretation of DNNs, which is helpful for subsequent research.

KEYWORDS
attention module, brain decoding, deep learning, functional magnetic resonance imaging, neuroimaging

1 | INTRODUCTION

For many years, decoding the brain's activities has been one of the major topics in neuroscience. Inferring brain states consists of predicting the tasks subjects performed and identifying brain regions related to specific cognitive functions (Friston et al., 1994; Lv et al., 2015; McKeown et al., 1998; Norman, Polyn, Detre, & Haxby, 2006). Deep learning (DL) methods based on a variety of artificial neural networks have gained considerable attention in the scientific community for more than a decade, breaking benchmark records in several domains, including vision, speech, and natural language processing (Krizhevsky, Sutskever, & Hinton, 2017; LeCun, Bengio, & Hinton, 2015). In this context, deep neural networks (DNNs), especially convolutional neural networks (CNNs), have been recruited for brain decoding (Huang et al., 2018; Li & Fan, 2018; H. Wang et al., 2019; Yin, Li, & Wu, 2020; Zhang, Tetrel, Thirion, & Bellec, 2021), and achieved high accuracy (>90%) in brain multiple state decoding (Nguyen, Ng, Kaplan, & Ray, 2020; X. Wang et al., 2020). It is important to note, however, several open challenges still need to be addressed while using deep learning to investigate functional magnetic resonance imaging (fMRI) data.

The first challenge is the abstraction of complex temporo-spatial features within the fMRI time series. A fMRI time series is a four-dimensional (4D) data that consists of three-dimensional (3D) spatial and one-dimensional (1D) temporal information, which means brain regions engage and disengage in time during coherent cognitive activity (Chen, Kreutz-Delgado, Sereno, & Huang, 2019; Shine et al., 2016). Inspired by this, Mao et al. (2019) developed a model of 3D CNN stacks and a long short-term memory (LSTM) for spatial and temporal feature abstraction, respectively. A bit more reasonable approach would be to jointly leverage the inherent spatial–temporal information in fMRI data (Ismail Fawaz, Forestier, Weber, Idoumghar, & Muller, 2019). However, designing and optimizing architectures for 4D fMRI decoding is difficult due to the lack of systematic comparisons of various spatiotemporal processing and the substantial explosion of computational and memory requirements.

The second challenge is the researchers' requirement for a higher degree of accountability of the model, which is the core of the feasibility and reproducibility of brain decoding (Lindsay, 2020). Deep learning is regarded as a black-box model, and recent efforts have been made to develop an interpretable brain decoding model through feature ranking (Li & Fan, 2019), visualizing the convolutional kernels (Vu, Kim, Jung, & Lee, 2020), guided back-propagation (X. Wang et al., 2020), and so on. Improved DNN interpretability in fMRI analysis could lead to more accountable usage, better algorithm maintenance and improvement, and more open science (Tjoa & Guan, 2021).

Another challenge is the conflict between the DNNs' requirement for large amounts of data and the relatively modest quantity of datasets in typical cognitive research (Yotsutsuji, Lei, & Akama, 2021). Most fMRI experiments comprise tens to hundreds of participants due to experimental costs or participant selection. It is natural to use transfer learning to alleviate the data scarcity problem in the target domain (e.g., small sample datasets) by utilizing the knowledge acquired in the source domain (e.g., large cohorts; Gao, Zhang, Wang, Guo, & Zhang, 2019; Svanera et al., 2019; Thomas, Müller, & Samek, 2019; X. Wang et al., 2020). The fMRI data vary across datasets (e.g., scanner, scanning parameters, task design, template space), so it remains an open question how far the DNN can transfer-learn in fMRI.

Inspired by these challenges, the main contributions to this article are threefold. First, we extended the problem of temporal modeling and spatial feature extraction to the 4D convolution module and compared various approaches to fMRI data processing. Second, we employed the mixed attention modules to improve the decoding performance, which not only enhanced the ability to distinguish and focus on specific features but also presented an in-depth interpretation of CNN. Third, we explored the benefits of transfer learning in fMRI analysis under different problem definitions and task design, demonstrating that the model that captures cognitive similarities can extend to distinguish individual trait differences.

2 | MATERIALS AND METHODS

2.1 | Dataset

2.1.1 | Human Connectome Project dataset

The minimally preprocessed 3T data from the S1200 release of the Human Connectome Project (HCP; Glasser et al., 2013) were used in this research. The present study included task fMRI of 1,034 subjects during seven tasks: emotion, gambling, language, motor, relational, social, and working memory (WM). The seven tasks, which lasted for about 20–30 frames under different conditions during each block, provided a high degree of brain activation coverage (Barch et al., 2013). Thus, the parameter estimates of the model trained on this dataset contained similarities to multiple cognitive domains and were utilized as the source domain in the transfer learning experiment. The HCP S1200 dataset has been preprocessed with the HCP functional pipeline and normalized to the Montreal Neurological Institute's (MNI) 152 space. According to the previous studies (Nguyen et al., 2020; X. Wang et al., 2020), only one condition was selected for each task (Table 1) and resulted in 14,821 fMRI 4D instances across all subjects and tasks. To save computing memory, a bounding box with the size of [80, 96, 88] voxels was applied to each fMRI volume, and the blank parts that did not contain brain tissues were cropped out.

2.1.2 | BOLD5000 dataset

The BOLD5000 (Chang et al., 2019) dataset was also used for transfer learning of the proposed model. The dataset selected event-related design paradigms to investigate visual perception, which collected the fMRI data of four participants while viewing 5,000 real-world images.

TABLE 1 Details of the selected HCP time series

Task	Selected condition	Frames of the block
Emotion	Fear	26
Gambling	Loss	39
Language	Present story	29
Motor	Right hand	17
Relational	Relational	23
Social	Mental	32
Working memory (WM)	2-Back places	39

Each image was presented for 1 s and followed by a 9 s blank screen with a fixation cross. Thus, a single trial lasted five frames (repetition time, TR = 2 s). Two conditions of stimulus images were employed in this study: Scene containing whole scenes and ImageNet focusing on a single object. Implicit image attributes can provide category selectivity in high-level visual regions. Using fMRIPrep (Esteban et al., 2017), the preprocessing including motion correction, distortion correction, and co-registration to the corresponding T1w of the fMRI data was applied. Then each volume was also cropped to the size of [80, 96, 88] voxels, and each segmented fMRI input covered the entire trial and included two extra TRs extended forward and backward. Thus, the size of the input data was [80, 96, 88, 7].

2.2 | The proposed neural network

The proposed model consists of a 4D convolution layer and four 3D attention modules, followed by a fully-connected layer (Figure 1a).

2.2.1 | 4D convolution

The 4D convolution kernel $K \in \mathbb{R}^{k_l \times k_h \times k_w \times k_d \times k_c}$ was applied to the input $x \in \mathbb{R}^{l \times h \times w \times d \times c}$, where l is the temporal length, h is the height, w is the width, d is the depth, and c is the length of the channels. The 4D convolution operation, Conv4D, was implemented by two loops of the native 3D convolution operation, Conv3D, of the Pytorch (Paszke et al., 2019):

$$(K * x) = \sum_{i}^{k_l} \sum_{j}^{(l-k_l)/s_t+1} Conv3D_{s=2}(K(i), x(j \cdot s_t + i)),$$

where s_t is the temporal strides ($s_t = 1, 2, \ldots$) and Conv3D employed 3D convolution with a spatial stride of $s = 2$. A stride of >1 leads to a down-sample in the designated dimension. After the 4D convolution, the temporal dimension was squeezed and flattened to channel dimension of the subsequent 3D attention module.

2.2.2 | The attention module

The attention mechanism in the DNN selects focused regions and thus enhances the discriminative representation of objects (Vaswani et al., 2017). The attention module is also beneficial for optimizing by serving as a gradient update filter to prevent gradients from noisy regions. Inspired by previous researches (F. Wang et al., 2017; Woo, Park, Lee, & Kweon, 2018), we developed a 3D mixed attention module (Figure 1b), where the processing flow was split into the main branch and the attention branch. The main branch serves for feature extraction and retains effective back-propagation. The feature processing in the main branch may be any convolution network

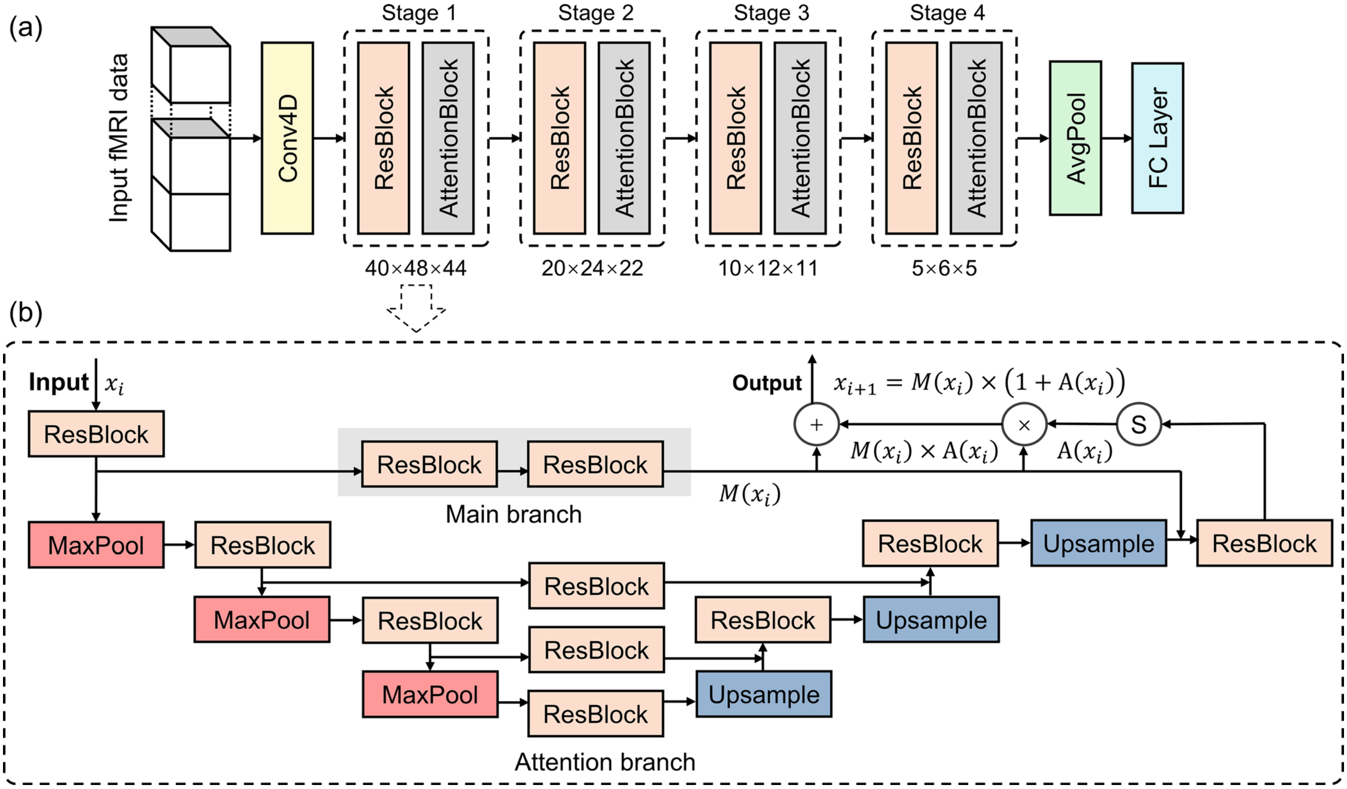

FIGURE 1 The proposed neural network. (a) The model consists of a 4D convolution layer, four 3D attention modules, and a fully-connected layer to provide labeled task classes. (b) The attention module, which includes the main branch and an attention branch composed of down-sample and up-sample paths, was connected by a shortcut skip

structure, and the ResNet block (K. He, Zhang, Ren, & Sun, 2016) was used in the present work. Formally, the output of the main branch is denoted as $M(x)$ with an input feature map x. The attention branch is a U-shaped architecture (Ronneberger, Fischer, & Brox, 2015) to mimic the feedforward and feedback attention processes. The down-sample path is built by several stacks of a 3D MaxPoll and a ResBlock to capture valuable context at multiple scales. The symmetric up-sample path consists of the same amount of trilinear interpolation and ResBlock (Figure 1b). Finally, the output was normalized by a Sigmoid function to obtain the $A(x)$.

Naive dot production of two branches degrades the value of features. Attention residual learning is used to ease this problem by constructing the attention branch as an identical mapping. Formally, the output of attention module x_{i+1} serving as the input of the next layer is modified as:

$$x_{i+1} = M(x_i) \times (1 + A(x_i)).$$

What's more, the attention mask branch can be viewed as an identical mapping that changes adaptively as layers go deeper. What the neural network learns at each level can be demonstrated by the distribution of attention. The attention masks of each channel were visualized to present an in-depth interpretation of the network by upsampling the feature map corresponding to $A(x)$ and mapping it to T1w.

2.3 | Training and evaluation

The implementation of the different model variants is based on the PyTorch framework. Training was performed on an NVIDIA GTX 1080Ti graphic card. To conduct a fair comparison, the batch size was set to 16 and each model was trained for 60 epochs using the Adam algorithm with the standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The learning rate was initialized at 0.0001 and decayed by a factor of 5 when the validation loss plateaued after 15 epochs. The loss converged well and overfitting was not observed during validation experiments. Our validation strategy employed a fivefold cross-validation across subjects and the dataset was categorized into subsets as follows: training set (70%), validating set (10%), and testing set (20%). Control experiments were conducted on various model variants (Table 2) to verify whether the 4D convolution and attention modules brought a substantial improvement. We also analyzed a set of 4DResNet consisting of different sizes of 4D kernels and presented comparison results using different frames as input. A segment of k continuous frames, which was randomly split from each instance, was used as input for training. During the testing stages, the predictions for all segmentations of one instance are summed up, and the task label with the majority vote is predicted to represent the final class of the instance.

2.4 | Transfer learning

Transfer learning describes a process in which a network is trained on a source dataset and subsequently reuses the parameters of the pretrained network that contained knowledge about the source domain on the target dataset. Transferability is an important advantage of deep learning methods compared with traditional methods in fMRI decoding. To this end, the transfer learning strategy was applied to evaluate the general use representation of the trained model.

2.4.1 | Inter-task (same dataset, different task) transfers

Since fluid intelligence (gF) measures the intelligence-related score which reflects inherent cognitive ability, there is great interest in inferring gF from fMRI data (Greene, Gao, Scheinost, & Constable, 2018; T. He et al., 2018, 2020). In the HCP data set, gF was quantified using a 24-item version of the Penn Progressive Matrices test. Here, we used the WM-trans-set, which is a subset of the HCP dataset and only contains 2-back condition of WM task data for inferring gF. The parameters of the low-level layers were adapted from the pretrained model on HCP seven tasks, and the fully connected layers were redefined and initialized. Besides, the loss function was changed to MSE Loss. To avoid leakage of individual information, the subjects which were split to pretrain/validate/test the model on the pretraining on HCP dataset were also split to the same train/validation/test partition for transfer learning on the WM-trans-set. In other words, the regression of gF should be validated and tested on new, unseen subjects that could not belong to train partitions both of pretrain and transfer learning. We evaluate the performance of transferability by comparing Spearman's correlation coefficient between the predicted gF and the observed gF of the initial model, the transferred model, and the previous work (Greene et al., 2018).

2.4.2 | Inter-datasets (different dataset, different task) transfers

BOLD5000 that selected event-related design paradigms is another small sample target dataset including with four participants. The source and target datasets are different in data statistics and distributions. The key idea of this workflow is similar to that mentioned above. We fine-tuned the model to decode binary types of stimulus images (scene vs. object) seen by subjects and employed the leave-one-subject-out (LOSO) cross-validation, which means that the data from three subjects was used to train and one to test.

3 | RESULTS

3.1 | Performance evaluation on HCP dataset

The performance of various models was compared by the mean and SD of accuracy (Table 2). All of the proposed models effectively distinguished seven tasks, with the 4DResNet-Att outperforming the others with an accuracy of 97.4% ± 0.4% (mean ± SD). Figure 2a shows the decoding performance of 4DResNet-Att on seven cognitive tasks, and the confusion matrix shows a nice block diagonal architecture. The cognitive tasks were accurately identified with the accuracy of: Emotion (96.2 ± 0.2%), gambling (99.4 ± 0.3%), language (98.7 ± 0.4%), motor (96.0 ± 0.4%), relational (93.6 ± 0.9%), social (99.4 ± 0.3%), and WM (98.9 ± 0.4%). Furthermore, the confusion matrix showed misclassifications of the relational and the gambling, the emotion and the gambling, the motor and the gambling, and the relational and the WM.

The superior performance of the 4DResNet-Att model in comparison to the 3DResNet (X. Wang et al., 2020) and other recent researchers (Nguyen et al., 2020) is possibly due to the capability to handle complex spatiotemporal dynamics in fMRI series via 4D convolution operations and the use of the attention mechanism to adaptively select a focused location.

TABLE 2 Comparisons with previous methods on the HCP dataset

Authors	Model	Accuracy ± SD
X. Wang et al. (2020)	3DResNet	93.7 ± 1.9%
Nguyen et al. (2020)	3DResNet-TF	95.1 ± 0.6%
	3DResNet-LSTM++	97.0 ± 0.4%
	3DResNet-TF++	97.2 ± 0.6%
Ours	3DResNet-Att	96.3 ± 1.1%
	4DResNet	96.1 ± 0.8%
	4DResNet-Att	**97.4 ± 0.4%**

The bolded values indicate the highest accuracy of different models.

Specifically, the 4DResNet is able to capture dynamic changes in hemodynamic response on temporal dimension and to integrate these representations from interconnected brain regions on spatial dimension. To evaluate whether 4DCNN brings a substantial improvement over 3DCNN, the 4DResNet-Att model was compared with the 3DResNet-Att model on the same brain decoding tasks using different lengths of frames as input (Figure 2b). Overall, the 4DResNet substantially enhanced classification performance compared to the 3DResNet, except for the 7-frame condition. The low performance at shorter fMRI input could be caused by two factors: (1) few information in short input, especially in series shorter than a hemodynamic response; (2) the 4DResNet tends to measure the relative dynamic change over a long range. Besides, we also evaluated a set of 4DResNet consisting of different sizes of 4D kernels to decode brain activity. Our results revealed that decoders with a short 4D-kernel size achieved lower decoding performance than decoders using a relatively longer 4D-kernel (Figure 2c).

Furthermore, to establish whether the use of attention mechanisms could enhance fMRI decoding, we compared the 4DResNet with attention modules and the naive 4DResNet. Figure 2c shows the results. The 4DResNet-Att outperformed the naive 4DResNet on the HCP dataset under different sizes of 4D kernel. In addition, the 4DResNet-Att network (about 12 hr) reduced nearly 1/3 of the training time compared with the naive 4DResNet (about 19 hr) while achieving 90% accuracy. As expected, the capability of the attention mechanism to adaptively learn the focused location brings increased performance while reducing training time.

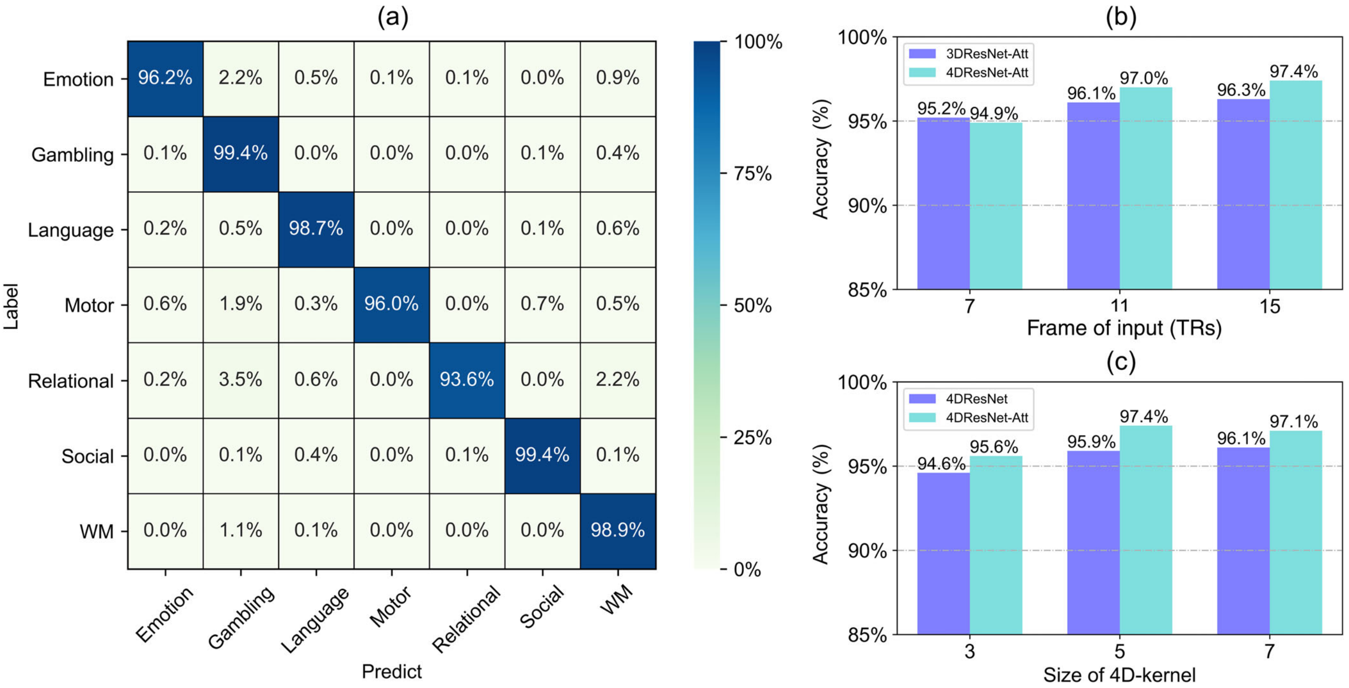

FIGURE 2 Performance evaluation on the HCP dataset. (a) The average confusion matrix showed a nice block diagonal architecture. (b) The 3DCNN and 4DCNN comparisons used different frames as input (frames = 7, 11, and 15). In terms of dynamic change over a long range, 4DCNN outperformed. (c) The classification performance with or without the attention module (frame = 15). Decoders with attention and a relatively longer 4D-kernel performed better

3.2 | Visualization of attention mask on the HCP dataset

Previous studies have employed some visualizations to build an interpretable brain decoding model in fMRI analysis (Vu et al., 2020; X. Wang et al., 2020; Yin et al., 2020). Here, we visualized the focused regions of the attention module in each convolution layer to present an in-depth interpretation of the DNN. Each channel obtained seven attention masks for different tasks, which were averaged across all of the input samples from all of the subjects.

Overall, the resulting attention masks at the low-level (first and second stages) have excellent coverage of the brain and prefer to highlight the areas containing the useful BOLD signal, such as the whole brain structure (Figure 3a), and diminish the noise areas like the brainstem or cerebrospinal fluid areas (Figure S1b,c). The masks also focused on some functional networks and cerebral cortex related to different cognitive functions (Figure S1), such as the default mode network, sensorimotor network, temporal lobe, and occipital lobe. The enhancement of gray matter areas helped to preserve the important features that could be further refined to distinguish between different cognitive states at high-level.

The attention masks at the high-level (third and fourth stages) are getting more focused to cover task-specific brain areas (Figure 3c). It is notable, however, the focused layouts of the attention masks varied across different tasks and were remarkably task-specific. A channel could generate specific focused regions for different tasks, such as the left motor cortex areas in motor task, the ventral lateral prefrontal cortex and both superior and inferior temporal cortex in language task, the prefrontal cortex in relational task, and the temporal parietal junction and superior temporal cortex regions in social task (Figures S2 and S3). At the fourth stage, the attention masks become more abstract due to the stride in the convolution operation (Figure 3d), and the weights of attention have a narrower range, which could be due to the fact that the masks also serve as gradient update filters. A small range of attention weights in the high-level feature map could prevent some gradient problems.

3.3 | Transfer learning

Two different approaches were used to explore the benefits of transfer learning in fMRI analysis under different problem definitions or task design.

First, we evaluated the general use of representation of the trained model between different problems, from cognitive similarities of group to individual trait differences in subjects. Recent research has demonstrated that connectome-based predictive modeling built from task-based fMRI data improve prediction of individual traits (Greene et al., 2018). Here, the knowledge about similarities and differences between intrinsic and task-induced brain states contained in a

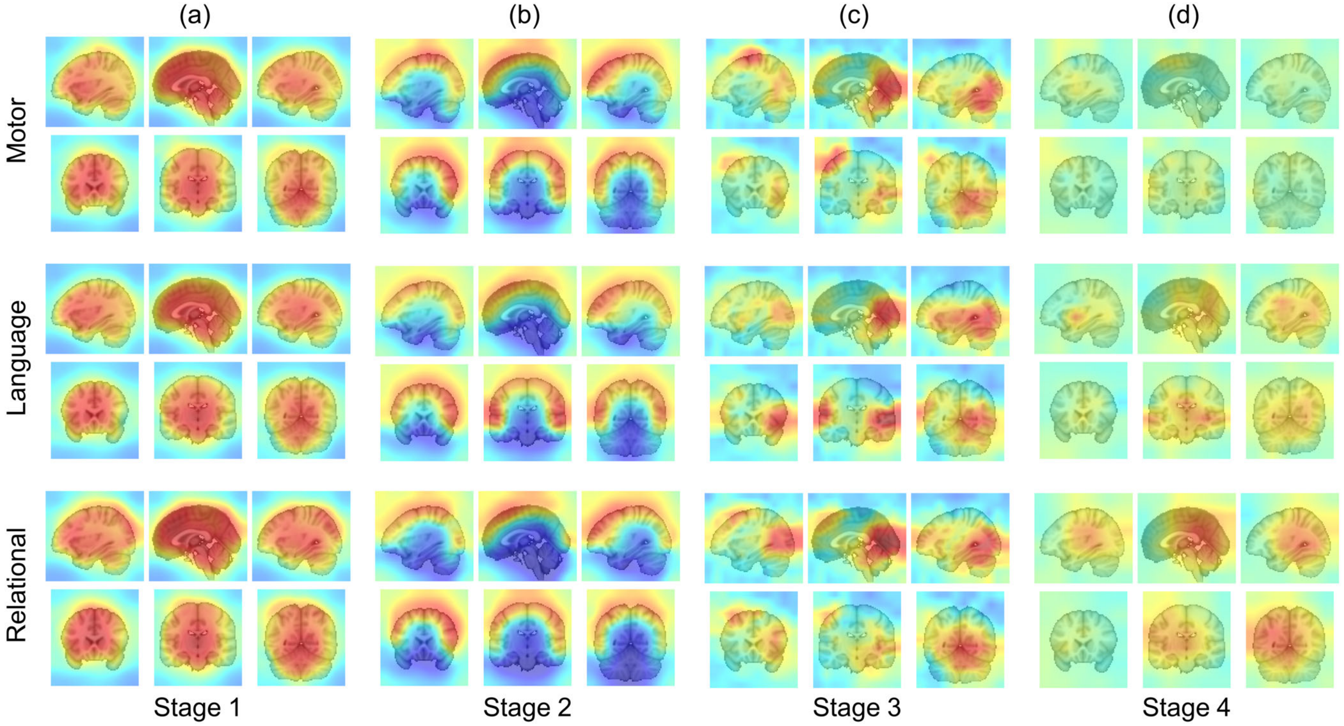

FIGURE 3 Visualization of attention masks on the HCP dataset. (a)–(d) Examples show the average focused regions on four attention stages (from low-level to high-level) of different tasks (language, motor, and relational). Each of the attention masks was color-coded with a color gradient indicating the enhancement (positive with red) or diminishment (negative with blue) of the feature maps. [Correction added on March 11, 2022, after first online publication: Figure 3 has been updated to correct the task labels in 3c.]

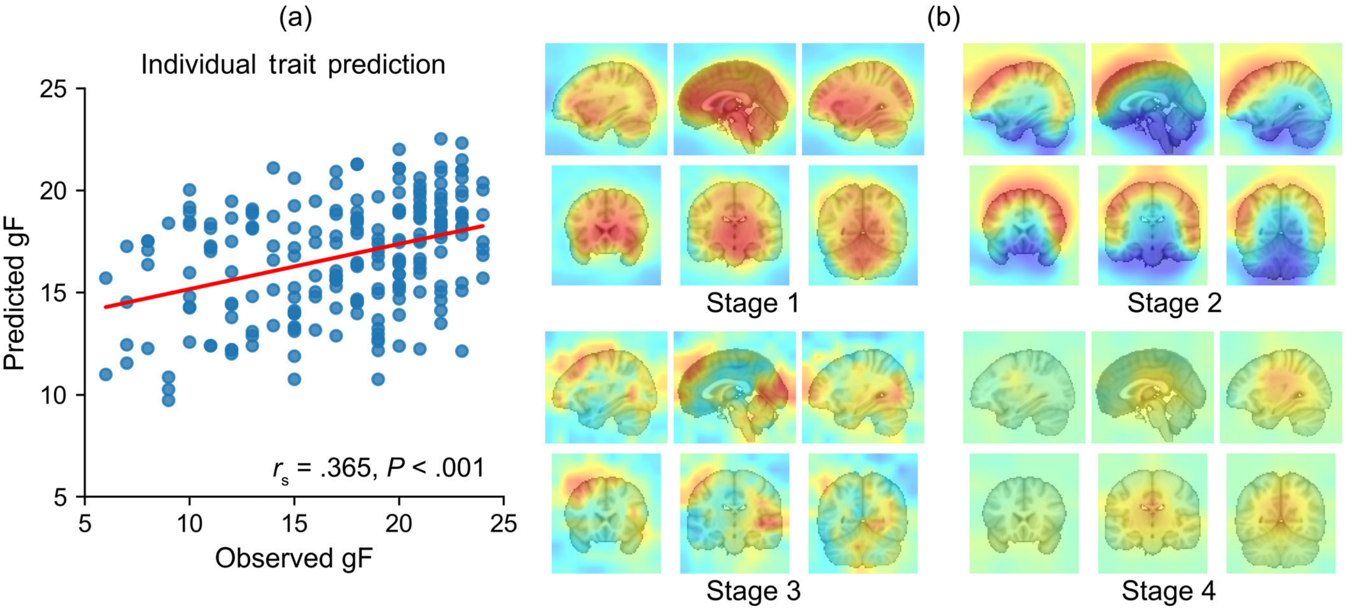

FIGURE 4 Prediction of individual traits. (a) An example showing that the transfer learning model yielded significant predictions of gF. (b) The attention masks from low-level to high-level after transfer learning. The focused regions of high-level change adaptively

pretrained model was transferred to the WM-trans-set, which is a dataset including the WM task, to predict individual trait differences. Figure 4a shows that the transferred regression model yielded significant predictions of gF. The average performance of 4DResNet-Att after transfer learning ($r_s = .354$, $p < .001$) evaluated by the average Spearman's correlation coefficient is better than the previous study that used the same dataset (Greene et al., 2018; $r_s = .325$, $p = .001$). What's more, the initial model, which used the same architecture and was trained from scratch by initializing random weights achieved a lower correlation coefficient in prediction ($r_s = .306$, $p < .001$). The comparisons of predictions between different models were shown in Table 3. Furthermore, the visualization analysis shows that low-level attention masks remained distributed similarly to the source domain, whereas high-level attention masks changed adaptively as knowledge transferred from group similarities to individual differences (Figure 4b).

Second, the pretrained model from the HCP dataset was fine-tuned to decode different types of stimulus images on BOLD5000. The knowledge learned from the source domain is highly applicable to the target domain, and the transferred model achieved 77.6 ± 3.4% (4DResNet-Att), 73.5 ± 2.1% (4DResNet), and 64.3 ± 3.8% (3DResNet-Att) accuracy. However, all initial models trained from scratch failed to converge to satisfactory accuracy (<60%) across a wide range of choices of hyper-parameters. Furthermore, the visualizations demonstrated that the attention masks changed adaptively to fit individual subjects' brain structures, despite the fact that the fMRI data were registered to the corresponding T1w space rather than the standard MNI152 space (Figure 5). As the model was fine-tuned to decode visual tasks, the attention masks from the high-levels also changed adaptively to reweight task-related brain regions.

TABLE 3 Prediction of individual traits between different model

Model	Initial training	Transfer learning
3DResNet-Att	$r_s < .3$, $p < .001$	$r_s = .329$, $p < .001$
4DResNet	$r_s < .3$, $p < .001$	$r_s = .335$, $p < .001$
4DResNet-Att	$r_s = .306$, $p < .001$	$r_s = .354$, $p < .001$

4 | DISCUSSION

4.1 | 4D convolution

Brain decoding has been a popular topic in neuroscience for decades. Recently, DNNs have gained considerable attention in the scientific community and shown promising performance in brain decoding. The fMRI data are a 4D data consisting of a time series of 3D brain volumes. 4D CNN has shown the feasibility of 4D medical applications, such as 4D computed tomography (CT; Clark & Badea, 2019) and OCT-based force estimation (Gessert, Bengs, Schluter, & Schlaefer, 2020). However, the fMRI data are big and a full 4D DNN is too large to be applied and efficiently trained. Thus, (X. Wang et al., 2020) proposed a model of 1D convolution in the first layer for the abstraction of temporal features, followed by stacks of 3D CNNs for spatial features. Mao et al. (2019) developed a network architecture that extracted spatial features out of each fMRI frame using 3D CNNs and passed these latent features to an LSTM network to take into account the temporal dependencies within task-evoked brain activity. The model we proposed includes a 4D convolution layer to detect temporo-spatial features, and puts the features into the channel dimension of the following 3D layers to reduce memory consumption. The above results suggest that the proposed model has a good balance

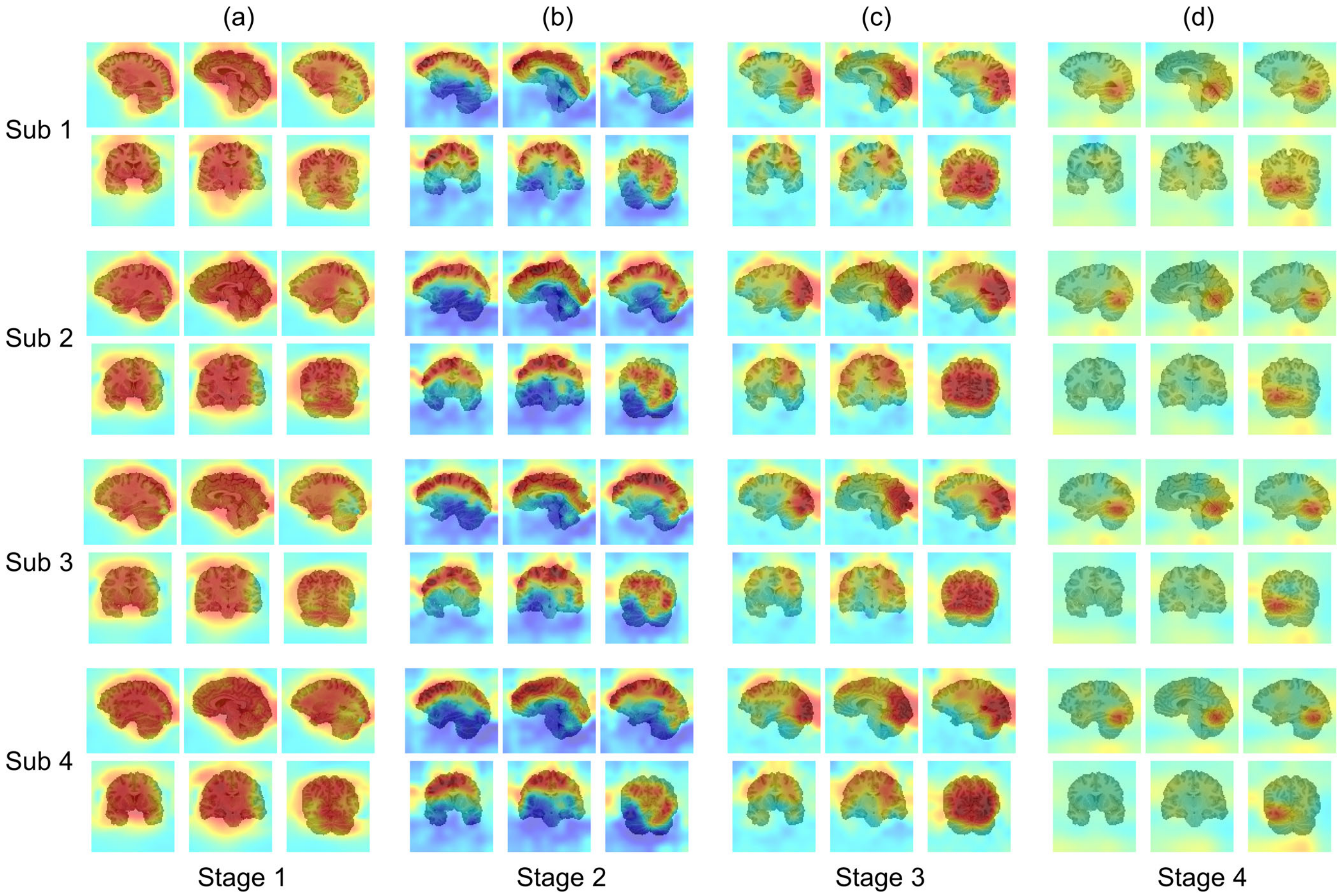

FIGURE 5 Visualization of attention masks on the BOLD5000 dataset. (a)–(d) Attention masks from low-level to high-level after transfer learning. The examples show the attention masks of four participants, which employed LOSO cross-validation. The masks adaptively change to fit different subjects' brain structures

of accuracy and efficiency. Our model could achieve better performance while taking less time than the previous state-of-art works.

4.2 | Attention module and interpretation of networks

The attention mechanism helps humans to mainly focus on the most useful information in the human perception process. Inspired by this, attention mechanisms have been studied extensively in many deep learning fields (Vaswani et al., 2017; F. Wang et al., 2017; Woo et al., 2018). In this research, the proposed 3D mixed attention module consisted of a main branch and an attention branch and considered both channel and spatial features. The experimental results demonstrate that attention modules have many advantages. For example, the architecture with attention modules was trained to converge faster and more easily and achieve better performance, which could be due to the attention mechanism reweighting the focused areas to enhance discriminative features. The attention module is also beneficial for optimizing during back-propagation, which serves as a gradient update filter to prevent noisy gradients and enhance gradients from important regions.

What's more, the attention modules not only improve decoding performance but also serve as a visualization tool to investigate how neural networks work in fMRI decoding. Cognitive neuroscience research requires a higher degree of accountability, while an end-to-end trainable network has always been regarded as a black-box in neuroscience. Presenting an in-depth interpretation of a method can demonstrate the feasibility and reproducibility of fMRI studies (Li & Fan, 2019; Vu et al., 2020). A good visual explanation should not only be treated as a localization method but also allow researchers to investigate how the neural network works. The analysis shows that the low-level masks provide excellent coverage of the brain to highlight useful structures while pruning noisy areas. As the layers go deeper, the attention masks get finer to cover various specific cortexes. The high-level attention masks varied across different tasks, re-weighting more attention to the areas related to the specific target task. What's more, the attention masks adapted to fit different subjects' brain structures. This also suggests that our architecture could be a suitable approach to avoid individual variability across subjects in the raw and minimally preprocessed fMRI series without spatial normalization. Besides, the attention areas that could present biologically meaningful interpretations of cognitive neuroscience demonstrated that the proposed CNN decoded states from task-related activations but not from nuisance variables.

4.3 | Transfer learning

Transferability has been demonstrated to be a significant advantage of DL methods over traditional methods in fMRI decoding (Gao et al., 2019; X. Wang et al., 2020). To this end, we explored the benefits of transfer learning under various conditions. The transferred regression model yielded significant predictions of individual trait differences and achieved better Spearman's correlation coefficient than the previous study (Greene et al., 2018). This could be due that the previous study relied on the discriminative power of feature selections, and not all connectivity parameters are relevant for prediction, while the transferred model could automatically capture the full range of individual trait differences. This also suggests that the group cognitive similarities among intrinsic brain states could generally be reused to predict individual differences, which is important for precision medicine in clinical research. Furthermore, previous studies most commonly applied transfer learning between the block-design dataset. On the BOLD5000, the pretrained model from the HCP dataset was fine-tuned to decode different visual tasks and obtained 77.6%. Despite the fact that the model was trained using the block-design dataset, the internal properties of human hemodynamic responses contained in the parameters are consistent and could be reused in the event-design dataset.

4.4 | Limitations and future applications

In this project, the proposed model outperformed other architectures. Despite the 4D convolution processing dynamic changes more efficiently, some limits remain, such as a substantial increase in computational and memory requirements. What's more, we only chose one condition for each cognitive domain in order to be comparable to previous studies, while the BOLD signals might be a mixture of hemodynamic responses evoked by different task events. A decoding model with fine cognitive granularity would generalize similarities and differences among task-induced brain states from multiple cognitive domains, which is important for transfer learning. The visualization result demonstrated that the high decoding performance was driven by the response of biologically meaningful brain regions. However, the statistical property of the attention mask remains unclear. We could have the results of qualitative analysis and should be cautious until further investigations into its reliability and statistical properties. The transfer learning method, which successfully extended similarities in brain activity to individual differences, showed potential for research in psychiatry and neurology. The pretrained model based on cognitive state can serve as a brain information retrieval system to distinguish differences in neurologic diseases and classify different psychiatric categories.

5 | CONCLUSION

In this study, we designed a 4DResNet with attention module for brain decoding. After investigating the efficacy of some alternative classifiers, the proposed 4DResNet-Att achieved 97.4% on the HCP dataset. We further demonstrated the model's transferability to a variety of tasks and datasets and presented an in-depth interpretation of the network. The visualization analysis of attention distributions illustrated the hierarchical emergence of task-specific masks with depth. After transfer learning, the adaptively changed attention distribution demonstrated the representation could be general extended from cognitive similarities to individual differences.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant nos. 81701665, 21876041), the Fundamental Research Funds for the Central Universities (WK5290000002), and the University Synergy Innovation Program of Anhui Province (GXXT-2021-003).

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The code and data supporting the findings of this study are available from the corresponding author upon reasonable request. The pretrained model is available at https://github.com/ustc-bmec/fMRI-Conv-Att

ORCID

Xiaoxiao Wang https://orcid.org/0000-0002-8498-7388
Bensheng Qiu https://orcid.org/0000-0003-2987-7378

REFERENCES

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., … Consortium, W. U.-M. H. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, *80*, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Chang, N., Pyles, J. A., Marcus, A., Gupta, A., Tarr, M. J., & Aminoff, E. M. (2019). BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific Data*, *6*(1), 49. https://doi.org/10.1038/s41597-019-0052-3

Chen, C.-F., Kreutz-Delgado, K., Sereno, M. I., & Huang, R.-S. (2019). Unraveling the spatiotemporal brain dynamics during a simulated reach-to-eat task. *NeuroImage*, *185*, 58–71. https://doi.org/10.1016/j.neuroimage.2018.10.028

Clark, D., & Badea, C. (2019). Convolutional regularization methods for 4D, x-ray CT reconstruction. *Medical Imaging 2019: Physics of Medical Imaging*. Proceedings of SPIE, San Diego, CA. 10948, 574–585. https://doi.org/10.1117/12.2512816

Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. *PLoS One*, *12*(9), e0184661. https://doi.org/10.1371/journal.pone.0184661

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. https://doi.org/10.1002/hbm.460020402

Gao, Y., Zhang, Y., Wang, H., Guo, X., & Zhang, J. (2019). Decoding behavior tasks from brain activity using deep transfer learning. *IEEE Access*, *7*, 43222–43232. https://doi.org/10.1109/ACCESS.2019.2907040

Gessert, N., Bengs, M., Schluter, M., & Schlaefer, A. (2020). Deep learning with 4D spatio-temporal data representations for OCT-based force estimation. *Medical Image Analysis*, *64*, 101730. https://doi.org/10.1016/j.media.2020.101730

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … Consortium, W. U.-M. H. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Greene, A. S., Gao, S., Scheinost, D., & Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, *9*(1), 2807. https://doi.org/10.1038/s41467-018-04920-3

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition (CVPR), 27–30 June 2016.

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., … Yeo, B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, *206*, 116276. https://doi.org/10.1016/j.neuroimage.2019.116276

He, T., Kong, R., Holmes, A. J., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D. & Yeo, B. T. T. … (2018). Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? International workshop on pattern recognition in neuroimaging (PRNI), 12–14 June 2018.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., … Liu, T. (2018). Modeling task fMRI data via deep convolutional autoencoder. *IEEE Transactions on Medical Imaging*, *37*(7), 1551–1561. https://doi.org/10.1109/TMI.2017.2715285

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. (2019). Deep learning for time series classification: A review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, H., & Fan, Y. (2018). Brain decoding from functional MRI using long short-term memory recurrent neural networks. *Medical Image Computing and Computer-Assisted Intervention*, *11072*, 320–328. https://doi.org/10.1007/978-3-030-00931-1_37

Li, H., & Fan, Y. (2019). Interpretable, highly accurate brain decoding of subtly distinct brain states from functional MRI using intrinsic functional networks and long short-term memory recurrent neural networks. *NeuroImage*, *202*, 116059. https://doi.org/10.1016/j.neuroimage.2019.116059

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *1–15*, 2017–2031. https://doi.org/10.1162/jocn_a_01544

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., … Liu, T. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Medical Image Analysis*, *20*(1), 112–134. https://doi.org/10.1016/j.media.2014.10.011

Mao, Z., Su, Y., Xu, G., Wang, X., Huang, Y., Yue, W., … Xiong, N. (2019). Spatio-temporal deep learning method for ADHD fMRI classification. *Information Sciences*, *499*, 1–11. https://doi.org/10.1016/j.ins.2019.05.043

McKeown, M. J., Jung, T.-P., Makeig, S., Brown, G., Kindermann, S. S., Lee, T.-W., & Sejnowski, T. J. (1998). Spatially independent activity patterns in functional MRI data during the Stroop color-naming task. *Proceedings of the National Academy of Sciences*, *95*(3), 803–810. https://doi.org/10.1073/pnas.95.3.803

Nguyen, S., Ng, B., Kaplan, A. K., & Ray, P. (2020). *Attend and decode: 4D fMRI task state decoding using attention models*. arXiv. Retrieved from https://deepai.org/publication/attend-and-decode-4d-fmri-task-state-decoding-using-attention-models

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*, 8026–8037.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. (pp.234–241). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., … Poldrack, R. A. (2016). The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron*, *92*(2), 544–554. https://doi.org/10.1016/j.neuron.2016.09.018

Svanera, M., Savardi, M., Benini, S., Signoroni, A., Raz, G., Hendler, T., … Valente, G. (2019). Transfer learning of deep neural network representations for fMRI decoding. *Journal of Neuroscience Methods*, *328*, 108319. https://doi.org/10.1016/j.jneumeth.2019.108319

Thomas, A. W., Müller, K.-R., & Samek, W. (2019). Deep transfer learning for whole-brain FMRI analyses. In *OR 2.0 context-aware operating theaters and machine learning in clinical neuroimaging*. (pp.59–67). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-030-32695-1_7

Tjoa, E. & Guan, C. (2021). Convolutional neural network interpretability with general pattern theory. arXiv: abs/2102.04247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. & Polosukhin, I. … (2017). Attention is all you need. Proceedings of the 31st international conference on neural information processing systems, Long Beach, CA.

Vu, H., Kim, H.-C., Jung, M., & Lee, J.-H. (2020). fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations. *NeuroImage*, *223*, 117328. https://doi.org/10.1016/j.neuroimage.2020.117328

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., … Tang, X. (2017). *Residual attention network for image classification*. IEEE conference on computer vision and pattern recognition (CVPR), 21–26 July 2017.

Wang, H., Zhao, S., Dong, Q., Cui, Y., Chen, Y., Han, J., … Liu, T. (2019). Recognizing brain states using deep sparse recurrent neural network. *IEEE Transactions on Medical Imaging*, *38*(4), 1058–1068. https://doi.org/10.1109/TMI.2018.2877576

Wang, X., Liang, X., Jiang, Z., Nguchu, B. A., Zhou, Y., Wang, Y., … Qiu, B. (2020). Decoding and mapping task states of the human brain via deep learning. *Human Brain Mapping*, *41*(6), 1505–1519. https://doi.org/10.1002/hbm.24891

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). *CBAM: Convolutional block attention module*. Computer vision—ECCV 2018, Cham.

Yin, W., Li, L., & Wu, F.-X. (2020). Deep learning for brain disorder diagnosis based on fMRI images. *Neurocomputing*, *469*, 332–345. https://doi.org/10.1016/j.neucom.2020.05.113

Yotsutsuji, S., Lei, M., & Akama, H. (2021). Evaluation of task fMRI decoding with deep learning on a small sample dataset. *Frontiers in Neuroinformatics*, *15*, 577451–577451. https://doi.org/10.3389/fninf.2021.577451

Zhang, Y., Tetrel, L., Thirion, B., & Bellec, P. (2021). Functional annotation of human cognitive states using deep graph convolution. *NeuroImage*, *231*, 117847. https://doi.org/10.1016/j.neuroimage.2021.117847

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Jiang, Z., Wang, Y., Shi, C., Wu, Y., Hu, R., Chen, S., Hu, S., Wang, X., & Qiu, B. (2022). Attention module improves both performance and interpretability of four-dimensional functional magnetic resonance imaging decoding neural network. *Human Brain Mapping*, *43*(8), 2683–2692. https://doi.org/10.1002/hbm.25813