Concept-Level Causal Explanation Method for Brain Function Network Classification

Jinduo Liu[1] , **Feipeng Wang**[1] and **Junzhong Ji**[1*]
[1] Beijing University of Technology
jinduo@bjut.edu.cn, wfp19981125@gmail.com, jjz01@bjut.edu.cn

Abstract

Using deep models to classify brain functional networks (BFNs) for the auxiliary diagnosis and treatment of brain diseases has become increasingly popular. However, the unexplainability of deep models has seriously hindered their applications in computer-aided diagnosis. In addition, current explanation methods mostly focus on natural images, which cannot be directly used to explain the deep model for BFN classification. In this paper, we propose a novel concept-level causal explanation method for BFN classification called CLCEM. First, CLCEM employs the causal learning method to extract concepts that are meaningful to humans from BFNs. Second, it aggregates the same concepts to obtain the contribution of each concept to the model output. Finally, CLCEM adds the contribution of each concept to make a diagnosis. The experimental results show that our CLCEM can not only accurately identify brain regions related to specific brain diseases but also make decisions based on the concepts of these brain regions, which enables humans to understand the decision-making process without performance degradation.

1 Introduction

The brain functional network (BFN) is usually constructed from functional magnetic resonance imaging (fMRI) data, which can reveal the patterns of brain functional activity. Current studies have shown that functional connectivity (FC) is an effective biomarker for the diagnosis of mental disorders, so it is of great significance and value to diagnose mental disorders based on the BFN [Seguin *et al.*2023, Steward *et al.*2023, Liu *et al.*2022]. In recent years, using deep learning methods to classify BFN has realized the auxiliary diagnosis and treatment of brain diseases, which has become a research hotspot in the medical imaging area. However, the black-box nature (uninterpretability) of deep learning methods has led to a fatal flaw in real applications [Gu *et al.*2020]. Therefore, it is necessary to explore a novel explanation method for deep learning methods in BFN classification [Hu *et al.*2021].

*Corresponding Author

Current explanation methods can be roughly divided into two categories: post-hoc methods, and built-in methods. The post-hoc methods aim to explain the decision-making process of a well-trained model by some means, including saliency map methods [Selvaraju *et al.*2019], example-based methods [Kim *et al.*2016], and perturbation-based methods [Wagner *et al.*2019]. Among them, the saliency map methods reverse-propagate the model to obtain the derivatives of input features and measure the contributions of input features to model outputs. The example-based methods pick a specific example from the dataset to explain the predictions and data distribution of models. The perturbation-based methods observe changes in the output of the model by perturbing some input data, determine the weight of the perturbation data based on the degree of changes, and explain by the perturbation data and its weight. Overall, these methods have difficulty in automatically explaining how a deep model makes decisions.

By comparison, built-in methods are apt to reveal the reasoning process of models through some customized structures or restrictions in training, which can automatically produce an explanation that humans can understand. The existing built-in methods can be roughly divided into three types: activation map-based methods [Woo *et al.*2018], protype-based methods [Li *et al.*2018] and concept-based methods [Ghorbani *et al.*2019]. Activation map-based methods use feature upsampling to obtain a weighted combination of activation maps to indicate the important pixels. Essentially, activation map-based methods belong to the pixel-level explanation, which is difficult for humans to understand. The prototype-based methods use special structures to extract prototypes in the training process, and make a decision by comparing the input samples with each prototype. However, their prototype extraction abilities may poor. Thus, they always deal with relatively simple gray image datasets, e.g., MNIST. In contrast, concept-based methods can explain the reasoning process and result in a way that humans easily understand, and the explanation effects for these methods have nothing to do with the complexity of the data. Therefore, from the perspective of convenience, the concept-based methods are better than other built-in methods in real applications [Yao *et al.*2023]. This provides a clue for explaining BFN classification. Even so, current concept-based methods usually bind extracted features with concepts by visual means, which are suitable for natural pictures. In fact, BFN data have no vi-

sual meaning, so there is an urgent need for a new method, which can bind the extracted features with concepts that are meaningful to clinicians without visual means and provide a reasonable explanation for the BFN classification process.

To accommodate this need, we propose a novel concept-level causal explanation method for BFN classification called CLCEM in this paper. The proposed method can bind the extracted features from BFN with the concepts of BFN by using the strength of causal relationships between them, and give a reasonable explanation from the view of concepts. In detail, CLCEM includes three phases: concept extraction process, concept aggregation process and decision process. CLCEM first extracts human-understandable concepts (regions of interest) from the BFN by means of causal concept loss in the concept extraction process. After that, it utilizes the location connection to aggregate the same concepts to obtain the contribution of every concept to the output of the model in the concept aggregation process. Finally, CLCEM computes the contribution of each concept for outputs to obtain a probability of whether the subject is a patient or not in the decision process. The main contributions of our paper can be summarized as follows:

- To the best of our knowledge, the proposed CLCEM is the first work that develops a concept-level causal explanation method to explain the inference process of deep learning models in the brain network classification task.
- The proposed method employs the causal concept loss, which can automatically extract human-understandable concepts (regions of interest) from BFN without the need for manual intervention.
- Systematic experiments have been conducted to verify that CLCEM can effectively explain the decision process of the BFN classification model without losing the classification performance of original deep learning models.

2 Related Works

2.1 BFN Classification Methods

Current studies have shown that the connection between brain regions is an effective biomarker for the diagnosis of brain disorders [Liu *et al.*2024, Zhang *et al.*2024]. Therefore, an increasing number of BFN classification methods have been proposed for the computer-aided diagnosis (CAD) of mental illness in the past several years. These methods can be roughly divided into two categories: traditional machine learning methods and deep learning methods. The traditional machine learning methods including linear regression (LR), support vector machine (SVM), random forest (RF), often achieve low performance without manually extracting features [de Vos *et al.*2018, Sen and Parhi2021]. Different from traditional machine learning methods, deep learning methods can automatically extract high-level features to obtain good classification performance. Therefore, using the deep learning method to classify BFNs has received increasing popularity from related researchers. Henaff [Henaff *et al.*2015] pointed out that a convolution neural network (CNN) can effectively extract the topology information of BFN with fewer parameters. Kawahara [Kawahara *et al.*2017] proposed a new CNN called BrainNetCNN that is more suitable for the special topology of BFNs by using three new convolution layers: e2e, e2n, and n2g. Ji [Ji *et al.*2021] proposed a new convolution kernel CNNEW with an elementwise weighting mechanism to extract hierarchical topological features of brain networks. Each weight is assigned to an element with a unique neuroscience significance. Although deep learning methods tend to obtain better performance more easily than traditional machine learning methods, there is a fatal defect in them called uninterpretability, which makes people unable to understand the decision process of deep models and hinders their widespread adoption in CAD [Kwon *et al.*2019].

2.2 Explanation Methods

At present, the explanation methods can be divided into two types: post-hoc, and built-in. post-hoc methods treat the entire deep model as a black box and try to explain its decisions by analyzing the external behavior of the model. However, the explanations given by the post-hoc methods illustrate particular areas valued by deep models [Selvaraju *et al.*2019, Kim *et al.*2016, Wagner *et al.*2019, Byrne2023]. These post-hoc methods regard the deep model as a black box and cannot explain the decision-making process of the models. By comparison, built-in methods [Woo *et al.*2018, Li *et al.*2018, Ghorbani *et al.*2019], which are integrated into deep models, can explain their decision-making processes. In recent years, a multitude of concept-based built-in methods have been introduced [Goyal *et al.*2019, O'Shaughnessy *et al.*2020, Heskes *et al.*2020], owing to their advanced functionality in comparison to other methods. However, current concept-based methods focus on tasks related to natural pictures and are not applicable to BFN classification [Huai *et al.*2022]. Therefore, there is an urgent need to develop novel concept-based explaination methods for BFN classification.

3 Preliminaries

3.1 Problem Description

BFN can be described as a two-dimensional matrix, where each element (feature) represents the functional connectivity strength between two brain regions. Since most existing human-brain medical studies focus on brain regions, which are also known as regions of interest (ROIs), we define the ROI as the concept to facilitate understanding the decision-making process of deep models for BFN classification. The corresponding elements of each ROI in the BFN are shown in Figure. 1, where ROI i corresponds to the elements in the i-th row and i-th column of the BFN.

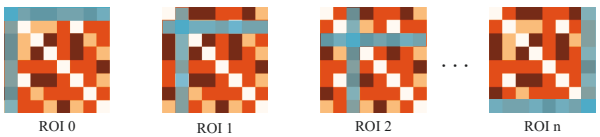

Figure 1: The corresponding elements of each ROI in BFN.

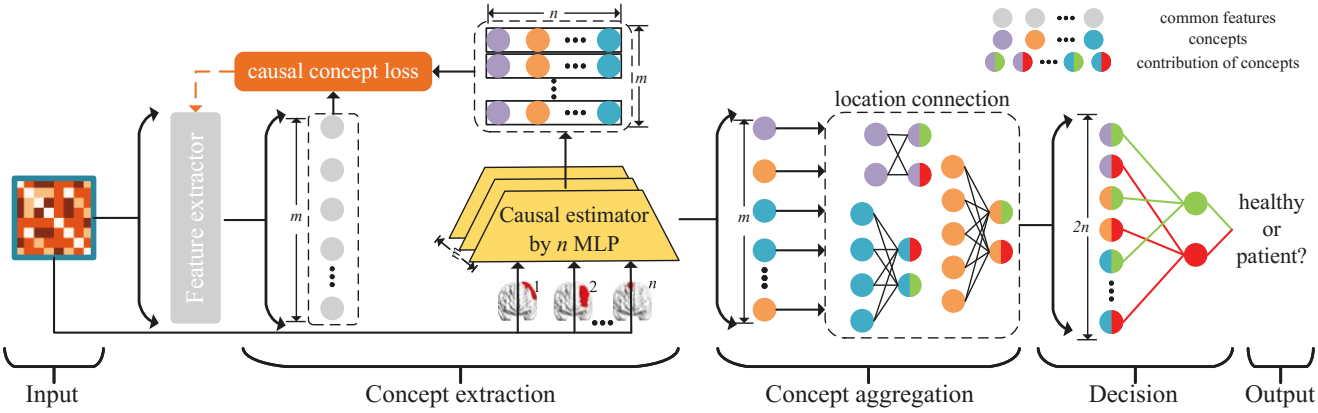

Figure 2: The schematic diagram of CLCEM. The CLCEM first utilizes the causal concept loss to extract m concepts from the BFN in the concept extraction process, subsequently aggregates the same concepts to obtain the contribution of each concept to the output of the model in the concept aggregation process by the location connection, and finally makes decisions by computing these contributions of concepts to the output in the decision process. Notably, grey balls represent common features (The features may be extracted from any deep learning model that needs to be explained. In the experiment section, we employed the most commonly used convolutional neural network (CNN) model), each color ball represents one of n concepts, and bicolor balls represent the contributions of concepts to the output, in which the left color represents the concept and the right color represents the category.

Current explanation works mostly focus on traditional natural image classification tasks, and the pixel set is easy for humans to understand, such as with a bird's head and feathers. However, for BFN classification, each element has a unique neurological significance. Limited by current human brain medical studies, the element set may be difficult for people to understand. This has also caused great obstacles to the explanation works. Therefore, it is necessary to define the concept to provide an understandable explanation in BFN classification tasks.

3.2 Notations

Let $\mathcal{X} \in \mathbb{R}^{n \times n}$ be the matrix of a BFN, where n is the number of ROIs. The label of the BFN is given by $\mathcal{Y}$. The concept set of a BFN can be denoted as $\{C_0, C_1, ..., C_{n-1}\}$, where n is the total number of ROIs. The feature set of the concept extraction process can be represented as $\{v_0, v_1, ..., v_{m-1}\}$, where m is the number of concepts which are extracted by the method. According to the definition, n and m may be different. Note that we treat ROIs (brain regions) as human-understandable concepts, so numerically they are considered equal in this paper ($m = n$).

4 Methodology

4.1 Main Idea

In this section, we propose a concept-level explanation method based on causality called CLCEM to provide an explanation of deep models for BFN classification. The proposed method can automatically extract human-understandable concepts from BFN through depth structure and use these concepts to make decisions to explain deep models. Figure. 2 shows the schematic diagram of the CLCEM. The feature extractor only plays the role of a deep feature extractor without any special requirement, which can be artificial neural network (ANN), CNN and so on. Therefore, in this paper, we focus on the concept extraction process, concept aggregation and decision process. Next, we will introduce them in detail.

4.2 Concept Extraction

From Figure. 2, we can see that the concept extraction process takes the deep features extracted from BFN as inputs and outputs human-understandable concepts. In detail, the concept extraction process first uses the causal estimator to evaluate the causal relationship strength between the deep features extracted from BFN and the FC strength of n concepts in BFN. Then, it utilizes the causal relationship strength to calculate the causal concept loss. By optimizing it, the concept extraction process can automatically extract the concepts that are meaningful to humans. Next, we introduce the causal estimator and causal concept loss, respectively.

Causal Estimator

The causal estimator is consists of n multilayer perceptrons (MLPs), which can be regarded as nonlinear structural equation modeling (SEM). They can evaluate the causal strength between n concepts of BFN and each of m features in the concept extraction process. Next, we explain how to evaluate causal strength by causal estimators.

The SEM used by the causal estimator is a special form of a Bayesian network model, and is skilled in verifying the causal relationship between continuous variables. Generally, SEM can be roughly divided into two categories: linear and nonlinear. Considering the complex nonlinear relationship between the input features and the outputs of hidden layer neurons, we choose nonlinear SEM to evaluate causal strength in the causal estimator.

For concept C_i and feature v_j, we can use nonlinear SEM

to evaluate the causal strength between them as

$$\epsilon_{i,j} = (f_{C_i,v_j}(C_i) - v_j)^2, \tag{1}$$

where $f_{C_i,v_j}(\cdot)$ is a nonlinear fitting function for v_j by C_i, and $\epsilon_{i,j}$ is the fitting error. By Eq. (1), the loss for the causal estimator can be written as

$$CEL = \sum_{i}^{n-1} \sum_{j}^{m-1} \epsilon_{i,j}, \tag{2}$$

where m is the quantity of features in the concept extraction process and n is the number of ROIs or concepts in the BFN.

After fitting well by minimizing CEL, $\epsilon_{i,j}$ is inversely proportional to the causal strength from C_i to v_j. By Eq. (1), we can obtain the causal strength from C_i to v_j as

$$s_{i,j} = \frac{1}{\epsilon_{i,j}} = \frac{1}{(f_{C_i,v_j}(C_i) - v_j)^2}. \tag{3}$$

And the $s_{i,j}$ can be used to calculate the causal concept loss for CLCEM.

Causal Concept Loss

The causal concept loss can be calculated by the causal relationship strength evaluated by the causal estimator. By minimizing it, the concept extraction process can automatically extract m concepts from BFN. Next, we will introduce how to calculate the causal concept loss.

Through the causal estimator, we have obtained the causal relationship strength from the ith concept of BFN to the jth feature in the concept extraction process as $s_{i,j}$. We can obtain the causal concept loss as

$$\begin{aligned} CCL &= \textstyle\sum_{j=0}^{m-1}(\frac{\mathcal{S}_j[1]}{\mathcal{S}_j[0]}) \\ s.t. \mathcal{S}_j &= \grave{sort}(s_{i,j}|i \in [0,n)), \end{aligned} \tag{4}$$

where $\grave{sort}(.)$ is the descending sort function. According to Eqs.(3) and (4), we can finally obtain the causal concept loss as

$$\begin{aligned} CCL &= \textstyle\sum_{j=0}^{m-1}(\frac{\mathcal{E}_j[0]}{\mathcal{E}_j[1]}) \\ s.t. \mathcal{E}_j &= \acute{sort}(\epsilon_{i,j}|i \in [0,n)), \end{aligned} \tag{5}$$

where $\acute{sort}(.)$ is the ascending sort function. By minimizing CCL, each feature of the concept extraction process is forced to maintain a strong causal relationship with one of the n concepts in BFN and weaker causal relationships with the other concepts. Therefore, we can think that there are m concepts automatically extracted from BFN by minimizing the causal concept loss in the concept extraction process.

4.3 Concept Aggregation and Decision

Through the concept extraction process, we obtained m concepts extracted from the deep features of BFN. The causal concept loss CCL cannot guarantee that there are no duplicates in m concepts. This will lead to a misunderstanding if the model uses these concepts to make decisions directly. We cannot explain the difference among the duplicate concepts. So we need the concept aggregation process to aggregate the same concepts to obtain the contribution of every concept to outputs. Then the deep model can make decisions based on these contributions of concepts in the decision process.

In the concept aggregation process, CLCEM uses location connections to aggregate the same concepts. There are $2n$ aggregated concepts in it as $\{\alpha_{i,0}, \alpha_{i,1}|i \in [0,n)\}$, and each kind of concept corresponds to two aggregated concepts, which represent its contribution to the subject as a healthy person or as a patient. When aggregating, CLCEM connects only the neurons representing the ith concept in the concept extraction process to $\alpha_{i,0}$ and $\alpha_{i,1}$.

During the decision process, CLCEM can calculate the probability that the subject is a patient or healthy person through addition as follows:

$$\begin{aligned} P(y=0) &= \textstyle\sum_{i=0}^{n-1}(\alpha_{i,0}) \\ P(y=1) &= \textstyle\sum_{i=0}^{n-1}(\alpha_{i,1}), \end{aligned} \tag{6}$$

where $y = 0$ or 1 represents that the subject is a healthy person or patient, respectively.

In other words, through the concept extraction process, concept aggregation process and decision process, we can say that the CLCEM has a transparent decision-making process, reflects the contribution of various concepts to decision-making and is highly explanatory.

4.4 Model Training

In training, the causal estimator attempts to use n concepts to fit the deep features extracted from BFN by the feature extractor. This increases the accuracy of the estimation of causal relationship strength. The causal concept loss for the feature extractor tries to extract m concepts by making each of the m features in the concept extraction process fit well by only one of the n concepts using the causal estimator. Therefore, we can also regard this training process as adversarial training, and the loss for the causal estimator and feature extractor can be separately written as $Loss_1$ and $Loss_2$

$$\begin{aligned} Loss_1 &= CEL, \\ Loss_2 &= \lambda CCL + CE(\mathcal{Y}', \mathcal{Y}), \end{aligned} \tag{7}$$

where λ (set to 0.1) is the weight of causal concept loss for concept extraction, $CE(.,.)$ is the cross entropy loss for classification, and $\mathcal{Y}'$ or $\mathcal{Y}$ represents the output label or real label vector, respectively.

The whole process of CLCEM's training is shown in Algorithm 1, where $nbepochs$ is the number of training epochs, lr is the learning rate and wd represents the $L2$ regularity coefficient.

In detail, the algorithm takes the training set ($\mathcal{X}_t$, $\mathcal{Y}_t$) and verification set ($\mathcal{X}_v$, $\mathcal{Y}_v$) as inputs and the parameters of the total network Θ as the output. $\mathcal{X}_t$ and $\mathcal{X}_v$ are the BFNs in the training and verification sets, respectively. $\mathcal{Y}_t$ or $\mathcal{Y}_v$ is the corresponding label set in the training or verification set, respectively. $\{v_0, v_1, ..., v_{m-1}\}$ is the output of the feature extractor, which can be bound to concepts by the concept extraction process. $f_{C_i,v_j}(C_i)$ is the output of the causal estimator in the concept extraction process, which tries to use concept i of BFN to fit v_j.

Algorithm 1 can be roughly divided into three stages in each epoch: first, train the causal estimator in the concept

Algorithm 1 Model Training of CLCEM

Input: $\mathcal{X}_t, \mathcal{Y}_t, \mathcal{X}_v, \mathcal{Y}_v$
Parameter: $nbepochs$, lr, wd, λ
Output: Θ

1: Initialize Adam optimizer.
2: Let $minloss = 1e^5$.
3: **for** _ in $range(nepochs)$ **do**
4: **for** _ in $range(2)$ **do**
5: Input $\mathcal{X}_t$ into the feature extractor to get the output as $\{v_0, v_1, ..., v_{m-1}\}$.
6: Input $\mathcal{X}_t$ into the causal estimator to get the output as $\{f_{C_i,v_j}(C_i), i \in [0,n), j \in [0,m)\}$.
7: Calculate $Loss_1$ by Eqs.(1), (2) and (7).
8: Minimize $Loss_1$ by $optimizer_1$.
9: **end for**
10: Input $\mathcal{X}_t$ into the causal estimator to get the output as $\{f_{C_i,v_j}(C_i), i \in [0,n), j \in [0,m)\}$.
11: Input $\mathcal{X}_t$ into the feature extractor to get the output as $\{v_0, v_1, ..., v_{m-1}\}$.
12: Calculate $Loss_2$ by Eqs.(1), (5) and (7).
13: Minimize $Loss_2$ by $optimizer_2$.
14: Input $\mathcal{X}_v$ into the causal estimator and feature extractor separately.
15: Calculate $Loss_2$ by Eqs.(1), (5) and (7).
16: **if** $minloss > Loss_2$ **then**
17: $minloss = Loss_2$.
18: Save the parameters of the total network as Θ.
19: **end if**
20: **end for**
21: **return** Θ.

extraction process on the training set by minimizing $Loss_1$ (5-10). Then, calculate $Loss_2$ by the output of the causal estimator and feature extractor on the training set and minimize it to train the feature extractor (11-14). Finally, keep the best parameters of the total network with minimal $Loss_2$ on the verification set (15-20). To clarify the training process of CLCEM, we have drawn it as Figure. 3. In practical usage, CLCEM as an explanation method, is trained end-to-end together with a BFN classification model (e.g., CNN).

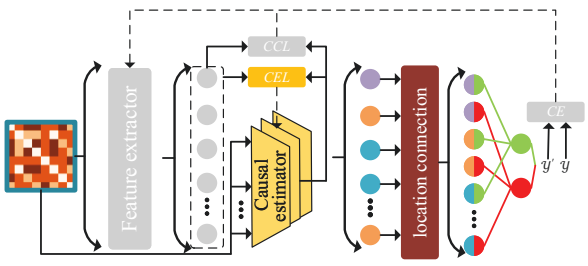

Figure 3: The training process of CLCEM.

5 Results and Discussion

In the experiments, our initial evaluation focuses on assessing the strength of the causal relationship between BFN concepts and the features of the concept extraction process. This evaluation aims to verify whether CLCEM can automatically extract human-understandable concepts based on the causal concept loss. Subsequently, the explanation's accuracy is confirmed through an assessment of the contribution of ROIs to classification. This involves a comparison of evaluation results with existing medical studies, followed by the identification of key ROIs to guide the classification process. The code is available at https://github.com/bjutAILab/CLCEM.

5.1 Configuration of the Device Operating Environment

This section outlines the computational framework used for the experiments, which includes both hardware and software components. The hardware setup features a high-performance NVIDIA GeForce RTX 3090 and RTX 3080 Ti GPUs, paired with an AMD Ryzen 9 5950X 16-Core Processor. This combination forms a powerful computing environment. The system operates on Ubuntu 20.04 LTS, providing a stable foundation for our projects. It is equipped with 64GB of high-speed DDR4 RAM, ensuring exceptional responsiveness and efficiency. For deep learning operations, we use the PyTorch framework, version 1.12.1, which includes GPU support via CUDA version 11.4.

5.2 Dataset Description

The dataset used in this paper is obtained from the ABIDE I database. It has both functional and structural brain imaging data of 1112 individuals including 539 autism spectrum disorder (ASD) patients and 573 typically developing (TD) controls, which were collected from 16 different sites around the world. Since the data processing of fMRI is very flexible, the preprocessed connectomes project provides the preprocessed fMRI data of ABIDE I, and is available at http://preprocessed-connectomes-project.org/abide/.

In this study, we chose 505 ASD patients and 530 TD controls from the dataset after removing subjects who did not have complete phenotypic information. The data employs automated anatomical labeling (AAL, 90 ROIs except for cerebellum) template and regard the mean value of all voxels in an ROI as its signal.

5.3 Metric Description

Concept Extraction Metric

In this paper, the metrics we use can be divided into two categories: causal metrics and performance metrics. For the causal metric, we mainly use it to evaluate the causal relationship strength between the concepts of BFN and the features of the concept extraction process. By this, we can verify whether the causal concept loss in CLCEM works. For a given causal structure $\mathcal{G}$, the causal metric M_{cl} can be calculated as

$$M_{cl} = -2 \log p(X; \hat{\theta}, \mathcal{G}), \tag{8}$$

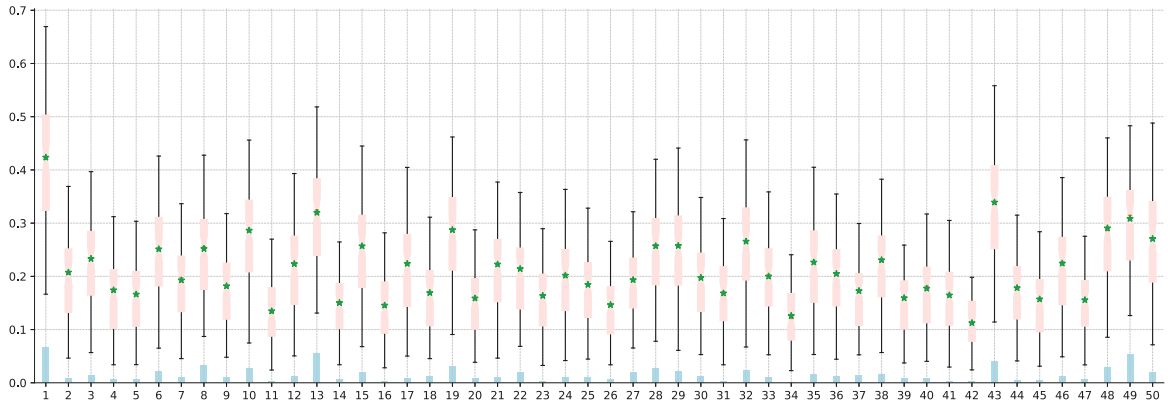

Figure 4: The M_{cl} between the randomly selected 50 concepts of BFN and the features of concept extraction process. Note that the 50 concepts represent the causal relationships between features in the concept extraction process, not the final selection of 90 concepts (ROIs).

where $X \in \mathcal{R}^{\tau \times d_\theta}$ represents the given data, $\hat{\theta}$ is the maximum likelihood estimation of $\mathcal{G}$'s parameter θ, d_θ is the dimensionality of θ or the amount of variables in $\mathcal{G}$ and τ denotes the amount of samples in X.

Then, we choose regression models to describe every causal relationship and assign the $\hat{x}_i^k$ as the corresponding estimate for x_i^k, which is the ith variable in the kth observed sample. Thus we can obtain M_{cl} as

$$M_{cl} = \sum_{i=0}^{d_\theta - 1} (\tau \log(\frac{RSS_i}{\tau})), \tag{9}$$

where $RSS_i = \sum_{k=0}^{\tau-1}(\hat{x}_i^k - x_i^k)^2$ represents the fitting error of the regression models. Moreover, the Eq. (9) is equivalent to the log-likelihood objective used by GraN-DAG, and we further assume that the variances of noise are equal. Thus, the M_{cl} can finally be written as

$$M_{cl} = d_\theta \tau \log(\frac{\sum_{i=0}^{d_\theta - 1}(RSS_i)}{d_\theta \tau}), \tag{10}$$

where d_θ is the dimensionality of θ or the amount of variables in $\mathcal{G}$ and τ denotes the amount of samples in X. $RSS_i = \sum_{k=0}^{\tau-1}(\hat{x}_i^k - x_i^k)^2$ represents the fitting error of the regression models. The smaller the M_{cl}, the stronger the causal relationship.

Classification Performance Metric

We use six performance metrics primarily to verify the correctness of the explanation and evaluate the influence of CLCEM on classification performance. Four common performance metrics include accuracy (Acc), recall (Rec), precision (Pre) and F1-score (F1). In addition, we also utilize balanced accuracy (b-Acc) and the area under the curve (AUC) to further evaluate the expected performance.

b-Acc is a metric used to evaluate the performance of classification models, particularly when dealing with imbalanced datasets where the number of instances in each class is not evenly distributed. It can be calculated by:

$$b\text{-}Acc = \frac{1}{2} \times (\frac{TP}{TP + FN} + \frac{TN}{FP + TN}), \tag{11}$$

where TP, FP, TN and FN are the number of true positive subjects, false positive subjects, true negative subjects and false negative subjects, respectively.

AUC refers specifically to the area under ROC curve in this paper, which takes false positive rate (FPR) as abscissa and true positive rate (TPR) as ordinate. They can be calculated by:

$$FPR = \frac{FP}{FP + TN}, \tag{12}$$

$$TPR = \frac{TP}{TP + FN}. \tag{13}$$

5.4 Effectiveness of Causal Concept Loss in CLCEM for Concept Extraction

To verify that CLCEM can automatically extract human-understandable concepts by the causal concept loss to explain the decision process of deep models, we evaluate the causal relationship strength between the concepts of BFN and features of the concept extraction process. In this paper, we use M_{cl} to accomplish this goal.

The M_{cl} between the concepts of the input BFN and the features of the concept extraction process are shown in Figure. 4, where the abscissa represents the serial number of features, and the ordinate represents the value of M_{cl}. In detail, the M_{cl} is calculated by Eq. (10), and then we standardized them by

$$M_{cl} = \frac{M_{cl} - min(M_{cl})}{max(M_{cl}) - min(M_{cl})}). \tag{14}$$

For the demonstration purposes, we randomly selected only 50 features from all features. Because too many features will make the graph too dense and difficult to understand. There are two different types of graphs in Figure. 4. The histogram represents the M_{cl} of the strongest causal relationship between the corresponding feature and n BFN concepts, and the box diagram represents the distribution of M_{cl} between the corresponding feature and the other $n-1$ BFN concepts. We find that all bar charts are much lower than the lower edge

of the corresponding box chart. Because M_{cl} is inversely proportional to the strength of causal relationship, we can assert that every feature maintains a strong causal relationship with one of n BFN concepts and weaker ones with the others. It is worth noting that there are thousands of features in brain network classification involving 90 brain regions. We randomly displayed 50 to convey that the relationships extracted by our method are effective.

Therefore, we can conclude that by causal concept loss, CLCEM can automatically extract concept-level causal explanations for deep models' decision processes.

5.5 Explanation Analysis of CLCEM in ASD Classification

To verify the correctness of the explanation generated by CLCEM, we used the explanation to locate the ROI highly related to ASD. Then, we checked these ROIs based on existing brain science research results. Finally, we relied only on these ROI for classification to more intuitively show the correctness of the explanation.

Next, we introduce how to use the explanation generated by CLCEM to locate the ROI highly related to ASD or contributing most to the CAD of ASD. The concept aggregation process uses location connections to aggregate the same concepts to obtain the contribution of each concept to the output. Let us abstract the concept extraction process and concept aggregation process of CLCEM into a nonlinear function Γ. Then, the output $\alpha \in \mathbb{R}^{2n}$ of the concept aggregation process for an input can be written as

$$\alpha = \Gamma(x), \tag{15}$$

where $\alpha = \{\alpha_{i,j}\}(i \in [0, n), j \in [0, 2)$, and x represents an input subject. $\alpha_{i,j}$ is the probability of the input subject's label equaling j according to the concept i.

From the perspective of classification, we think that the concept whose probability differs greatly for different classes is important. Therefore, we can use α to locate the ROI that contributes greatly to classification by

$$s_i = \frac{1}{q}\sum_{k=0}^{q-1} |\alpha_{i,0}^k - \alpha_{i,1}^k|, \tag{16}$$

where s_i is the score of ROI i and is positively correlated with the contribution to classification, $\alpha_{i,j}^k$ is the probability of the kth input subject's label equaling to j according to the concept i and q is the quantity of input subjects.

Using Eq. (16), we can obtain all ROI contributions to classification. For ease of presentation, we standardized them according to Eq. (14). For a comprehensive and detailed presentation, we use Table 1 and the Figure. 5 to display the location of the top 10 important ROIs in the human brain, which can show the important ROI more figuratively. In Table 1, "No." is the default order of ROI in the AAL template, and "score" indicates scientific counting.

To more clearly illustrate the important ROIs, Figure 5 is presented to display the locations of these ROIs. From Figure. 5 we can find that most of the functions of the ROI are closely related to some typical manifestations of ASD patients, such as lack of social contact, difficulty in communicating, stereotyped or repeated behavior and interest, lack of sensitivity to facial features and slow development of perception. Correspondingly, the LING.R is related to visual signal processing, logic analysis and visual memory. The STG.R is the language center and plays an important role in social cognition. The ANG.L is mainly involved in the language and digit processing of human brain and in processes related to attention. The CAL.L is related to visual function. The ORBsupmed.R is believed to contribute to cognitive function. The MTG.R is associated with different processes that recognize known faces, semantic memory processing, integrate multimodal sensory, etc. Finally, the FFG.R is important in object and face recognition.

Several existing medical studies showed that changes in these ROIs are directly and significantly associated with the onset of ASD. Keehn [Jao Keehn *et al.*2017] found that compared with TD, the ASD group showed a bilateral cluster of

No.	ROI	abbreviation	score
48	Lingual gyrus	LING.R	$1.00e^{0}$
82	Superior temporal gyrus	STG.R	$9.99e^{-1}$
21	Olfactory cortex	OLF.L	$4.37e^{-1}$
65	Angular gyrus	ANG.L	$1.47e^{-1}$
43	Calcarine fissure and surrounding cortex	CAL.L	$1.46e^{-1}$
26	Superior frontal gyrus (medial orbital)	ORBsupmed.R	$1.14e^{-1}$
86	Middle temporal gyrus	MTG.R	$7.80e^{-2}$
56	Fusiform gyrus	FFG.R	$7.24e^{-2}$
44	Calcarine fissure and surrounding cortex	CAL.R	$6.83e^{-2}$
88	Temporal pole: middle temporal gyrus	TPOmid.R	$6.36e^{-2}$

Table 1: The score of the top 10 ROIs evaluated by CLCEM.

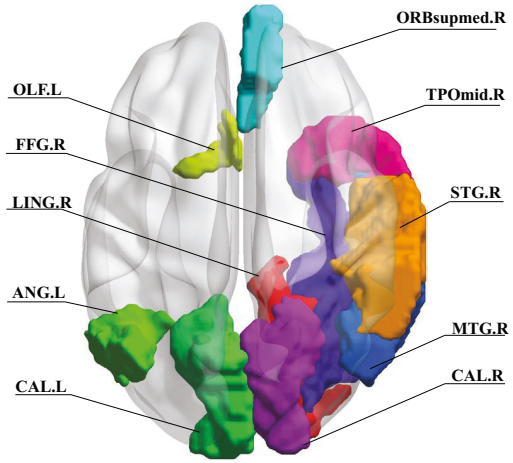

Figure 5: Visualization of top 10 important ROIs selected by CLCEM in human brain.

ROI Selection Methods	Acc(%)↑	Rec(%)↑	Pre(%)↑	F1(%)↑	b-Acc(%)↑	AUC(%)↑
AAL (All regions)	65.02±4.47	64.05±8.84	64.62±6.44	63.82±5.43	65.25±4.45	69.45±4.58
LIME (2016)	66.09±3.87	69.80±11.26	64.92±3.31	66.84±5.86	65.95±4.11	70.23±4.78
Grad-CAM (2019)	65.80±3.81	66.52±12.88	64.98±5.46	64.85±6.35	66.09±3.98	69.63±4.62
FGVis (2019)	65.26±2.92	65.11±8.98	66.87±4.87	65.39±3.35	65.33±3.11	69.78±4.16
AutoRMI (2022)	65.96±2.43	64.08±8.02	66.15±4.05	64.71±4.14	66.04±2.45	69.98±2.88
CMIC (2023)	66.18±3.78	65.92±9.97	66.06±5.06	65.09±6.82	66.31±3.81	69.94±5.15
CLCEM (Ours)	**68.37±2.37**	**71.27±6.07**	**67.62±3.73**	**69.20±3.03**	**68.16±2.57**	**71.75±2.67**

Table 2: The BFN classification performance of CNN using various ROIs selected by different explanation methods. The AAL atlas comprises all 90 ROIs, while other explanation methods select the top 30 most important ROIs from 90 ROIs.

activation with peak intensity in the right lingual gyrus, which has a more significant effect on the activity of the left lingual gyrus. Kana [Kana *et al.*2016] studied the FC difference in age- and IQ-matched adults with and without ASD and found that the brain activation level of the superior temporal gyrus in ASD patients decreased significantly during implicit emotion processing. David [Menassa *et al.*2017] believes that ASD is characterized by sensory abnormalities, including impaired olfactory recognition, and changes in the cellular structure of the olfactory cortex may be the basis for the olfactory differences in ASD. Li [Li *et al.*2014] thinks that the local connectivity of the angular gyrus in ASD patients is stronger to support their cognitive function. Libero [Libero *et al.*2014] collected fMRI data from ASD patients and TD individuals when judging the model's actions and intentions, and found that patients with ASD had significantly reduced activation in the calcarine sulcus compared with TD individuals when attending to the intentions of actions. Yerys [Yerys *et al.*2015] found that children with ASD demonstrated increased activation in the superior frontal gyrus during the switch-vs-stay contrast compared to controls. Cauda [Cauda *et al.*2011] analyzed the voxel-based morphometry findings and found that gray matter in the middle temporal gyrus was significantly increased in ASD patients. Van [van Kooten *et al.*2008] found that the activity of the fusiform gyrus and other cortical regions supporting face processing in ASD patients decreased.

5.6 Comparative Analysis of Explanation Methods

Through the above demonstration, we have proven the correctness of the explanation given by CLCEM from the perspective of the brain mechanism. However, it is still uncertain whether CLCEM is more effective in aligning with deep classification models of BFN when elucidating significant brain regions, compared to other explanation methods. Indeed, CLCEM is a general explanation method for BFN classification (suitable for most deep models), rendering direct comparisons with established methods challenging. To validate its advantages in explaining BFN classification tasks, we indirectly assess whether the features explained by CLCEM are the most significant for model decisions, in comparison with other explanation methods.

In our experiments, we utilized extensively adopted and high-performing deep learning models CNN [Henaff *et al.*2015] for brain network classification. The explanation methods encompass: LIME [Ribeiro *et al.*2016], Grad-CAM [Selvaraju *et al.*2019], FGVis [Wagner *et al.*2019], AutoRMI [Huai *et al.*2022], and CMIC [Yao *et al.*2023]. In detail, we use the explanations provided by CLCEM and other state-of-the-art methods to identify the top 30 most important ROIs, and then employ CNN to classify BFNs using only these ROIs. We choose the top 30 ROIs, because most explanation methods could yield approximately optimal results. The number of brain regions selected can be 10, 30, 70, or any number less than the total of 90 brain regions. We chose 30 for the presentation of results because preliminary testing with different quantities showed that selecting 30 regions provides a balance where the number is not too high, yet the performance is significantly improved. In particular, the dataset was randomly divided into 10 equal parts, and we conducted a 10-fold cross validation. For fair comparison, different numbers of ROIs have the same network structure except that the input layer was finely tuned to accommodate different scales of BFN. The results are shown in Table 2 by $mean \pm stddev$.

In Table 2, we can see that the performance of models using the top 30 most important ROIs is better than the performance of models using all ROIs for most of the methods. And CLCEM demonstrates the most outstanding performance, achieving optimal results across all metrics. Therefore, we can conclude that the explanation (important ROIs) obtained from CLCEM is consistent with the characteristics that the deep classification model actually values.

In summary, through a series of experiments and analysis, we have established the correctness and advantages of CLCEM's exposition across three dimensions: brain neural mechanism, existing medical research, and classification model performance. Considering these factors, we posit that CLCEM's explanation is comprehensible to humans and aligns with the decision logic of the deep model.

6 Conclusion and Future Work

To explain deep models for BFN classification, we proposed a novel concept-level causal explanation method called CLCEM. The proposed method can automatically extract human-understandable concepts from BFN under the constraints of causal concept loss and use these concepts to explain the decision logic of deep models. CLCEM provides a better explanation than other kinds of explanation methods and does not require the input data to be human-understandable natural pictures.

In the future, we will study the theoretical aspects and extensibility of CLCEM in BFN classification, and explore the use of CLCEM in multi-class brain disease diagnosis tasks. Additionally, we will also attempt to incorporate other easily comprehensible brain information as concepts.

Acknowledgements

Thanks for the professional and constructive suggestions from the reviewers. This work was partly supported by National Natural Science Foundation of China Research Program (62106009, 62276010), in part by R&D Program of Beijing Municipal Education Commission (KM202210005030, KZ202210005009).

References

[Byrne, 2023] Ruth MJ Byrne. Good explanations in explainable artificial intelligence (xai): Evidence from human explanatory reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6536–6544, 2023.

[Cauda *et al.*, 2011] Franco Cauda, Elisabetta Geda, Katiuscia Sacco, Federico D'Agata, Sergio Duca, Giuliano Geminiani, and Roberto Keller. Grey matter abnormality in autism spectrum disorder: an activation likelihood estimation meta-analysis study. *Journal of Neurology, Neurosurgery & Psychiatry*, 82(12):1304–1313, 2011.

[de Vos *et al.*, 2018] Frank de Vos, Marisa Koini, Tijn M Schouten, Stephan Seiler, Jeroen van der Grond, Anita Lechner, Reinhold Schmidt, Mark de Rooij, and Serge ARB Rombouts. A comprehensive analysis of resting state fmri measures to classify individual patients with alzheimer's disease. *Neuroimage*, 167:62–72, 2018.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32:9277–9286, 2019.

[Goyal *et al.*, 2019] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

[Gu *et al.*, 2020] Donghao Gu, Yaowei Li, Feng Jiang, Zhaojing Wen, Shaohui Liu, Wuzhen Shi, Guangming Lu, and Changsheng Zhou. Vinet: A visually interpretable image diagnosis network. *IEEE Transactions on Multimedia*, 22(7):1720–1729, 2020.

[Henaff *et al.*, 2015] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

[Heskes *et al.*, 2020] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

[Hu *et al.*, 2021] Wenxing Hu, Xianghe Meng, Yuntong Bai, Aiying Zhang, Gang Qu, Biao Cai, Gemeng Zhang, Tony W Wilson, Julia M Stephen, Vince D Calhoun, et al. Interpretable multimodal fusion networks reveal mechanisms of brain cognition. *IEEE transactions on medical imaging*, 40(5):1474–1483, 2021.

[Huai *et al.*, 2022] Mengdi Huai, Jinduo Liu, Chenglin Miao, Liuyi Yao, and Aidong Zhang. Towards automating model explanations with certified robustness guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6935–6943, 2022.

[Jao Keehn *et al.*, 2017] R Joanne Jao Keehn, Sandra S Sanchez, Claire R Stewart, Weiqi Zhao, Emily L Grenesko-Stevens, Brandon Keehn, and Ralph-Axel Müller. Impaired downregulation of visual cortex during auditory processing is associated with autism symptomatology in children and adolescents with autism spectrum disorder. *Autism Research*, 10(1):130–143, 2017.

[Ji *et al.*, 2021] Junzhong Ji, Xinying Xing, Yao Yao, Junwei Li, and Xiaodan Zhang. Convolutional kernels with an element-wise weighting mechanism for identifying abnormal brain connectivity patterns. *Pattern Recognition*, 109:107570, 2021.

[Kana *et al.*, 2016] Rajesh K Kana, Michelle A Patriquin, Briley S Black, Marie M Channell, and Bruno Wicker. Altered medial frontal and superior temporal response to implicit processing of emotions in autism. *Autism Research*, 9(1):55–66, 2016.

[Kawahara *et al.*, 2017] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[Kim *et al.*, 2016] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29:2288–2296, 2016.

[Kwon *et al.*, 2019] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, 25(1):299–309, 2019.

[Li *et al.*, 2014] Hai Li, Zhong Xue, Timothy M Ellmore, Richard E Frye, and Stephen TC Wong. Network-based analysis reveals stronger local diffusion-based connectivity and different correlations with oral language skills in brains of children with high functioning autism spectrum disorders. *Human brain mapping*, 35(2):396–413, 2014.

[Li *et al.*, 2018] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Libero *et al.*, 2014] Lauren E Libero, Jose O Maximo, Hrishikesh D Deshpande, Laura G Klinger, Mark R Klinger, and Rajesh K Kana. The role of mirroring and mentalizing networks in mediating action intentions in autism. *Molecular autism*, 5(1):1–13, 2014.

[Liu *et al.*, 2022] Jinduo Liu, Junzhong Ji, Guangxu Xun, and Aidong Zhang. Inferring effective connectivity networks from fmri time series with a temporal entropy-score.

IEEE transactions on neural networks and learning systems, 33(10):5993–6006, 2022.

[Liu *et al.*, 2024] Jinduo Liu, Lu Han, and Junzhong Ji. Mcan: Multimodal causal adversarial networks for dynamic effective connectivity learning from fmri and eeg data. *IEEE Transactions on Medical Imaging*, 2024.

[Menassa *et al.*, 2017] David A Menassa, Carolyn Sloan, and Steven A Chance. Primary olfactory cortex in autism and epilepsy: increased glial cells in autism. *Brain Pathology*, 27(4):437–448, 2017.

[O'Shaughnessy *et al.*, 2020] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 33:5453–5467, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Seguin *et al.*, 2023] Caio Seguin, Olaf Sporns, and Andrew Zalesky. Brain network communication: concepts, models and applications. *Nature reviews neuroscience*, 24(9):557–574, 2023.

[Selvaraju *et al.*, 2019] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019.

[Sen and Parhi, 2021] Bhaskar Sen and Keshab K Parhi. Predicting biological gender and intelligence from fmri via dynamic functional connectivity. *IEEE Transactions on Biomedical Engineering*, 68(3):815–825, 2021.

[Steward *et al.*, 2023] Anna Steward, Davina Biel, Matthias Brendel, Anna Dewenter, Sebastian Roemer, Anna Rubinski, Ying Luan, Martin Dichgans, Michael Ewers, Nicolai Franzmeier, et al. Functional network segregation is associated with attenuated tau spreading in alzheimer's disease. *Alzheimer's & Dementia*, 19(5):2034–2046, 2023.

[van Kooten *et al.*, 2008] Imke AJ van Kooten, Saskia JMC Palmen, Patricia von Cappeln, Harry WM Steinbusch, Hubert Korr, Helmut Heinsen, Patrick R Hof, Herman van Engeland, and Christoph Schmitz. Neurons in the fusiform gyrus are fewer and smaller in autism. *Brain*, 131(4):987–999, 2008.

[Wagner *et al.*, 2019] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Yao *et al.*, 2023] Liuyi Yao, Yaliang Li, Sheng Li, Jinduo Liu, Mengdi Huai, Aidong Zhang, and Jing Gao. Concept-level model interpretation from the causal aspect. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8799–8810, 2023.

[Yerys *et al.*, 2015] Benjamin E Yerys, Ligia Antezana, Rachel Weinblatt, Kathryn F Jankowski, John Strang, Chandan J Vaidya, Robert T Schultz, William D Gaillard, and Lauren Kenworthy. Neural correlates of set-shifting in children with autism. *Autism Research*, 8(4):386–397, 2015.

[Zhang *et al.*, 2024] Zuozhen Zhang, Junzhong Ji, and Jinduo Liu. Metarlec: Meta-reinforcement learning for discovery of brain effective connectivity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10261–10269, 2024.