From Brains to Artificial Intelligence: Exploring Brain-Inspired AI

by

Lin Zhao

(Under the Direction of Tianming Liu)

Abstract

Inspired by biological neural networks (BNNs), artificial neural networks (ANNs) have achieved great success in revolutionizing a wide range of tasks and scenarios from computer vision (CV) to natural language processing (NLP). Given their powerful representation capabilities, ANNs have also been widely used in the brain science community to represent the organization and dynamics of the human brain from the perspective of BNNs, such as functional brain networks (FBNs). Despite this, the connections between ANNs and BNNs remain largely unexplored due to the lack of effective tools to bridge and connect two different domains, i.e., the brain and artificial intelligence. Furthermore, how to leverage the prior knowledge of BNNs to inspire the design of ANNs and boost their performance is still an open question. To overcome these challenges, we proposed a series of computational frameworks to bridge the gap mentioned above. Our approaches involve exploring the hierarchical organization of brain activities, representing the brain structure and function as embeddings, connecting them with ANNs to couple the semantics of two domains, and utilizing the prior knowledge from the human brain to inspire and guide the design of ANNs. Extensive experiments demonstrated that the proposed computational frameworks could effectively explore the connection between ANNs and BNNs, yielding neuroscientifically meaningful interpretations. Additionally, our brain-inspired design of ANNs, informed by prior knowledge from human brains, achieved comparable and state-of-the-art performances in several tasks. Overall, this study provides novel insights from brains toward brain-inspired artificial intelligence.

Index words: [Brain-inspired AI, Brain, Biological Neural Network, Artificial Neural Network, Functional Brain Network]

From Brains to Artificial Intelligence: Exploring Brain-Inspired AI

by

Lin Zhao

B.S., Northwestern Polytechnical University, China, 2017

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2023

From Brains to Artificial Intelligence: Exploring Brain-Inspired AI

by

Lin Zhao

Major Professor: Tianming Liu

Committee: Sheng Li
Ninghao Liu

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
May 2023

Dedication

This dissertation is dedicated to my loving parents, Jiangpu Zhao and Aijun Xing, whose endless love and unwavering support have been a constant source of power throughout my life and academic journey. I am forever grateful for their sacrifices, guidance, and encouragement that have shaped me into the person I am today. I would also like to dedicate this dissertation to my wife, Qiaohui Gao, for her love, companionship, and unwavering support during my Ph.D. study.

Acknowledgments

I would like to express my most sincere gratitude to Professor Tianming Liu, for his invaluable guidance, support, and encouragement throughout my doctoral study. Prof. Liu's vast knowledge, insightful comments, and constructive criticism have been essential in shaping my research and helping me grow as a researcher. His passion for research and his dedication to his students are truly inspiring. I am fortunate to have had the opportunity to grow under his mentorship for both research and life over the past five years. I will always be grateful for the knowledge and skills he has imparted to me, the resources and opportunities he provided, and all he has done for me.

I would like to express my special thanks to my Ph.D. advisory committee: Dr. Sheng Li and Dr. Ninghao Liu, for their valuable contributions and insightful comments throughout my doctoral journey. Their constructive feedback, encouragement, and support have helped me overcome the obstacles that I faced along the way. I am honored and privileged to have had them as members of my Ph.D. advisory committee, and I will always be grateful for their mentorship and friendship.

I would like to express my gratitude to Dr. Dajiang Zhu, Dr. Tuo Zhang, and Dr. Xi Jiang, for their unwavering encouragement and insightful feedback, which have been instrumental in helping me grow as a researcher and scholar. Furthermore, I could not accomplish all of my achievements without the assistance of my colleagues in our laboratory. It has been a pleasure to work with such a creative and supportive group, which makes my doctoral life nothing but remarkable.

Specifically, I would like to express my sincere appreciation to my mentors during the internship at UII America and Alibaba. They selflessly shared their expertise in both academic and industrial fields, providing me with invaluable insights and guidance that will serve as a foundation for my future endeavors. They are Dr. Xiao Chen, Dr. Shanhui Sun, Dr. Yuxing Tang, Dr. Ling Zhang, and Dr. Le Lu. I believe the mentorship they provided will help me make significant contributions to society in the future.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Connection of Brain and Artificial Intelligence

Inspired by biological neural networks (BNNs), artificial neural networks (ANNs) have achieved great success in revolutionizing a wide range of tasks and scenarios from computer vision (CV) to natural language processing (NLP) (LeCun et al., 2015). With a biologically plausible inspiration from the cat visual cortexHubel and Wiesel, 1959; LeCun, Bengio, et al., 1995, convolutional neural networks (CNNs) (LeCun, Bengio, et al., 1995) hierarchically learn the visual representations of images/videos as low-level to high-level features and have been widely used in many real-word CV applications (Khan et al., 2020; LeCun et al., 2015). Inspired by human visual attention, a large group of deep learning approaches successfully integrated attention mechanisms into their deep neural networks to improve the performance and the interpretability (Hassabis et al., 2017; Vaswani et al., 2017). For example, in the CV field, a lightweight attention module was introduced into CNNs and demonstrated consistent improvements on both image classification and object detection tasks (Woo et al., 2018). In the NLP field, based on the self-attention mechanism, the Transformer model (Vaswani et al., 2017) has been widely adopted in NLP tasks such as machine translation, text classification, and question answering. Based on the Transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) model achieved state-of-the-art results on a wide range of benchmarks, becoming a popular choice for NLP tasks due to its strong performance and ease of use with pre-trained models available for many languages. Recently, by adapting self-attention mechanism into image processing, Vision Transformer (ViT) Dosovitskiy et al., 2020 model demonstrated state-of-the-

art performance in various CV tasks by representing the image as a sequence of patches.

Concurrently, ANNs have also been widely adopted in the brain science community to model the brain structure, represent the organization and dynamics of brain function, to help us understand the mechanism of the human brain. For example, deep convolutional autoencoder (DCAE) model was used to extract dozens of features from functional magnetic resonance imaging (fMRI) time series to explore the temporal characteristics of brain activities (H. Huang et al., 2017; L. Zhao, Dai, et al., 2021). Deep sparse recurrent autoencoder (DSRAE) (Q. Li et al., 2019) was employed to simultaneously decompose BNNs, i.e., the spatiotemporal FBNs, from 4D fMRI volumes. CNNs were utilized to characterize the rhythm of brain activity in gyri and sulci (M. Jiang et al., 2020; H. Liu et al., 2019; Q. Wang et al., 2022), and it was found that gyral/sulcal signals exhibit different frequency characteristics across different brain regions. Recently, a group of studies employed Transformer-based models and self-attention mechanism to model brain activity and function given the powerful ability to model sequential data (Dong et al., 2020; M. He et al., 2022; L. Zhao, Wu, et al., 2022). Despite the successes of the aforementioned ANNs, the architectures of those ANNs are manually crafted, which may not be optimized for addressing domain-specific questions such as representing brain functional activity. Neural Architecture Search (NAS) provided feasible approaches for automatically selecting the optimal architecture that may be congruent with the human brain. For example, W. Zhang et al., 2019 adopted an evolutionary optimization method to search the optimal deep belief networks (DBNs) for identifying spatiotemporal FBNs. Q. Li, Wu, et al., 2021 take advantage of the differentiable method to search the optimal spatial/temporal brain function network decomposition.

1.2 Challenges in Brain-inspired AI

Despite the progress mentioned above, the connections between ANNs and BNNs remain largely unexplored due to the lack of effective tools to bridge and connect two different domains. For example, whether the visual representation space of ANNs such as CNNs or ViT retains biologically meaningful semantics as in the initial inspiration, BNNs, is still an open question. The challenges for answering this question are multi-fold. Even though the natural stimulus fMRI paradigm provides a powerful tool for investigating visual perception, current approaches for representing high-dimensional fMRI data are limited in interpreting and describing the semantics perceived by the human brain given the

complexity of dynamic brain activities. Meanwhile, the brain responses evoked by naturalistic stimuli exhibit great inter-subjects variability (Golland et al., 2007; Ren et al., 2017), while the existing methods do not encode the regularity and variability of different brains, and thus do not offer a general, comparable, and stereotyped embedding space for representing the brain activity and semantics. In addition, the development of computational methods for linking the functional embedding and semantic representation of the human brain with external natural stimuli remains highly unexplored and desirable.

How we can utilize the prior knowledge from human brains to inspire and optimize the ANNs for better performances is another challenging problem. Recently, a group of studies suggested that artificial neural networks (ANNs) and biological neural networks (BNNs) may share common principles in optimizing the network architecture. For example, the property of small-world in brain structural and functional networks are recognized and extensively studied in the literature (Bassett & Bullmore, 2017; Bassett & Bullmore, 2006; Bullmore & Sporns, 2009). Interestingly, in (S. Xie et al., 2019), the neural networks based on Watts-Strogatz (WS) random graphs with small-world properties yield competitive performances compared with hand-designed and NAS-optimized models. Through quantitative post-hoc analysis, (You et al., 2020) found that the graph structure of top-performing ANNs such as CNNs and multilayer perceptron (MLP) is similar to those of real BNNs such as the network in macaque cortex. These studies suggest the potential of taking advantage of prior knowledge from brain science to guide the model architecture design. However, the practical way of incorporating prior knowledge of the brain into neural network design remains uncharted territory, i.e., there is significant potential for exploration and advancement in the area of brain-inspired AI.

1.3 Contributions

To address these challenges and explore brain-inspired AI, we proposed a series of computational frameworks to bridge the gap between the brain and artificial intelligence. Our approaches involve exploring the hierarchical organization of brain activities, representing the brain structure and function as embeddings, coupling the semantics of two domains, and utilizing the prior knowledge from the human brain to inspire the design of ANNs. The contributions of this dissertation are summarized as follows:

- We proposed a novel hierarchical interpretable autoencoder (HIAE) model for representing the brain functional activities as hierarchical spatiotemporal features across different scales, which is interpretable and meaning-

ful. The analysis based on it indicated hierarchical functional differences between gyri and sulci and suggested a potential core-periphery organization of human brain function.

- We developed a novel approach to encode the cortical region as a dense embedding vector. The derived embeddings can simultaneously profile the structural, connectional and functional information of a brain region and provide a feasible solution for representing the regularity and variability of human brain architectures.

- A novel and generic embedding framework was proposed for representing the human brain function in a general, comparable, and stereotyped space. The learned embeddings are neuroscientifically meaningful and interpretable, paving the road for bridging the gap of semantic spaces between human brain function and ANNs.

- We proposed a synchronized activation (Sync-ACT) framework to connect the visual representation and semantics between FBNs and ANNs. Significant correlations were found between the two domains' semantics and the relationship with the CNNs performance in image classification tasks.

- Inspired by the biased competition process in the brain's visual system, we developed a brain-inspired adversarial visual attention network (BI-AVAN) to characterize and decode the visual attention in movie watching. Experimental results show that BI-ANAN achieves robust and promising results when inferring meaningful human visual attention and mapping the relationship between brain activities and visual stimuli.

- We proposed a novel eye-gaze guided vision transformer (EG-ViT) model to infuse the human expert's visual attention to guide the model focusing on the region with potential pathology, avoiding the harmful shortcut learning and improving models' interpretability with much higher performance.

- We incorporated core-periphery organization recognized in the human brain to design a core-periphery principle guided convolutional neural network (CP-CNN) for image classification. The evaluation on various datasets demonstrated consistent improvement compared with the baselines.

Overall, this dissertation paves the road for future studies in brain-inspired AI and provides novel insights for a better understanding of the connections between brains and artificial intelligence.

1.4 Dissertation Outline

This dissertation contains 7 chapters.

Chapter 2 introduces the investigation of human brain function and the inspirations for neural architectures. A novel hierarchical interpretable autoencoder (HIAE) model was proposed to extract and interpret hierarchical spatiotemporal features of brain activities. The analyses based on the extracted features demonstrated the hierarchical functional differences between gyri and sulci. A potential core-periphery organization of human brain function was also discussed.

Chapter 3 covers the two embedding methods for representing brain structure and function. First, a novel approach was proposed to encode each cortical region as a dense embedding vector which simultaneously profiles the structural, connectional and functional information of a brain region. Second, a generic framework for embedding human brain function was introduced to lay the foundation for the following works.

Chapter 4 details the Sync-ACT framework for connecting the visual representation and semantics between FBNs and ANNs. We introduce of details of Sync-ACT framework and discuss the finding of correlations between the two domains' semantics and the relationship with the ANNs performance in image classification tasks.

Chapter 5 covers the two studies related to human visual attention. An eye-gaze guided vision transformer model was developed to infuse radiologists' visual attention into model training for rectifying the shortcut learning and improving the performances. A brain-inspired adversarial visual attention network (BI-AVAN) was also introduced to characterize and decode visual attention directly from brain activity.

Chapter 6 introduces the brain-inspired CP-CNN model for image classification. We designed a novel core-periphery graph generator and introduced the details of CP-CNN's wiring patterns and convolution operation.

Chapter 7 concludes the whole dissertation and discusses future works.

Chapter 2

Investigation of Human Brain Function

2.1 Overview

Gyri and sulci are two basic cortical folding patterns of the human brain. Recent studies suggest that gyri and sulci may play different functional roles given the heterogeneity in both structural substrates and functional organization. However, our understanding of gyri/sulci functional differences is still limited in the sense that a) previous studies focused on either spatial domain or temporal domain, while brain functions are intrinsically spatiotemporal; b) the analyses are limited to either local scale or global scale. Whether hierarchical functional differences exist remains unclear; c) lack of suitable analytical tools to interpret the hierarchical spatiotemporal features that may answer the question. To address those limitations, in this section, we proposed a novel Hierarchical Interpretable Autoencoder (HIAE) to explore the hierarchical functional difference between gyri and sulci.

2.2 Background

The convex gyri and concave sulci are recognized as the prominent features of the human cerebral cortex and have attracted the interests of the neuroscience community for decades (Armstrong et al., 1995; Llinares-Benadero & Borrell, 2019; Welker, 1990; Zilles et al., 1988; Zilles et al., 2013). Previous neuroimaging studies have demonstrated the structural differences of gyri and sulci using multi-modality neuroimages (H. Chen et al., 2013; X. Jiang et al., 2021; S. Liu et al., 2022; Nie et al., 2012; Yang et al., 2019; T. Zhang et al., 2014). For example, it was found that the wiring diagrams of white matter axonal fibers are

significantly different between gyri and sulci at macro-scale (Nie et al., 2012; G. Xu et al., 2010). The axonal fiber bundles course around the sulcal regions in a U-shape and align radially inside gyri (G. Xu et al., 2010; T. Zhang et al., 2014). Such structural differences were also distinguished in cyto- and myeloarchitecture at micro-scale (Essen, 1997; Goldman-Rakic et al., 1984; Hilgetag & Barbas, 2005; Richman et al., 1975), and across species such as mouse, macaque, and chimpanzee cerebrum (H. Chen et al., 2013; Nie et al., 2012; T. Zhang et al., 2014), suggesting the different functional roles that gyri and sulci may play.

Inspired by this, various computational tools have been recruited to explore the functional differences between gyri and sulci and the connections with their structural substrates based on functional magnetic resonance imaging (fMRI). These approaches can be roughly categorized into functional connectivity based methods, sparse dictionary learning (SDL) based methods and deep learning based methods. By constructing the functional connectivity matrix, F. Deng et al., 2014 found that gyri-gyri pair have the strongest functional connectivity than gyri-sulci pair (modest) and sulci-sulci-pair (weakest). Based on this observation, a functional model was proposed that gyri could be functional centers exchanging information remotely while sulci communicate locally with the neighboring gyri (F. Deng et al., 2014). SDL was employed in(X. Jiang et al., 2015; Lv et al., 2015; Lv et al., 2014; L. Zhao et al., 2019) to decompose the whole brain function as functional brain networks and examined the functional differences from a network perspective (X. Jiang et al., 2015; H. Liu, Jiang, et al., 2017; L. Zhao, Zhang, et al., 2021). For instance, it was found that gyri have strong overlap patterns regarding those functional networks than sulci (X. Jiang et al., 2015; X. Jiang et al., 2016; X. Jiang et al., 2018). H. Liu, Jiang, et al., 2017 investigated the interactions within and across gyral and sulcal functional networks and found gyri are more functionally integrated while sulci are more functionally segregated. A recent work studied the signal representation residual of SDL and suggested that gyri were more involved in global functions and interregional communications of the human brain across different task conditions and intrinsic functional networks (L. Zhao, Zhang, et al., 2021). Recently, deep neural networks have been employed to study the functional differences between gyri and sulci due to their powerful representation ability (M. Jiang et al., 2020; H. Liu et al., 2019; Q. Wang et al., 2022; S. Zhang et al., 2018). In (H. Liu et al., 2019), a convolutional neural network (CNN) was utilized to classify the gyral/sulcal fMRI signals. By examining the corresponding convolutional filters, it was found that gyral signals are of low frequency while sulcal signals are of high frequency. (M. Jiang et al., 2020) showed that gyral/sulcal signals exhibit different frequency characteristics across different regions based on a regional-specific

1D CNN model. An intrinsic connectivity network (ICN)-guided pooling-trimmed convolutional neural network(I-ptFCN) was proposed in (Q. Wang et al., 2022), and it was demonstrated that sulcal signals show the heterogeneous frequency features across different ICNs while gyral signals are homogeneous.

Despite the remarkable progresses made in the aforementioned studies, our understanding of gyral/sulcal functional differences is still limited by the lack of appropriate analytical tools (L. Zhao, Dai, et al., 2021). First, the current analytical tools mainly act on either spatial domain such as SDL for the spatially distributed functional networks, or temporal domain such as CNN for temporal characteristics of fMRI signals. However, the brain activities recorded in the 4D fMRI data are intrinsically spatiotemporal, and previous studies have suggested that the spontaneous brain function exhibit a spatiotemporal organization and dynamics (Gutierrez-Barragan et al., 2022; Kourtzi & Huberle, 2005; Moon et al., 2013). The investigations from either spatial or temporal perspective cannot provide comprehensive characterizations of the gyral/sulcal functional differences. Whether there exist spatiotemporal functional differences between gyri and sulci remains to be elucidated. Moreover, both the analytical tools and result interpretation approaches are limited in a single-scale functional architecture. For example, the SDL-based studies were confined at a global scale as the functional brain networks are globally decomposed from whole-brain fMRI signals (X. Jiang et al., 2015; X. Jiang et al., 2018; H. Liu, Jiang, et al., 2017; L. Zhao, Zhang, et al., 2021). CNN used in previous works focused on a local scale due to the limited receptive field and the shallow structure (M. Jiang et al., 2020; H. Liu et al., 2019; Q. Wang et al., 2022). Hence, our understanding of gyral/sulcal functional differences is also limited to a single scale. In fact, the external stimuli span across different time scales and over extended regions, and consequently, the neuronal activity patterns of human brain function were suggested to be organized in a hierarchical manner both spatially and temporally (Deco & Kringelbach, 2017; Friston, 2008; Golestani, 2014; Kiebel et al., 2008), where the higher levels process the information from the lower levels Golestani, 2014. In this sense, with an appropriate approach, the gyral/sulcal functional differences can also be deciphered from the local scale to the global scale in a hierarchical manner, compensating the aforementioned scale gaps. To explore the hierarchical differences between gyri and sulci, a deep neural network with multiple layers and hierarchical organization seems suitable for this objective, e.g., deep convolutional autoencoder (DCAE) model (H. Huang et al., 2017). However, the features from deep layers of neural networks are usually abstractive, multi-dimensional and hard to be interpreted or to have a straightforward neuroscientific meaning. Previous studies did not addressed

it directly and systematically, while trying to avoid it. For example, the CNN model adopted in aforementioned studies consists of only one or two convolutional layers for the ease of interpretation (M. Jiang et al., 2020; H. Liu et al., 2019; Q. Wang et al., 2022). The DCAE model in (H. Huang et al., 2017) only interpreted the features picked from one dimension while those from other dimensions remained agnostic to human perception.

2.3 Hierarchical Interpretable Autoencoder

To facilitate our understanding of gyri/sulci functional roles beyond these limitations, in this subsection, we develop a novel analytical tool, hierarchical interpretable autoencoder (HIAE) to conduct an in-depth investigation of the hierarchical differences between gyri and sulci. We adopted an unsupervised convolutional autoencoder framework (H. Huang et al., 2017) as the backbone to extract the spatiotemporal features hierarchically from the fMRI signals at different scales. In the HIAE model, features from the previous layer are the inputs for the next layer, and thus a hierarchy is naturally formed between layers. Meanwhile, layers in HIAE have different receptive fields, and the corresponding features are also at different scales. We represent the extracted spatiotemporal hierarchical features through a carefully designed feature interpreter (FI) which embeds the features corresponding to the fMRI time series in a voxel as a one-dimensional vector. The vectors of all voxels can be stacked and each digit is interpreted as the spatial distribution of the learned temporal features. The corresponding temporal activity patterns can be obtained by regressing the spatial distributions to the fMRI signals. In this way, the spatiotemporal features of brain function at different scales can be hierarchically extracted and interpreted, and the gyri/sulci functional differences can be then contrasted and analyzed.

2.3.1 Materials and Methods

The main idea is to utilize deep neural networks to extract hierarchical features from the fMRI time series of gyri/sulci and interpret those features for analyzing the gyri/sulci functional differences. First, in Section 2.3.1, we introduce the data and dataset for validating the proposed method. In Section 2.3.1, we propose a novel hierarchical interpretable autoencoder (HIAE) based on a convolutional autoencoder (CAE) and a carefully-designed feature interpreter (FI). We illustrate the parcellation of gyri and sulci in Section 2.3.1 and finally introduce the activation ratio (AR) for contrasting the gyri and sulci in Section 2.3.1.

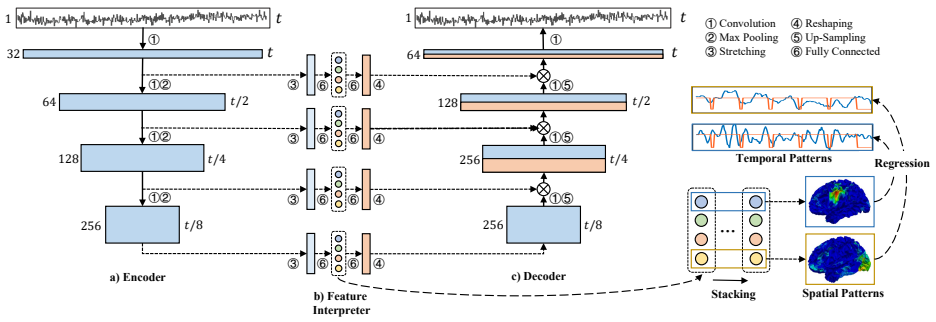

Figure 2.1: The architecture of the proposed HIAE model. a) The encoder consists of 4 convolution layers. The number on the left side of each rectangular (i.e., feature map) denotes the number of channels. The number on the right side indicates the length of each feature map. b) The feature interpreter consists of two fully-connected layers. By stacking the embedded vectors along the spatial dimension, the digit values of embedded vectors can be mapped back to the cortical surface to reveal the spatial distributions. The details of this layer are on shown in the right-most inset. c) The decoder for reconstructing the input fMRI time series.

Data Acquisition and Pre-Processing

In this work, we adopt HCP grayordinate-based task fMRI (tfMRI) data from Human Connectome Project (HCP, S1200 release) (Barch et al., 2013; Van Essen et al., 2013) to validate the proposed framework. HCP grayordinate-based data consist of high-resolution cortical surface mesh and accurate mapping of fMRI time series from volume space to surface space, which greatly facilitates our exploitation of gyral/sulcal functional differences. Specifically, HCP tfMRI dataset has seven different task paradigms including Emotion (176 frames), Gambling (253 frames), Language (316 frames), Motor (284frames), Relational (232 frames), Social (274 frames) and Working Memory (405 frames). We use all seven task paradigms as our testing beds. The acquisition parameters of tfMRI data are as follows: 90×104 matrix, 72 slices, TR=0.72 s, TE=33.1 ms, 220 mm FOV, flip angle = 52°, BW =2290 Hz/Px, in-plane FOV = 208 × 180 mm, 2.0 mm isotropic voxels. More details of seven task designs and data acquisition are referred to (Barch et al., 2013).

The HCP tfMRI data were preprocessed by HCP minimal preprocessing pipelines including spatial artefacts and distortions removal, tissue segmentation and cortical surfaces generation, within-subject cross-modal registration

(between T1-weighted MRI and fMRI), cross-subject registration to standard volume and surface spaces, cortical ribbon-constrained volume to surface mapping, and Gaussian surface smoothing (Glasser et al., 2013; Robinson et al., 2014). In addition, we normalize all fMRI time series with zero mean and standard deviation one. For all healthy adult participants in HCP S1200 release, we randomly select 300 subjects for each task and extract fMRI time series only from cortical surface vertices (64,984 vertices for each subject). In the model training phase, the training/validation dataset comprises 80%/20% of the extracted tfMRI time series, respectively. After the model training, all subjects are included with the whole-brain fMRI time series for analysis.

Hierarchical Interpretable Autoencoder (HIAE)

The technical limitations of the previous studies are: a) employing the shallow model, such as SDL in (L. Zhao, Zhang, et al., 2021); b) the employed neural network only consists of one layer for the ease of analyses, e.g., 1D-CNN in (H. Liu et al., 2019; Q. Wang et al., 2022). In this study, we adopt an autoencoder model with multiple layers to extract the hierarchical features and then analyze those features based on feature interpreter (FI). The overall architecture of the proposed HIAE model is illustrated in Figure 2.1, which has a 4-layer autoencoder and 4 corresponding FIs.

Convolutional Autoencoder We propose a convolutional autoencoder (CAE) model to extract the hierarchical features from fMRI time series. The CAE model is composed of an encoder (Figure 2.1(a)) which encodes the input fMRI time series into high-level latent space in a step-wise manner and a decoder (Figure 2.1(b)) which reconstructs the fMRI series from latent representations. There are three major advantages of CAE for feature extraction in our study: a) The training process is completely unsupervised, which eliminates the class-specific biases of learnt features compared with supervised learning methods like the CNN model in (H. Liu et al., 2019); b) It inherits the powerful ability of CNN for hierarchical feature abstraction. In CAE, features from the previous layer are the inputs for the next layer, and a hierarchy is thus naturally formed between layers; 3) Different layers inherently capture features with specific characteristics. The low-level layers capture local/high-frequency features while the high-level layers capture global/low-frequency features.

In detail, as illustrated in Figure 2.1(a), the encoder consists of 4 one-dimensional convolution layers which extract features of neural activities from low-level to high-level with rectified nonlinearity unit (ReLU) as the activation function. ReLU helps interpret feature representation in a more neuroscientifically meaningful way given its intrinsic sparsity (H. Huang et al., 2017). After each convo-

lution operation (except the last one), max-pooling is recruited to down-sample the feature maps with a stride of 2. The effects of max-pooling are two folds: 1) The computational costs for deep and high-level layers are reduced. 2) It enlarges the receptive field and reduces the sampling frequency of fMRI time series. In this sense, features from subsequent convolution layers are of low frequency compared with those from the previous layers. Feature maps from each convolution layer are also input into a Feature Interpreter (Figure 2.1(b)). In the decoder (Figure 2.1(c)), the deconvolutions are implemented as up-sampling and convolution instead of transposed convolution to eliminate potential Checkerboard Artifacts induced by hyperparameter settings (Odena et al., 2016). The activation function of deconvolutions is tanh for reconstructing the original fMRI time series. For the first deconvolution layer, it operates directly on the feature maps reconstructed by the highest-level FI For the rest layers, up-sampled features from the previous deconvolution layer are firstly concatenated with those reconstructed by the corresponding FI and then fed for deconvolution. The reconstructed fMRI time series are output by the last deconvolution layer.

Feature Interpreter

The CAE model can extract features of fMRI time series from shallow layer to deep layer hierarchically. However, the interpretability of those extracted features is relatively tricky considering: 1) Except for those in the first convolution layer, features from deep layers are much more abstract and complex without a straightforward neuroscientific meaning; 2) Feature maps for whole brain fMRI time series at each convolution layer are usually multi-dimensional, and it is consequently difficult to derive the one-dimensional spatial/temporal patterns of those learnt features for region-based analysis. To address these issues, we proposed a novel Feature Interpreter. As illustrated in Figure 2.1(b) (as well as the inset), each convolution layer and its symmetrical deconvolution layer are connected by a Feature Interpreter which consists of two fully connected (FC) layers: the first one is for embedding the feature maps as a n-digit embedded vector; the second one is for reconstructing the feature maps from the embedded vector. For a specific feature map $\mathbf{f} \in \mathbb{R}^{t \times c}$ with t time points and c channels, it is firstly stretched into a one-dimensional vector $\mathbf{z} \in \mathbb{R}^{1 \times tc}$ and then embedded as a n digits vector $\mathbf{h} \in \mathbb{R}^{1 \times n}$ by:

$$\mathbf{h} = tanh(\mathbf{z}\mathbf{W}_e + \mathbf{b}_e) \tag{2.1}$$

where $\mathbf{W}_e \in \mathbb{R}^{tc \times n}$ and $\mathbf{b}_e \in \mathbb{R}^{1 \times n}$ are the trainable weight and bias of the first FC layer, respectively. $tanh(\cdot)$ is the tanh activation function. In the reconstruction process, the embedded vector $\mathbf{h}$ is firstly up-sampled to a

one-dimensional vector $\mathbf{z}' \in \mathbb{R}^{1\times tc}$ as defined in:

$$\mathbf{z}' = tanh(\mathbf{h}\mathbf{W}_d + \mathbf{b}_d) \tag{2.2}$$

where $\mathbf{W}_d \in \mathbb{R}^{n\times tc}$ and $\mathbf{b}_d \in \mathbb{R}^{1\times tc}$ denote the trainable weight and bias of the second FC layer, respectively. Then, we can simply reshape $\mathbf{z}'$ to yield the reconstructed feature map $\mathbf{f}' \in \mathbb{R}^{t\times c}$.

Generally, FI performs the task of embedding multi-dimensional feature maps into a one-dimensional vector. For the first FC layer, it actually learns n independent nonlinear functions over the input feature map, each of which distills the information to a specific value (the digit value in embedded vector) representing an activation degree regarding the input feature. Because those functions are the same for all fMRI time series in the whole brain, the differences of this value reveal the differences in cortical regions with respect to feature activation degree and corresponding intrinsic properties (e.g., frequency characteristic). Thus, we can simply map the digit value to the corresponding vertex in the cortical surface to visualize spatial heterogeneity or distribution pattern. The corresponding temporal pattern can be obtained by regressing the spatial distribution pattern to the whole-brain fMRI signals as in (Lv et al., 2014). In this way, the complex and multi-dimensional features extracted by the autoencoder are interpreted in a more straightforward and familiar way as decoupled spatial and temporal patterns, based on which the neuroscientific investigation and discussion throughout our work are performed.

The parameters of the proposed model are optimized by minimizing the Mean Square Error (MSE) between the original fMRI time series $X \in \mathbb{R}^{k\times t}$ and the reconstructions $X' \in \mathbb{R}^{k\times t}$:

$$\min \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}'\|_F^2 \tag{2.3}$$

where k is the number of fMRI time series in each training batch and t is the number of time points/frames. The proposed framework is implemented by PyTorch (https://pytorch.org/). We use Adam optimizer (Kingma & Ba, 2014) to minimize the loss function in Eq. (2.3). Training is performed for 100 epochs with a batch size of 128 for each task on a single NVIDIA GTX 1080Ti GPU. Early stopping strategy is also employed to terminate the training process when overfitting starts.

Parcellation of Gyri and Sulci on Cortical Surface

To divide gyral/sulcal regions on the cortical surface, we follow the method in (H. Liu et al., 2019) to segment gyri and sulci based on the geometric informa-

tion of convoluted cortical surfaces. Specifically, each surface vertex of HCP grayordinate system has fMRI time series and is associated with geometric information such as principal curvature (i.e., "curv" map in FreeSurfer) and average convexity (i.e., "sulc" map in FreeSurfer). Average convexity is the signed distance that a vertex moves during the inflation process, which mainly depicts the primary folding patterns (Destrieux et al., 2010; Fischl, 2012). In this study, we take advantages of it to annotate the region that each vertex belongs to (Figure 2.2). In order to minimize the wrong annotations, we mainly focus on the crown of primary gyri and fundi of primary sulci as well as the relatively small neighborhood around them. In particular, the 30% vertices with the most positive convexity value were considered gyral vertices; the 30% vertices with the most negative convexity value were considered sulcal vertices. The remaining 40% vertices were excluded to guarantee sufficient geodesic distances between gyral/sulcal regions and thus enhance the reliability of the subsequent region-based analysis.

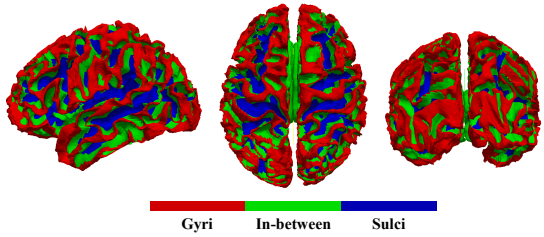

Figure 2.2: The cortical surface mesh of one randomly selected subject to illustrate the parcellation of gyri/sulci on cortical surfaces. Areas with red/blue/green color are gyral/sulcal/in-between regions, respectively.

Activation Ratio for Contrasting Gyri and Sulci

As discussed in Section 2.3.1, the digit value can indicate the differences of cortical regions regarding the activation degree of the input feature, and further indicate the activation degree of the intrinsic characteristics that the input feature retains. Considering that the low-level convolution layers in CAE extract high-frequency/local features whereas high-level convolution layers extract low-frequency/global features, the digit value can be used to investigate the hierarchical differences of gyri and sulci regarding the frequency and scale across different layers. To this end, we define an Activation Ratio (AR) for each layer of each subject:

$$AR = r \times \frac{1}{n} \sum_{i=1}^{n} \frac{g_i}{s_i} \tag{2.4}$$

where g_i/s_i denotes the number of gyral/sulcal vertices with digit value larger than zero with respect to the i^{th} digit, n is the number of digits at each layer and r is the ratio of the total number of sulcal vertices to the total number of gyral vertices (around one in our parcellation method in Section 2.3.1). If AR is greater than one, it means that gyri are more activated at this layer and tend to have more corresponding characteristics (e.g., high/low frequency) than sulci, and vice versa. The differences of AR across layers also reveal the hierarchical differences of those characteristics in gyri and sulci. In this way, the functional difference and organization of gyri and sulci can be contrasted and measured for each data modality.

2.3.2 Experimental Results

In this section, we first visualize and analyze the interpreted spatial/temporal features from the proposed HIAE model in Section 2.3.2. Then we investigate the differences of activation degree and AR metric of gyri and sulci in Section 2.3.2.

Visualization of Interpreted Spatiotemporal Features

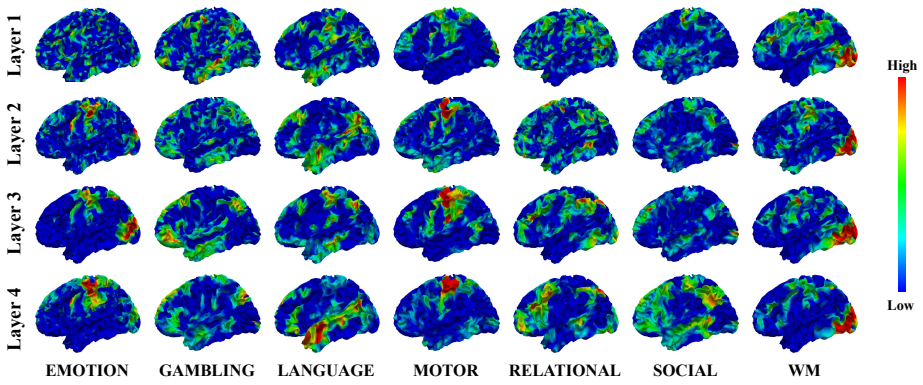

Figure 2.3: The group-averaged spatial distribution pattern with its preceding counterparts of a randomly selected digit in Layer #4 for all seven tasks. The regions with red color have larger digit value than regions with blue color. These spatial patterns are rescaled for different layers and tasks for ease of visualization.

The interpreted spatial patterns can be obtained by stacking the embedding vectors from the FI as a matrix and then mapping each row (corresponding to a specific digit in embedding vector) to the cortical surface. In Figure 2.3, we randomly select one digit in the Layer #4 and visualize its spatial pattern as well as the most similar counterparts in preceding layers. It is observed that

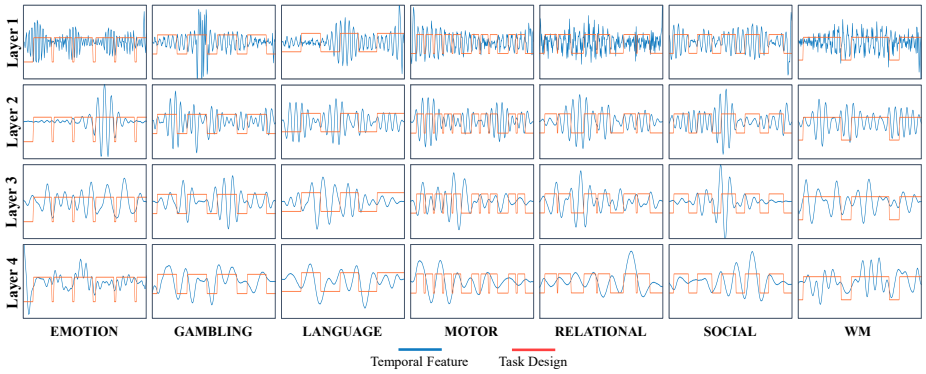

Figure 2.4: The averaged temporal pattern over all subjects from one randomly selected digit at each layer. The blue and orange curves represent the temporal pattern and task design, respectively.

the activated areas (regions in green and red colors) for the first two layers are randomly distributed in the whole brain as several small and disjunct regions. Conversely, in the deep layers such as Layer #4 , the activated areas form several larger and connected regions. These observations are well reproduced across different spatial patterns of each layer. Notably, the activated regions in deep layers are meaningful and relevant to some well-recognized functional networks. For example, the spatial patterns of Layer #3 and Layer #4 for Working Memory task in Figure 2.3 could be Visual Network locating on the occipital lobe (Smith et al., 2009). For Motor task, the spatial patterns of Layer #4 are relevant to Sensorimotor Network in Smith et al., 2009.

With the concatenated embedding vectors as a matrix for each layer, we regress the original fMRI time series matrix on it, and obtain the estimated coefficients of linear regression as the temporal patterns for all subjects. We randomly select one temporal pattern corresponding to a digit in embedding vector for each layer, and demonstrate the averaged pattern over all subjects in Figure 2.4. It is observed that the temporal patterns in shallow layers (e.g., Layer #1 and Layer #2) are in faster or high-frequency oscillation than those in deep layers (e.g. Layer #3 and Layer #4) for all seven tasks. And the frequency of oscillation seems to decrease with deeper layers. To further verify such observation, we perform the Fast Fourier Transform (FFT) for all temporal patterns, and, in Figure 2.5, report the averaged power spectrum of all temporal patterns at each layer. We find that the power spectrum of deep layers is more concentrated in low-frequency bands, and decreases quickly against increasing frequency. Com-

pared with the following layer, the previous layer has the power spectrum more concentrated in higher-frequency band.

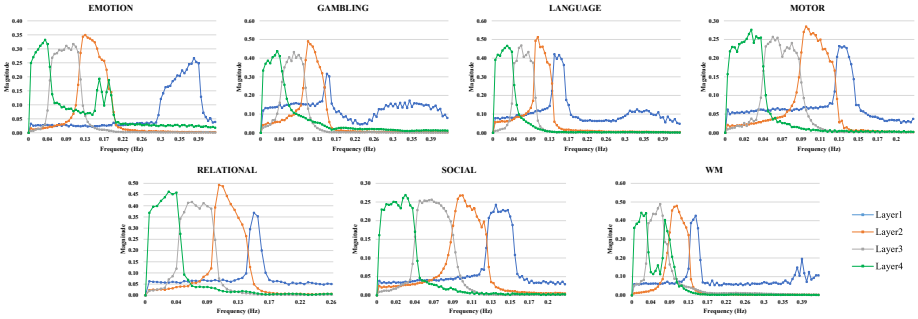

Figure 2.5: The averaged power spectrum over all temporal patterns at each layer for seven tasks, respectively

Table 2.1: Mean ($\pm$ standard deviation) digit value for gyri and sulci averaged over all subjects at each layer, and the proportion of subjects (%) with significant digit value differences (two-sample t-test, $p < 0.025$, corrected) between gyri and sulci (gyri<sulci for Layer #1 and #2; gyri>sulci for Layer #3 and #4).

Tasks		Layer #1	Layer #2	Layer #3	Layer #4
Emotion	Gyri	$29.86\pm2.81\times10^{-2}$	$24.97\pm1.89\times10^{-2}$	$18.78\pm1.77\times10^{-2}$	$11.52\pm1.10\times10^{-2}$
	Sulci	$31.72\pm2.31\times10^{-2}$	$25.70\pm1.62\times10^{-2}$	$18.29\pm1.59\times10^{-2}$	$10.73\pm1.07\times10^{-2}$
	Proportion	97.33%	85.67%	77.33%	99.00%
Gambling	Gyri	$29.19\pm2.35\times10^{-2}$	$19.81\pm1.69\times10^{-2}$	$16.39\pm1.39\times10^{-2}$	$10.36\pm1.29\times10^{-2}$
	Sulci	$29.31\pm1.95\times10^{-2}$	$20.08\pm1.44\times10^{-2}$	$16.12\pm1.18\times10^{-2}$	$9.56\pm1.27\times10^{-2}$
	Proportion	50.33%	62.33%	62.67%	99.33%
Language	Gyri	$34.38\pm2.66\times10^{-2}$	$22.84\pm1.89\times10^{-2}$	$16.75\pm1.44\times10^{-2}$	$12.42\pm1.07\times10^{-2}$
	Sulci	$34.95\pm2.37\times10^{-2}$	$22.82\pm1.70\times10^{-2}$	$16.63\pm1.24\times10^{-2}$	$11.64\pm1.10\times10^{-2}$
	Proportion	72.00%	41.33%	53.33%	96.00%
Motor	Gyri	$32.76\pm2.50\times10^{-2}$	$21.45\pm1.77\times10^{-2}$	$16.29\pm1.33\times10^{-2}$	$9.39\pm0.89\times10^{-2}$
	Sulci	$33.37\pm2.40\times10^{-2}$	$21.56\pm1.62\times10^{-2}$	$15.84\pm1.16\times10^{-2}$	$8.68\pm0.92\times10^{-2}$
	Proportion	73.33%	48.67%	78.33%	99.00%
Relational	Gyri	$33.58\pm2.25\times10^{-2}$	$17.14\pm1.59\times10^{-2}$	$15.71\pm1.38\times10^{-2}$	$11.95\pm1.32\times10^{-2}$
	Sulci	$34.61\pm1.98\times10^{-2}$	$17.47\pm1.42\times10^{-2}$	$15.39\pm1.20\times10^{-2}$	$11.06\pm1.28\times10^{-2}$
	Proportion	90.00%	69.67%	70.33%	98.33%
Social	Gyri	$32.59\pm2.82\times10^{-2}$	$26.73\pm2.65\times10^{-2}$	$21.61\pm1.92\times10^{-2}$	$16.10\pm1.61\times10^{-2}$
	Sulci	$33.01\pm2.50\times10^{-2}$	$27.37\pm2.46\times10^{-2}$	$21.51\pm1.73\times10^{-2}$	$15.00\pm1.61\times10^{-2}$
	Proportion	64.33%	76.33%	51.00%	98.00%
WM	Gyri	$24.98\pm2.57\times10^{-2}$	$15.65\pm1.74\times10^{-2}$	$12.31\pm1.14\times10^{-2}$	$8.60\pm1.02\times10^{-2}$
	Sulci	$26.24\pm2.27\times10^{-2}$	$16.03\pm1.49\times10^{-2}$	$11.95\pm1.01\times10^{-2}$	$7.96\pm0.97\times10^{-2}$
	Proportion	93.33%	74.67%	81.00%	99.00%

In general, all of those observations are consistent with what we assumed in Section 2.3.1, i.e., the shallow and low-level layers in CAE capture local/high-frequency features whereas the deep and high-level layers capture global/low-

frequency features. This also indicates that the designed Feature Interpreter is capable of interpreting the extracted features of the CAE model into meaningful spatial/temporal patterns.

The Differences of Digit Values and Activation Ratio

In this subsection, we investigate the differences of digit values and AR between gyri and sulci across different layers. In Table 2.1, we report the averaged digit values of all subjects at each layer for gyri and sulci, respectively.

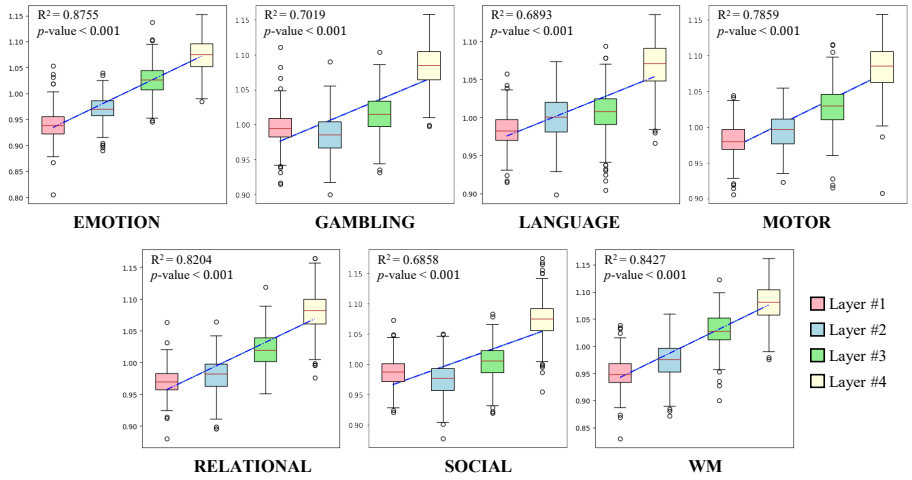

Figure 2.6: The linear regression modeling the relationship between AR values and different layers. The distribution of AR values across different layers is also demonstrated as box plots.

In general, the average digit value of gyri is smaller than that of sulci in Layer #1 and Layer #2 (except the Layer #2 of Language task), while the average digit value of gyri is larger than that of sulci in Layer #3 and Layer #4. From Layer #2 to Layer #3, it demonstrates a transition that the digit value of gyri becomes larger than that of sulci. Such transition of digit values becomes more evident in Layer #4 with a larger variation between gyri and sulci. We further confirm our observation by two-sample one-tailed un-pair-wise t-test with the alternative hypothesis that the digit value of gyri is smaller/larger than that of sulci ($p < 0.025$, corrected) for Layer #1, #2/Layer #3, #4, respectively. The proportion of subjects which reject the null hypothesis is also reported in Table 2.1. It is observed that except for the Gambling and Social tasks, the proportion Layer #2 is smaller than that of Layer #1. For all seven tasks, the proportion of Layer #4 exceeds the proportion of Layer #3. This is probably because of the consistency that gyri tend to have larger digit values becomes more pronounced in the deep

layers. For Layer #1, more than 50% subjects have smaller digit values for gyri, and for Layer #4, more than 95% subjects have larger digit values for gyri. The proportion of Layer #2 and Layer #3 varies across different tasks. We assume that the activation degree of gyri and sulci varies across different task conditions, and thus the degree of aforementioned transition also varies in Layer #2 and Layer #3.

Table 2.2: Mean ($\pm$ standard deviation) AR (Activation Ratio) of each layer averaged over all subjects for all seven tasks of HCP tfMRI dataset.

Tasks	Layer #1	Layer #2	Layer #3	Layer #4
Emotion	$94.01\pm2.88\times10^{-2**}$	$97.11\pm2.41\times10^{-2**}$	102.64 ± 2.87^{-2} **	$107.45\pm2.90\times10^{-2**}$
Gambling	$99.55\pm2.35\times10^{-2**}$	$98.60\pm2.77\times10^{-2**}$	$101.63\pm2.95\times10^{-2**}$	$108.43\pm2.91\times10^{-2**}$
Language	$98.33\pm2.18\times10^{-2**}$	$100.07\pm 2.87\times10^{-2}$	$100.66\pm 2.79\times10^{-2**}$	$106.79\pm 3.08\times10^{-2**}$
Motor	$98.17\pm2.16\times10^{-2**}$	$99.46\pm2.56\times10^{-2**}$	$102.81\pm3.05\times10^{-2}$ **	$108.40\pm3.28\times10^{-2**}$
Relational	$97.00\pm2.06\times10^{-2**}$	$98.09\pm2.64\times10^{-2**}$	$102.07\pm2.70\times10^{-2**}$	$108.09\pm3.01\times10^{-2**}$
Social	$98.68\pm2.44\times10^{-2**}$	$97.60\pm2.80\times10^{-2**}$	$100.42\pm2.90\times10^{-2*}$	$107.48\pm3.05\times10^{-2**}$
WM	$95.11\pm2.81\times10^{-2**}$	$97.49\pm3.16\times10^{-2**}$	$102.94\pm3.08\times10^{-2}$ **	$108.11\pm3.23\times10^{-2**}$

Table 2.2 demonstrate the mean AR value over all subjects at different layers across different tasks. We observe that the AR is smaller than 1 (except the Layer #2 of Language task) in Layer #1 and Layer #2 while larger than 1 in Layer #3 and Layer #4. It demonstrates a gradient that the AR increases along with deeper layers and a transition between Layer #2 and Layer #3 that the AR values increase from less than 1 to larger than 1. To confirm the existence of such gradient, we conduct linear regression to model the relationship between the AR values of all subjects and different layers. The results are reported in Figure 2.6 along with the distribution of AR values. It is found that the relationship can be well represented by the linear model (blue lines in Figure 2.6) with R^2 larger than 0.65 and p-value smaller than 0.001 for all seven tasks. This is consistent with what we observe in Table 2.2, indicating the gradient that gyri become more activated in the deep layers exists among different task conditions.

Similar to the differences of digit value in Tab. 2.1, the differences of AR are also confirmed by employing a one-sample one-tailed t-test under the alternative hypothesis that AR is smaller/larger than 1 for Layer #1 and #2/Layer #3 and #4, respectively. The significant difference is indicated by a star ($p < 0.025$) or double stars ($p < 0.005$) in Table 2.2. In general, except for the Layer #2 in Language task, most of the AR have a significant result. This suggests that sulci are more activated in Layer #1 and Layer #2 while gyri are more activated in Layer #3 and Layer #4.

Overall, these experimental results suggest that sulci are more activated with a bigger activation value in shallow layers while gyri are more activated with a bigger activation value in deep layers.

2.3.3 Conclusion

In this section, we proposed a novel computational framework for representing the brain functional activities in a hierarchical and interpretable manner, based on which we investigated the spatiotemporal functional difference of gyri and sulci at different scales. We found that gyri are of low-frequency and have more global features while sulci are of high-frequency and have more local features when the scale increases. A hypothesis was then inferred that gyri are the core while sulci are the periphery of brain function. We also demonstrated the connection of our observation and hypothesis to existing neuroscientific findings. Overall, our study offers a novel computational tool for studying the functional differences across different regions, and provides novel insights about advancing our understanding of gyral/sulcal functional differences.

2.4 Discussion

We propose a novel computational framework based on CAE and FI in this study to explore the functional differences between gyri and sulci. We demonstrate that our framework was capable of extracting hierarchical features and embedding those features into a vector, the digit value of which can be further transformed as neuroscientifically meaningful spatial and temporal patterns. The differences of digit value and AR defined as the contrast of activation degree between gyri and sulci are systematically investigated. We find that gyri are more activated in deep layers while sulci are more activated in shallow layers. Considering the differences of feature characteristics in shallow and deep layers, such gyro-sulcal contrasts indicate that gyral signals are more global and of low-frequency than sulci while sulcal signals are more local and of high-frequency. This finding agrees the previous study based on supervised CNN arguing that sulcal fMRI signals are more diverse and of more high frequency than gyral signals (H. Liu et al., 2019). Our framework exhibited similar findings but in an unsupervised manner, eliminating the task-specific bias in (H. Liu et al., 2019). It is also consistent with the previous studies showing that high-frequency brain activity reflects local domains of cortical processing, while low-frequency brain activity synchronizes across distributed brain regions (Buzsáki et al., 2013).

Compared with previous methods such as SDL which only has a global receptive field (L. Zhao, Zhang, et al., 2021) and CNN which only has a local receptive field (H. Liu et al., 2019; Q. Wang et al., 2022), our framework extract meaningful neural activities with different receptive filed ranging from local one to global one. We find that with a small and local receptive field in

the temporal domain, the corresponding spatial patterns also demonstrate local property as dozens of small and disjunct regions distributed over the whole cortex. Meanwhile, those small regions with larger digit value seem to be located more in sulci than gyri. When the receptive field becomes larger and global, the activated regions in spatial patterns begin to concentrate and merge as several large and global regions. And at this time, the most activated region lies more on gyri than sulci. Such observation along with the activation degree contrasts in Table 2.1 and Table 2.2 suggested that gyri are more involved in long-time neural activities and the global functional communication and sulci are more involved in short-time neural activities and the local information integration. More importantly, with the increasing receptive field, the transitions of activation degree and spatial distribution of gyri and sulci form a gradient from the local side to the global side. This finding is in line with the literature arguing that gyri are more functionally integrated while sulci were more functionally segregated in the organizational architecture of cerebral cortex (H. Liu, Jiang, et al., 2017), and the study suggesting that gyri are the global functional connection centers which perform interregional neural communication among distinct regions on emotion processing (X. Jiang et al., 2018). It is also congruent with (L. Zhao, Zhang, et al., 2021) proposing that gyral regions are more involved in global functions of the brain. Our findings supplemented those studies from a hierarchical and spatiotemporal perspective.

Based on these works and the findings in this section, a hypothesis can be inferred that the brain function follows a bisectional core-periphery organization, i.e., gyri are the core of brain function which accounts for the global and interregional communication and information integration, while sulci are the periphery for local brain function and information processing. Previous studies have demonstrated that the core-periphery structure can effectively boost the efficiency of information communication and processing (Everett & Borgatti, 1999; Rombach et al., 2014). The core-periphery organization also exists in the brain functional networks of human and other mammals(Bassett et al., 2013; Gu et al., 2020). Our hypothesis for brain functional organization is from a cortical folding perspective, and is coherent with structural studies finding that gyri possess denser axonal connections than sulci (H. Chen et al., 2013; Nie et al., 2012), and that the axonal fiber bundles course around the sulcal regions in a U-shape to connect the neighboring gyri (T. Zhang et al., 2014). We believe this hypothesis deserves more independent works to thoroughly investigate and exploit.

From the perspective of technique, the proposed HIAE framework actually performs an embedding task, i.e., representing the fMRI time series as embed-

ding vectors at different scales which embody both temporal and spatial information. Recently, representing the brain function as embeddings has attracted more interests because of the advantages over traditional matrix decomposition method (L. Zhao, Wu, et al., 2022), e.g. SDL, and the convenience for multi-modality studies of brain function and computer vision (H. Huang et al., 2022; L. Zhao, Dai, et al., 2022). For example, in (L. Zhao, Wu, et al., 2022), an autoencoder model based on Transformer model was proposed to represent a 3D fMRI volume at a time point as an embedding vector. This method can be considered as performing the embedding from the spatial perspective. Our method represents fMRI time series as a set of embedding vectors. It can be viewed as performing the embedding task from the temporal perspective, and supplementing the current fMRI embedding methods.

A potential limitation of the current HIAE model is that it does not explicitly model the spatial correlations of all time series. Meanwhile, the hyperparameters of our framework are manually defined. Although we did sensitivity analysis regarding those hyperparameters in previous work (L. Zhao, Dai, et al., 2021) and showed that the results and findings are not produced by a specific setting, integrating a data-driven approach to determine those hyperparameters is supposed to lead to more reasonable receptive fields which may match the actual neural activities better at different scales. Moreover, the functional organization and differences of gyri and sulci are explored in this study at a macro-imaging level. Whether our findings and conclusions still hold for cyto- and myelo- architecture at micro-scale are still unanswered and worth a further attention.

Chapter 3

Embedding Brain Structure and Function

3.1 Embedding Human Brain Architecture

3.1.1 Overview

Representing human brain architectures and constructing their correspondence across individual brains have been a long-standing challenge in the brain mapping field. Current brain image registration and cortical surface parcellation methods build brain's correspondence at voxel or region level, which can be viewed as representing each voxel or region as an "one-hot" vector for the matching algorithms to seek their correspondence across individuals. However, these "one-hot" vector representations do not encode the regularity and variability across different brain regions and individuals, thus degenerating the mapping accuracy or even causing the mismatch. In this section, we develop a novel approach to encoding each cortical region as a dense embedding vector. The derived embeddings can simultaneously profile the structural, connectional and functional information of a brain region.

3.1.2 Background

Delineating the human brain architectures and mapping their correspondence across different individual brains have attracted the interests of neuroscientists for more than a century (Brodmann, 1909; Fischl et al., 1999; Huntenburg et al., 2018; D. Shen & Davatzikos, 2002). To match the common and corresponding anatomical/functional regions across individuals, image registration has been one of the dominant methodologies in the brain mapping field by aligning

neuroimaging data geometrically into a common reference space (D. Shen & Davatzikos, 2002; Toga & Thompson, 2001). One of the challenges for brain image registration is dealing with the huge structural and functional variability of human brain, and many efforts including the recent deep learning ones have been made to remedy the mapping inaccuracy and even mismatch due to the intrinsic human brain variability (Cao et al., 2017; T. Liu et al., 2004; Shi et al., 2010; Wu et al., 2011; Z. Xu & Niethammer, 2019). Besides registration, parcellation of cerebral cortex is another important approach for human brain mapping, which can be broadly classified into model-driven and data-driven categories. The first group usually warps the brain atlases onto the cortex for parcellation (Fischl et al., 2004), while the second group relies on the discriminative features such as morphological, structural and functional features (T. Zhang et al., 2016). Some recent works also employ deep learning models to learn from labeled regions and directly perform the parcellation(R. He et al., 2020; Tang et al., 2020), thus achieving promising results.

Despite the remarkable progresses in those works, the image registration and surface parcellation methods build the correspondence across different subjects in a voxel-to-voxel or region-to-region manner. From our perspective, each voxel/region is actually viewed as a discrete "one-hot" vector and the correspondence is built by matching the “one-hot” vector across subjects. However, the discrete and sparse "one-hot" representations do not encode the regularity and variability of different voxels/regions across different subjects. In addition to those discrete methods, an increasing interest in presenting brain structures and functions as large-scale gradients, on which the respective spatial relationship between cortical regions and across subjects is encoded and aligned (Huntenburg et al., 2018). By far, the gradient and alignment based on it, however, were obtained and implemented in a group-wise manner, still regardless of the inter-individual variability. To address this critical problem, an intuitive way is to profile the regularity and variability into a continuous high-dimensional space where the cortical region is represented as a dense vector and the relationship between different regions and across individuals can be measured by the geometric distance in that space. Recent embedding approaches seem suitable to our aim. For example, in the natural language processing (NLP) filed, word embedding captures both semantic and syntactic information of words and represents them as dense vectors in a continuous space where the words have similar meaning are closer to each other (Mikolov et al., 2013). For graph representation, graph convolutional network (GCN) learns the continuous embedding of nodes in a graph with both local graph structure and features (Kipf & Welling, 2016),

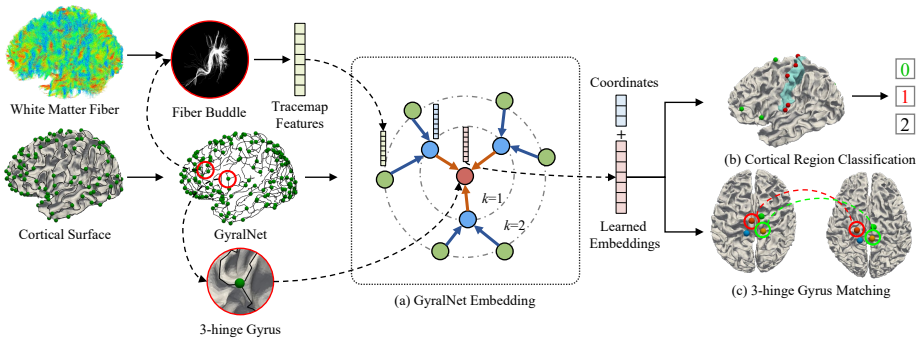

Figure 3.1: Illustration of the proposed embedding framework. GyralNet models the human brain architecture as a connected graph where the 3-hinge gyri (the conjunction of gyri, circled by red color) are represented as nodes. (a) The GyralNet embedding based on graph embedding approach with Tracemap that encodes the connectional patterns as input features. After k iterations, the embedding vector is concatenated with coordinates of 3-hinge gyrus for (b) cortical region classification and (c) 3-hinge gyrus matching downstream tasks.

which is used for downstream tasks such as node classification with significant improvements.

3.1.3 Methods

We represent the human brain architecture in a novel way using embedding. The key idea is that each brain region, regardless of its size or location, can be represented as an embedding vector which profiles anatomy, connectivity, function of that region from the corresponding features (e.g., structural,functional, and connectivity patterns). Within the embedding space, the regions with similar profiles are closer while individual variance is still retained. In this way, the mapping of corresponding regions can be conducted by computing the similarity of embedding vectors. The variance of regions/individuals can be also quantitatively measured by the calculation of distance. In this section, the GyralNet (H. Chen et al., 2017) is adopted as a test bed which models the cortical architecture as a graph of gyral crests. We learn the embeddings for the nodes in the GyralNet, i.e., the conjunction of gyri (T. Zhang et al., 2018), based on structural and connectional features with graph embedding approaches (Hamilton et al., 2017).

GyralNet Embedding

As illustrated in Figure 3.1, GyralNet models the cortical architecture as a connected graph in which the gyral crests are considered as edges and the conjunctions of gyri with the degree more than 2 are regarded as nodes (H. Chen et al., 2017). Most of the nodes in GyralNet have the degree equal to 3, which are defined as 3-hinge gyri and have been shown to have unique and consistent anatomical, structural, and functional patterns across subjects (T. Zhang et al., 2018). In this study, we aim to learn the embeddings for 3-hinge gyri in GyralNet.

Intuitively, graph embedding methods such as graph convolutional network (GCN) (Kipf & Welling, 2016) are suitable for this task. However, due to the complexity and variability of cortical folding patterns, the number of nodes and edges in GyralNet varies significantly across subjects, and thus GCN based transductive methods which focus on a single fixed graph are not feasible in our task. Inspired by GraphSAGE in (Hamilton et al., 2017), we adopt an inductive approach to learn an embedding function which aggregates features from a node's neighbors to generate the embeddings, iteratively. Specifically, given a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and node features$\{h_v^k, \forall v \in \mathcal{V}\}$ in the $k-1^{th}$ iteration, the aggregation of neighborhood nodes' features $h_{\mathcal{N}(v)}^k$ can be represented as:

$$h_{\mathcal{N}(v)}^k \leftarrow \mathrm{AGG}\left(h_u^{k-1}, \forall u \in \mathcal{N}(v)\right) \tag{3.1}$$

where $\mathrm{AGG}(\cdot)$ is an aggregation function and $\mathcal{N}(v)$ represents the neighborhood nodes of node v. The node feature h_v^k in the k^{th} iteration are then updated as:

$$h_v^k \leftarrow \sigma\left(W^k \cdot \mathrm{CONCAT}(h_v^{k-1}, h_{\mathcal{N}(v)}^k)\right) \tag{3.2}$$

where $\sigma(\cdot)$ is an activation function, W^k denotes a trainable parameter in the k^{th} iteration and $\mathrm{CONCAT}(\cdot)$ represents a concatenation operation. For the first iteration, the h_v^0 is initialized as the input features of nodes. The h_v^n in the last iteration (totally n iterations) is considered as the learned node embedding. It is noted that for each iteration, the node only aggregates the information from its directly connected neighbors. As the process iterates, the nodes can aggregate the information from further nodes indirectly. Given that the embedding function is learned based on a single node and its neighbors, it can be trained in a batch-wise manner and can be easily extended to new nodes and new graphs. So with this approach, we can learn a more general embedding function for 3-hinge gyri across different cortical regions and subjects.

Cortical Region Classification

In this subsection, we formulate the learning of embedding function for a cortical region as a classification task. Specifically, the cortical surface is firstly parcellated into 74 anatomical regions with no overlaps based on (Destrieux et al., 2010). Then, in a specific region, e.g., precentral gyrus, we obtain 3 different classes by assigning label 0 to the 3-hinge gyri nodes in the left hemisphere, 1 to those in the right hemisphere, and selecting the same number of nodes randomly in other cortical regions with label 2. We perform the classification task by employing an additional fully-connected (FC) layer with $softmax$ activation function after the embedding layer. Considering that the embeddings are heavily based on the input features which might be similar among spatially distinct but structurally connected regions, we incorporate the coordinates of each 3-hinge gyrus by concatenating it with its GyralNet embeddings. So, the concatenated embeddings will have 3 additional dimension. The FC layer takes the concatenated embedding as input and outputs the probability of each class. The whole framework is optimized in an end-to-end manner by minimizing the cross-entropy between predictions and labels.

Matching the corresponding 3-hinge Gyrus

We formulate the matching of corresponding 3-hinge gyrus across different subjects as a classification task. Our hypothesis is that the corresponding gyrus of different subjects have the similar structure, connectivity and function, such that their embeddings will be similar and can be easily classified into the same group. Here, we firstly choose a target 3-hinge gyrus and obtain its GyralNet embeddings. The GyralNet embedding is concatenated with the coordinates, and then fed into a FC layer for classification where the target 3-hinge gyrus is recognized as positive sample while other randomly selected 3-hinge gyri in GyralNet are considered as negative samples. After the training, the correspondence of target 3-hinge gyrus can be matched by selecting the one classified into the target group.

3.1.4 Experiments

Dataset and Pre-Processing

In this study, we adopted the structural MRI (T1-weighted) and diffusion tensor imaging (DTI) data of 300 subjects from Human Connectome Project (HCP) S1200 release (https://www.humanconnectome.org/) (Van Essen et al., 2013). The training, validation and testing dataset have 200/50/50 subjects, respec-

tively. The T1-weighted MRI data was preprocessed by the HCP minimal preprocessing pipeline (Glasser et al., 2013) including brain skull removal, tissue segmentation and the reconstruction of cortical surface mesh. The Destrieux Atlas (Destrieux et al., 2010) was used to parcellate the cortical surface into 74 anatomical regions, among which 8 different regions at each hemisphere were selected as testing beds for cortical region classification task.

The GyralNet and 3-hinge gyri of each subject are extracted on the reconstructed cortical surface according to the pipeline in (H. Chen et al., 2017). More details are referred to (H. Chen et al., 2017). The preprocessing of DTI data included skull removal, motion correction and eddy current correct. Fiber tracking was then performed by MedInria (https://med.inria.fr/). We extracted the white matter fiber bundles around each 3-hinge gyrus and projected them to a standard sphere space called Tracemap (Zhu et al., 2011). Tracemap can be represented by a vector of 144 dimensions and encodes the fiber density information and the connectivity patterns. The Tracemap was used as the initial input features for 3-hinge gyrus in GyralNet. We also obtained the coordinates of 3-hinge gyrus as additional 3 dimensions of embedding. The coordinates are normalized by deducting the centroid of all surface mesh points.

Implementation Details

In our experiments, we use the mean function as our aggregation function in Eq. (3.1), i.e., the features from all neighborhood nodes are averaged. For the GyralNet embedding model, in cortical region classification task, the embedding size of the first/last iteration is 128/64, respectively. As for the middle iterations, the embedding size is 96. For 3-hinge gyrus matching, the embedding size is 128 for all iterations. The framework is implemented with PyTorch (https://pytorch.org/) deep learning library. We used the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is 16 and the model is trained for 150 epochs with a learning rate 0.001 on a single GTX 1080Ti GPU.

Cortical Region Classification Results

In this subsection, we selected 8 regions to evaluate the embeddings learned from our framework: precentral (PRC), postcentral (PTC), superiorfrontal (SF), rostralmiddlefrontal (RMF), lateraloccipital (LO), superiortemporal (ST), middletemporal (MT), inferiortemporal (IT). The division of these regions on the cortical surface is shown in Figure 3.2.

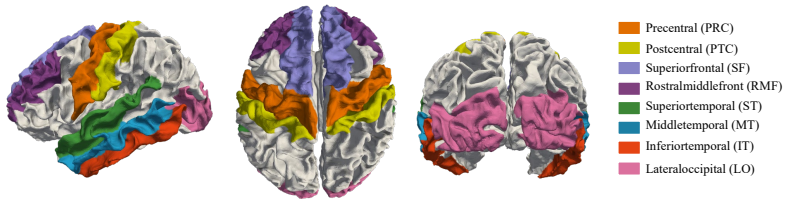

Figure 3.2: The illustration of the selected 8 different regions (totally 16 in two hemispheres) on cortical surface. Different regions are denoted by different colors.

Table 3.1: The classification results of several baselines and the proposed method in 8 different cortical regions. Red and blue denotes the best and the second-best results, respectively. Abbreviations: PRC: precentral; PTC: postcentral; SF: superiorfrontal; RMF: rostralmiddlefrontal; LO: lateraloccipital; ST: superiortemporal; MT: middletemporal; IT: inferiortemporal.

Methods	PRC	PTC	SF	RMF	LO	ST	MT	IT
Coor	0.5173	0.6925	0.8420	0.6023	0.5907	0.7926	0.8274	0.8245
Identity	0.5539	0.5542	0.5518	0.5594	0.5644	0.6210	0.5987	0.5777
GN-Embed^1	0.6084	0.6217	0.6370	0.6441	0.6198	0.6466	0.6323	0.6340
GN-Embed^1+Coor	0.7906	0.7765	0.9568	0.9020	0.9181	0.8997	0.9271	0.9372
GN-Embed^2+Coor	0.8220	0.8518	0.9670	0.9362	0.9494	0.9008	0.9305	0.9436
GN-Embed^3+Coor	0.8314	0.8816	0.9655	0.9432	0.9628	0.9030	0.9350	0.9404

We introduce two baselines for cortical region classification task: Coor and Identity. Here, Coor means the only the 3 dimension coordinates are used as the embeddings for classification; Identity means the input features, i.e., Tracemap features, are directly used as the embeddings for classification. We report the classification results in Table 3.1.

It is observed that the baseline Coor performes better than the baseline Identity in some regions. This is consistent with our expectation that the Tracemap features (structural connectivity) are similar among some distinct regions and the coordinates can provided additional information to differentiate these regions. The GyralNet embedding model with only one iteration, i.e., GN-Embed^1, outperforms the baseline Identity, suggesting the effectiveness of embeddings to encoding the structural and connectional information. By concatenating the learned embeddings with the coordinates (e.g., GN-Embed^1+Coor), the classification accuracy is significantly improved among all regions. We also observed that increasing the number of iterations in GyralNet embedding contributes to the classification performance. This is probably because with more iterations,

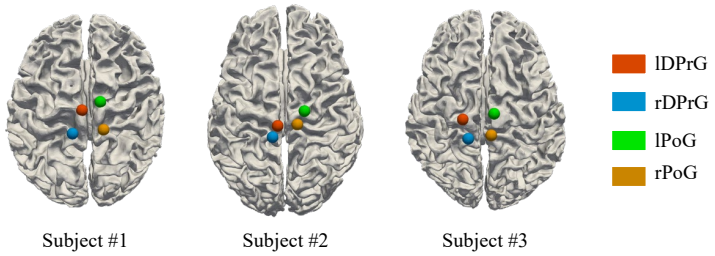

Figure 3.3: The illustration of the 4 3-hinge gyri from 3 randomly selected subjects on cortical surface. Different 3-hinge gyri are denoted by different colors.

the embedding of a 3-hinge gyrus has more information from its further neighbors, which makes the delineation of a 3-hinge gyrus more precise. But the improvements become saturated with 3 iterations. Overall, these results suggest that the concatenated embeddings with structural, connectional and stereotaxic information is more powerful in representing the variability of human brain.

Correspondence of 3-Hinge Gyri

We manually labeled 4 3-hinge gyri: left/right postcentral gyrus (lPoG/rPoG) and dorsal left/right precentral gyrus (lDPrG/rDPrG), to evaluate whether the corresponding 3-hinge gyrus can be matched by our framework. These 3-hinge gyri are reported to be consistently preserved across different subjects (X. Li et al., 2017). We randomly select 3 subjects and visualize the gyri locations in Figure 3.3.

We include two additional classifications between lDPrG/rDPrG and between lPoG/rPoG to evaluate whether our framework is capable in learning discriminative embeddings for gyri which are spatially close in the same region. The results are reported the in Table 3.2. It is noted that in testing dataset (50 subjects), we have 100 samples for each classification in Table 3.2.

For the classification of target 3-hinge gyrus with others, we observed that the performance of Identity and GN-Embed^1 model are comparable for all 3-hinge gyri while the baseline Coor is slightly better than those two. When combining the learned embeddings with coordinates, the accuracy is improved significantly, especially for GN-Embed^2+Coor with two iterations, indicating the effectiveness of our model in matching the corresponding 3-hinge gyrus. However, the performance degenerate in GN-Embed^3 model. For the classification between lDPrG/rDPrG and between lPoG/rPoG, it is observed that the accuracy of baseline Coor is 0.5 which is closed to random guess. This is

Table 3.2: The 3-hinge gyri classification accuracy of several baselines and the proposed method. Red and blue denotes the best and the second-best results, respectively. Abbreviations: lDPrG: dorsal left precentral gyrus; rDPrG: dorsal right precentral gyrus; lPoG: left postcentral gyrus; rPoG: right postcentral gyrus.

Methods	lDPrG	rDPrG	lPoG	rPoG	lDPrG/rDPrG	lPoG/rPoG
Coor	0.74	0.75	0.70	0.76	0.50	0.50
Identity	0.59	0.77	0.66	0.66	0.81	0.76
GN-Embed^1	0.62	0.75	0.66	0.67	0.85	0.77
GN-Embed^1+Coor	0.85	0.85	0.88	0.83	0.97	1.00
GN-Embed^2+Coor	0.95	0.87	0.94	0.93	0.90	0.98
GN-Embed^3+Coor	0.86	0.83	0.85	0.72	1.00	0.98

probably because the two 3-hinge gyri are spatially close to each other. Given the variability of coordinates for different subjects, they cannot be simply differentiated by coordinates. The GN-Embed^1 model is slightly better than Identity with an accuracy over 0.7. However, when concatenating the embeddings with coordinates, the accuracy is very closed to 1.0. This suggests that the concatenated embeddings can provide enough information to differentiate the gyri accurately.

3.1.5 Conclusion

In this section, we proposed a novel framework for delineating the human brain architecture using embeddings. Our method represents the brain architecture as a connected graph on gyral crests and embeds the structural and connectional information of 3-hinge gyrus into a dense vector, which can be used for downstream tasks such as cortical region classification and correspondence matching. The key advantage of our approach is that the regularity and variability of human brain are encoded as embeddings in a general, comparable, and stereotyped space. Future works include incorporating the functional features in the embedding process to improve the accuracy and validating the framework in more downstream tasks.

3.2 Embedding Human Brain Function

3.2.1 Overview

Learning an effective and compact representation of human brain function from high-dimensional fMRI data is crucial for studying the brain's functional organization. Traditional representation methods such as independent component analysis (ICA) and sparse dictionary learning (SDL) mainly rely on matrix decomposition which represents the brain function as spatial brain networks and the corresponding temporal patterns. The correspondence of those brain networks across individuals are built by viewing them as one-hot vectors and then performing the matching. However, those one-hot vectors do not encode the regularity and/or variability of different brains very well, and thus are limited in effectively representing the functional brain activities across individuals and among different time points. To address this problem, in this section, we formulate the human brain functional representation as an embedding problem, and propose a novel embedding framework based on the Transformer model to encode the brain function in a compact, stereotyped and comparable latent space where the brain activities are represented as dense embedding vectors.

3.2.2 Background

Functional magnetic resonance imaging (fMRI) has been widely used in studying the human brain's functional organization and the responses to external stimuli (Engel et al., 1994; Heeger & Ress, 2002; Huettel et al., 2004; Logothetis, 2008). However, the fMRI-based imaging studies face a major challenge that the number of voxels in 4D spatiotemporal fMRI data is far beyond the number of subject brains included in the study. (Mwangi et al., 2014), i.e., the "curse-of-dimensionality" (Bellman, 2015). To diminish the negative impacts of this intrinsic imbalance, various computational tools have been recruited to extract the representative features from the raw voxels and dump the redundant information as well as the noises. (Andersen et al., 1999; Calhoun & Adali, 2006; Lv et al., 2014). For example, principal component analysis (PCA) extracted the relevant features by linearly transforming the correlated voxels into several uncorrelated variables, i.e., the principal components which capture most of the data variance (Andersen et al., 1999). Independent component analysis (ICA) based studies assumed that the raw fMRI signals are linear mixtures of independent and relevant components (e.g., paradigm-related patterns in task fMRI). (Calhoun & Adali, 2006; Calhoun et al., 2001; Calhoun et al., 2009). By decomposing those independent components as temporal and spatial patterns,

the following analysis can be performed with a much more compact representation rather than within the vast volume space. In addition, sparse dictionary learning (SDL) has also been applied to decoupling the spatial and temporal patterns of brain functional activities from the 4D spatiotemporal data (X. Jiang et al., 2015; X. Jiang et al., 2018; Lv et al., 2015; Lv et al., 2014; L. Zhao et al., 2019; L. Zhao, Zhang, et al., 2021), where the learned dictionary is regarded as temporal activation patterns and the sparse coefficient matrix is recognized as the spatial distributions of the corresponding temporal patterns, i.e., spatial brain networks.

Despite that the aforementioned matrix decomposition techniques are widely adopted and applied, the temporal and/or spatial patterns decomposed by those approaches are not intrinsically comparable across different subject brains. The correspondence of the spatial/temporal patterns from different brains does not exist even with the same hyper-parameter setting in those matrix decomposition methods such as ICA or SDL. To build such correspondence, group-wised ICA and SDL methods were proposed by concatenating the fMRI signal matrices from different subjects spatially or temporally, resulting in common spatial brain networks/temporal activation patterns and individual temporal/spatial features (Calhoun et al., 2009; Ge et al., 2018; H. Liu, Zhang, et al., 2017; L. Zhao et al., 2019). However, a fundamental problem of those group-wised methods is that the huge spatial/temporal variability of brain function is overlooked and not encoded. From our perspective, these methods represent the brain functional organization as functional brain networks and then try to map the correspondence of those networks across individuals and populations. In this process, the functional brain networks are viewed as one-hot vectors (i.e., common spatial/temporal patterns), based on which the mapping is performed. Actually, the one-hot vectors do not represent the spatial/temporal variability very well, while the individual features do not encode the temporal/spatial regularity of different brains. In other words, those methods represent the regularity in one dimension but the variability in another, failing to offer a general, comparable, and stereotyped representation encoding both regularity and variability of different brains. To address this intrinsic limitation, an intuitive and potential way is to represent the human brain function in a general, comparable, and stereotyped space where the brain functional activities can be meaningfully and compactly represented as embeddings.

As an effective embedding method for high-dimensional data, deep learning has achieved great successes and superior performances than traditional matrix decomposition methods in modeling fMRI data (Dong et al., 2020; Q. Li, Dong, Ge, Qiang, et al., 2021; Qiang et al., 2021; L. Zhao, Dai, et al., 2021).

However, to the best of our knowledge, current deep learning based fMRI representation studies were not specifically designed for constructing a compact embedding space where the regularity and variability of different brains can be effectively represented. Instead, the prior methods targeted specific tasks and applications such as time series classification (M. Jiang et al., 2020; H. Liu et al., 2019), brain network decomposition (Q. Li, Dong, Ge, Qiang, et al., 2021; Q. Li, Zhang, et al., 2021; Qiang et al., 2020; Yu et al., 2022; W. Zhang et al., 2019), brain state differentiation (H. Wang et al., 2018), among others. Therefore, the current deep learning based methods still do not offer a general, comparable, and stereotyped space for encoding human brain function. Importantly, a generic embedding framework can be easily integrated with other representation learning methods for multi-model representation learning, providing a potential way for connecting the semantic space of human brain function with the one in Natural Language Processing (NLP) or Computer Vision (CV).

3.2.3 Methods

We formulate the representation learning of human brain function as an embedding problem. The regularity and variability of brain function across individual brains and at different time points are represented in a general, comparable, and stereotyped embedding space, where the 3D volumes of fMRI data that record functional brain activities at different time points are profiled as dense vectors. Specifically, we design a novel unsupervised Temporally Correlated Autoencoder (TCAE) based on the Transformer model (Vaswani et al., 2017) and self-attention mechanism to construct an effective embedding space. The major theoretical and practical advantage of the proposed framework is that it explicitly models the temporal correlations of different time points and implicitly attends to the information of different representation sub-spaces.

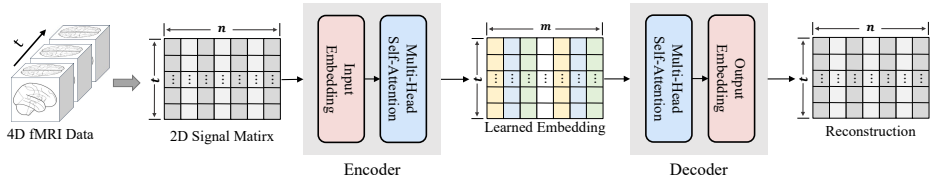

Figure 3.4: Illustration of the proposed TCAE embedding framework. The 4D fMRI data are firstly rearranged into a 2D signal matrix. Then the matrix are input into the encoder consisting of an input embedding layer and a multi-head self-attention module. The output of the encoder is recognized as the learned embedding, which can be used for downstream tasks and for reconstructing the input signal matrix with the decoder.

Temporally Correlated Autoencoder

The traditional matrix decomposition methods represent the human brain function as functional brain networks and the corresponding temporal activations, which are not intrinsically comparable across individuals. Thus, we aim to construct a unified and stereotyped embedding space, where the brain function can be compactly represented and compared. Given that fMRI data are multi-dimensional with both spatial/temporal information and numerous voxels, the spatial dimensionality should be compressed and the temporal correlations should be explored in order to obtain a compact embedding with representative spatiotemporal features. Meanwhile, the learning of such embedding should be done in an unsupervised manner considering the lack of "group truth".

Therefore, we introduce a novel computational framework, Temporally Correlated Autoencoder (TCAE), based on the Transformer model (Vaswani et al., 2017). As demonstrated in Figure 3.4, TCAE has an encoder-decoder architecture. The encoder and decoder both consist of an embedding layer and a multi-head self-attention module. In the encoder, the embedding layer represents the 3D fMRI volumes as 1D feature vectors through a linear transformation for compressing the spatial dimensionality. The features vectors of all time points are then input into the multi-head self-attention module for modeling the temporal correlations of different time points and generating the final embedding aggregating both spatial and temporal information. Compared with recurrent neural networks (RNNs), the self-attention mechanism is more capable in capturing the global and long-distance temporal dependencies and thus improves the quality of the learned embedding.

Specifically, the 4D fMRI data are firstly rearranged into a 2D fMRI signal matrix $\boldsymbol{S} \in \mathbb{R}^{t \times n}$, where t is the number of time points and n is the number of voxels, by extracting the time series of each voxel and concatenating them together. Then, the rearranged 2D fMRI signal matrix $\boldsymbol{S}$ is embedded as a feature matrix $\boldsymbol{S}_f \in \mathbb{R}^{t \times m}$ by the embedding layer for spatial dimensionality reduction, where m is the reduced feature dimension ($m \ll n$). The feature matrix $\boldsymbol{S}_f$ is then input into the multi-head self-attention module to model the iterations of all time points. For each attention head i, the self-attention map that captures the temporal correlations of different time points is computed as:

$$ATTN_i = \boldsymbol{S}_f \boldsymbol{W}_i^Q (\boldsymbol{S}_f \boldsymbol{W}_i^K)^T \tag{3.3}$$

where $\boldsymbol{W}_i^Q \in \mathbb{R}^{m \times k}$ and $\boldsymbol{W}_i^K \in \mathbb{R}^{m \times k}$ are projection matrices and k is the feature dimension of the self-attention operation. With the self-attention maps,

the output of attention head i can be computed as:

$$\boldsymbol{S}_{attn_i} = softmax(\frac{ATTN_i}{\sqrt{k}})\boldsymbol{S}_f\boldsymbol{W}_i^V \tag{3.4}$$

where $\boldsymbol{W}_i^V \in \mathbb{R}^{m\times v}$ and v is the feature dimension for the output of attention heads. The output of each head $\boldsymbol{S}_{attn_i}$ is then concatenated along the feature dimension and transformed into a new feature matrix $\boldsymbol{S}_{heads} \in \mathbb{R}^{t\times m}$ as:

$$\boldsymbol{S}_{heads} = Concat(\boldsymbol{S}_{attn_1}, \ldots, \boldsymbol{S}_{attn_h})\boldsymbol{W}^O \tag{3.5}$$

where $Concat(\cdot)$ represents the concatenation operation. h is the number of heads and $\boldsymbol{W}^O \in \mathbb{R}^{hv\times m}$ is the projection matrix. $\boldsymbol{S}_{heads}$ is further fed into the feed forward layer to obtain the learned embedding $\boldsymbol{E} \in \mathbb{R}^{t\times m}$:

$$\boldsymbol{E} = ReLU(\boldsymbol{S}_{heads}\boldsymbol{W}_{ff1} + b_{ff1})\boldsymbol{W}_{ff2} + b_{ff2} \tag{3.6}$$

where $\boldsymbol{W}_{ff1} \in \mathbb{R}^{m\times d_{ff}}$, $\boldsymbol{W}_{ff2} \in \mathbb{R}^{d_{ff}\times m}$, b_{ff1} and b_{ff2} are biases. d_{ff} represents the inner feature dimension of the feed-forward layer.

To facilitate an end-to-end training, we also include a decoder to reconstruct the 2D fMRI signal matrix. The hypothesis here is that if the learned embedding is informative and representative, the original signal matrix can be better reconstructed from it. In this way, by maximizing the similarity of original signal matrices and the reconstructed ones, the embedding framework can be optimized in an unsupervised manner. The decoder in our framework also consists of a multi-head self-attention module and an embedding layer which increases the feature dimension from m to n to match the input signal matrix.

The whole framework is optimized by minimizing the Mean Square Error (MSE) between the original fMRI signals $\boldsymbol{S} \in \mathbb{R}^{t\times n}$ and their corresponding reconstruction $\boldsymbol{S}' \in \mathbb{R}^{t\times n}$:

$$\min \ \frac{1}{2}\|\boldsymbol{S} - \boldsymbol{S}'\|_F^2 \tag{3.7}$$

Prediction of Brain State

In this subsection, we introduce a brain state prediction downstream task to evaluate the learned embedding. During the task fMRI acquisition, each participant is required to perform a specific task according to the block-design paradigm. Accordingly, the fMRI data at each time point can be classified into a specific brain state according to the task that the subject participated in, e.g., math calculation or listening to a story. Here, we use the learned embedding of

each time point to predict the brain state for each subject. Specifically, we pre-train a TCAE model to construct an embedding space. Then, the pre-trained model is fixed and the embeddings are obtained through the inference. The embeddings are input into a classifier $f(\cdot)$ to derive the prediction of brain state $\hat{\boldsymbol{y}}$. Here, we implement the classifier as a two-layer multi layer perceptron (MLP):

$$\hat{\boldsymbol{y}} = softmax(\boldsymbol{W}_2(tanh(\boldsymbol{W}_1\boldsymbol{E} + \boldsymbol{b}_1)) + \boldsymbol{b}_2) \tag{3.8}$$

The MLP can be optimized by minimizing the cross-entropy between predictions $\hat{\boldsymbol{y}}$ and labels $\boldsymbol{y}$:

$$\min -\sum_i \boldsymbol{y}_i log(\hat{\boldsymbol{y}}_i) \tag{3.9}$$

The classification performance indicates whether the learned embedding represents the current brain activity well. Also, the learned embedding is not taks-specific, and thus the effectiveness and generalizability of our embedding framework can be fairly evaluated.

Interpretation of Embedding Space

Besides the performance on downstream tasks, the interpretability of learned embedding and embedding space is another important criteria for a comprehensive evaluation of our framework. Here, we explore the spatial distribution of voxels mixed into a digit in embedding vector and the temporal variance of the digit value, i.e., the temporal pattern.

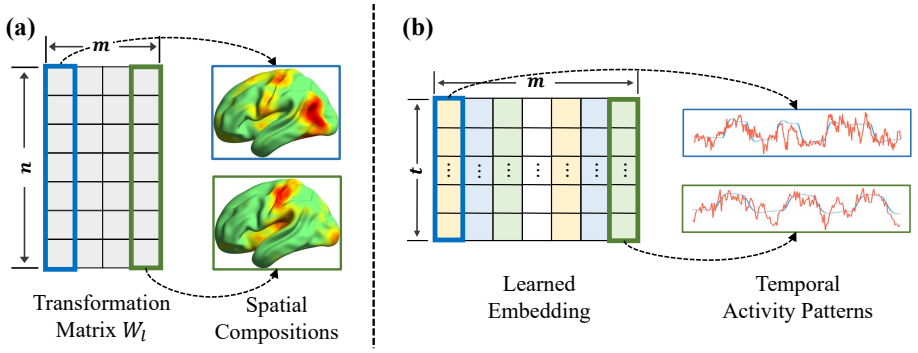

Figure 3.5: Interpretation of learned embedding and embedding space. (a) Mapping the spatial compositions of digits to 3D volume space. (b) Digit values of different time points naturally form a time series, i.e., temporal activity patterns

As illustrated in Figure 3.5(a), the linear transformation matrix $W_l \in \mathbb{R}^{n \times m}$ in the encoder embedding layer maps the all voxels as a embedding vector. Each

column of W_l contributes to a specific digit of embedding vector, which can be mapped back to 3D volume space to visualize the spatial composition of this digit. Meanwhile, the digit values of different time points naturally form a time series (Figure 3.5(b)). Similar to (Q. Li, Dong, Ge, Qiang, et al., 2021), we assume that it reflects the temporal activity pattern of that digit. It is noted that digits in the embedding vector correspond to subspaces spanning the whole embedding space. By computing the Pearson Correlation Coefficient (PCC) between the temporal pattern and the Hemodynamic Response Function (HRF) responses of task stimuli, we can examine if there exists subspaces corresponding to the task stimuli and their correlations. In this way, the embedding space can be interpreted from both spatial and temporal perspectives.

3.2.4 Experiments

Dataset and Pre-Processing

We adopt the publicly available HCP task fMRI (tfMRI) dataset of S1200 release (https://www.humanconnectome.org/) (Barch et al., 2013). In this section, we select Emotion, Motor, Language and Working Memory (WM) tasks as testbeds. The acquisition parameters of HCP tfMRI data are as follows: 90 × 104 matrix, 72 slices, 220 mm FOV, in-plane FOV = 208 × 180 mm, 2.0 mm isotropic voxels, TR=0.72 s, TE=33.1 ms, flip angle = 52°, BW =2290 Hz/Px. The preprocessing pipelines of tfMRI data are implemented by FSL FEAT (Woolrich et al., 2001), including skull removal, motion correction, slice time correction, spatial smoothing, global drift removal (high-pass filtering) and registration to the standard MNI 152 4 mm space for reducing the computational overhead. Besides, the time series from each voxel are normalized with zero mean and standard deviation one before rearranging into 2D signal matrix. For a total of more than 1000 subjects in HCP S1200 release, we randomly select 600 subjects as training set, 200 subjects as validation set, and another 200 subjects as testing set. All the experimental results in the following sections are reported on the testing set.

Brain State Prediction Results

We report the brain state prediction downstream task performance is this subsection. Here, we introduce several baseline models with various architectures for comparison: Autoencoder (AE), deep sparse recurrent autoencoder (DSRAE) (Q. Li, Dong, Ge, Qiang, et al., 2021), deep recurrent variational autoencoder (DRVAE) (Qiang et al., 2021), spatiotemporal attention autoencoder (STAAE) (Dong et al., 2020). AE represents an autoencoder model composed of one

embedding layer with $tanh$ activation function for both encoder and decoder, which can be assumed as our baseline model without the multi-head self-attention module. Notably, DSRAE was designed for decomposing the spatial/temporal brain networks; DRVAE model was designed for the augmentation of fMRI data; STAAE was proposed for the classification of Attention Deficit Hyperactivity Disorder (ADHD) based on resting state fMRI. We re-implement these frameworks, and take the latent representation from the encoder as the learned embedding. More details about the configuration of those baselines refer to supplemental materials.

Table 3.3: The averaged prediction accuracy (ACC), F1 score (F1), Precision, Recall and Area Under the Curve (AUC) of baselines and the proposed framework on brain state prediction task. For each cognitive task, the results are averaged among all subjects in the test dataset. Red and blue denotes the best and the second-best results, respectively.

Tasks	Methods	ACC	F1	Precision	Recall	AUC
Emotion	AE	0.6499	0.5915	0.6297	0.5926	0.7701
	STAAE (Dong et al., 2020)	0.5341	0.4641	0.5292	0.4745	0.6705
	DRVAE (Qiang et al., 2021)	0.5623	0.5096	0.5819	0.5084	0.6939
	DSRAE (Q. Li, Dong, Ge, Qiang, et al., 2021)	0.6182	0.5655	0.6297	0.5633	0.7485
	TCAE (ours)	0.7315	0.6747	0.7465	0.6712	0.8261
Language	AE	0.8211	0.6303	0.7175	0.6227	0.8789
	STAAE (Dong et al., 2020)	0.7821	0.5383	0.5268	0.5564	0.8056
	DRVAE (Qiang et al., 2021)	0.8167	0.5939	0.6849	0.5976	0.8598
	DSRAE (Q. Li, Dong, Ge, Qiang, et al., 2021)	0.8216	0.5957	0.6829	0.6002	0.8542
	TCAE	0.8550	0.5881	0.5713	0.6079	0.8758
Motor	AE	0.5545	0.5212	0.5367	0.5363	0.8416
	STAAE (Dong et al., 2020)	0.3638	0.2764	0.3231	0.3067	0.7007
	DRVAE (Qiang et al., 2021)	0.4384	0.3729	0.4152	0.3938	0.7599
	DSRAE (Q. Li, Dong, Ge, Qiang, et al., 2021)	0.5039	0.4555	0.4845	0.4753	0.8126
	TCAE	0.6426	0.6136	0.6347	0.6262	0.8908
WM	AE	0.4795	0.4330	0.4533	0.4294	0.7901
	STAAE (Dong et al., 2020)	0.3000	0.1852	0.2133	0.2093	0.6629
	DRVAE (Qiang et al., 2021)	0.3924	0.3217	0.3407	0.3251	0.7378
	DSRAE (Q. Li, Dong, Ge, Qiang, et al., 2021)	0.4176	0.3545	0.3723	0.3574	0.7589
	TCAE	0.5822	0.5522	0.5785	0.5476	0.8456

In Table 3.3, we report the averaged brain state classification results over all subjects including accuracy (ACC), F1 scores (F1), precision, recall, area under the curve (AUC) for each task, respectively. In Table 3.4, we report the number fo parameters and the number of multiply–accumulate operations (MACs) of those baselines and the proposed method. The dimension of embedding vector is set as 64 for all mehtods. It is observed that the proposed TCAE embedding framework significantly outperforms all other baselines in terms of all metrics

Table 3.4: The number of parameters (Params) and the number of multiply–accumulate operations (MACs) of different baselines and the proposed framework.

Methods	Params (M)	MACs (G)
AE	3.68	10.29
STAAE	14.70	41.32
DRVAE	15.13	42.54
DSRAE	15.12	42.52
TCAE	3.75	10.47

(two-sample one-tailed un-pair-wise t-test ($p < 0.025$, corrected)) except on Language task. On Emotion, Motor and WM tasks, TCAE demonstrates superior performance far beyond the baselines with much less parameters and MACs (Table 3.4), indicating the effectiveness of introducing self-attention mechanism to explore the temporal correlations. It is also noted that the baseline AE exceed other compared methods with the second best performances. This is probably because other baselines such as DRVAE are designed for a specific task which requires a specially designed architecture with more parameters. However, an architecture with more parameters may not be generalizable for our embedding task and thus degenerates the performance. On Language task, the performances of all methods are comparable with an accuracy around 0.8. This is probably because the brain states in Language task (3 classes) is relatively easy to be recognized compared with other tasks (e.g., Motor task have 7 classes and WM task have 9 classes), so the performance differences become smaller. Overall, these experimental results suggest that the proposed TCAE framework is compact and able to learn a more generalizable and effective embedding compared with other baselines.

Interpretation of the Embedding Space

In this subsection, we firstly explore the temporal activity patterns of each digit as illustrated in Section 3.2.3. With the extracted temporal pattern of each digit, we compute the Pearson Correlation Coefficient (PCC) between each temporal pattern and the Hemodynamic Response Function (HRF) responses of task stimuli. Among all digits, the one with the highest PCC value are selected as the task-relevant digit which is an indicator of the embedding' relevance to task stimuli. Here, we randomly select four subjects as examples to show the

temporal pattern of task-relevant digit from the TCAE model as well as the corresponding HRF responses in Figure 3.6.

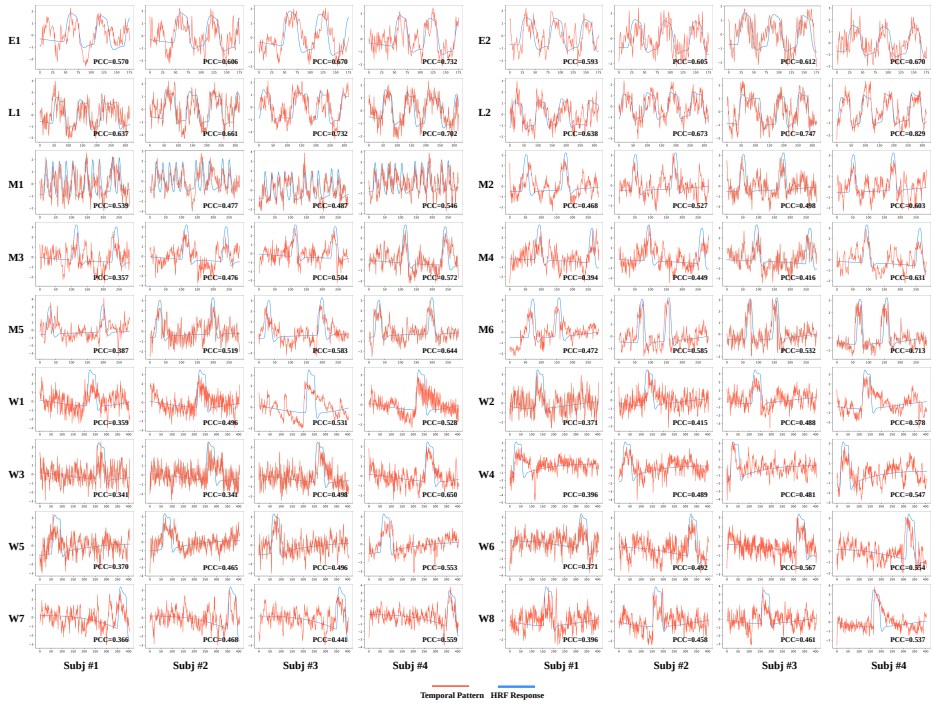

Figure 3.6: The temporal pattern of task-relevant digit from TCAE model compared with HRF responses of 4 randomly selected subjects for each task stimulus, respectively. Abbreviations: E1: 'Faces' stimulus; E2: 'Shapes' stimulus; L1: 'Math' stimulus; L2: 'Story' stimulus; M1: 'Task cues' stimulus; M2: 'Left foot' stimulus; M3: 'Right foot' stimulus; M4: 'Left hand' stimulus; M5: 'Right hand' stimulus; M6: 'Tongue' stimulus; W1: '0Back body' stimulus; W2: '0Back faces' stimulus; W3: '0Back places' stimulus; W4: '0Back tools' stimulus; W5: '2Back body' stimulus; W6: '2Back faces' stimulus; W7: '2Back places' stimulus; W8: '2Back tools' stimulus.

Generally, the temporal pattern of task-relevant digit matches the corresponding HRF response well, indicating that the digits of the learned embedding may be correlated with task stimuli. To quantitatively measure such correlation, we average the PCC of all subjects and compare it with those from baseline models in Table 3.5. It is observed that in Motor and WM task, the averaged PCC of TCAE model is larger than all compared baselines. However, in the Language task, AE and STAAE have the highest PCC for two task designs, respectively. In Emotion task, AE has the highest PCC in 'Shapes' stimulus while the TCAE has the highest PCC in 'Faces' stimulus. A possible reason is that for Motor and WM task, the responses of task stimuli are more complex and harder to be decoded and decomposed from the raw fMRI data. Our em-

beddings from TCAE model can better characterize such responses, which is in alignment with the highest brain state prediction accuracy in Table 3.3. While in the Language task, the responses are quite straightforward and can be easily captured by other deep learning models. It is consistent with overall higher brain state prediction accuracy in Table 3.3 for all compared methods. The TCAE embedding model may focus on more intrinsic responses and patterns which are still task-relevant but discriminative, resulting in a comparable accuracy in brain state prediction but relatively lower PCC than baselines.

Table 3.5: Mean (± standard deviation) PCC (Pearson Correlation Coefficient) between the temporal pattern of task-relevant digit and the HRF response. Abbreviations: E1: 'Faces' stimulus; E2: 'Shapes' stimulus; L1: 'Math' stimulus; L2: 'Story' stimulus; M1: 'Task cues' stimulus; M2: 'Left foot' stimulus; M3: 'Right foot' stimulus; M4: 'Left hand' stimulus; M5: 'Right hand' stimulus; M6: 'Tongue' stimulus; W1: '0Back body' stimulus; W2: '0Back faces' stimulus; W3: '0Back places' stimulus; W4: '0Back tools' stimulus; W5: '2Back body' stimulus; W6: '2Back faces' stimulus; W7: '2Back places' stimulus; W8: '2Back tools' stimulus. Red and blue denotes the highest and the second-highest PCC, respectively.

Tasks	Stimulus	Methods				
		AE	STAAE	DRVAE	DSRAE	TCAE
Emotion	E1	0.45±0.09	0.48±0.11	0.45±0.11	0.48±0.11	0.53±0.09
	E2	0.51±0.11	0.40±0.11	0.40±0.10	0.45±0.10	0.49±0.09
Language	L1	0.76±0.10	0.73±0.09	0.59±0.08	0.68±0.08	0.60±0.08
	L2	0.61±0.10	0.72±0.09	0.68±0.10	0.71±0.08	0.59±0.08
Motor	M1	0.41±0.09	0.38±0.10	0.39±0.08	0.39±0.08	0.42±0.08
	M2	0.34±0.05	0.30±0.09	0.31±0.07	0.34±0.08	0.38±0.07
	M3	0.35±0.05	0.28±0.08	0.30±0.07	0.34±0.07	0.40±0.07
	M4	0.34±0.05	0.30±0.08	0.33±0.07	0.34±0.07	0.37±0.07
	M5	0.33±0.04	0.29±0.08	0.32±0.07	0.34±0.07	0.40±0.07
	M6	0.40±0.05	0.47±0.09	0.42±0.08	0.45±0.10	0.52±0.09
WM	W1	0.29±0.04	0.26±0.07	0.31±0.07	0.31±0.08	0.36±0.07
	W2	0.27±0.05	0.27±0.07	0.30±0.07	0.32±0.07	0.34±0.08
	W3	0.28±0.04	0.28±0.07	0.31±0.08	0.32±0.08	0.35±0.07
	W4	0.28±0.05	0.30±0.06	0.34±0.07	0.36±0.06	0.37±0.07
	W5	0.29±0.05	0.25±0.09	0.30±0.08	0.32±0.08	0.35±0.08
	W6	0.28±0.05	0.23±0.08	0.29±0.07	0.29±0.08	0.35±0.08
	W7	0.28±0.04	0.27±0.07	0.31±0.07	0.31±0.06	0.35±0.07
	W8	0.27±0.05	0.25±0.08	0.32±0.08	0.31±0.08	0.34±0.08

We also demonstrate the spatial composition of digits and compare them with spatial maps from the general linear model (GLM) (Monti, 2011) and resting state networks (RSNs) reported in (Smith et al., 2009). For each task stimulus, we select the most similar spatial composition from our model and visu-

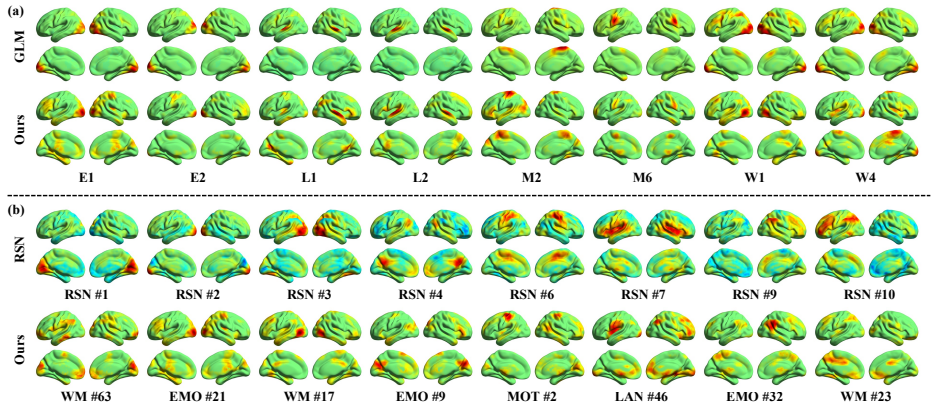

Figure 3.7: Visualizations of spatial compositions from the selected digits. (a) The comparison of the GLM maps and the most similar spatial composition for each task stimulus. (b) The comparison of the RSNs and the most similar spatial composition selected from different tasks. Abbreviations: E1: 'Faces' stimulus; E2: 'Shapes' stimulus; L1: 'Math' stimulus; L2: 'Story' stimulus; M2: 'Left foot' stimulus; M6: 'Tongue' stimulus; W1: '0Back body' stimulus; W4: '0Back tools' stimulus; EMO: Emotion; LAN: Language; MOT: Motor WM: Working Memory; RSN: Resting State Network.

alize them in Figure 3.7(a). It is observed that for each stimulus, we can find a similar spatial composition from a digit. This indicates that our embedding framework extracts the features that are in close relationship with task stimuli and the learned embedding has encoded the corresponding information. In Figure 3.7(b), we select and visualize the spatial compositions similar to RSNs. RSN #1-#3 correspond to visual networks locating on occipital lobe, and the visual system of human brain is involved in Emotion and Working Memory task. RSN #4 is the Default Mode Network (DMN). It is widely reported that DMN plays a crucial role for the emotion process (Satpute & Lindquist, 2019; X. Xie et al., 2016; J. Zhao et al., 2017). RSN #6 is the somatomotor network and we successfully match a spatial composition with it from the Motor task. RSN #7 corresponds to Auditory Network. In the acquisition of Language task fMRI data, the participants were presented with brief auditory stories. RSN #9 and RSN #10 are frontoparietal networks located on frontal lobe and parietal lobe. The involvements of these networks in emotion regulation and working memory were reported widely in literature (Braunlich et al., 2015; Harding et al., 2015; Lindquist & Barrett, 2012; Sabatinelli et al., 2014; Salazar et al., 2012). Overall, these results indicate that the learned embedding and the embedding

space are neuroscientifically meaningful and interpretable, and also coincide with the task designs.

Implementation Details

In our experiments, we uniformly set the embedding size as 64 for the proposed model and all compared baselines. In TCAE model, the m, k, v in TCAE model are set to 64 and d_{ff} is set to 128. For the MLP for brain state prediction task, the dimension of two layers are 64/32, respectively. The framework is implemented with PyTorch (https://pytorch.org/) deep learning library. We use the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is 16 and the model is trained for 100 epochs with a learning rate 0.005 for each tasks on a single GTX 1080Ti GPU. It is noted that all experiments were performed on testing dataset based on the model with the lowest loss on validation dataset.

3.2.5 Conclusion

In this section, we represented the human brain function in a general, comparable, and stereotyped space through a novel transformer-based embedding framework, with which the brain activities measured by fMRI data at different time points and across populations are meaningfully and compactly represented. The experimental results on brain state prediction downstream task demonstrated the effectiveness and generalizability of learned embedding. It was also found that the embedding and embedding space is interpretable and neuroscientifically meaningful. Our future works include employing the proposed framework to explore the connections between the semantic spaces of human brain function and those in deep visual models or language models such as ViT (Dosovitskiy et al., 2020) or BERT (Devlin et al., 2018). We will also apply the embedding for disease diagnosis such as ADHD and Alzheimer's disease with resting state fMRI data.

Chapter 4

Coupling the Semantics of ANNs and BNNs

4.1 Overview

Artificial neural networks (ANNs), originally inspired by biological neural networks (BNNs), have achieved remarkable successes in many tasks such as visual representation learning. However, whether there exists semantic correlations/connections between the visual representations in ANNs and those in BNNs remains largely unexplored due to both the lack of an effective tool to link and couple two different domains, and the lack of a general and effective framework for representing the visual semantics in BNNs such as human functional brain networks (FBNs). To answer this question, we propose a novel computational framework, Synchronized Activations (Sync-ACT), to couple the visual representation spaces and semantics between ANNs and BNNs in human brain based on naturalistic functional magnetic resonance imaging (nfMRI) data.

4.2 Background

Inspired by the biological neural networks (BNNs), artificial neural networks (ANNs) have achieved great success in a variety of tasks and scenarios due to their powerful representation ability (LeCun et al., 2015). In computer vision (CV) field, convolutional neural networks (CNNs) (LeCun, Bengio, et al., 1995) hierarchically learn the visual representations of images/videos as low-level to high-level features in embedding space and have been widely used in many real-word applications (Khan et al., 2020; LeCun et al., 2015). Recent Vision Transformer (ViT) (Dosovitskiy et al., 2020) demonstrates promising performance by representing the image as a sequence of patches and embedding the image

patches as latent vectors, based on which the dependencies/correlations among those vectors are modeled. However, the semantics of those embedding spaces for visual representation are not manifest for human perception and challenge us for a comprehensive understanding of representation learning of ANNs. To unveil and describe the semantics of latent space of ANNs for visual representation, increasing efforts have been devoted to interpret the ANNs' behaviors and annotating their neurons with semantic concepts (Bau et al., 2017; Zhou et al., 2016). For example, (Bau et al., 2017) proposed to label the hidden units of convolutional layer with visual concepts from a broad dataset. A recent study employed fine-grained natural language description to annotate the semantics of neurons in various ANNs (Hernandez et al., 2021). Despite the remarkable progresses achieved by these methods, whether the visual representation space of ANNs retains biologically meaningful semantics as in the initial inspiration, BNNs, is still an open question.

In the field, researchers now have employed naturalistic functional magnetic resonance imaging (nfMRI) to assess the activity and functional mechanism of BNNs (Hu et al., 2010; T. Liu et al., 2014; Ren et al., 2021; J. Wang et al., 2017), e.g., functional brain networks (FBNs), under the naturalistic stimuli such as real-life images and video streams. This natural stimulus fMRI paradigm provides a powerful tool for investigating the visual perception of human brain and representing the corresponding semantics (T. Liu et al., 2014), allowing us to answer the aforementioned question and annotate the neurons in ANNs with biological description even further. However, the current approaches for representing high-dimensional fMRI data, e.g., matrix decomposition based on independent component analysis (ICA) (Calhoun & Adali, 2006) and sparse dictionary learning (SDL) (Lv et al., 2014), are commonly used for task and resting state fMRI. Considering the brain activities encoded by nfMRI are dynamic and complex, it is quite challenging to interpret and describe the semantics perceived by the human brain. In addition, the brain responses evoked by naturalistic stimuli exhibit great inter-subjects variability (Golland et al., 2007; Ren et al., 2017), while those existing methods do not encode the regularity and variability of different brains, and thus do not offer a general, comparable, and stereotyped embedding space for representing the brain activity and functional semantics. Recently, deep learning approaches demonstrated superior performance in modeling fMRI data (Dong et al., 2019; Q. Li, Zhang, et al., 2021; H. Liu et al., 2019; H. Wang et al., 2018; W. Zhang et al., 2019; L. Zhao, Dai, et al., 2021). However, as far as we know, these deep learning methods were designed for specific tasks. A more general and effective framework of embedding brain function and representing semantics under naturalistic stimuli is still

much needed. In parallel, linking such human brain's functional embedding and semantics representation with external natural stimulus is very desirable and significant.

4.2.1 Related Works

4.2.2 Visual Representation Interpretation

Interpreting the behavior of deep neural networks and the learned visual representation has attracted growing interest in the CV field. As the semantics of visual representation in deep networks are not manifest for human perception, a possible approach is to visualize the activation of neurons and characterize the visual concept they recognize (Bau et al., 2017; Dalvi et al., 2019; Morcos et al., 2018; Mu & Andreas, 2020). For example, (Bau et al., 2017) labels the neurons (i.e., convolutional filters) of CNNs with visual concepts by aligning the activation of neurons with a set of images with semantic concepts. Mu & Andreas (Mu & Andreas, 2020) searches the compositional logical concepts defined on primitive visual concepts that closely approximate neuron behavior. A recent study employed fine-grained natural language description to annotate the semantics of neurons in various ANNs (Hernandez et al., 2021) by maximizing the mutual information between the language description and imaging regions in which the neuron is activated. Our work is inspired by and in line with the aforementioned interpretation studies. The neuron's behaviors are measured by the maximum activation over time, forming a temporal activation series with which the temporal activation of FBNs is compared and correlated for alignment. In this way, we contribute a biologically meaningful description of the neurons in ANNs.

4.2.3 fMRI Data Representation

A major challenge for fMRI data representation learning is that the number of voxels in 4D spatiotemporal fMRI data is greatly larger than the number of subject brains (Mwangi et al., 2014). To deal with this imbalance, a variety of computational tools have been proposed to select the task-related features and discard the redundant ones as well as the noises (Calhoun & Adali, 2006; Lv et al., 2014). For example, independent component analysis (ICA) (Calhoun & Adali, 2006) and sparse dictionary learning (SDL) were employed to decompose the fMRI as two compact matrices (temporal and spatial patterns). However, the temporal and/or spatial patterns obtained from ICA or SDL based methods are not intrinsically comparable across different individual brains. Recently,

deep learning has been widely employed in fMRI data modeling and achieved superior results over the traditional matrix decomposition methods (Dong et al., 2019; Q. Li, Zhang, et al., 2021; H. Liu et al., 2019; H. Wang et al., 2018; W. Zhang et al., 2019; L. Zhao, Dai, et al., 2021). However, as far as we know, prior deep learning models of fMRI data were not specifically designed towards a general, comparable and compact representation of brain function. Instead, prior methods were designed for some specific tasks, such as fMRI time series classification (H. Liu et al., 2019), brain network decomposition (Dong et al., 2019; Q. Li, Dong, Ge, Qiang, et al., 2021), brain state differentiation(H. Wang et al., 2018), among others. Even though some methods derive comparable temporal patterns (Q. Li, Dong, Ge, Qiang, et al., 2021; Q. Li, Zhang, et al., 2021), which might be suitable for our objective, they still rely on matrix decomposition to obtain spatial patterns that are not comparable across different individuals. In this work, we proposed a more general and unified framework to represent the fMRI data from different subjects as functional brain networks and their temporal activations in a general, comparable and stereotyped latent space. This design enables us to explore the correlation between the semantics of this latent space and those in CNNs.

4.2.4 Connection of ANNs and BNNs

Current ANNs are inspired by the BNNs at the beginning. For example, CNNs are inspired by the hierarchical organization of the vision systems in the human's primary vision cortex (Kim et al., 2016). Recently, there is a growing interest in exploring the potential connections between ANNs and BNNs. For instance, the receptive field analysis reveals that the receptive fields of filters in CNNs become progressively larger (W. Luo et al., 2016) and more complex, which is similar to the ventral pathway in cerebral cortex (Barrett et al., 2019). The filters in the last convolutional layer have class-specific receptive fields akin to concept-cells in the visual cortex (Mahendran & Vedaldi, 2016). (Yamins & DiCarlo, 2016; Yamins et al., 2014) synthesized CNNs outputs by linear regression to predict the neural responses in both the V4 and inferior temporal (IT) cortex. This shows a strong correlation between a CNN's categorization performance and its ability to predict individual IT neural responses, implicitly indicating the potential representation similarity between ANNs and BNNs. (You et al., 2020) proposed a graph-based representation of ANNs called relational graph, and found that top-performing ANNs have graph structures similar to those of BNNs. Inspired by these studies, we explore the semantic similarity of visual representations in CNNs and the functional representation of the human brain, providing a novel insight into the connection between ANNs and BNNs.

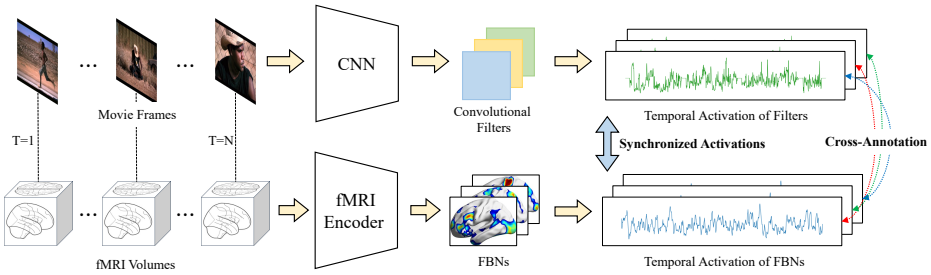

Figure 4.1: The proposed Sync-ACT framework. The temporal activation of FBNs and convolutional filters are synchronized for matching the embedding space and cross-annotation.

4.3 Methods

In this section, we propose a novel computational framework, Synchronized Activations (Sync-ACT) (Figure 4.1), to explore the connections of the visual representation space and semantics between ANNs and BNNs in human brain. Based on Sync-ACT, we describe and annotate the neurons in ANNs with biologically meaningful descriptions for the first time, bridging the gaps between these two drastically different domains.

4.3.1 Formulation of Sync-ACT Framework

Even though the ANNs are originally inspired by the BNNs, the input/output, operating, and reasoning processes of neural networks in the two domains are quite different and not comparable. Our intuition is to avoid being trapped by the remarkable differences but focus on their responses, such as the activation of neurons, to the external stimuli. In this way, the behavior of the neural networks measured by the responses (i.e., the temporal activation of neurons) in two domains can be directly compared if the stimuli are synchronized, and thus the most similar neurons in two domains can be easily identified and paired. Let $\mathcal{F} : X_a \rightarrow Y_a$ represents an artificial neural network, and $f_i(\mathbf{x_a})$ represents the temporal activation of neuron f_i with respect to stimulus sequence $\mathbf{x_a}$. Similarly, let $\mathcal{G} : X_b \rightarrow Y_b$ represent a biological neural network, and $g_j(\mathbf{x_b})$ denotes the temporal activation of neuron g_j with the stimulus sequence $\mathbf{x_b}$. If the stimuli $\mathbf{x_a}$ and $\mathbf{x_b}$ are synchronized, the paired neuron of f_i in the biological

neural network $\mathcal{G}$ can be defined by:

$$\text{Sync-ACT}(f_i, \mathcal{G}) = \arg\max_{g_j \in \mathcal{G}} \ \delta(f_i(\mathbf{x_a}), g_j(\mathbf{x_b})), \tag{4.1}$$

where $\delta(\cdot)$ is the measurement of similarity between two temporal activations. Similarly, we could define the paired neuron of g_i in the artificial neural network $\mathcal{F}$ as:

$$\text{Sync-ACT}(g_i, \mathcal{F}) = \arg\max_{f_j \in \mathcal{F}} \ \delta(g_i(\mathbf{x_b}), f_j(\mathbf{x_a})). \tag{4.2}$$

In this work, we adopt the Pearson correlation coefficient (PCC) for similarity measurement $\delta(\cdot)$. Based on Eq. (4.1) and Eq. (4.2), we can obtain the paired neuron g_j/f_j for any f_i/g_i by choosing the one with the most significant similarity value. With the neuron pairs, the semantics of one neural network can be used to annotate the other, i.e., the cross-annotation. We define the semantic description of neurons f_i/ g_i as d_{f_i}/d_{g_i}. The cross-annotation of paired neurons is then denoted as $d_{f_i} \rightarrow d_{g_j}$ and $d_{g_i} \rightarrow d_{f_j}$.

We adopted nfMRI data to evaluate the Sync-ACT framework by leveraging the fact that, during nfMRI scan, video frames (stimuli for both ANNs and BNNs) and functional brain responses measured by fMRI are temporally aligned in an intrinsic fashion.

4.3.2 fMRI Embedding Framework

In the brain imaging field, it is common to represent the brain function as interactions of FBNs and the corresponding temporal patterns. Thus, the FBNs can be viewed as the neurons of BNN in the human brain, and temporal patterns represent the activations of those FBNs. However, the previous methods including the deep learning ones do not offer a general and stereotyped space in modeling FBNs. The FBNs and corresponding temporal activations are not intrinsically comparable across different individual brains.

So, in this section, we propose a general fMRI embedding framework to represent brain function as FBNs and derive the temporal activations in a unified and comparable embedding space. Specifically, the fMRI embedding framework has an encoder-decoder architecture. Figure 4.2 illustrates the major components in the encoder. The rearranged 2D fMRI signal matrix $\mathbf{S} \in \mathbb{R}^{t \times n}$, where t is the number of time points and n is the number of voxels, is firstly embedded as a new feature matrix $\mathbf{S}_f \in \mathbb{R}^{t \times m}$ through a learnable transformation matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$, where m is the reduced feature dimension ($m \ll n$). This transformation can be viewed as compressing the voxels in 3D volume space into m components, i.e., m functional brain networks, by linear combination. The

columns in the transformation matrix $\mathbf{W}$ recorded the contributions of the voxels to each FBN, i.e., the composition of each FBN, which can be mapped back to 3D volume space for visualizing the spatial pattern of FBN. It is noted that the linear transformation in the encoder parameterized by $\mathbf{W}$ is optimized in a data-driven manner and consistent for all subjects, which guarantees the comparability of $\mathbf{S}_f$ for all subjects.

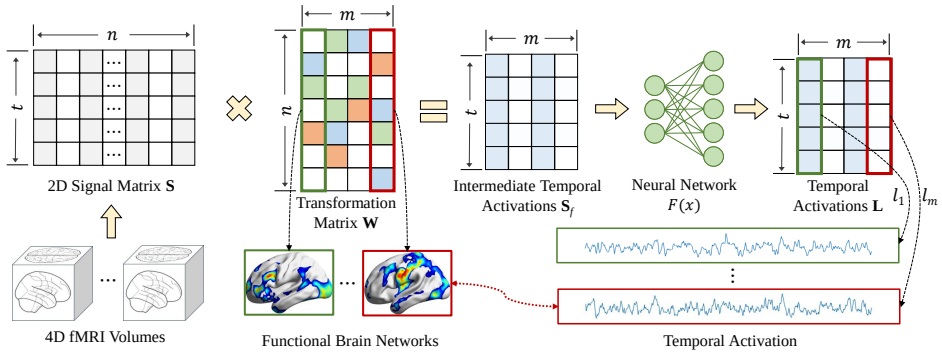

Figure 4.2: The illustration of the encoder in fMRI embedding framework. The green and red boxes correspond to the first and last FBNs and their temporal activations.

The row vectors in matrix $\mathbf{S}_f$ recorded the activation of all resulted FBNs at different time points, and the column vector represents the temporal activation of a specific FBN. We further model the temporal correlations of the column vectors with a neural network $F(x)$. Here, we explore two popular neural networks for modeling temporal data, long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) and multi-head self-attention (MSA) module in the Transformer model (Vaswani et al., 2017). The column vector l_i in the resulted matrix $\mathbf{L} = F(\mathbf{S}_f), \mathbf{L} \in \mathbb{R}^{t \times m}$ is the temporal activation of the i^{th} FBNs, which encodes the regularity and variability of different brains in the same latent space. We average the vector l_i from all subjects in the testing dataset as the temporal activation for neuron g_i

The decoder has a symmetrical architecture as the encoder. The whole framework is optimized in an unsupervised manner by minimizing the Mean Square Error (MSE) between the original fMRI signals matrix $\mathbf{S} \in \mathbb{R}^{t \times n}$ and their corresponding reconstruction $\mathbf{S}' \in \mathbb{R}^{t \times n}$.

4.3.3 Neurons and Activations in CNNs

We adopt CNNs as the representative ANNs in this work because of CNNs' powerful visual representation ability and wide application in many computer vision tasks. We recognize the convolutional filters as the neurons in ANN. To

derive the temporal activation of CNN's filters, we adopt a simple but effective strategy by collecting the feature map $A_{f_i}(x_t)$ of each CNN filter f_i with image x_t from the image sequence $\mathbf{x}$ at time point t. Then the maximum value in feature map $\max(A_{f_i}(x_t))$ is extracted to represent the activation degree of filter f_i at time point t, resulting in the temporal activation $f_i(\mathbf{x})$. When the image sequence $\boldsymbol{x}$ is the corresponding movie frames of nfMRI, the derived temporal activations are automatically synchronized with the ones in FBNs. This strategy can be easily applied to any pre-trained CNN model.

With the obtained temporal activations $f_i(\mathbf{x_a})$ and $g_j(\mathbf{x_b})$, given synchronized stimuli $\mathbf{x_a}$ and $\mathbf{x_b}$, we will be able to pair the neurons between ANNs and BNNs and perform cross-annotation following Eq. (4.1) and Eq. (4.2). It is noted that the Sync-ACT is a general framework that is also compatible with temporal activations derived from other representation methods, such as those potentially from other fMRI embedding methods or ViT.

4.4 Experiments

Datasets. In this study, we adopt the publicly available HCP 7T movie-watching fMRI dataset (http://www.humanconnectomeproject.org/) of S1200 release (Barch et al., 2013). The dataset contains 184 subjects who were scanned in 4 runs while watching short independent films and Hollywood movie excerpts concatenated into .mp4 files. The important fMRI acquisition parameters are as follows: 130×130 matrix, 85 slices, TR=1.0 s, TE=22.2 s, 208 mm FOV, flip angle = 45°, 1.6 mm isotropic voxels. The fMRI data are preprocessed by HCP minimal preprocessing pipeline (Glasser et al., 2013). Then, we downsample and register the preprocessed fMRI data into the standard MNI 152 4 mm space for reducing the computational overhead. The movie clips have a resolution of 1024×720 pixels (24fps). We extracted the last movie frame in each second as the corresponding image for fMRI data and resized it with a resolution of 256×180 pixels.

In addition, we adopt the StudyForrest movie-watching fMRI dataset (https://www.studyforrest.org/) (Hanke et al., 2016) with 15 subjects watching 2 hours of Forrest Gump movie. The important acquisition parameters are as follows: 80×80 matrix, 35 slices, TR=2.0 s, TE=30.0 ms, 240 mm FOV, flip angle = 90°, 3.0 mm isotropic voxels. The fMRI data in StudyForrest dataset are preprocessed using fMRIPrep (Esteban et al., 2019). The movie clips have a resolution of 1280×720 pixels (25fps). We extracted the last movie frame every two seconds as the corresponding image for fMRI data and resized it with a resolution of 320×180 pixels. It is noted that the data quality and spatial/temporal resolution

of the StudyForrest dataset are relatively low and only 15 subjects are available, so we just use it for validation.

For both datasets, the time series from the voxels of preprocessed fMRI data is rearranged into a 2D array with zero mean and standard deviation one. We used 60%/10% of the subjects for model training/validation and the rest 30% for testing. Unless we specifically mentioned, all the experimental results are based on testing data of HCP 7T movie-watching dataset.

Implementation Details. In our experiments, we uniformly set the number of derived FBNs from our fMRI embedding framework as 64. The investigation of its influences can be found in supplementary materials. For StudyForrest fMRI dataset, we cut the fMRI data from run #1 to run #7 with 430 time points and consider them as the same samples for training due to the lack of subjects. The inference is conducted for each run with uncut data. The framework is implemented with PyTorch (https://pytorch.org/) deep learning library. We use the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is 16 and the model is trained for 100 epochs with an initial learning rate 0.01 for both tasks on a single GTX 1080Ti GPU.

4.4.1 Correlations of Representations in Two Spaces

Correlations with different CNN Models. We firstly explored the correlation of temporal activations between the FBNs and convolutional filters in a variety of CNN models pre-trained on the ImageNet dataset (J. Deng et al., 2009) and Places365 dataset (Zhou et al., 2017). The filters in the last convolutional layer for all CNNs are selected and paired with FBNs. In Table 4.1, the averaged PCC values over all pairs across different CNN models and different runs of HCP 7T movie-watching dataset are reported. It is observed that for almost all CNN models across different runs (except the AlexNet, VGG-16, ResNet-18 at Run #1 and/or Run #2), the correlation measured by PCC is statistically significant with values larger than 0.2. The significance is well reproduced on the models pre-trained on different datasets (ImageNet and Places365) though the averaged PCC values may vary on different pre-trained models and runs.

We perform a similar analysis on the StudyForrest movie-watching fMRI dataset for validation, and the results are reported in Table 4.2 with CNN models pre-trained on ImageNet dataset. However, the averaged PCC values are smaller than those on HCP 7T movie-watching dataset. This might be due to that the number of subjects in StudyForrest dataset (15) is smaller than those in HCP 7T fMRI dataset (184); the spatial/temporal resolution, image quality and signal-to-noise ratio of nfMRI data are worse than those in HCP 7T fMRI

Table 4.1: The averaged PCC on HCP dataset for the pairs of FBNs and the filters on CNN models pre-trained on ImageNet and Places365 dataset. The correlations measured by PCC in this table are all statistically significant(p-value≤ 0.05) for different runs.

Methods	ImageNet (J. Deng et al., 2009)				Places365 (Zhou et al., 2017)			
	Run #1	Run #2	Run #3	Run #4	Run #1	Run #2	Run #3	Run #4
AlexNet (Krizhevsky et al., 2012)	0.2323	0.2223	0.2558	0.2607	0.2651	0.2374	0.2788	0.2774
VGG-16 (Simonyan & Zisserman, 2014)	0.2376	0.2176	0.2654	0.2617	-	-	-	-
ResNet-18 (K. He et al., 2016b)	0.2415	0.2267	0.2516	0.2530	0.2703	0.2536	0.2896	0.2931
ResNet-50 (K. He et al., 2016b)	0.2862	0.2660	0.2942	0.3022	0.3008	0.2745	0.3076	0.3159
DenseNet-161 (G. Huang et al., 2017)	0.3031	0.2767	0.3051	0.3152	0.3052	0.2879	0.3134	0.3199
Inception V3 (Szegedy et al., 2016)	0.2720	0.2615	0.2747	0.2895	-	-	-	-
ShuffleNet V2 (Ma et al., 2018)	0.2663	0.2515	0.2742	0.2831	-	-	-	-
MobileNet V2 (Sandler et al., 2018)	0.2628	0.2517	0.2635	0.2870	-	-	-	-
ResNeXt-50 (S. Xie et al., 2017)	0.2774	0.2607	0.2904	0.2940	-	-	-	-
MNASNet (Tan et al., 2019)	0.2612	0.2422	0.2669	0.2699	-	-	-	-

Table 4.2: The averaged PCC on StudyForrest dataset for the pairs of FBNs and the filters on CNN models pre-trained on ImageNet dataset. The PCC value with a * marker indicates that the corresponding pairs (less than 10) are not statistically significant(p-value≤ 0.05), otherwise, it is significant.

Methods	Run #1	Run #2	Run #3	Run #4	Run #5	Run #6	Run #7	Run #8
AlexNet (Krizhevsky et al., 2012)	0.1369*	0.1659*	0.1712	0.1540	0.1609	0.1724	0.1590	0.1961
VGG-16 (Simonyan & Zisserman, 2014)	0.1398*	0.1836	0.1822	0.1777	0.1752	0.1839	0.1732	0.2112
ResNet-18 (K. He et al., 2016b)	0.1474*	0.1854	0.1851	0.1708	0.1821	0.1840	0.1823	0.2134
ResNet-50 (K. He et al., 2016b)	0.1584	0.2075	0.2111	0.1956	0.1983	0.2113	0.1983	0.2344
DenseNet-161 (G. Huang et al., 2017)	0.1673	0.2143	0.2154	0.1985	0.2049	0.2158	0.2049	0.2497
Inception V3 (Szegedy et al., 2016)	0.1617	0.2015	0.2070	0.1967	0.1979	0.2024	0.1956	0.2371
ShuffleNet V2 (Ma et al., 2018)	0.1488	0.1922	0.2004	0.1900	0.1898	0.1986	0.1920	0.2324
MobileNet V2 (Sandler et al., 2018)	0.1499	0.1961	0.2001	0.1898	0.1906	0.1958	0.1929	0.2350
ResNeXt-50 (S. Xie et al., 2017)	0.1624	0.2083	0.2046	0.1886	0.1982	0.2070	0.1936	0.2365
MNASNet (Tan et al., 2019)	0.1553	0.2036	0.2015	0.1907	0.2013	0.2029	0.1974	0.2387

dataset. However, it is still found that the correlations are significant for almost all models and runs of fMRI except the AlexNet, VGG-16 and ResNet-18 on run #1 and/or run #2. Overall, these results consistently suggest that there exists a significant correlation between the convolutional filters in CNN model and FBNs in the human brain.

Correlations with different CNN Layers. We further assess the correlations of FBNs with convolutional filters in different layers of 4 different CNN models. The PCC values averaged over all FBN-filter pairs in each layer are reported in Table 4.3. We observe that the correlations are significant for pairs in the last two blocks/layers while some of them in the first two blocks/layers are not significant. The PCC values in different layers also show a trend that it

increases and reaches a peak at the third layer, which is in line with the literature study (Yamins et al., 2014) reporting that the model's intermediate layers are highly predictive of the brain's neural responses.

Table 4.3: The averaged PCC for the pairs of FBNs and filters in different convolutional layers of CNN model and the ratio of **NOT** statistically significant pairs. The colors red and blue denote the highest and the second-highest PCC value among different layers, respectively.

Methods	Layer	Run 1		Run 2		Run 3		Run 4	
		PCC	Ratio	PCC	Ratio	PCC	Ratio	PCC	Ratio
ResNet-18 (K. He et al., 2016b)	Block #1	0.2401	1/64	0.2012	2/64	0.2189	6/64	0.2286	0/64
	Block #2	0.2534	0/64	0.2150	0/64	0.2513	1/64	0.2480	0/64
	Block #3	0.2743	0/64	0.2483	0/64	0.2821	0/64	0.2791	0/64
	Block #4	0.2415	0/64	0.2267	0/64	0.2516	0/64	0.2530	0/64
ResNet-50 (K. He et al., 2016b)	Block #1	0.2607	0/64	0.2276	0/64	0.2664	0/64	0.2504	0/64
	Block #2	0.2876	0/64	0.2498	0/64	0.2851	0/64	0.2768	0/64
	Block #3	0.2962	0/64	0.2684	0/64	0.3059	0/64	0.3060	0/64
	Block #4	0.2862	0/64	0.2660	0/64	0.2942	0/64	0.3022	0/64
ShuffleNet V2 (Ma et al., 2018)	Stage #2	0.2634	0/64	0.2226	0/64	0.2511	1/64	0.2460	0/64
	Stage #3	0.2633	0/64	0.2365	0/64	0.2640	0/64	0.2647	0/64
	Stage #4	0.2749	0/64	0.2511	0/64	0.2861	0/64	0.2968	0/64
	Conv #5	0.2663	0/64	0.2515	0/64	0.2742	0/64	0.2831	0/64
ResNeXt-50 (S. Xie et al., 2017)	Block #1	0.2633	0/64	0.2280	0/64	0.2525	2/64	0.2478	0/64
	Block #2	0.2803	0/64	0.2501	0/64	0.2840	0/64	0.2769	0/64
	Block #3	0.2954	0/64	0.2635	0/64	0.3026	0/64	0.3030	0/64
	Block #4	0.2774	0/64	0.2607	0/64	0.2904	0/64	0.2940	0/64

4.4.2 PCC Variance in Different CNN Models

From Table 4.1 and Table 4.2, we can observe that the ResNet-50 and DenseNet-161 have higher PCC values than the VGG-16 and ResNet-18, which suggests that the PCC values may have correlations with the model's representation ability and performance. To verify this hypothesis, we conduct linear regression to model the relationship between the PCC values and the CNN model's top-1 accuracy on ImageNet classification task. The results are reported in Figure 4.3. We found that the relationship can be represented well by the linear model with R^2 larger than 0.7 and p-value smaller than 0.05 across all 4 runs of HCP 7T movie-watching task fMRI dataset. This result indicates that if the visual representations of CNN models are more similar to those of human brain function, its performance on image classification tasks will be better. This is consistent with the study (You et al., 2020) finding that the top-performing ANNs have a graph structure similar to those of BNNs.

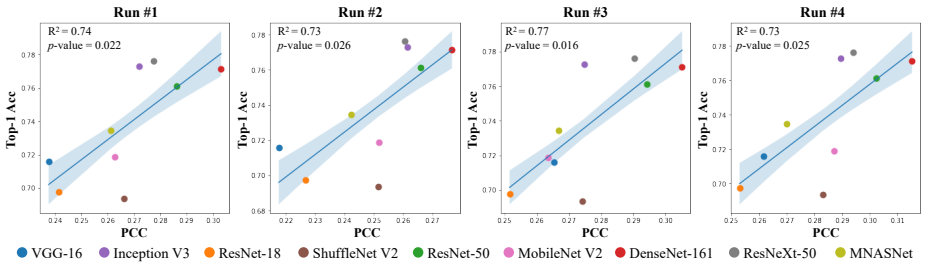

Figure 4.3: The linear regression modeling the relationship between PCC and CNN's top-1 image classification accuracy on ImageNet. Different CNN models are marked as circle with different color.

4.4.3 Visualizations of the Cross-Annotation

We conduct the cross-annotation based on Eq. (4.2) to pair each FBN with a filter at the last convolutional layer of ResNet-18 and visualize several sample pairs in Figure 4.4. The left panel in Figure 4.4 shows the FBNs to be paired and the corresponding semantic description from fMRI meta-analysis. The right panel shows the most activated images obtained by (Bau et al., 2017) from the movie frame sequence of paired CNN filters. The filter's corresponding semantic description and representative images are also demonstrated. In Figure 4.4, we found some interesting connections between the semantic description of the paired FBN and filters. For example, the description of FBN #25 is related to place and navigation, while the paired filters are labeled as rock and the representative images are related to some natural scenes. Such observation is obvious on some pairs, which is congruent with the results in Section 4.4.1. We provided more samples in supplementary materials for comparison.

4.4.4 Ablation Studies of fMRI Embedding Framework

We conduct the ablation studies for our fMRI embedding framework on three variants: a) encoder/decoder only has one linear transformation (LT) layer; b) encoder/decoder has one LT layer and two LSTM layers with tanh activation function; c) encoder/decoder has one LT layer followed by multi-head self-attention module. We measure the similarity of temporal activations of FBN-filter pairs identified by PCC. The averaged values for different metrics over all pairs and runs are reported in Table 4.4. Overall, the similarity of LT+LSTM and LT+MSA is larger than the linear transformation baseline. The LT+LSTM and LT+MSA have comparable performances in terms of similarity. However,

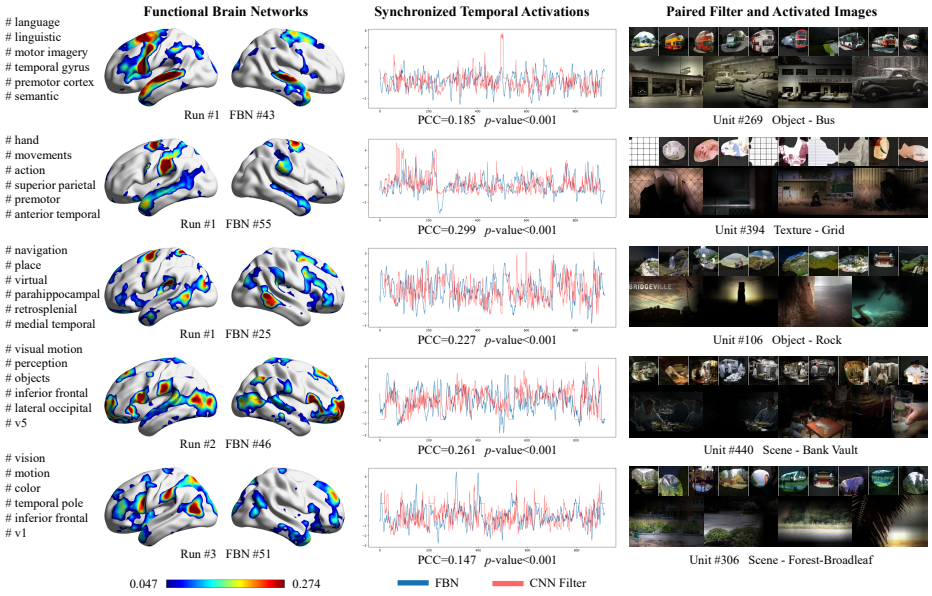

Figure 4.4: The visualization of FBN-Filter pairs obtained from our model. The left panel is the FBNs to be paired and semantic description from fMRI meta-analysis. The middle panel shows the synchronized activations from FBN and paired CNN filter. The right panel shows the most activated frames and the corresponding semantic description and filter's representative images in (Bau et al., 2017).

the LT+MSA has better FBNs quality. We provide the details in the supplementary materials. The results and analysis of our work are based on LT+MSA.

4.5 Conclusion

In this section, we proposed a novel computational framework, Sync-ACT, to couple the visual representation spaces and semantics between ANNs and BNNs in the human brain by synchronizing their activations to visual stimuli. We found a significant correlation in the semantics between the visual representations in CNNs and those in the human brain. Also, CNN's visual representation similarity to the human brain is closely related to its performance on the image classification tasks. In the future, our Sync-ACT model can be easily generalized to other naturalistic stimuli such as natural language and/or audio to explore the connection of the model's semantic space with the one in the human brain. Overall, our study introduces a general and effective paradigm to couple the ANNs and BNNs and provides novel insight into their connections.

Table 4.4: The comparison of temporal activations similarity of FBN-filter pairs identified by PCC. The colors red and blue denote the best and the second-best results, respectively. Abbreviations: LT: linear transformation; MSA: multi-head self-attention.

Methods	MAE↓	MSE↓	RMSE↓	DTW↓	PCC↑
a) LT	0.9781	1.5309	1.2360	18.0688	0.2346
b) LT+LSTM	0.9392	1.4411	1.1986	18.4683	0.2794
c) LT+MSA	0.9658	1.5136	1.2289	18.7328	0.2432

4.6 Discussions

Interpretability. The proposed Sync-ACT framework matches and pairs the neurons in ANNs and BNNs, based on which the cross-annotation is performed to annotate the neurons in one domain with the semantic description in the other. The Sync-ACT framework opens a new paradigm for the interpretability studies of ANN by using the prior knowledge in neuroscience to interpret ANNs. In parallel, we can understand the dynamic function of FBNs with visual and language descriptions from the paired ANNs' neurons in a direct way, providing a novel way for unveiling the complex brain function.

Neural architecture search (NAS). One important finding of this study is that the performance of CNNs on image classification tasks is closely related to its visual representation similarity with the human brain. In the literature (Elsken et al., 2019; H. Liu et al., 2018; Zoph & Le, 2016), the typical evaluation criteria for NAS is the performance of the searched neural network on a specific task. Our Sync-ACT framework provides new inspirations: the ANN's representation similarity to the human brain could be a reliable and meaningful criterion for NAS and thus guide the NAS approaches to improve interpretability and performance. Our Sync-ACT framework contributes to the emerging field of brain-inspired AI, i.e., using domain knowledge of brain science to inspire and guide the design of AI models.

Limitations. Our approach has several potential limitations. a) We use the maximum value in the feature map to represent the activation degree of convolutional filters. Currently, how to characterize the activation degree of filters is still an open question. b) The semantics descriptions from meta-analysis (Yarkoni et al., 2011) and (Bau et al., 2017) for neurons in FBNs and CNNs are coarse-grained and ad-hoc (e.g., Figure 4.4, unit #440, Bank Vault) due to their intrinsic

limitations. More fine-grained descriptions can be explored and adopted in the future. c) We mainly focus on CNNs for image classification. CNNs and ViTs for other tasks should be investigated in the future.

Chapter 5

Exploring Human Visual Attention

5.1 Eye-gaze Guided Vision Transformer

5.1.1 Introduction

Deep neural networks have been widely used and achieved remarkable successes in many fields including natural language processing, computer vision, and medical image analysis (LeCun et al., 2015), etc. Recent studies suggest that deep neural networks may be prone to learn the shortcut knowledge (Geirhos et al., 2020) such as the spurious correlations between the background and objects in the image (e.g., cows usually stand on the grass land) rather than the intended relevant features. Recent studies (X. Luo et al., 2021; Xiao et al., 2021) revealed that background is a harmful shortcut which drastically impacts the deep learning model's performance in a negative way. The harmful shortcut knowledge, on the one hand, may not be able to generalize to new domains and tasks, and thus degenerates the performance in some scenarios such as few-shot learning (FSL). On the other hand, it jeopardizes the interpretability of the model and prevents humans from validating its underlying reasoning which is crucial in many applications, e.g., disease diagnosis with medical images.

Medical image analysis is a representative scenario where the harmful shortcut learning should be rectified because the generalizability and interpretability are highly desired and required, considering the scarcity of the clinical data (e.g., MR images with pathology) and the importance of reliability and transparency in clinical applications. The literature has already reported the existence of shortcuts in medical image analysis (L. Luo et al., 2021; Zech et al., 2018). For example, in (Zech et al., 2018), convolutional neural networks (CNNs) were em-

ployed to detect pneumonia and performed well with extremely high accuracy on the chest X-rays from a group of hospitals. However, it failed to generalize to the X-rays from other external hospitals with much lower performance: CNNs unexpectedly learned to detect a hospital-specific metal token at the corner of scans and utilized it for disease prediction indirectly (Geirhos et al., 2020; Zech et al., 2018). To motivate the work in this section, in Figure 5.1, we also visualize four samples of harmful shortcuts learned by vision transformer (ViT) (Dosovitskiy et al., 2020) model which are the medical images' background.

To solve this problem, one possible way is to enforce the model to concentrate on task-related objects or features rather than the harmful shortcuts by using prior knowledge (X. Luo et al., 2021). For example, the bounding box and voxel/pixel-level segmentation mask of medical image directly indicate the location of the lesion on which the model should focus. However, accurate manual annotation/segmentation requires experienced radiologists and the devotion of their additional time, which is costly and not easily accessible. On the other hand, radiologists read dozens of patients' images and write diagnosis reports on average in their routine work. This means that there is a huge amount of valuable data that is not collected and fully exploited. For example, the eye-gaze information can indicate the regions-of-interest (ROIs) of radiologists, which might be highly related to potential pathology and is easily accessible by installing the eye-tracker. Such domain knowledge embedded in ROIs of an expert is naturally interpretable and generalizable because it reaches the professional standard and has been validated and widely used in longstanding clinical practice. Some recent deep learning studies have already integrated eye-gaze of radiologists to improve the performance of medical image applications (Karargyris et al., 2021; S. Wang et al., 2022), suggesting the potential usage and convenience of using eye-gaze rather than precious annotation in avoiding harmful shortcut learning.

Inspired by this, we propose an intuitive and effective method to infuse the domain knowledge of an expert, i.e., eye-gaze, with the training of deep learning models for rectifying the harmful shortcut learning in medical image analysis. Specifically, based on vision transformer (ViT) (Dosovitskiy et al., 2020), we introduce a novel eye-gaze-guided vision transformer (EG-ViT) model which applies an eye-gaze mask to input image patches to screen out those irrelevant to radiologist's visual attention and guide the model to focus on patches that are highly related to potential pathology during the model training/fine-tuning. Meanwhile, a residual connection between the unmasked input and the last ViT encoder layer is intentionally added to maintain the interactions and relationships of all patches. In the testing stage, the mask operation and residual

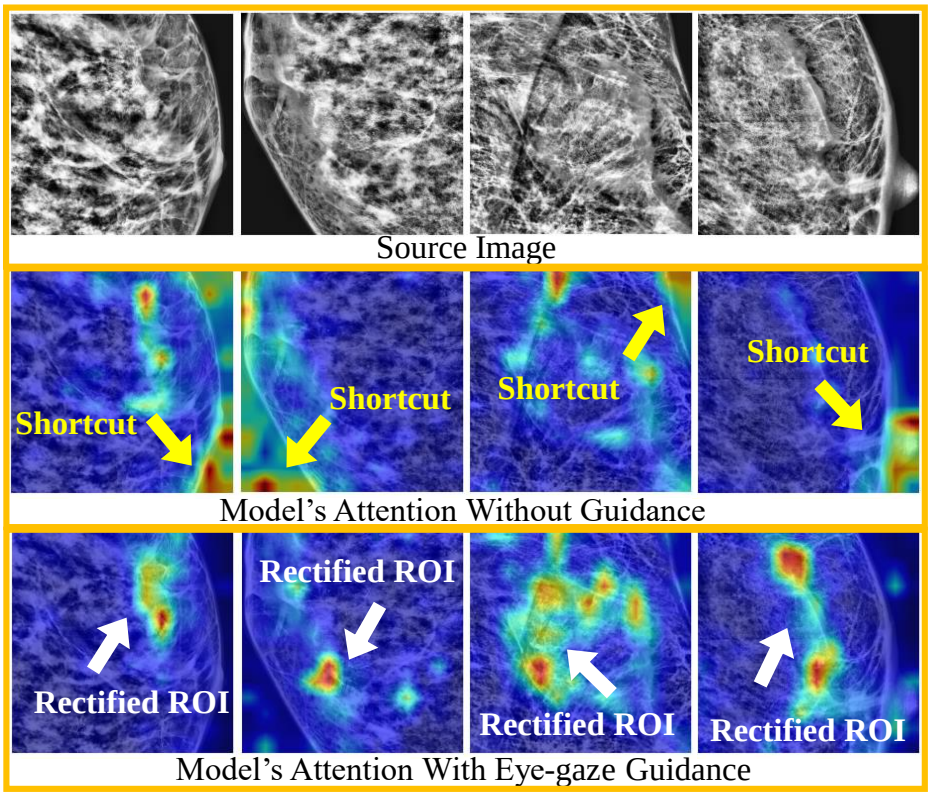

Figure 5.1: Illustration of the shortcuts learned by ViT model. The first row is the enhanced source image from the public INbreast dataset. The second row corresponds to the model's attention derived by Grad-CAM without the guidance of eye-gaze. It is observed that the model focuses on background shortcuts (yellow arrows) rather than the breast tissues. The third row is the Grad-CAM from our eye-gaze-guided vision transformer (EG-ViT) model. The regions-of-interest (ROIs) of EG-ViT are denoted by white arrows.

connection are removed to maintain the original structure of ViT model. In this way, the EG-ViT model infuses the expert's domain knowledge to enforce the model to avoid learning harmful shortcut while takes the power of data-intensive ViT model in a more effective manner. We evaluate the proposed EG-ViT on disease diagnosis with two publicly available datasets, namely, INbreast (I. Moreira et al., 2012) and SIIM-ACR ("SIIM-ACR Pneumothorax Segmentation", 2020). Our extensive experiments demonstrate that the proposed EG-ViT model effectively avoids the harmful shortcut learning (Figure 5.1). The

diagnostic accuracy is also improved, compared with CNN (about 4%) and ViT (about 2%) baselines with limited data.

In general, the main contributions of our work are as follows:

1. We propose a novel EG-ViT model to infuse the human expert's prior knowledge to guide the model focusing on the region with potential pathology, avoiding the harmful shortcut learning and improving models' interpretability with much higher performance.

2. The proposed EG-ViT model only includes an additional mask operation and a residual connection compared with vanilla ViT, thus allowing the inheritance of the parameters from a pre-trained ViT model without any additional cost.

3. We introduce a novel evaluation metric for quantifying the degree of shortcuts in models and measuring the improvement in rectifying the shortcut learning, which can also be generalized to other scenarios and tasks.

5.1.2 Related Works

Shortcut Learning

Deep neural networks often solve the task-specific problem, e.g., image classification, by learning the shortcuts such as the correlations between cows and grass instead of the intended solution, e.g., the features from cows (Geirhos et al., 2020). Recently, the shortcut in deep learning models gains increasing attention across the deep learning field from computer vision (CV) (Dancette et al., 2021; Minderer et al., 2020; Xiao et al., 2021), natural language processing (NLP) (McCoy et al., 2019; Niven & Kao, 2019) to reinforcement learning (Amodei et al., 2016). Various methods have been devised to mitigate the negative effects of shortcuts (Du et al., 2021; X. Luo et al., 2021; X. Shen & Lam, 2021). For example, in (X. Luo et al., 2021), a framework named COSOC was proposed to tackle this shortcut problem by extracting the foreground objects in images to get rid of background-related shortcuts based on a contrastive learning approach. (Du et al., 2021) proposed a measurement for quantifying the shortcut degree, with which a shortcut mitigation framework was introduced for natural language understanding (NLU). (X. Shen & Lam, 2021) forced the network to learn the necessary features for all the words in the input to alleviate the shortcut learning problem in supervised Paraphrase Identification (PI).

In medical image analysis, shortcut learning not only has a negative impact on the models' generalizability, e.g, the same type of medical images the model

performance often varies greatly between images from different vendors (C. Chen et al., 2020), but also degenerates the interpretability and the reliability of the applications. Recently, more studies begin to scrutinise the shortcut learning in different scenarios of medical imaging applications. For example, (L. Luo et al., 2021) demonstrated that models for the localization task are less prone to shortcut learning than models for the classification task, because the data annotation for the localization task is more fine-grained. (Mahapatra et al., 2022) mitigated the shortcut learning in medical image classification and segmentation by introducing an interpretability-guided inductive bias loss function which is composed of the class-distinctiveness and spatial coherent loss between the attention maps. (Nauta et al., 2022) made the model more focusing on the lesion area by replacing some image patches of the original image.

However, the aforementioned methods still need sufficient samples (more than 10,000) and even with fine-grained annotations such as pixel-level segmentation mask. In this section, we rectify possible shortcut learning on small-scale medical image datasets (around 1,000) by infusing the accessible coarse eye-gaze data from radiologists.

Eye Tracking in Radiology

Visual diagnosis plays a central role in radiology, and eye-tracking procedures have proven to be a valuable tool in the study of visual diagnostic processes in radiology for decades (Krupinski, 2010). A group of early studies have found that experts can quickly locate potential lesions with a global search and use a larger functional field of view and more conceptual knowledge than novices to find abnormalities (Drew et al., 2013; Kundel et al., 2007; Swensson, 1980). (Kok & Jarodzka, 2017) showed that experts searched normal CXR more systematically than novices. With the rise of deep learning in computer aided diagnosis (CAD), the integration of radiologists' eye movement into deep learning models becomes more popular. For example, (Mall et al., 2018) modeled the visual search behavior of radiologists and their interpretation of mammography with CNNs. Furthermore, they (Mall et al., 2019) investigated the relationship between human visual attention and CNNs in finding lesions in mammography. Recently, (Karargyris et al., 2021) developed a dataset with CXR, eye-gaze, and text diagnosis reports. They proposed a multi-task framework which predicted eye-gaze and diagnosed diseases at the same time. (S. Wang et al., 2022) used radiologists' visual attention to supervise the CNN's attention via an attention consistency module, thus improving the diagnosis performance in osteoarthritis assessment of knee X-ray images.

Despite the successes of combining the radiologists' eye-gaze information with CNNs, how to integrate eye-gaze information with powerful ViT model to further boost its performance in medical imaging applications still needs investigations.

Vision Transformer

Since ViT (Dosovitskiy et al., 2020) was introduced, transformer structure has been receiving increasing attention in the computer vision community (K. Han et al., 2022). Several effective strategies have been proposed to improve model performance and efficiency in image classification, such as knowledge distillation in DeiT (Touvron, Cord, Douze, et al., 2021), depth-wise convolution in CeiT (Yuan et al., 2021), shifted windows in Swin Transformer (Z. Liu et al., 2021b), and tree-like structure in NesT (Z. Zhang et al., 2021). However, the data-intensive characteristic of ViT makes it challenging to adapt to the target domain quickly with limited amount of data. To address this problem, several methods with distillation approach (Touvron, Cord, Douze, et al., 2021), smoothing the loss landscapes at convergence (X. Chen et al., 2021), incorporating CNNs like CCT (Hassani et al., 2021) and locality information (Y. Li et al., 2021) have been proposed to reduce the demand for extensive training data to a certain extent. Nonetheless, fast adaption to the target domain still requires more innovative and effective methods to further reduce the demand for training data.

In medical image analysis, for models trained on large datasets, ViT-style models have been explored in CAD tasks on the chest X-ray (CXR) images (Shamshad et al., 2022). For example, (Krishnan & Krishnan, 2021) and (Park et al., 2021) utilize ViT-based models to achieve higher COVID-19 classification accuracy through CXR images. COVID-Transformer (Shome et al., 2021) and xViTCOS (Mondal et al., 2021) have been proposed to improve classification accuracy and focus on diagnosis-related regions. (Bhattacharya et al., 2022) combined the radiologists' eye gaze information using a transformer model based on the teacher-student approach to effectively improve the diagnostic performance of chest X-ray disease. However, the aforementioned methods still requires large amount of training data. In the scenarios with limited amount of data, it is more likely to trigger shortcut learning by exploiting the redundant information such as backgrounds, which leads to serious consequences. In addition, the small scale of medical image datasets also limits the performance of ViT. In this section, we guide the transformer model to focus on the important regions directly by introducing the radiologists' eye-gaze as auxiliary information and demon-

strate the vanilla ViT model can handle the diagnosis task even on small scale datasets.

5.1.3 Method

The main idea of our EG-ViT model is to utilize the eye-gaze data to mask the patches out of radiologists ROIs for ViT model. We first illustrate the collection of eye-gaze data and generation of eye-gaze heatmap in Section 5.1.3. Then, we introduce the generation of eye-gaze mask used in EG-ViT model in Section 5.1.3. Finally, we elaborate the architecture of our EG-ViT model.

Eye-Gaze Data Collection and Heatmap

Compared with fine-grained annotations such as pixel-level segmentation masks, coarse eye-gaze data is much more accessible during the radiologists routine works. However, we are still lacking a eye-gaze collection systems specially designed for radiologists for the diagnosis with the minimal interruption. To this end, we design a new collection system specifically for capturing eye-gaze data from radiologists. Specifically, we use Tobii pro Nano as the hardware platform to collect eye-gaze data. In addition, we develop a software system that support radiologists to manipulate images freely while collecting the eye-gaze and mouse data simultaneously. To the best of our knowledge, this is the first eye-gaze collection software that supports free image adjustment. More details about the collection system can be found in the project page.

After the collection of the raw eye-gaze data, we apply two pre-processing steps. The first step is to remove the noises. As a result of blinking or turning the head, the radiologist's eye-gaze points can fall into the areas outside of the breast tissue (such as the black background in a mammogram image), which will be filtered out. The second step is to extract the effective fixation points. Eye movements mainly consist of fixation and saccade, where saccade is a rapid movement process between two gaze points and thus does not reflect the region of attention. So we filter the saccade part and only keep the fixations of radiologists ("Processed Gaze Points" in Figure 5.3) by using the well-established I2MC method (Nyström & Holmqvist, 2010). Finally, the eye-gaze heatmap ("Gaze Heatmap" in Figure 5.3) of whole image is generated by smoothing the binary eye-gaze points map through a two-dimensional Gaussian kernel with the radius of 150 pixels and the sigma 25.

Generation of Eye-Gaze Guided Mask

Image Cropping Strategy

Medical images often have a large resolution, for example, the INbreast (I. Moreira et al., 2012) dataset used in this study consists of images with a size of 3000×4000 pixels. Direct utilization of images with the original resolution requires huge amount of computational resources which is infeasible in practice especially for model training. However, compressing the whole image with a lower resolution can cause the loss of details and even miss smaller lesions. We introduce a strategy to crop the original image into image patches and use the cropped image patches as samples for training and testing. Specifically, in the training stage, we adopt random cropping to generate image patches for each image as well as the corresponding heatmap. If a cropped image contains the lesion area, we assign a corresponding label for it. Then, we balance the number of cropped images with different labels. In the testing stage, we apply a large view sliding window to crop the whole image as patches with overlap for model evaluation.

Eye-Gaze Guided Mask With the cropped eye-gaze heatmap as in previous section, we can generate the eye-gaze guided mask for the corresponding cropped image during the model training. We introduce two types of eye-gaze guided masks: Focused Eye-Gaze Mask and Separated Eye-Gaze Mask. As shown in Figure 5.2, the focused eye-gaze mask is defined as a rectangular binary mask centered at the pixel with the largest value in heatmap. The separated eye-gaze mask is obtained by selecting a certain percentage of pixels according to the values in heatmap in a descending order and setting the selected positions of the mask to 1. The focused mask only keeps the greatest interest of radiologists while the separated mask tends to include all sub-regions of radiologists' interest. If the cropped image does not have eye-gaze data, mask is not used in the training process. For comparison, we also include a random mask and generate a self-supervised mask based on model's own attention by using Grad-CAM (Selvaraju et al., 2017) method.

Eye-Gaze-Guided Vision Transformer

Compared with natural images, medical images usually have a higher resolution while pathology such as lesions locates in a small region with a noisy background, which makes model prone to learning background shortcuts rather than the intended meaningful features. To avoid learning harmful shortcuts, an intuitive idea is to guide the model to focus on the regions that are potentially related to pathology based on some prior knowledge. As discussed in Section 5.1.2, the visual attention from a radiologist during the diagnosis can serve as such prior knowledge as the guidance for the model training. We implement this idea by introducing an eye-gaze guided mask on input image patches of ViT model to

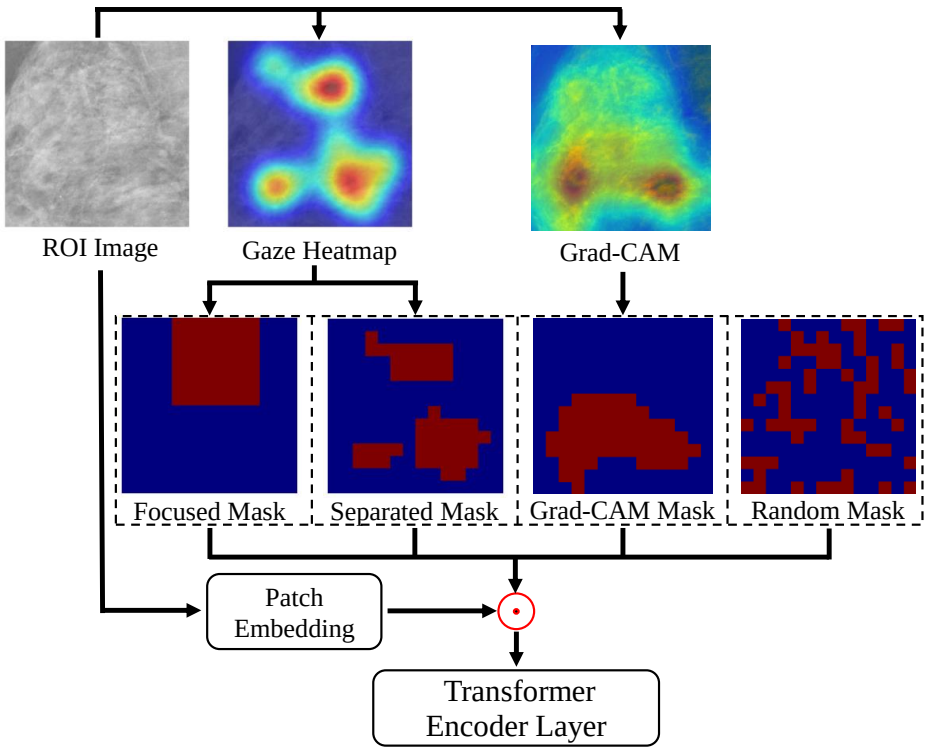

Figure 5.2: Illustration of different masks: the focused/separated eye-gaze masks are generated by using eye-gaze heatmap; the grad-cam mask is generated by binarizing the Grad-CAM of the model; random mask is also included for comparison. These masks are used to mask the corresponding patch embedding, which is the input of the encoder layer in EG-ViT model.

screen out the background patches. The overall architecture of EG-ViT model is shown in Figure 5.3. Specifically, we first pre-process the collected radiologists' eye-gaze data. Then, we generate the eye-gaze heatmap and randomly crop the original image and corresponding heatmap into a smaller size for model training. After the patch embedding, we mask out the regions out of radiologists' ROI based on the mask generated by heatmap to make the network only focus on specific regions (i.e, ROI of radiologists). Meanwhile, to maintain the information and interaction of all patches, a residual connection is introduced in the last layer of the EG-ViT model.

Eye-Gaze Guided Mask Operation With the eye-gaze guided mask, we can perform a mask operation on the input patches of the ViT model. Specifically, the input cropped image can be divided into N patches where $N = (H \times$

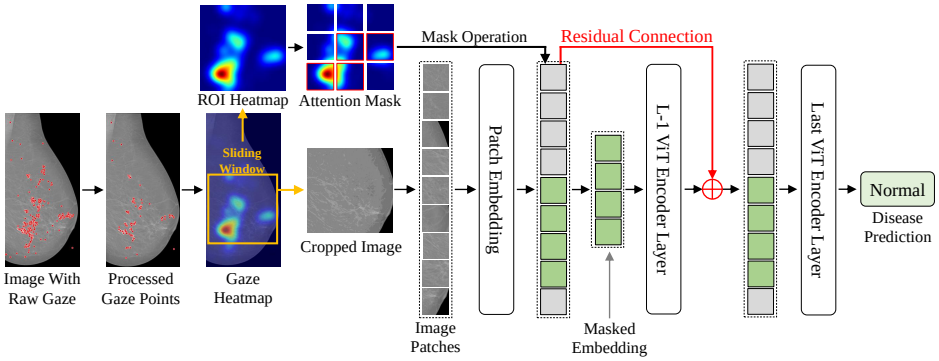

Figure 5.3: The architecture of the proposed EG-ViT model. The eye-gaze points are collected and pre-processed to generate the eye-gaze heatmap. Then, the original image and the corresponding heatmap are randomly cropped with a smaller size. The cropped image is divided into several image patches for patch embedding. The eye-gaze mask is then applied to screen out the patches that may be unrelated to pathology and radiologists' interest. The masked image patches (green rectangle) are treated as input to the transformer encoder layer. Note that to maintain the information of all patches including those been masked, we add an additional residual connection (highlighted by the red arrow) from the input and the last encoder layer.

$W)/P^2$ is the patch number, H and W are the height and weight of images, P is the patch size. The ViT model maps the images patches x_p^i ($i = 1, 2, \cdots$,N) to D dimension patch embedding $z_0 \in \mathbb{R}^{(N+1)\times D}$ (contacted with a class token) with a trainable linear projection $E \in \mathbb{R}^{P^2C\times D}$ where C is the number of channels of the images:

$$z_0 = [x_{class}; x_p^1 E; x_p^2 E; \cdots ; x_p^N E] + E_{pos} \tag{5.1}$$

where $z_0^0 = x_{class} \in \mathbb{R}^N$ is the $class$ token for classification and $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ is the learnable position embedding. Then, the embedding of the input image patch z_0 is masked as:

$$\tilde{z}_0 = [z_0^0; z_0^{1:N} \odot mask] \tag{5.2}$$

where $mask \in \mathbb{R}^N$ is the binary eye-gaze mask detailed in Section 5.1.3 and $z_0^{1:N} = [x_p^1 E; x_p^2 E; \cdots ; x_p^N E]$ is the embedding of image patches. The masked patch embedding $\tilde{z}_0$ is then input into the first layer of ViT encoder, forcing the model only exploiting the patches with potential pathology.

Residual Connections Preserving Global Features For the EG-ViT model, the forward propagation of each transformer encoder layer can be written as:

$$\tilde{z}'_l = MSA(LN(\tilde{z}_{l-1})) + \tilde{z}_{l-1} \tag{5.3}$$

$$\tilde{z}_l = MLP(LN(\tilde{z}'_l)) + \tilde{z}'_l \tag{5.4}$$

where $\tilde{z}'_l$ is the l-th layer's embedding of masked patches. MSA, MLP, and LN are the multiheaded self-attention, multilayer perceptron, and layer norm in each block, respectively.

However, masking some patches in the first layer results in a risk of missing useful background information and positional relationships among all patches. So we add the initial patch embedding to the last layer's embedding through a residual connection to maintain the information from all patches and the correlations among them. Therefore, the input of the last transformer encoder layer $\hat{z}^i_{l-1}$ (i=0,1,2 $\cdots$ N) can be written as:

$$\hat{z}^i_{l-1} = \begin{cases} \tilde{z}^0_{l-1}, & if\ i = 0 \\ z^i_0, & if\ mask_i = 0 \\ \tilde{z}^i_{l-1} + z^i_0, & otherwise \end{cases} \tag{5.5}$$

where $\tilde{z}_{l-1}$ and $\hat{z}_{l-1}$ are the embeddings before and after additive operations.

Pre-training and Fine-tuning Style As mentioned in Section 5.1.2, although the ViT model has a strong feature representation ability, it relies heavily on large amount of training data. However, medical images are often limited and the ViT models trained from scratch on small-scale medical images datasets often perform poorly. So instead of training from scratch, we initialize the EG-ViT parameters with the weights pre-trained on ImageNet-1K (Russakovsky et al., 2015) and fine-tune on the small-scale medical image datasets. It is noted that EG-ViT does not introduce any addition parameters to the vanilla ViT model, thus allowing our model to directly inherit the parameters of the pre-trained models. In this way, the time required for model training is greatly reduced while the performance can be also guaranteed. Notably, the computational overhead of the transformer model is related to the number of patches, so by adding a mask, the resources for training EG-ViT model are further reduced.

5.1.4 Experiments

In this section, we conduct detailed experiments to demonstrate the effectiveness and advantages of EG-ViT model in rectifying the shortcut learning and improving the accuracy in diseases diagnosis. We firstly introduce the datasets

used in this study for training and evaluation (Section 5.1.4) and then propose a new metric for quantifying the degree of shortcut learning in disease diagnosis (Section 5.1.4). We demonstrate the performance of EG-ViT model in shortcut rectification in Section 5.1.4. In Section 5.1.4, we compare the EG-ViT model with several baselines with and without eye-gaze data in disease diagnosis. Finally, we evaluate the effects of different masks for EG-ViT model in Section 5.1.4.

Datasets

To evaluate the proposed EG-ViT model, we adopt two different public clinical datasets with approval: INbreast (I. Moreira et al., 2012) and SIIM-ACR (Saab et al., 2021; "SIIM-ACR Pneumothorax Segmentation", 2020). The INbreast dataset (I. Moreira et al., 2012) includes 410 full-field digital mammography images which were collected during low-dose X-ray irradiation of the breast. We invited a radiologist with 10 years of experience to diagnose the images in this dataset and collected the complete eye movement data using the aforementioned collection system. According to BI-RADS (Liberman & Menell, 2002) assessment of masses, these images can be classified into three groups: normal (302 cases), benign (37 cases), and malignant (71 cases), respectively. As for SIIM-ACR dataset ("SIIM-ACR Pneumothorax Segmentation", 2020), (Saab et al., 2021) randomly selected 1,170 images with 268 cases of Pneumothorax and collected gaze data from three experienced radiologists. More details refer to (Saab et al., 2021).

For INbreast dataset (I. Moreira et al., 2012), we apply the following three experimental setups. 1). In the training stage, we adopt a random cropping strategy instead of sliding window in order to ensure the balance and diversity of different samples. Specifically, we first randomly split the patients into 80% and 20% as training and testing datasets. To balance the training dataset, we perform several random cropping as well as the contrast-related augmentation for each image. Finally, our training set consists of 482 normal samples, 512 benign mass samples, and 472 malignant mass samples. 2). In the testing phase, the images of the remaining 20% patients were cropped by using a sliding window as described in Section 5.1.3. The windows size is set as 1024 and the stride is set as the half of windows size, i.e., 512. For the SIIM-ACR dataset ("SIIM-ACR Pneumothorax Segmentation", 2020), the size of the original images are all 1024×1024, so we directly use the original image as the input for the model.

Evaluation Metrics

For the evaluation of model's performance in disease diagnosis, we report the accuracy (ACC), area under curve (AUC), and F1-score (F1) on testing dataset. For evaluating the performance in rectifying the shortcut learning, we adopt the Structure Similarity Index Measure (Z. Wang et al., 2004) (SSIM) to assess the similarity between model's attention and the radiologists' attention heatmap during the testing stage. In addition, we propose a new metric for assessing the degree of shortcut learning. As shown in Figure 5.4, we first generate model's attention heatmap by using Grad-CAM (Selvaraju et al., 2017). Then we select the region with highest values in heatmap and make an intersection with the mask region of the lesion A_M to get A_U. Next, we set the single sample evaluation score $O^{mm} = A_U/A_M$. If O^{mm} is greater than 0.9, it is counted as 1, if O^{mm} is between 0.3 and 0.9, it is counted as 0.5, and O^{mm} below 0.3 is not counted. Finally, the average counted score is used as the measure score. Note that the O^{mm} here is the score generated by the model's own attention during the testing phase. For comparison, we also compute the radiologist's eye gaze heatmap as described above and finally obtain a score of 0.53 in the INbreast (I. Moreira et al., 2012) dataset and 0.56 in the SIIM-ACR ("SIIM-ACR Pneumothorax Segmentation", 2020) dataset. Therefore, $O^{mm}_{0.53}$ and $O^{mm}_{0.56}$ represent a standard for INbreast and SIIM-ACR, respectively.

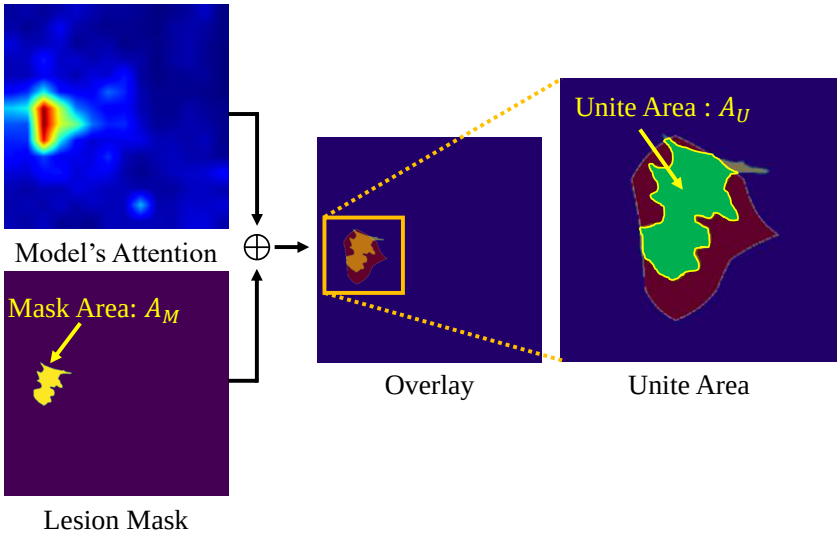

Figure 5.4: Generation of unite area between model's attention heatmap and the lesion area.

Evaluation of Shortcut Rectification

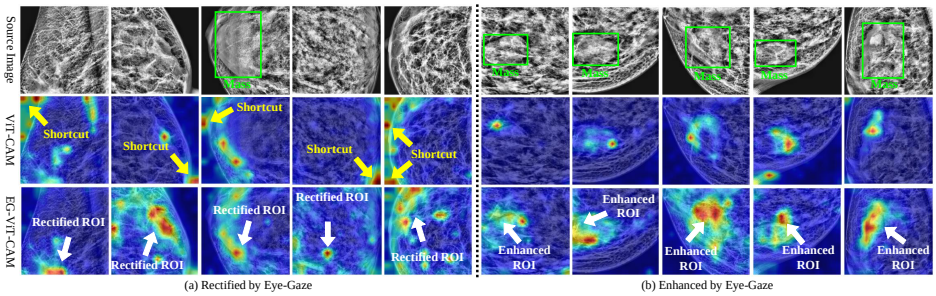

Figure 5.5: (a) Harmful shortcut learning rectified by eye gaze guidance. (b) Useful feature learning enhanced by eye gaze guidance. In each panel of (a) and (b), the first row is the enhanced source image, the second row is the attention map of ViT obtained using Grad-CAM, and the third row is the attention map of EG-ViT. Each column corresponds to the same example.

Table 5.1: Comparison results with other baselines with eye-gaze (the bottom half) and without eye-gaze (the top half) in terms of Accuracy, F1, and AUC scores. The number of parameters in each model is also reported. And the SSIM and O^{mm} are our metrics for shortcut learning evaluation. Red and blue denote the best and the second-best results, respectively.

Method	Params	INbreast					SIIM-ACR				
		Acc. ↑	AUC ↑	F1 ↑	SSIM ↑	$O^{mm}_{0.53}$ ↑	Acc. ↑	AUC ↑	F1 ↑	SSIM ↑	$O^{mm}_{0.56}$ ↑
ResNet-18 (K. He et al., 2016a)	11M	87.84	83.73	86.28	0.283	0.50	82.40	68.33	81.33	0.161	0.28
ResNet-50 (K. He et al., 2016a)	24M	89.19	89.62	87.59	0.276	0.40	84.80	71.97	83.88	0.153	0.14
ResNet-101 (K. He et al., 2016a)	43M	90.54	89.85	87.73	0.302	0.17	84.80	70.55	83.54	0.152	0.28
EfficientNet b0 (Tan & Le, 2019)	5.3M	87.84	79.59	86.65	0.258	0.43	80.40	70.65	79.31	0.147	0.22
EfficientNet b4 (Tan & Le, 2019)	19M	90.54	89.08	87.71	0.390	0.13	81.00	71.14	79.57	0.196	0.13
EfficientNet b7 (Tan & Le, 2019)	66M	91.45	85.79	88.73	0.341	0.27	83.20	72.12	81.81	0.215	0.19
SwinT v1 (Z. Liu et al., 2021a)	49M	91.80	88.10	90.27	0.227	0.23	85.20	73.64	84.52	0.149	0.06
ViT-S (Dosovitskiy et al., 2020)	22M	91.89	86.94	90.16	0.395	0.43	84.00	70.76	83.03	0.205	0.17
ResNet-18+Gaze (S. Wang et al., 2022)	11M	87.84	86.56	87.47	0.208	0.50	84.80	71.26	83.71	0.266	0.27
ResNet-50+Gaze (S. Wang et al., 2022)	24M	89.19	82.12	86.79	0.209	0.40	83.20	70.25	82.35	0.220	0.14
ResNet-101+Gaze (S. Wang et al., 2022)	43M	91.88	89.63	89.75	0.212	0.31	84.80	72.68	84.03	0.254	0.19
U-Net+Gaze (Karargyris et al., 2021)	6M	86.25	83.01	85.33	0.331	0.19	81.88	72.27	81.10	0.205	0.07
EG-ViT (ours)	22M	93.24	93.32	92.92	0.402	0.53	85.60	75.30	85.14	0.280	0.29

In this subsection, we evaluate the performance of EG-ViT model in rectifying the shortcut learning both qualitatively and quantitatively. For a better visualization, we employ the Grad-CAM (Selvaraju et al., 2017) to generate the model's attention map. Grad-CAM uses gradient to calculate the attention map of the model, which does not require any changes to the model structure and thus can be easily deployed to the ViT model.

Figure 5.5 shows two ways of rectification by our EG-ViT model for qualitative comparison. In Figure 5.5(a), the enhanced source images are shown in the

first row. The Grad-CAM maps from fine-tuned vanilla ViT model and from EG-ViT model are demonstrated in the second and third rows, respectively. It is observed that without the expert's domain knowledge, ViT model is likely to make classification decisions from the areas that are related to unrelated regions, such as background edge, rather than valid human tissues. But with the help of visual attention from radiologists, the EG-ViT model focuses on disease-related areas, such as the inner mammary region. Figure 5.5(b) demonstrates the cases that EG-ViT model enhances its attention to be congruent with the radiologist's attention. The regions with mass are more emphasized by EG-ViT model compared with the vanilla ViT model, which makes the decision of the EG-ViT model more interpretable.

In Table 5.1, we compare the results of different models with and without eye-gaze guidance. The comparison results in terms of SSIM and O^{mm} of ViT-S and EG-ViT are also consistent with our observation that the attention heatmap generated by the EG-ViT model is more similar to the radiologist's attention and it focuses more on the lesion area. We also manually count the number of samples with differences in EG-ViT model's attention map compared with the ViT model's attention. We found that with 64% (264 cases) of all 410 cases have the significant differences compared with ViT in INbreast dataset. Among them, 151 cases with shortcut learning are rectified and 113 cases are enhanced. Overall, these experiments demonstrate that the proposed EG-ViT model can effective rectify the shortcut learning in medical imaging classification task.

Comparison with Baselines in Disease Diagnosis

Here, we adopt ResNet (K. He et al., 2016a), Efficientnet (Tan & Le, 2019), the vanilla Vision Transformer (Dosovitskiy et al., 2020) and Swin Transformer (Z. Liu et al., 2021a) as baselines for comparison. The baseline models were pretrained on ImageNet dataset (J. Deng et al., 2009) and fine-tuned on the two clinical datasets. The experimental results of each model are reported in the top half of Table 5.1. It is observed that the proposed EG-ViT model outperforms all compared baselines in terms of all evaluation metrics, with a relatively small number of parameters. We also observed that the ViT-S is inferior to those of CNN based models for some metrics. One possible reason is that ViT has a larger model capacity than CNN based models while lacks some inductive biases, making it difficult to be pretrained and fine-tuned on a small dataset like the INbreast and SIIM-ACR datasets. However, with the guidance of eye-gaze from the radiologist, the performance of ViT-based EG-ViT model is significantly improved, which suggests that eye-gaze serves as a strong prior guidance

Table 5.2: Comparison of performance using different masks and different degree of mask operation. Red and blue denote the best and the second-best results, respectively.

Method	INbreast					SIIM-ACR				
	Acc. ↑	AUC ↑	F1 ↑	SSIM ↑	$O^{mm}_{0.53}$ ↑	Acc. ↑	AUC ↑	F1 ↑	SSIM ↑	$O^{mm}_{0.56}$ ↑
ViT-S (Dosovitskiy et al., 2020) (Baseline)	91.89	86.94	90.16	0.395	0.43	84.00	70.76	83.03	0.205	0.17
Grad-CAM Mask	90.54	88.06	88.79	0.387	0.47	84.00	71.79	83.58	0.303	0.05
Random Mask	89.92	84.53	88.94	0.365	0.33	84.20	71.51	84.06	0.232	0.03
Focused Gaze Mask	91.18	83.33	89.42	0.392	0.50	83.60	68.70	82.10	0.241	0.13
Separated Gaze Mask	93.24	93.32	92.92	0.402	0.53	85.60	75.30	85.14	0.280	0.29
80% Eye Gaze Mask	91.91	87.67	90.19	0.388	0.47	85.20	72.13	84.06	0.252	0.24
75% Eye Gaze Mask	93.24	93.32	92.92	0.402	0.53	85.60	75.30	85.14	0.280	0.29
70% Eye Gaze Mask	91.40	89.09	90.59	0.372	0.57	84.40	73.62	82.37	0.239	0.20
65% Eye Gaze Mask	92.61	91.64	90.95	0.393	0.40	85.60	73.31	83.42	0.263	0.13
60% Eye Gaze Mask	90.74	88.91	88.72	0.384	0.53	83.60	71.83	81.98	0.238	0.27

to assist the model training and reduces the potential overfitting problem induced by insufficient samples.

In the bottom half of Table 5.1, we compare our EG-ViT model with two recent studies that also utilized eye-gaze for medical image classification tasks (Karargyris et al., 2021; S. Wang et al., 2022) . In (S. Wang et al., 2022), ResNet (K. He et al., 2016a) was used as the classification backbone, with which visual attention from eye-gaze data was incorporated to enhance osteoarthritis assessment on knee X-ray images. (Karargyris et al., 2021) used a U-Net structure to classify three chest diseases and output the attention map to compare with human attention. We train these two methods for the disease diagnosis on INbreast (I. Moreira et al., 2012) and SIIM-ACR ("SIIM-ACR Pneumothorax Segmentation", 2020), respectively. As shown in Table 5.1, our proposed EG-ViT model outperforms the compared methods on both INbreast (I. Moreira et al., 2012) and SIIM-ACR datasets ("SIIM-ACR Pneumothorax Segmentation", 2020) in terms of all metrics. It is also observed that using eye-gaze guidance can improve the performance of ViT model on small datasets, which is even beyond the CNNs with inductive biases.

Ablation Study

In this subsection, we first discuss the effect of different types of masks, and then we investigate the degree of mask operation. The comparison of performance using four types of masks (Figure 5.2) is shown in the top part of Table 5.2. The first row of the table shows the results of the vanilla ViT-S (Dosovitskiy et al., 2020) model. The second to fifth rows correspond to four types of masks. We

found that separated mask is better than the other ones, especially for separated eye-gaze mask. This might be attributed to the advantage of separated regions in guiding the model to learn the relationship between features that are far apart in larger images. We also observed that the focused gaze position mask is worse than other masks. This may be related to the radiologists' individualized reading habits. If the radiologists' gaze points are spread out and the saccade path is long, the use of a focused position mask will ignore the features at other locations within radiologists' ROI.

For the separated eye-gaze mask, we also explore different degrees of mask operation. The degree of mask operation is the percentage of the masked region with respect to original image. Specifically, the pixels of the original image are firstly sorted according to their values in the eye-gaze heatmap in a descending order. Then the pixels with the top 20%, 25% or 30% value are selected as the areas that the radiologist focuses on, which means 80%, 75%, or 70% region of the original image are masked. In the bottom part of the Table 5.2, we demonstrate the model's performance with different degree of mask operation. It is observed that using a 75% degree of mask has a better result, except for the $O^{mm}_{0.53}$ in INbreast dataset. We also found that the metrics of the model decrease wit a smaller degree. The performance of the model with a degree of 60% is even inferior than the Vanilla ViT-S. A potential explanation for such observation is that the lesion area in two datasets is relatively small, so masking out more irrelevant or redundant areas is an effective way to improve the performance of the model. This is also in parallel with the findings in Masked Autoencoder (MAE) (K. He et al., 2022) that masking a high proportion of the original image, e.g., 75% yield a meaningful self-supervised image representation. Meanwhile, for eye-gaze data, except for the most concerned regions, the rest of the regions may contain redundant information that misleads the model training, resulting in a lower performance.

Implementation Details

We fine-tune the model for 60 epochs based on a cosine decay learning rate scheduler with an initial learning rate of 10^{-4} and 8 warm-up epochs. An Adam optimizer (Kingma & Ba, 2014) with a batch size of 64 are used for optimization in our study. The cropped images are resized to 224×224 pixels. For all models in our experiment, we use the weights pre-trained on ImageNet (J. Deng et al., 2009) and fine-tune on each dataset above. It should be noted that our EG-ViT model only uses eye-gaze data in the training stage. In the testing stage, we use the vanilla ViT architecture to load the trained weights for inference. All models were trained on an internal server with 10 NVIDIA GeForce RTX 1080Ti GPUs

(11GB). All experiments used the PyTorch deep learning framework (Paszke et al., 2019).

5.1.5 Conclusion

In this section, we proposed a novel eye-gaze-guided vision transformer (EG-ViT) to infuse human expert's intelligence and domain knowledge into the training of deep neural networks. This EG-ViT model is designed and implemented via the combination of eye-gaze guided mask generation and mask-guided vision transformer. The experiments on the INbreast (I. Moreira et al., 2012) and SIIM-ACR ("SIIM-ACR Pneumothorax Segmentation", 2020) datasets demonstrated that the radiologist's visual attention can effectively guide the model to concentrate on regions with potential pathology and achieve better performance. In particular, our EG-ViT model successfully rectifies the harmful shortcut learning and effectively improves the model's interpretability.

Overall, this work contributes a feasible solution for rectifying the harmful shortcuts in medial imaging application. It also provides a novel insight towards advancing current artificial intelligence paradigms by infusing human intelligence. Our future works include extending and evaluating the EG-ViT framework on other types of images, e.g., natural images, with eye-tracking data for few-shot learning problems and various downstream tasks.

5.2 Brain-inspired Adversarial Visual Attention Network

5.2.1 Introduction

Visual attention refers to the capability of selectively focusing on part of a visual scene rather than the entirety, given the limited processing capacity of the human visual system (Kastner & Ungerleider, 2000). Inspired by this fundamental biological process of the brain, a large group of deep learning studies successfully integrated attention mechanism into their deep neural networks for improving the performance and the interpretability (Hassabis et al., 2017; Vaswani et al., 2017). For example, in the computer vision (CV) field, a lightweight attention module has been introduced into convolutional neural networks (CNNs) and demonstrated consistent improvements on both image classification and object detection tasks (Woo et al., 2018). Also, those attention based neural networks have been employed to predict and study the human visual attention (N. Liu et al., 2015; W. Wang & Shen, 2017). In general, bridging the gap between brain

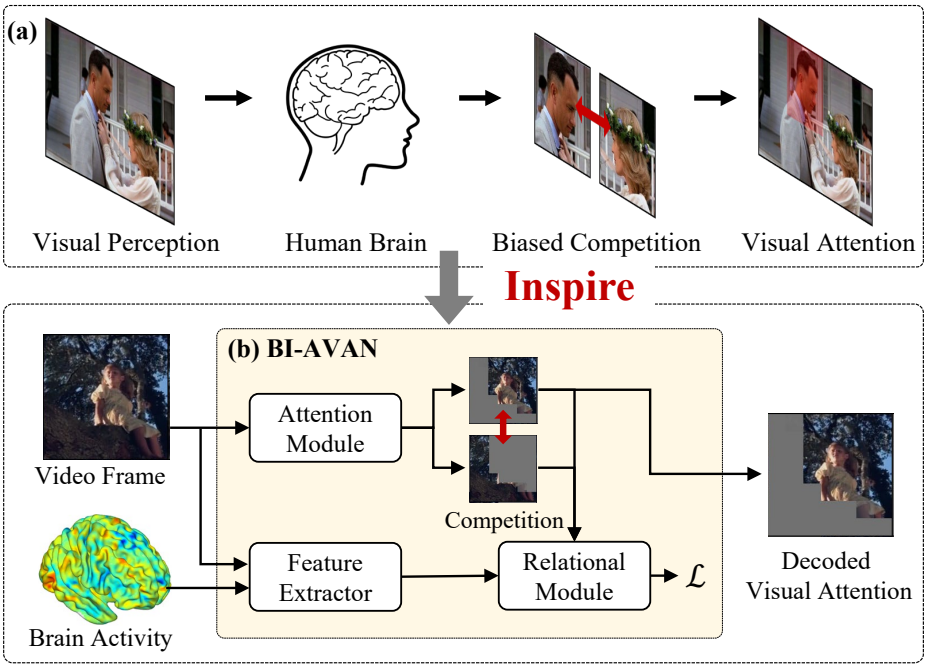

Figure 5.6: (a) An illustration of biased visual competition in human brain. (b) Overview of the proposed BI-AVAN model. Inspired by the biased competition in human brain, the attention module outputs the attention-related/neglected contents for visual competition to decode the human visual attention from brain activity.

science and artificial intelligence, e.g., incorporating the attention mechanism into deep learning can not only inspire and guide the design of neural networks with better performance/interpretability but also facilitate the understanding of our human brain.

In previous visual attention studies, eye-tracking based technology has been the dominant method to characterize human visual attention(H. Liu & Heynderickx, 2011; N. Liu et al., 2015; Sood et al., 2020; W. Wang & Shen, 2017). Researchers have demonstrated the close relationship between eye movements and visual attention (S. Liu et al., 2016; Sheliga et al., 1994). However, as an indirect way to study human visual attention, e.g., by linking the locations of eye gaze and brain attention, eye-tracking based methods may be limited in comprehensively and robustly characterizing human visual attention. For instance, some studies have reported that eye-tracking based methods may ignore the complex selective process of human brains including visual recognition, memory retrieving, and social perception (Kastner & Ungerleider, 2000). Therefore,

some studies turned to a more natural way by using functional magnetic resonance imaging (fMRI) to directly study visual attention from brain activity (Hu et al., 2015; Kanwisher & Wojciulik, 2000; Parhizi et al., 2018), and their results indicate the importance and superiority of fMRI in studying visual attention. But these studies only focus on the neuroscientific findings of visual attention, such as which brain regions (mainly visual cortex) are related to specific visual attention tasks (visual stimulus). Using functional brain activities to characterize and represent attention remains largely unexplored. Meanwhile, many brain science studies suggested that human visual attention has a biased competition (the top-down mechanisms of biased competition theory) (Beck & Kastner, 2009; Duncan et al., 1997). That is, the capacity of an individual's visual system is limited, and visual objects have to compete for the limited brain resources (Figure 5.6(a)). Hence, the attention-related objects and the neglected background always have an adversarial relationship: at the neural level, whenever there is a widespread maintenance of the attention related object's representation, there is a widespread suppression response to neglected background objects at the same time (Duncan et al., 1997). Unfortunately, this adversarial relationship has not been fully exploited, neither in neural networks design nor in characterizing the human visual attention using deep learning methods.

To bridge these gaps, in this section, we developed a Brain-inspired Adversarial Visual Attention Network (BI-AVAN) to characterize and decode the visual attention in a more real-world and complex scene (movie watching) with fMRI-derived brain activities. Specifically, in the visual attention decoding process, our BI-AVAN model imitates the biased competition process between attention-related and neglected objects (Figure 5.6(b)). To do so, we introduced an attention module to divide the outer environment (e.g., an image or a frame of the movie) into attention-related and neglected parts in an adversarial manner, in which each visual object in the image can only belongs to one of them. A relational module is then employed to maximize/minimize the relation between the attention-related/neglected parts and the brain activities. The rationale behind our BI-AVAN model design is that attention-related parts gain dominance in brain activities while neglected parts are suppressed in the brain. Thus, we can use the brain activities to guide the BI-AVAN model in the training stage. In the inference stage the attention module can locate the attention-related objects which are more coherent with brain activities. We adopted an fMRI dataset with simultaneous eye-tracking data to evaluate the performance of proposed BI-AVAN model. The experimental results demonstrate that the BI-AVAN model can effectively and robustly decode group-wise and individual-specific human visual attention. The brain networks identified from our BI-AVAN

model are meaningful and have a close relationship with the biased competition in the human brain. The objects of interests in human visual attention are also analyzed based on the inferred attention-related content. Overall, our BI-AVAN model provides novel insights on the computational aspects of the visual attention mechanism and contributes to the emerging field of brain-inspired AI.

5.2.2 Methods

Overview

Our proposed BI-AVAN model is an imitation of the biased competition process in human visual attention. Figure 5.7 illustrates the major modules of the BI-AVAN model, which consists of an attention module for locating the visual attention (Figure 5.7(a)), a feature encoding module for extracting both image and brain activity features (Figure 5.7(b)), and a relational module for discriminating attention-related and attention-neglected content (Figure 5.7(c)).

Attention Module with Adversarial Learning

The attention module is designed to imitate the biased competition process in human visual attention for characterizing the attention-related and attention-neglected parts. In order to maintain the adversarial relationship between these two parts in the attention module, the pixels of the input image are enforced to belong to only one of the two parts: mathematically, we use $F(x)$ to map pixel x to a possibility value α which denotes the probability of the pixel x belonging to the attention-related part. The possibility of pixel x belonging to the attention-neglect part is then represented as $1-\alpha$ to maintain the adversarial relationship. In this section, we illustrate the architecture of residual network in the attention module of BI-AVAN.

In this study, we used a residual neural network (ResNet-18) (K. He et al., 2016b) as the core component of our attention module to learn the function F(x). We removed the fully connected layer in ResNet-18 and modified the last convolution layer to produce a probability matrix α which represents the current probability of attention-related content (Figure 5.7(a)). $1-\alpha$ represents the probability of attention-neglected content in the image. Specifically, as shown in Figure 5.8(a), the residual network contains one 2D convolution layer and four residual building blocks (Figure 5.8(b)) followed by a Sigmoid activation function. For an input image, the residual network will process it layer by layer and output the probabilistic matrices α and $1-\alpha$. The residual building blocks input the images/feature maps into a 2D convolution layer followed by

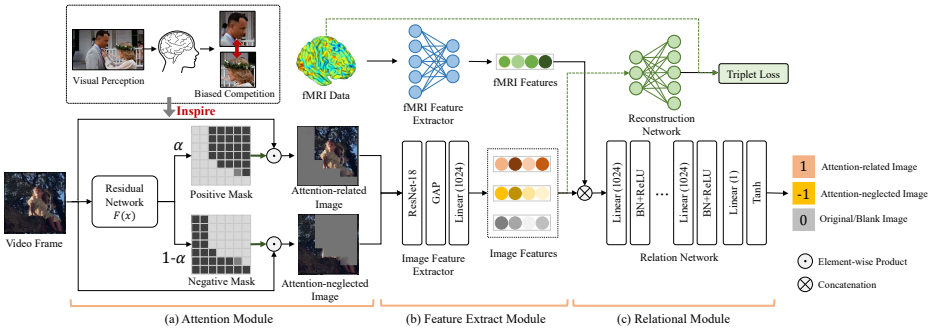

Figure 5.7: The proposed BI-AVAN framework. (a) shows the overall computational pipeline of the attention module. With an input image, the residual network generates two possibility matrices (denote as α and $1-\alpha$, respectively). Attention-related and attention-neglected content are obtained by dot product of the original image with the upsampled α and $1-\alpha$. (b) illustrates the feature extractor module which consists of an fMRI feature extractor and an image feature extractor. The concatenation of image features and fMRI features are input into (c) relational module to maximize the distance between attention-related/neglected contents.

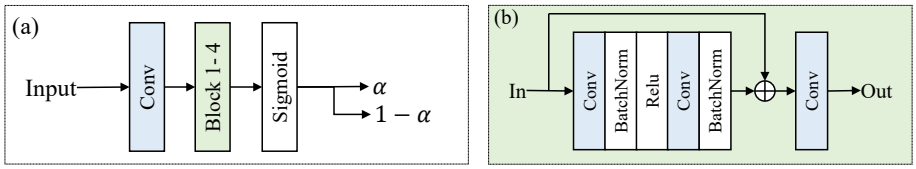

Figure 5.8: The architecture of the residual network in the attention module. (a) The residual network contains one 2D convolution layer and four residual building blocks, and the structure of which is shown in (b).

Batch Normalization (BN) and Relu activation function. Then, the output of another convolution layer with BN is concatenated with the input and passed into a final convolution layer to obtain the final output. Both α and $1-\alpha$ have the size of 7×7, which can be up-sampled to the original image size as two probability masks. The segmentation of attention-related/neglected parts are obtained by the dot product of the input image with the two probability masks, respectively. After segmentation, the resulted attention-related/neglected images will be the input of the feature encoding module (Section 5.2.2) for feature extraction.

Feature Encoding Module

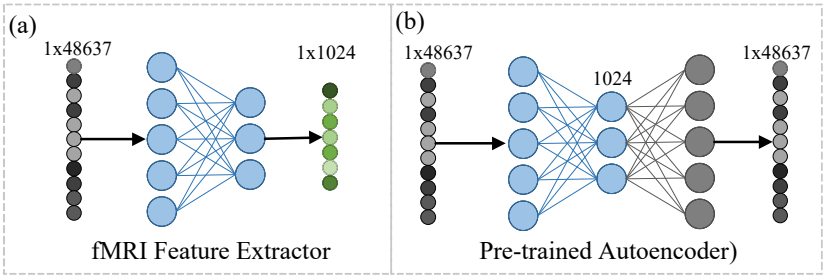

Figure 5.9: (a) The fMRI feature extractor is a regular feedforward neural network. Before the model training, it will be initialized with the encoder's weights in a pre-trained autoencoder (b).

The feature encoding module consists of an image feature extractor for attention-related/neglected images and an fMRI feature extractor. Both image and fMRI data will be encoded as feature vectors with sizes of 1×1024. For image feature extractor, we used ResNet-18 and replaced the last average pooling layer with a global average layer. The fully connected layer was correspondingly modified with 1024 units and tanh activation function. For fMRI feature extractor, we used a single fully connected layer to conduct a linear decomposition of fMRI data. A L1 regularization (with a penalty coefficient 5e-6) was applied to introduce the sparsity of the decomposition for the convenience of brain activity pattern visualization (Lv et al., 2015). It is noted that we adopted a pre-trained fMRI autoencoder to initialize the fMRI feature extractor. Specifically, before starting the training of the whole BI-AVAN model, we initialize the fMRI feature extractor with the encoder's weights in a pre-trained fMRI autoencoder (Figure 5.9). Considering that fMRI data is very high-dimensional, if we try to directly optimize the fMRI feature extractor, the number of parameters will be 48637×1024. Notably, the acquisition of fMRI is very expensive and time-consuming, and we only have 53,985 fMRI volumes in this study. So, such initialization can help reduce the overfitting and make model converge much faster. In our experiments, we found that by initializing the fMRI feature extractor with a pre-trained fMRI autoencoder, the model's performance can be significantly improved.

Relational Module

The relational module is the core component in our BI-AVAN model. Here, we assume the brain activities should show different patterns when focusing on the attention-related and attention-neglected content in the movie. Therefore,

the goal of the relational module is to discriminate the attention-related content from attention-neglected content by maximizing/minimizing the relationship between the brain activities (fMRI features) and the potential attention-related/neglected pixels in the images. Here, we formulate the relational module as a neural network $f_{rel}(\cdot)$. Figure 5.7(c) is an illustration of the relational module consisting of multiple fully connected layers followed by batch normalization. Specifically, the feature vectors (1×1024) of the attention-related/neglected images (v_a/v_n) and fMRI data (v_f) are concatenated as a single vector (1×2048). The relational module takes it as input and outputs a scalar, ranging from -1 (for attention-neglected image) to 1 (for attention-related image). In addition, we introduced two additional regularization terms in the relational module. The first regularization term makes the concatenation of the fMRI feature vector and the original image feature vector towards to 0 (original image = attention related contents + neglected contents). This regularization term will prevent the BI-AVAN model from identifying the entire image as attention related or neglected content. The second regularization term makes the concatenation of feature vectors from a blank image and fMRI data towards 0, which can effectively exclude the influence of fade in (fade out) movie frames. The training objective of the relational module is then formulated in Eq. (5.6):

$$\begin{aligned}\mathcal{L}_{rel} =&(1 - f_{rel}(v_{af}))^2 + (-1 - f_{rel}(v_{nf}))^2 \\ &+ f_{rel}(v_{anf})^2 + f_{rel}(v_{bf})^2\end{aligned} \tag{5.6}$$

where v_{af}, v_{nf} and v_{bf} represent the concatenation of fMRI feature vector, the feature vector of attention-related, attention-neglected and blank images, respectively. v_{anf} is the concatenation of fMRI feature vector and the vector of original image (equals to $v_a + v_n$). We also designed a triplet loss term to exclude some randomness of BI-AVAN model by introducing an fMRI reconstruction network $f_{rec}(\cdot)$ which is consisted of a fully connected layer followed by batch normalization. We use the feature vector of attention-related/neglected images to reconstruct the original fMRI data. The rationale behind this triplet loss is that the attention-related images should have a stronger relation to the brain activities and thus should have lower reconstruction error compared to neglected ones. By introducing the triplet loss, the order of attention related contents (neglected contents) will be fixed. The equation of triplet loss is shown in Eq. (5.7):

$$\mathcal{L}_{trip} = max\{d(s, f_{rec}(v_a)) - d(s, f_{rec}(v_n)) + m, 0\} \tag{5.7}$$

where the d is Euclidean distance function, s is the original fMRI data. We used a very small margin value ($m = 0.1$) to avoid the reconstruction network from

dominating the training of relational module. By combining the Eq. (5.6) and Eq. (5.7), the final loss function of BI-AVAN model is shown in Eq. (5.8):

$$\mathcal{L} = \mathcal{L}_{rel} + \mathcal{L}_{trip} \tag{5.8}$$

In general, the training objective of our BI-AVAN model is to minimize the loss function Eq. (5.8). In this study, we trained our model in a GeForce GTX 1080Ti graphics card with Adam optimizer, it takes us 51.5 hours to train the entire model.

Individual-specific Attention Related Content

It is worth noting that our attention module does not combine any individual brain activities or eye-tracking information to infer the visual attention, which means the derived attention related content is an inference of group interest. To obtain the individual-specific attention, first, we used the image feature extractor (the global average and dense layer are temporarily removed) to encode it as feature maps. As shown in Figure 5.10, for each movie frame we can obtain 512 feature maps (feature map size is 40x22), and each location in the feature maps corresponds to a 32x32 image block in the original movie frames (the size of receptive field is 32x32). Then, for each location in the feature maps, we applied a sliding window (window size is 3x3) to achieve a corresponding image codes vector (with the global average layer and dense layer). The code vector is then concatenated with individual's fMRI code vector to generate a relational value from the relational module. The relational values represent the correlation strength between image content and brain activities. The final individual-specific attention is obtained by dot product of the original image with individual's relational map (the relational map is up-sampled to the same size of the original image).

5.2.3 Experimental Results

Dataset

In this study we used a public movie dataset with simultaneous fMRI and eye-tracking data (Forrest Gump dataset, (Hanke et al., 2016)). The fMRI and eye-tracking data were recorded during watching a movie (2 hours of Forrest Gump). The dataset contains 15 subjects with 53,985 fMRI volumes (each subject has 3,599 fMRI volumes and each fMRI volume has 48,637 valid voxels). To preprocess the fMRI data, we used the standard and widely used fMRIprep preprocessing pipeline (Esteban et al., 2019). The preprocessing steps include: skull

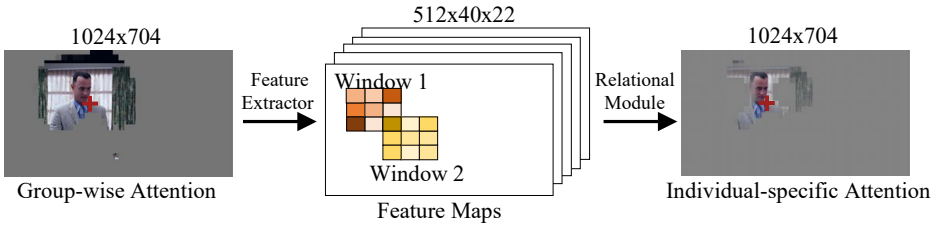

Figure 5.10: The computational pipeline of individual-specific attention. Each element on relational map represents the corresponding image blocks' relation to brain activities. In this figure, input image is center cropped for illustration purpose.

stripping, motion correction, slice time correlation, registration to standard the MNI space and global drift removal. More details are referred to (Esteban et al., 2019). After the standard preprocessing, we normalized voxels' signal with zero mean and standard deviation, and then applied a pre-defined brain mask to extract valid fMRI signals from these preprocessed 3D fMRI volumes. The extracted signals are flattened as 1D vectors for model training and validation. In our experiment, we have 15 subjects in total. Each subject has 3,599 fMRI volumes and each fMRI volume contains 48,637 valid fMRI time series signals.

The movie used in this study was encoded as H.264 video with a resolution of 1280×720 pixels (25fps). It contains 179,926 movie frames in total, and the viewing distance is 63cm which ensures that the participants can see the full screen (Hanke et al., 2016). For eye-tracking data preprocessing, we excluded the off-screen eye movements as well as the slow eye blinks (blink more than 300 milliseconds). Median filter with a window size of 40 was applied to reduce the noise in the eye-tracking data. According to the biased competition theory, the competition of visual objects only happens within the human visual field, which means for a large movie screen, only the objects around eye gaze point competes for the human attention. Thus, to obtain more accurate training data for every movie frame, we cropped the original movie frame around the eye-tracking points to a size of 224x224. By doing so, we implicitly introduced an assumption that all the cropped image contains both attention-related content and attention-neglected content. However, this assumption could be inaccurate considering the selective process of human brain. It is possible that the cropped image does not contain any attention-related content if it is not selected by the brain. Thus, to exclude the uncertainty, we used a regularization term in the relational module to force all original images to have zero relation

values. With this regularization term, the BI-AVAN model is forced to not build any relationships between the entire movie frame and the brain activities.

The eye movements of subjects were recorded at the frequency of 1000Hz. It is noted that the frequency for fMRI (0.5Hz), eye-tracking (1000Hz) and movie (25Hz) data are different. In order to eliminate such frequency difference and build the correspondence among different data sources, we downsampled the eye-tracking data to the same frequency as movie and up-sampled the fMRI data by interpolation.

In this study, we used 70% of the samples for model training and the rest 30% for testing. Unless we specifically mentioned, all the experimental results are based on testing data.

Evaluation of Group-wise Visual Attention

We first evaluated the performance of BI-AVAN model in generating the attention-related/neglected content. Since the attention-related content we directly obtained from the attention module is an inference from group interest, it contains all possible objects that the participants might be interested in. Figure 5.11 displays some random-selected results from training data (Figure 5.11(a)) and testing data (Figure 5.11(b)). More examples can be found in appendix. An interesting observation is that our attention module performs a semantic segmentation on image when identifying attention-related/neglected content: for most of the movie frames, the backgrounds (such as sky, ground, buildings, trees) are considered as attention-neglected content while other objects (such as people, poster, television, letters) are considered as attention-related content. We further compared the generated attention-related content to the results of eye-tracking. The corresponding eye-tracking points (the locations of the eye gaze) are marked as red dots in Figure 5.11. On both training and testing data, we found that most of the eye-tracking points are located within the attention-related regions instead of neglected regions. To quantitatively evaluate the performance of our attention module, we calculated the hit rates (hit rate equal to the ratio of eye-tracking points that are located within the attention-related region among all the eye-tracking points). For all 15 subjects, we have 1,586,459 eye-tracking points on the training data and 679,911 eye-tracking points on the testing data. We achieved 0.7793 and 0.5951 hit rate on training and testing data, respectively. The relatively high hit rates suggest that the attention-related part in our attention module has relatively higher possibility to draw participants' attention than the neglected part.

We also calculated the averaged values of the positive/negative/regularization terms in the relational module, as summarized in Table 5.3. In general, the out-

Table 5.3: The average output of relational module on training data and testing data

	Positive	Negative	Regularization
Target	1	-1	0
Training dataset	0.91	-0.94	0.023
Testing dataset	0.86	-0.83	0.12

puts of relational module indicate that the distance between attention-related and attention-neglected components, as expected, have been successfully maximized. Meanwhile, we observed the output values on training data are closer to targets than those on testing data, which suggests that over-fitting may exist due to the lack of fMRI data, given 53,985 fMRI volumes versus 179,926 movie frames. However, in our experiments, we found the over-fitting can be alleviated by initializing the parameters of fMRI feature extractor using a pre-trained fMRI autoencoder.

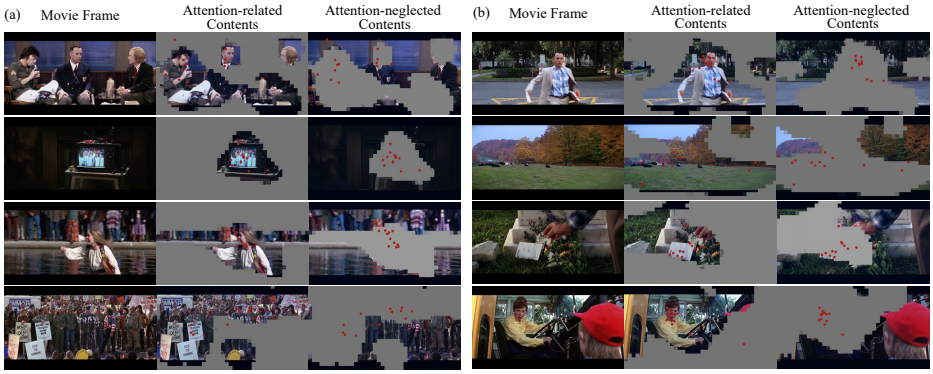

Figure 5.11: The group-wise attention-related content and attention-neglected content of BI-AVAN model from 4 randomly selected movie frames. Eye-tracking points are marked as red dots. (a) The segmentation results on training data; (b) The segmentation results on testing data.

Individual-specific Visual Attention and Eye Movements

In this section, we focus on individual-specific attention-related content and discuss its relationship with eye movements. Figure 5.12 shows an example of the obtained group-wise attention map as well as three individual-specific attention maps (participant #1-#3). Their corresponding eye-tracking points are highlighted in red, purple and blue color, respectively. More examples can be

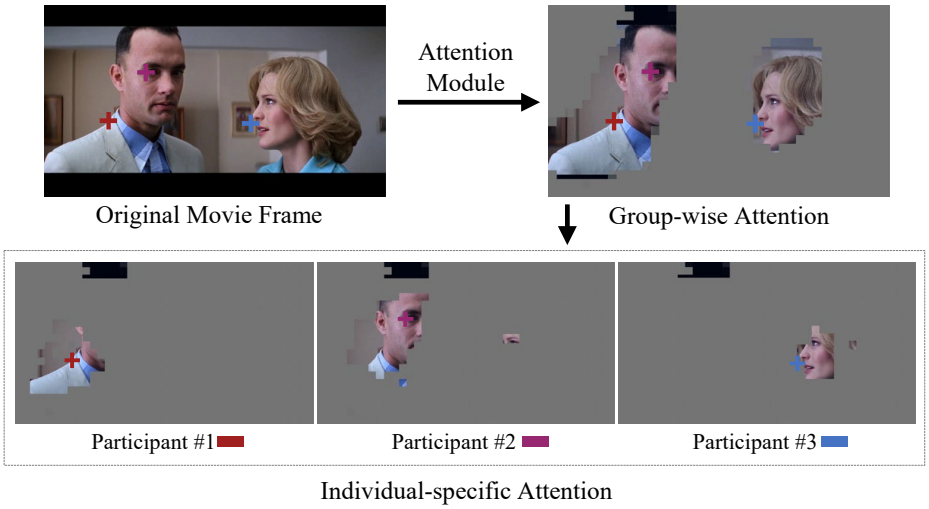

Figure 5.12: An example of group-wise attention and individual-specific attention. The eye-tracking points of different subjects are marked with different colors.

found in appendix. We can see that the individual-specific attention maps tend to be subsets of group-wise attention. The eye-tracking points are properly located in the individual-specific content, suggesting close relationship between individual attention and brain activities. Like the group-wise attention, we use the eye-tracking data to verify the individual results. Our model achieves 0.4189 (training data) and 0.3685 (testing data) hit rate on individual-specific attention. To investigate the reason of hit rate decrease, we visualized the individual-specific attention-related content maps and compared them to their corresponding eye-tracking points. We found that the decrease of hit rate is mainly caused by the random visual search of participants. Figure 5.13 illustrates a random case with three frames in a movie clip. We can see that during the movie watching, the participant quickly moved his/her eye gaze from Forrest to the sunset in Frame #2 and then moved back to Forrest in Frame #3. Although the eye movement happened in a real situation, our model will not capture it (highlight with green arrow in Figure 5.13) due to the sampling frequency of fMRI. The quick eye gaze movements here are usually related to human visual search, which is inevitable, since it is human instinct to scan visual environment for objects (Horowitz & Wolfe, 1998).

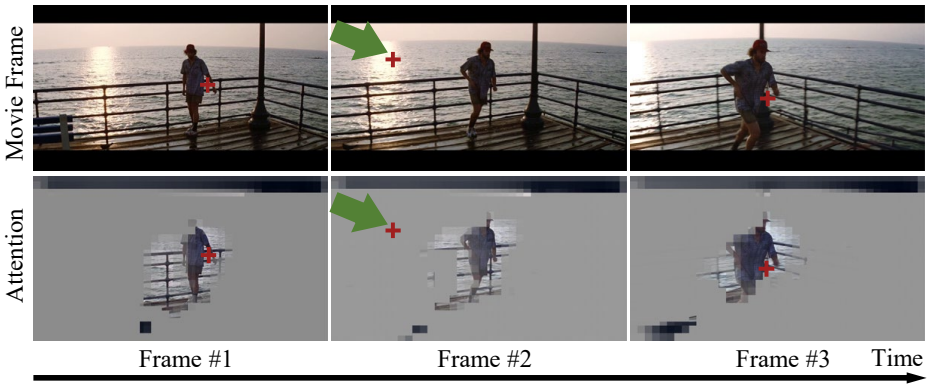

Figure 5.13: The illustration of human visual search. The images on top row are original movie frames, while those on the bottom row are the individual-specific attention. The eye-tracking points are highlighted as red cross-hair. The visual search happened in frame #2, where the individual-specific attention and eye-tracking point are mismatched.

Delay of Hemodynamic Response

It is noted that the brain activities measured by fMRI always lag behind the events (stimulus) due to the delay of hemodynamic response (HDR) (Aguirre et al., 1998). Therefore, we need to align the fMRI with the movie to eliminate the delay caused by HDR. We did several experiments to find the delay time for our case by assuming the delay time is 0s (no lag), 2s, 4s and 6s, respectively. For each assumption, we trained a corresponding BI-AVAN model and evaluated their performances according to their attention locating abilities (hit rate with eye-tracking points) as well as the relation value (output value) of attention-related components in the relational module. The comparison results are shown in Figure 5.14. From these experiments, we observed the best results when the delay time is 2s (in terms of the highest hit rate and the highest relation value, highlighted with black arrow) and the worst performance when delay time is 0s. In Figure 5.14, it is interesting to see relatively good results when the delay time is 4s and 6s. This might be related to the memory of the human brain (Nichols et al., 2006). Our experiment results are also consistent with previous fMRI study (DeYoe et al., 1994) which suggested that the fMRI response evoked by visual stimuli delayed 1-2s and reached 90% of peak in 5s. All experimental results in this study are based on the 2s delay time. However, we should point out that although we achieved reasonable results in this study, 2s delay time might not always be the optimal delay time for natural stimulus studies. We are not able to

continue narrowing down the range of the optimal delay time (a value between 0-2s) due to the physical limitations of fMRI. In the future, the results could be improved by utilize a higher temporal resolution fMRI scanner.

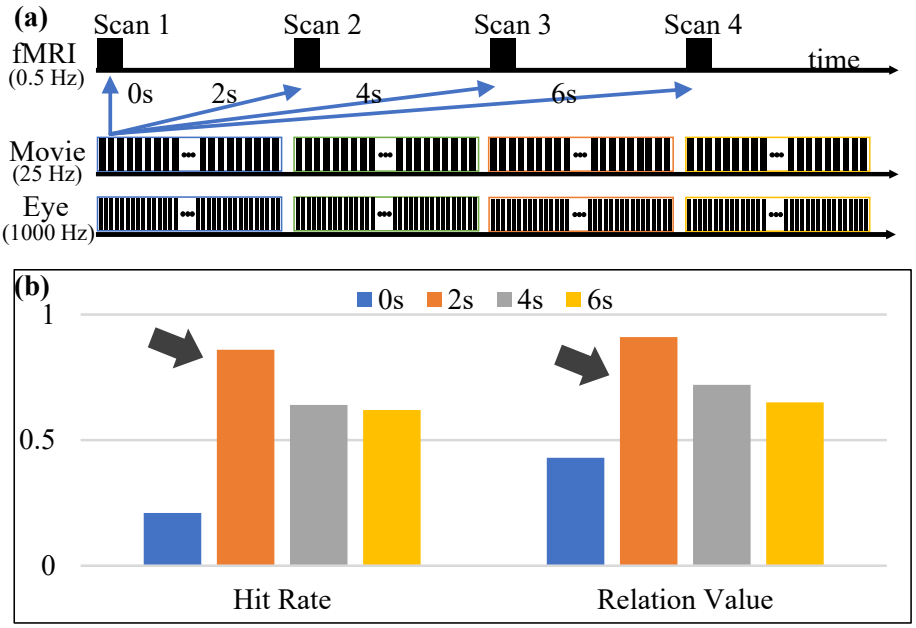

Figure 5.14: (a) An illustration of the frequency difference between different data sources. The black boxes represent the presence of samples. The movie frames and eye-tracking points between each two fMRI scans are highlighted in different colors. (b) The influence of different delay times with respect to the model's performance.

Brain Networks Learned by BI-AVAN Model

In this subsection, we evaluate if the BI-AVAN model learns meaningful brain networks from raw fMRI data for visual attention decoding. In BI-AVAN, the fMRI feature extractor is responsible for encoding the input fMRI signal to a feature vector. After the training process, the weight matrix of fMRI feature extractor contains the patterns learned from raw fMRI signals and each pattern can be interpreted as a specific brain network (Lv et al., 2015). The weight matrix has the size of 1024×48637, thus each row of the matrix represents a brain network. To visualize these brain networks, the values in each row are mapped back to the brain volume space and results in 1024 brain networks in total. The original fMRI data can be represented as a linear combination of these 1024 brain networks (Lv et al., 2015).

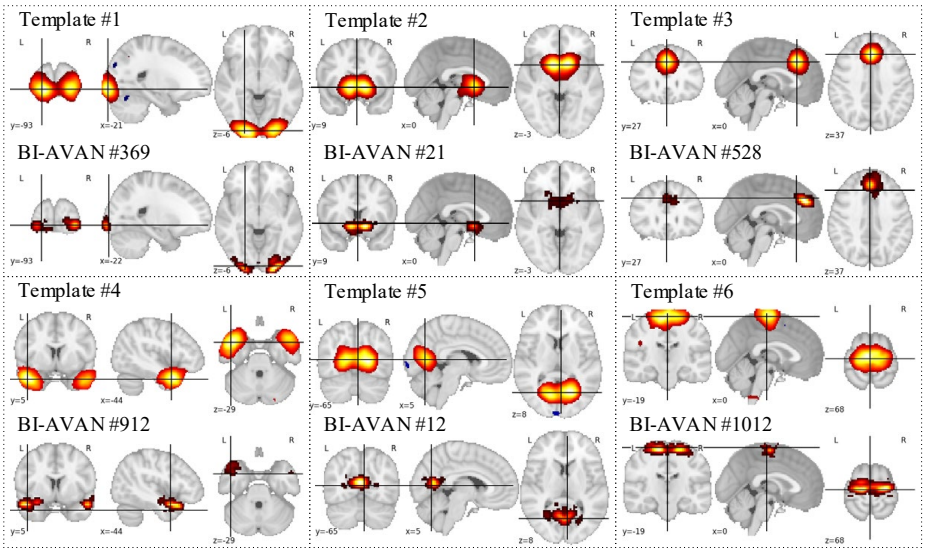

Figure 5.15: The comparison between the learned brain networks from BI-AVAN model and brain network templates shown in the same orthogonal slices with the same threshold value (3.0).

To verify if the learned brain networks are neurologically meaningful, we compared them with the widely used network templates (Smith et al., 2009) which include typical functional networks in human brain. Almost all the brain network templates can be found in our results (Figure 5.15). It is also inspiring that we found several brain networks which are highly related to the biasing attention of human brain. Previous fMRI studies (Henseler et al., 2011) demonstrated there is a control system located in rostral prefrontal cortex (rostral PFC). The rostral PFC acts as a gateway of human attention and plays a key role in balancing attentional orienting to external and internal information. In this study, the rostral PFC identified from BI-AVAN model are shown as three separate sub-regions in Figure 5.16. It is consistent with (Henseler et al., 2011) where they demonstrated the functional segregation exists between medial part and lateral part of rostral PFC. Specifically, the medial part (Rostromedial PFC) is responsible for processing external information by interacting with other parts of the brain, the lateral part (Rostrolateral PFC) processes the internally represented information and the anterior rostromedial PFC has relation to attentional biases generation.

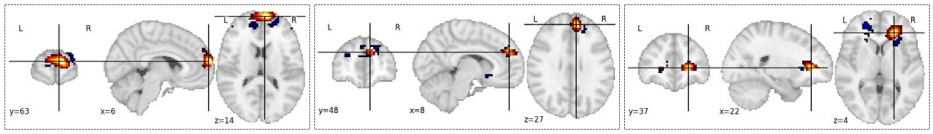

Figure 5.16: The rostral prefrontal cortex (rostral PFC) obtained from BI-AVAN model as three separate subareas. Only the right hemisphere is shown.

Objects of Interests in Visual Attention

We performed a statistical experiment to investigate which kind of movie objects have higher possibility to draw participants' attention. Figure 5.17 shows the statistical results of our experiment. We start from the main category (moving objects vs. stationary objects) and end up with specific sub-categories (facial features). During the movie watching, most subjects pay their attention to the moving objects rather than the stationary objects. The difference is significant (0.783 vs 0.217), indicating that when the moving and stationary objects present at the same time, the information of stationary objects is rarely processed by the participants (21.7% chance). This is probably because of the slow process of visual transduction, which has been suggested in (Berry et al., 1999) that visual stimulus evokes neural activity with a delay of 30-100 millisecond, therefore, the human brain may need to extrapolate the trajectory of a moving object in order to perceive its actual location. Among all the moving objects, participants are particularly interested in the human objects, and we are surprised to see that the human mouth has the highest possibility for drawing participants' attention (Figure 5.17). The underlying reasons still need to be investigated, but there are some studies suggest that the attention on mouth usually related to language learning or the intention of creating more opportunities for communication (Barenholtz et al., 2016). Because all the participants are Germans while the movie was filmed using English, so it is likely that the participants were trying to process language information of movie characters by focusing on their mouths.

5.2.4 Conclusion

In this work, we proposed a novel brain-inspired adversarial visual attention network (BI-AVAN) to characterize the human visual attention directly from functional brain activity. Our design of BI-AVAN model was inspired by the

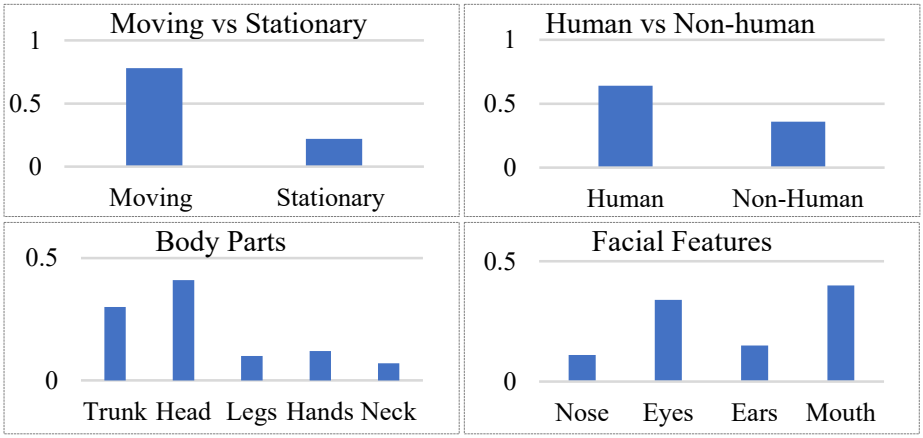

Figure 5.17: The objects of interests in visual attention. The y-axis denotes possibilities of objects drawing the human attention.

biased competition in the human visual system and can identify and locate the visual objects in a movie frame on which the human brain focuses. We evaluated the proposed BI-AVAN model with eye-tracking data and found that it achieved a high hit rate on both group-wise and individual-specific visual attentions. We also visualized the brain networks learned by the BI-AVAN model and discovered their strong correlations with the human visual attention. Finally, we studied the objects of interests in human visual attention statistically based on the proposed model. Overall, our BI-AVAN model contributes to the emerging field of leveraging the brain's functional architecture to inspire and guide the model design in AI, e.g., deep neural network. In our future work, we will try to use even larger scale natural stimulus fMRI data to further improve and evaluate the BI-AVAN model.

Chapter 6

Core-Periphery Principle Guided Convolutional Neural Network

6.1 Overview

The evolution of convolutional neural networks (CNNs) can be largely attributed to the design of its architecture, i.e., the network wiring pattern. Neural architecture search (NAS) advances this by automating the search for the optimal network architecture, but the resulting network instance may not generalize well in different tasks. To overcome this, exploring network design principles that are generalizable across tasks is a more practical solution. In this section, We explore a novel brain-inspired design principle based on the core-periphery property of the human brain network to guide the design of CNNs.

6.2 Background

Convolutional neural networks (CNNs) have greatly reshaped the paradigm of image processing with impressive performances rivaling human experts in the past decade (LeCun, Bengio, et al., 1995; LeCun et al., 2015; Q. Li et al., 2014). Though with a biologically plausible inspiration from the cat visual cortex(Hubel & Wiesel, 1959; LeCun, Bengio, et al., 1995), the evolution and success of CNNs can be largely attributed to the design of network architecture, i.e., the wiring pattern of neural network and the operation type of network nodes. Early CNNs such as AlexNet (Krizhevsky et al., 2017) and VGG (Simonyan & Zisserman, 2014) adopted a chain-like wiring pattern where the output of the preceding layer is the input of the next layer. Inception CNNs employ an In-

ception module that concatenates multiple branching pathways with different operations (Szegedy et al., 2015; Szegedy et al., 2016). ResNets propose a wiring pattern $x + F(x)$ aiming to learn a residual mapping that enables much deeper networks, and have been widely adapted for many scenarios such as medical imaging with superior performance and generalizability (K. He et al., 2016b). Orthogonally, depthwise separable convolution operation greatly reduces the number of parameters and enables extremely deeper CNNs (Howard et al., 2017). Recent studies also suggest that CNNs can benefit from adopting convolution operation with large kernels (e.g., 7×7) (Q. Han et al., 2021; Z. Liu et al., 2022) with comparable performance with Swin Transformer (Z. Liu et al., 2021b). By combining dilated convolution operation and large convolution kernel, a CNN-based architecture can achieve state-of-the-art in some visual tasks (Guo et al., 2022).

Neural Architecture Search (NAS) advances this trend by jointly optimizing the wiring pattern and the operation to perform. Basically, NAS methods sample from a series of possible architectures and operations through various optimization methods such as reinforcement learning (RL) (Zoph & Le, 2016), evolutionary methods (Real et al., 2019), gradient-based methods (H. Liu et al., 2018), weight-sharing (Pham et al., 2018), and random search (L. Li & Talwalkar, 2020). Despite its effectiveness, NAS does not offer a general principle for network architecture design. The outcome of NAS for each run is a neural network instance for a specific task, which may not be generalized to other tasks. For example, an optimal network architecture for natural image classification may not be optimal for X-ray image classification. Hence, some studies explored the design space of neural architectures (Radosavovic et al., 2020) and investigated the general design principles that can be applied to various scenarios.

Recently, a group of studies suggested that artificial neural networks (ANNs) and biological neural networks (BNNs) may share common principles in optimizing the network architecture. For example, the property of small-world in brain structural and functional networks are recognized and extensively studied in the literature (Bassett & Bullmore, 2017; Bassett & Bullmore, 2006; Bullmore & Sporns, 2009). In (S. Xie et al., 2019), the neural networks based on Watts-Strogatz (WS) random graphs with small-world properties yield competitive performances compared with hand-designed and NAS-optimized models. Through quantitative post-hoc analysis, (You et al., 2020) found that the graph structure of top-performing ANNs such as CNNs and multilayer perceptron (MLP) is similar to those of real BNNs such as the network in macaque cortex. (L. Zhao, Dai, et al., 2022) synchronized the activation of ANNs and BNNs and found that ANNs with higher performance are similar to BNNs in terms

of visual representation activation. Together, these studies suggest the potential of taking advantage of prior knowledge from brain science to guide the architecture design of neural networks.

6.3 Related Works

6.3.1 Neural Architecture of CNNs

Wiring Pattern. The development of the wiring pattern significantly contributes to CNN's performance. The early neural architecture of CNNs adopted chain-like wiring patterns, such as AlexNet (Krizhevsky et al., 2017) and VGG (Simonyan & Zisserman, 2014). Inception (Szegedy et al., 2015; Szegedy et al., 2016) concatenates several parallel branches with different operations together to "widen" the CNNs. ResNets(K. He et al., 2016b) propose a wiring pattern $x + F(x)$ for residual learning, which eliminates the gradient vanishing and makes the CNNs much deeper. DenseNet adopted a wiring pattern $[x, F(x)]$ which concatenates the feature maps from the previous layer. The wiring pattern of ResNet and DenseNet is well generalized in various scenarios and applications with improved performances.

Sparsity in Convolution Operation. Early CNNs used dense connectivity between input and output features, where every output feature is connected to every input feature. To reduce the parameter of such dense connectivity, depthwise separable convolution (Howard et al., 2017) was proposed to decompose the convolution operation as depthwise convolution and pointwise convolution, enabling much deeper CNNs. Another group of studies explored the pruning-based method to introduce sparsity in convolution operation, including channel pruning(Y. He et al., 2017), filter pruning (Q. Huang et al., 2018; J.-H. Luo et al., 2018), structured pruning (Z. Wang et al., 2021). The introduced sparsity reduced the number of parameters, making the networks easier to train, and also improved their performance on various tasks (Hoefler et al., 2021).

Neural Architecture Search. NAS jointly optimizes the wiring pattern and the operation to perform. NAS methods predefined a search space, and a series of possible architectures and operations are sampled and selected based on various optimization methods such as reinforcement learning (RL) (Zoph & Le, 2016), evolutionary methods (Real et al., 2019), gradient-based methods (H. Liu et al., 2018), weight-sharing (Pham et al., 2018), and random search (L. Li

& Talwalkar, 2020). However, the predefined search space still limited the feasible neural architectures to be sampled, regardless of the optimization methods. Meanwhile, the search process usually demands huge computational resources, while the searched architecture may not generalize well for different tasks.

6.3.2 Core-Periphery Structure

Core-periphery structure represents a relationship between nodes in a graph where the core nodes are densely connected with each other while periphery nodes are sparsely connected to the core nodes and among each other (Borgatti & Everett, 2000; Rombach et al., 2014). Core-periphery graph has been applied in a variety of fields, including social network analysis (Borgatti & Everett, 2000; Cattani & Ferriani, 2008), economics(Kostoska et al., 2020), biology such as modeling the structure of protein interaction networks(F. Luo et al., 2009). In the brain science field, it has been shown that brain dynamics has a core-periphery organization (Bassett et al., 2013). The functional brain networks also demonstrate a core-periphery structure(Gu et al., 2020). A recent study revealed the core-periphery characteristics of the human brain from a structural perspective (Yu et al., 2023). It is shown that gyri and sulci, two prominent cortical folding patterns, could cooperate as a core-periphery network which improves the efficiency of information transmission in the brain (Yu et al., 2023).

6.3.3 Connection of ANNs and BNNs

Recently, a group of studies suggested that artificial neural networks (ANNs) and biological neural networks (BNNs) may share some common principles in optimizing the network architecture. For example, the property of small-world in brain structural and functional networks are recognized and extensively studied in the literature (Bassett & Bullmore, 2017; Bassett & Bullmore, 2006; Bullmore & Sporns, 2009). Surprisingly, in (S. Xie et al., 2019), the neural networks based on Watts-Strogatz (WS) random graphs with small-world properties yield competitive performances compared with hand-designed and NAS-optimized models. Through quantitative post-hoc analysis, (You et al., 2020) found that the graph structure of top-performing ANNs such as CNNs and multilayer perceptron (MLP) is similar to those of real BNNs such as the network in macaque cortex. (L. Zhao, Dai, et al., 2022) synchronized the activation of ANNs and BNNs and found that ANNs with higher performance are similar to BNNs in terms of visual representation activation. Together, these studies suggest the

potential of taking advantage of prior knowledge from brain science to guide the model architecture design.

6.4 Method

We explore a brain-inspired Core-Periphery (CP) principle for guiding the architecture design of CNNs in this section. Core-Periphery organization is well-recognized in structural and functional brain networks of humans and other mammals (Bassett et al., 2013; Gu et al., 2020), which boosts the efficiency of information segregation, transmission and integration. In core-periphery graph, core-core node pairs have the strongest connection in comparison to core-periphery node pairs (moderate) and periphery-periphery node pairs (the lowest). We design a novel core-periphery graph generator according to this property and introduce a novel core-periphery principle guided CNN (CP-CNN). CP-CNN follows a typical hierarchical scheme of CNNs (e.g., ResNet (K. He et al., 2016b)) which consists of a convolutional stem and four consecutive blocks. For each block, we abandon the traditional chain-like wiring pattern but adopt a directed acyclic computational graph which is mapped from the generated core-periphery graph where each node corresponds to an operation such as convolution. In addition, we sparsify the convolution operation in a channel-wise manner and enforce it to follow a core-periphery graph constraint.

6.4.1 Generation of Core-periphery Graph

The core-periphery graph (CP graph) has a fundamental signature that the "core-core" node pairs have the strongest interconnections compared with the "core-periphery node" pairs (moderate) and "periphery-periphery" node pairs (weakest). According to this property, we introduce a novel CP graph generator to produce a wide spectrum of CP graphs in this subsection.

Specifically, the proposed CP graph generator is parameterized by the total number of nodes n, the number of "core" nodes n_c, and the wiring probabilities p_{cc}, p_{cp}, p_{pp} between "core-core", "core-periphery", "periphery-periphery" node pairs, respectively. The CP graph is generated based on the following process: for each "core-core" node pair, we sample a random number r from a uniform distribution on $[0, 1]$. If the wiring probability p_cc is greater than the random number r, the "core-core" node pair is connected. The same procedure is also applied to "core-periphery" node pairs and "periphery-periphery" node pairs with the wiring probability p_{cp} and p_{pp}, respectively. We summarize the whole generation process in Algorithm 1. With different combinations of n, n_c and

wiring probabilities p_{cc}, p_{cp}, p_{pp}, we can generate a wide range of CP graphs in the space, which are then used for constructing the CP-CNN framework introduced in the following subsections.

Algorithm 1: Generation of core-periphery graph

Input: n: number of nodes;
n_c: number of core nodes;
p_{cc}, p_{cp}, p_{pp}: wiring probabilities
Output: G: core-periphery graph

$G = \emptyset$;
// "core-core" node pairs
for $i \leftarrow 0$ **to** n_c **do**
  for $j \leftarrow i$ **to** n_c **do**
    Sample a uniform random number $r \in [0, 1)$ **if** $r < p_{cc}$ **then**
      $G \leftarrow (i, j)$
    end
  end
end
// "core-periphery" node pairs
for $i \leftarrow 0$ **to** n_c **do**
  for $j \leftarrow n_c$ **to** n **do**
    Sample a uniform random number $r \in [0, 1)$ **if** $r < p_{cp}$ **then**
      $G \leftarrow (i, j)$
    end
  end
end
// "periphery-periphery" node pairs
for $i \leftarrow n_c$ **to** n **do**
  for $j \leftarrow i$ **to** n **do**
    Sample a uniform random number $r \in [0, 1)$ **if** $r < p_{pp}$ **then**
      $G \leftarrow (i, j)$
    end
  end
end
return G

6.4.2 CP-CNN Framework

Our macro design of CP-CNN architecture follows a typical hierarchical scheme of CNNs (e.g., ResNet (K. He et al., 2016b)) with a convolutional stem and

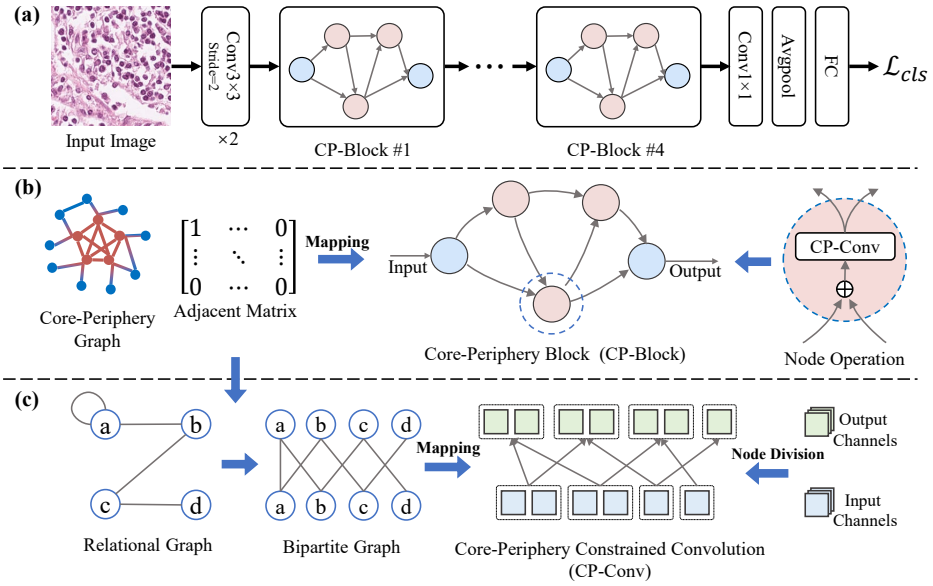

Figure 6.1: Illustration of the proposed CP-CNN framework. (a) The architecture of the CP-CNN with one convolution stem, four consecutive CP-Blocks, followed by one 1×1 convolution, one pooling and one fully-connected layer. (b)The construction of CP-Block and the illustration of the node in CP-Block. The core-periphery graph is mapped as a computational graph for CP-Block based on the node operation. (c) Utilizing core-periphery graph to constrain the convolution operation.

several convolution blocks (Figure 6.1(a)). Specifically, the input image is firstly input into a convolution stem which consists of two 3×3 convolutions with a stride of 2. The feature maps from the convolution stem are then processed by four consecutive core-periphery blocks (CP-Blocks, discussed in detail in Section 6.4.3 below). Within each CP-Block, the size of the feature map is decreased by $2\times$ while the number of channels is increased by $2\times$. A classification head with 1×1 convolution, global average pooling and a fully connected layer is added after the CP-Block to produce the final prediction.

6.4.3 Core-periphery Block

Unlike the traditional chain-like structure, our core-periphery block has a "graph" structure (Figure 6.1(b)) which is implemented based on the generated core-periphery graph. To construct the core-periphery block, we need to convert the generated core-periphery graph into computational graph in the neural network.

However, the generated core-periphery graphs are undirected while the computational graph in neural networks are directed and acyclic. So the first step is to convert the generated core-periphery graph into a directed acyclic graph (DAG), and then map the DAG into a computation graph for the CP-Block.

Specifically, we adopt a heuristic strategy to perform such conversion. For each node in the core-periphery graph, we randomly assign a unique label ranging from 1 to n (the number of nodes in the graph) to it. Then, for all undirected edges in the graph, we convert it into directed edges which always start from the node with the small label and end with the node having the large label. This approach guarantees that there are no circles in the resulting directed graph, i.e., the resulting graph is a DAG. The next step is to map the DAG into a computational graph in the neural network. To do so, we first need to define the node and edge in the computational graph.

Edges. Similar to edges in most computation graphs, we define that the directed edge in our implementation represents the direction of data flow, i.e., the node sends the data to another node along this flow.

Nodes. We define the nodes in our computational graph as processing units that aggregate and process the data from input edges and distribute the processed data to other nodes along the output edges. As illustrated in Figure 6.1(c), the data tensors along the input edges are firstly aggregated through a weighted sum. The weights of the aggregation are learnable. Then, the combined tensors are processed by an operation unit which consists of ReLU activation, 3×3 core-periphery convolution (discussed in detail in Section 6.4.4 below), and batch normalization. The unit's output is distributed as the same copies to other nodes along the output edges.

Using the defined nodes and edges, we obtain an intermediate computational graph. However, this graph may have several input nodes (those without input edges) and output nodes (those without output edges), while each block is expected to have only one input and one output. To address this, we introduce an additional input node that performs convolution with a stride of 2 on the previous block's output or the convolution stem, sending the same feature maps to all original input nodes. Similarly, we introduce an output node that aggregates the feature maps from all original output nodes using a learnable weighted sum, without performing any convolution within this node. This creates the CP-Block, which can be stacked in the CP-CNN as previously discussed.

6.4.4 Core-periphery Constrained Convolution

The CP-Block can also be constructed using conventional convolution in the nodes of the computational graph. However, traditional convolution is more "dense" whereas incorporating sparsity into the neural network can significantly lower its complexity and enhance its performance, especially in scenarios with limited training samples such as medical imaging.

Inspired by this, we propose a novel Core-Periphery Constrained Convolution that utilizes a core-periphery graph as a constraint to sparsify the convolution operation. Specifically, we divide the input and output channels of the convolution into n groups and represent the relationship between them as a bipartite graph (Figure 6.1(c)). In conventional convolution, the bipartite graph is densely connected, with all input channels in a filter contributing to the production of all output channels. For example, output channels in node $\#1$ integrate information from all input channels. In contrast, a sparse bipartite graph means that only a portion of input channels is used to generate output channels. As shown in Figure 6.1(c), the output channels in node $\#1$ only integrate information from input channels in node $\#1$ and node $\#2$. By sparsifying the convolution operation with a predefined bipartite graph, the convolution is constrained by a graph.S

We use the core-periphery graph as a constraint by converting the generated graph into a bipartite graph. The core-periphery graph is first represented as the relational graph proposed in (You et al., 2020) which represents the message passing between nodes. The relational graph is then transformed into a bipartite graph, where the nodes in two sets correspond to the divided sets of input and output channels, respectively. The edges in the bipartite graph represent message passing in the relational graph. We apply the resulting bipartite graph as a constraint to the convolution operation to obtain the core-periphery constrained convolution. It is worth noting that we apply the same core-periphery graph across the whole network, while the constrained convolution may vary among different nodes and blocks due to the varying number of channels.

6.5 Experiments

Datasets. We evaluate the proposed framework on three datasets, including one for natural images and two for medical images. **CIFAR-10** (Krizhevsky, Hinton, et al., 2009) consists of 60,000 32×32 images in 10 classes, with 50,000 images in the training set and 10,000 images in the test set. In our experiments, we upsample all original images in CIFAR-10 to 224×224. **NCT-CRC**

(Kather et al., 2018) contains 100,000 non-overlapping training image patches extracted from hematoxylin and eosin (H&E) stained histological images of human colorectal cancer (CRC) and normal tissue (Kather et al., 2018). Additional 7,180 image patches from 50 patients with no overlap with patients in the training set are used as a validation set. Both training and validation sets have 9 classes and size of 224×224 for each patch. **INbreast** dataset (I. C. Moreira et al., 2012) includes 410 full-field digital mammography images collected during low-dose X-ray irradiation of the breast. These images can be classified into normal (302 cases), benign (37 cases), and malignant (71 cases) classes. We randomly split the patients into 80% and 20% as training and testing datasets. To balance the training dataset, we perform several random cropping with a size of 1024×1024 as well as the contrast-related augmentation for each image, resulting in 482 normal samples, 512 benign mass samples, and 472 malignant mass samples. The images in both sets are downsized into 224×224.

Implementation Details. In our experiments, we set the number of nodes in the core-periphery graph to 16 and vary the number of core nodes. The three probabilities are set as $p_{cc} = 0.9$, $p_{cp} = 0.5$, $p_{cc} = 0.1$. The proposed model and all compared baselines are trained for 50 epochs with a batch size of 512. We use the AdamW optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a cosine annealing learning rate scheduler with initial learning 10^{-4} and 5 warm-up epochs. The framework is implemented with PyTorch (https://pytorch.org/) deep learning library and the model is trained on 4 NVIDIA A5000 GPU.

6.5.1 Comparison with Baselines

To validate the proposed CP-CNN, we compare the performance of CP-CNN with various state-of-the-art baselines, which can be roughly categorized as CNN-based methods and ViT-based methods. CNN-based category contains ResNet(K. He et al., 2016b), EfficientNet (Tan & Le, 2019), RegNet(Radosavovic et al., 2020), ConvNeXt(Z. Liu et al., 2022). The ViT-based class contains vanilla ViT(Dosovitskiy et al., 2020), CaiT(Touvron, Cord, Sablayrolles, et al., 2021) and Swin Transformer(Z. Liu et al., 2021b). Considering the amount of data, we set the number of nodes in the CP graph as 16, resulting CP-CNN model with around 22 million parameters. For the compared methods, we re-implement them and only report the tiny- or small-scale setting with comparable parameters as CP-CNN.

Table 6.1 demonstrates a comprehensive comparison of the Top-1 classification accuracy (%) achieved by different models on three datasets, as well as the

Table 6.1: Top-1 classification accuracy (%) of proposed and compared models on the CIFAR-10, NCT-CRC, and INBreast datasets, along with the number of parameters (M) and flops (G). The models with the highest accuracy are highlighted in **bold**. For some settings, the models do not get converged and are indicated by a slash (/) symbol.

Category	Models	CIFAR-10	NCT-CRC	INBreast	Param (M)	Flops (G)
CNNs	ResNet-18	90.35	95.96	83.56	11.18	1.8
	ResNet-50	90.55	95.11	82.19	23.53	4.1
	EfficientNet-B3	82.52	95.25	/	10.71	1.8
	EfficientNet-B4	81.73	95.21	/	17.56	4.4
	RegNetY-016	88.03	95.91	/	10.32	1.6
	RegNetX-032	88.78	95.91	/	14.30	3.2
	ConvNeXt-Nano	86.88	95.10	/	14.96	2.5
	ConvNeXt-Tiny	86.32	94.64	/	27.83	4.5
ViTs	ViT-Tiny	76.10	90.63	/	5.50	1.3
	ViT-Small	69.37	89.79	/	21.67	4.6
	CaiT-XXS-24	73.99	92.06	/	11.77	2.5
	CaiT-XXS-36	74.36	92.41	/	17.11	3.8
	SwinV2-T	81.76	95.61	/	27.57	5.9
CP-CNN	N=16, C=2	91.22	95.28	**85.75**	22.21	3.4
	N=16, C=4	91.71	95.43	82.19	22.21	3.4
	N=16, C=6	91.99	96.34	83.01	22.21	3.4
	N=16, C=8	92.41	**96.78**	83.01	22.21	3.4
	N=16, C=10	94.43	96.65	83.28	22.21	3.4
	N=16, C=12	92.54	96.29	83.56	22.21	3.4
	N=16, C=14	**92.65**	96.60	84.11	22.21	3.4

number of parameters and flops. It is observed that CNNs generally exhibit superior performance compared to ViTs. This can be attributed to the inductive biases in CNNs, which are essential in scenarios with a limited number of training samples. This is also suggested by the observation that SwinV2-T, which incorporates inductive biases, outperforms other ViT models.

Our proposed CP-CNN model achieves state-of-the-art performance compared to other CNN-based methods, demonstrating its superiority in terms of accuracy. Specifically, our CP-CNN outperforms baseline models in all settings on the CIFAR-10 dataset. For the NCT-CRC dataset, our CP-CNN model achieves higher accuracy compared to both CNNs and ViTs, except for sparse settings with only 2 or 4 core nodes. Furthermore, on the INBreast dataset, our sparse CP-CNN model with 2 core nodes achieves state-of-the-art performance. Importantly, our CP-CNN model's superior performance is achieved while requiring a comparable number of parameters and flops as other models. Thus,

the proposed CP-CNN can be a promising solution for image classification tasks, offering both high accuracy and efficiency.

It is also noted that our CP-CNN model outperforms the RegNet model which is also based on the exploration of design space of neural architecture. This indicates that the brain-inspired core-periphery design principle may be more generalized than the empirical design principles as those in RegNet.

6.5.2 Sparsity of CP Graph

The number of core nodes c controls the sparsity of the generated CP graph. More core nodes indicate the dense connections in the CP graph. In this subsection, we investigate the effects of CP graph sparsity on classification performances.

As illustrated in Table 6.1, we fix the number of total nodes to 16 and vary the number of core nodes from 2 to 14 in steps of 2, resulting in graph sparsity ranging from 0.125 to 0.875 with an interval of 0.125. For the CIFAR-10 dataset, we observed an increase in classification accuracy with the increase in the number of core nodes, reaching a peak with 14 core nodes (sparsity=0.875). It is probably because a dense graph increases the capacity of the CP-CNN model, so it can represent more complex relationships. In contrast, for the INBreast dataset, the sparsest CP-CNN model (2 core nodes, sparsity=0.125) yields the best performances. This may be due to the dataset having only thousands of training samples. A large and dense model may suffer from the overfitting problem, which reduces performance. For the NCT-CRC dataset, the performance increased with sparsity, with the highest accuracy achieved at a sparsity of 0.5. The accuracy slightly decreased with a more dense graph. This may be because a sparse model with low capacity may not be able to represent the complex relationships in the dataset, while a dense model may overfit. At a sparsity of 0.5, the right balance between model capacity and dataset complexity was achieved. Overall, the sparsity of the CP Graph can affect the capacity of the CP-CNN model and, thus, the performances on different datasets. Despite this, the CP-CNN model still has comparable and superior performances compared with baseline models.

6.5.3 Comparison with Random Graphs

To validate the effectiveness of the CP graph, we replace the CP graph in the CP-CNN model with two random graphs: Erdos-Renyi (ER) graph (Erdos, Rényi, et al., 1960) and Watts-Strogatz (WS) graph(Watts & Strogatz, 1998). ER graph is parameterized by P, which is the probability that two nodes are

connected by an edge. WS graph is considered to have the small-world property. In our experiment, we randomly generate 10 samples for ER, WS, and CP graphs with the same sparsity. In Figure 6.2, we demonstrate the average classification accuracy across the 10 samples for different graphs and sparsity.

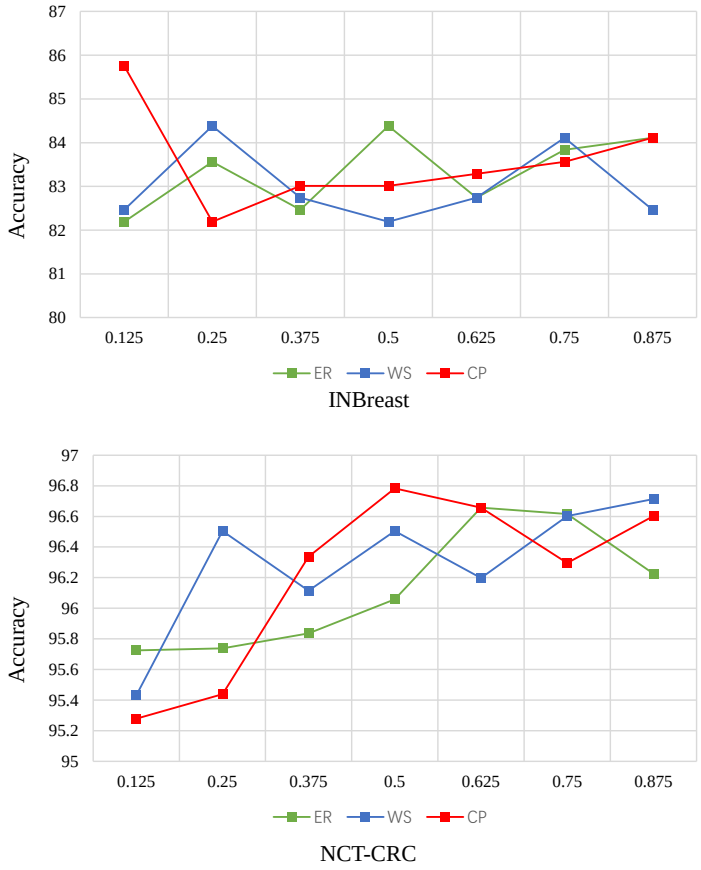

Figure 6.2: The comparison of ER, WS, and CP graph with varying sparsity based on the CP-CNN model, in terms of accuracy, using the INBreast and NCT-CRC datasets.

It is observed that the CP graph with a sparsity of 0.125 outperforms all other settings and graphs on the INBreast dataset, whereas on other sparsity settings, different graphs achieve the best accuracy. For the NCT-CRC dataset, the CP graph outperforms the ER and WS graphs with sparsity values of 0.375, 0.5, and 0.625, and achieves the highest accuracy among all settings and graphs with a sparsity of 0.5. These results suggest that the choice of graph type and sparsity can significantly affect the performance of the CP-CNN model on different datasets. However, with specific sparsity settings, CP graph can provide

superior performance compared to ER and WS graphs, i.e., the CP graph has an upper performance bound than ER and WS graphs.

In addition, the CP-CNN models based on ER and WS graphs also have competitive performances than the CNNs and ViTs in Table 6.1], highlighting the potential of incorporating graph structures in CNNs for improving their performance and generalization ability.

6.6 Discussion

Brain-inspired AI. The brain is a highly complex network of interconnected neurons that communicate with each other to process and transmit information. Core-periphery property is a representative signature of the brain network. The results reported in the study suggest that incorporating the properties/principles of brain networks can effectively improve the performance of CNNs. Our study provides a promising solution and contributes to brain-inspired AI by leveraging the prior knowledge of the human brain to inspire the design of ANNs.

Limitations. The sparsity of the CP graph can affect the capacity of CP-CNN model. The experiments demonstrated that the optimal capacity of the CP-CNN model may vary depending on the dataset and the specific problem being solved. Line and grid search may help us to determine the optimal sparsity for different datasets. However, how to effectively search the optimal sparsity is still an opening question. In addition, the proposed CP-CNN model is evaluated at a scale of 22 million parameters, which is suitable for relatively small datasets, especially those in medical imaging scenarios. The performances of a larger-scale CP-CNN model on a larger dataset, such as ImageNet-1K will be investigated in the future.

6.7 Conclusion

In this study, we explored a novel brain-inspired core-periphery design principle to guide the design of CNNs. The core-periphery principle was implemented in both the design of network wiring patterns and the sparsification of the convolution operation. The experiments demonstrate the effectiveness and superiority of the CP principle-guided CNNs compared to CNNs and ViT-based methods. In general, our study advances the growing field of brain-inspired artificial intelligence by integrating prior knowledge from the human brain to inspire the design of artificial neural networks.

Chapter 7

Conclusion and Future Works

This dissertation summarizes the major research works during my doctoral study. The investigation of human brain function in chapter 2 facilitates our understanding of the human brain's functional architecture, i.e., a potential core-periphery organization and the connections to its structural substrates. To leverage such prior knowledge for designing ANNs, the connection between ANNs and BNNs should be explored. The two embedding methods proposed in chapter 3 represent the brain structure and function in embedding spaces, providing a potential bridge to connect two areas. Based on the embedding of human brain function, a Sync-ACT framework was proposed to coupe the visual representation and semantics between FBNs and ANNs. We also explored human visual attention for training the neural network and proposed A brain-inspired adversarial visual attention network to decode the visual attention directly from brain activity. Finally, we introduce a brain-inspired core-periphery principle-guided CNN model which integrates prior knowledge from the human brain to inspire the design of artificial neural networks. To conclude, this dissertation contributes to the emerging field of brain-inspired artificial intelligence. Together, these contributions represent an important step forward in the field of brain-inspired artificial intelligence, and provide a foundation for future research and development in this exciting and rapidly evolving area.

The research works presented in this dissertation represent initial explorations in the field of brain-inspired artificial intelligence. Despite the progress made, there are still many promising avenues for future research. For instance, large language models (LLMs) like ChatGPT have recently shown remarkable performance in natural language processing (NLP) tasks, with 175 billion parameters, which is already comparable to the number of neurons in the human

brain (around 100 billion). Investigating the connection between LLMs and the human brain could prove to be a fruitful area of research. Additionally, further efforts are needed to better represent human brain function, which could lead to advances in the decoding of visual and linguistic information perceived by the brain. Attention and memory are also fundamental cognitive processes that are crucial to human brain function. Understanding how attention and memory work in the brain and incorporating these processes into artificial neural networks to improve their performance are promising directions for future research.

Bibliography

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). The variability of human, bold hemodynamic responses. *Neuroimage*, *8*(4), 360–369.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Andersen, A. H., Gash, D. M., & Avison, M. J. (1999). Principal component analysis of the dynamic response measured by fmri: A generalized linear systems framework. *Magnetic resonance imaging*, *17*(6), 795–815.

Armstrong, E., Schleicher, A., Omran, H., Curtis, M., & Zilles, K. (1995). The ontogeny of human gyrification. *Cerebral cortex*, *5*(1), 56–63.

Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., et al. (2013). Function in the human connectome: Task-fmri and individual differences in behavior. *Neuroimage*, *80*, 169–189.

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105.

Barrett, D. G., Morcos, A. S., & Macke, J. H. (2019). Analyzing biological and artificial neural networks: Challenges with opportunities for synergy? *Current opinion in neurobiology*, *55*, 55–64.

Bassett, D. S., & Bullmore, E. T. (2017). Small-world brain networks revisited. *The Neuroscientist*, *23*(5), 499–516.

Bassett, D. S., Wymbs, N. F., Rombach, M. P., Porter, M. A., Mucha, P. J., & Grafton, S. T. (2013). Task-based core-periphery organization of human brain dynamics. *PLoS computational biology*, *9*(9), e1003171.

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, *12*(6), 512–523.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, *49*(10), 1154–1165.

Bellman, R. E. (2015). *Adaptive control processes*. Princeton university press.

Berry, M. J., Brivanlou, I. H., Jordan, T. A., & Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature*, *398*(6725), 334–338.

Bhattacharya, M., Jain, S., & Prasanna, P. (2022). Radiotransformer: A cascaded global-focal transformer for visual attention-guided disease classification. *arXiv preprint arXiv:2202.11781*.

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, *21*(4), 375–395.

Braunlich, K., Gomez-Lavin, J., & Seger, C. A. (2015). Frontoparietal networks involved in categorization and item working memory. *NeuroImage*, *107*, 146–162.

Brodmann, K. (1909). *Vergleichende lokalisationslehre der grosshirnrinde in ihren prinzipien dargestellt auf grund des zellenbaues*. Barth.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, *10*(3), 186–198.

Buzsáki, G., Logothetis, N., & Singer, W. (2013). Scaling brain size, keeping timing: Evolutionary preservation of brain rhythms. *Neuron*, *80*(3), 751–764.

Calhoun, V. D., & Adali, T. (2006). Unmixing fmri with independent component analysis. *IEEE Engineering in Medicine and Biology Magazine*, *25*(2), 79–90.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional mri data using independent component analysis. *Human brain mapping*, *14*(3), 140–151.

Calhoun, V. D., Liu, J., & Adalı, T. (2009). A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, *45*(1), S163–S172.

Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., & Shen, D. (2017). Deformable image registration based on similarity-steered cnn regression. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 300–308.

Cattani, G., & Ferriani, S. (2008). A core/periphery perspective on individual creative performance: Social networks and cinematic achievements in the hollywood film industry. *Organization science*, *19*(6), 824–844.

Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep learning for cardiac image segmentation: A review. *Frontiers in Cardiovascular Medicine*, *7*, 25.

Chen, H., Li, Y., Ge, F., Li, G., Shen, D., & Liu, T. (2017). Gyral net: A new representation of cortical folding organization. *Medical image analysis*, *42*, 14–25.

Chen, H., Zhang, T., Guo, L., Li, K., Yu, X., Li, L., Hu, X., Han, J., Hu, X., & Liu, T. (2013). Coevolution of gyral folding and structural connection patterns in primate brains. *Cerebral Cortex*, *23*(5), 1208–1217.

Chen, X., Hsieh, C.-J., & Gong, B. (2021). When vision transformers outperform resnets without pre-training or strong data augmentations. *International Conference on Learning Representations*.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., & Glass, J. (2019). What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6309–6317.

Dancette, C., Cadene, R., Teney, D., & Cord, M. (2021). Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1574–1583.

Deco, G., & Kringelbach, M. L. (2017). Hierarchy of information processing in the brain: A novel 'intrinsic ignition'framework. *Neuron*, *94*(5), 961–968.

Deng, F., Jiang, X., Zhu, D., Zhang, T., Li, K., Guo, L., & Liu, T. (2014). A functional model of cortical gyri and sulci. *Brain structure and function*, *219*(4), 1473–1491.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, *53*(1), 1–15.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeYoe, E. A., Bandettini, P., Neitz, J., Miller, D., & Winans, P. (1994). Functional magnetic resonance imaging (fmri) of the human brain. *Journal of neuroscience methods*, *54*(2), 171–187.

Dong, Q., Ge, F., Ning, Q., Zhao, Y., Lv, J., Huang, H., Yuan, J., Jiang, X., Shen, D., & Liu, T. (2019). Modeling hierarchical brain networks via volumetric sparse deep belief network. *IEEE transactions on biomedical engineering*, *67*(6), 1739–1748.

Dong, Q., Qiang, N., Lv, J., Li, X., Liu, T., & Li, Q. (2020). Spatiotemporal attention autoencoder (staae) for adhd classification. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 508–517.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Drew, T., Evans, K. K., Võ, M. L.-H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: What can you see in a single glance and how might this guide visual search in medical images? *Radiographics*, *33*, 263–274.

Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T., & Hu, X. (2021). Towards interpreting and mitigating shortcut learning behavior of nlu models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 915–929.

Duncan, J., Humphreys, G., & Ward, R. (1997). Competitive brain activity in visual attention. *Current opinion in neurobiology*, *7*(2), 255–261.

Elsken, T., Metzen, J. H., & Hutter, F. (2019). Neural architecture search: A survey. *The Journal of Machine Learning Research*, *20*(1), 1997–2017.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., & Shadlen, M. N. (1994). Fmri of human visual cortex. *Nature*.

Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17–60.

Essen, D. C. v. (1997). A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature*, *385*(6614), 313–318.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). Fmriprep: A robust preprocessing pipeline for functional mri. *Nature methods*, *16*(1), 111–116.

Everett, M. G., & Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of mathematical sociology*, *23*(3), 181–201.

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, *9*(2), 195–207.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., et al. (2004). Automatically parcellating the human cerebral cortex. *Cerebral cortex*, *14*(1), 11–22.

Friston, K. (2008). Hierarchical models in the brain. *PLoS computational biology*, *4*(11), e1000211.

Ge, F., Lv, J., Hu, X., Guo, L., Han, J., Zhao, S., & Liu, T. (2018). Exploring intrinsic networks and their interactions using group wise temporal sparse coding. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 74–77.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, *80*, 105–124.

Goldman-Rakic, P., Selemon, L., & Schwartz, M. (1984). Dual pathways connecting the dorsolateral prefrontal cortex with the hippocampal formation and parahippocampal cortex in the rhesus monkey. *Neuroscience*, *12*(3), 719–743.

Golestani, N. (2014). Brain structural correlates of individual differences at low- to high-levels of the language processing hierarchy: A review of new approaches to imaging research. *International Journal of Bilingualism*, *18*(1), 6–34.

Golland, Y., Bentin, S., Gelbard, H., Benjamini, Y., Heller, R., Nir, Y., Hasson, U., & Malach, R. (2007). Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cerebral cortex*, *17*(4), 766–777.

Gu, S., Xia, C. H., Ciric, R., Moore, T. M., Gur, R. C., Gur, R. E., Satterthwaite, T. D., & Bassett, D. S. (2020). Unifying the notions of modularity and core–periphery structure in functional brain networks during youth. *Cerebral Cortex*, *30*(3), 1087–1102.

Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., & Hu, S.-M. (2022). Visual attention network. *arXiv preprint arXiv:2202.09741*.

Gutierrez-Barragan, D., Singh, N. A., Alvino, F. G., Coletta, L., Rocchi, F., De Guzman, E., Galbusera, A., Uboldi, M., Panzeri, S., & Gozzi, A. (2022). Unique spatiotemporal fmri dynamics in the awake mouse brain. *Current Biology*, *32*(3), 631–644.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, *30*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Han, Q., Fan, Z., Dai, Q., Sun, L., Cheng, M.-M., Liu, J., & Wang, J. (2021). Demystifying local vision transformer: Sparse connectivity, weight sharing, and dynamic weight. *arXiv preprint arXiv:2106.04263*.

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., & Stadler, J. (2016). A studyforrest extension, simultaneous fmri and eye gaze recordings during prolonged natural stimulation. *Scientific data*, *3*(1), 1–15.

Harding, I. H., Yücel, M., Harrison, B. J., Pantelis, C., & Breakspear, M. (2015). Effective connectivity within the frontoparietal control network differentiates cognitive control and working memory. *Neuroimage*, *106*, 144–153.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258.

Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., & Shi, H. (2021). Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M., Hou, X., Wang, Z., Kang, Z., Zhang, X., Qiang, N., & Ge, B. (2022). Multi-head attention-based masked sequence model for mapping functional brain networks. *Medical Image Computing and Computer As-*

sisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I, 295–304.

He, R., Gopinath, K., Desrosiers, C., & Lombaert, H. (2020). Spectral graph transformer networks for brain surface parcellation. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 372–376.

He, Y., Zhang, X., & Sun, J. (2017). Channel pruning for accelerating very deep neural networks. *Proceedings of the IEEE international conference on computer vision*, 1389–1397.

Heeger, D. J., & Ress, D. (2002). What does fmri tell us about neuronal activity? *Nature reviews neuroscience*, *3*(2), 142–151.

Henseler, I., Krüger, S., Dechent, P., & Gruber, O. (2011). A gateway system in rostral pfc? evidence from biasing attention to perceptual information and internal representations. *Neuroimage*, *56*(3), 1666–1676.

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2021). Natural language descriptions of deep features. *International Conference on Learning Representations*.

Hilgetag, C. C., & Barbas, H. (2005). Developmental mechanics of the primate cerebral cortex. *Anatomy and embryology*, *210*(5), 411–417.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *The Journal of Machine Learning Research*, *22*(1), 10882–11005.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature*, *394*(6693), 575–577.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, X., Deng, F., Li, K., Zhang, T., Chen, H., Jiang, X., Lv, J., Zhu, D., Faraco, C., Zhang, D., et al. (2010). Bridging low-level features and high-level semantics via fmri brain imaging for video classification. *Proceedings of the 18th ACM international conference on Multimedia*, 451–460.

Hu, X., Lv, C., Cheng, G., Lv, J., Guo, L., Han, J., & Liu, T. (2015). Sparsity-constrained fmri decoding of visual saliency in naturalistic video streams. *IEEE Transactions on Autonomous Mental Development*, *7*(2), 65–75.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., Guo, L., & Liu, T. (2017). Modeling task fmri data via deep convolutional autoencoder. *IEEE transactions on medical imaging*, *37*(7), 1551–1561.

Huang, H., Zhao, L., Hu, X., Dai, H., Zhang, L., Zhu, D., & Liu, T. (2022). Bi avan: Brain inspired adversarial visual attention network. *arXiv preprint arXiv:2210.15790*.

Huang, Q., Zhou, K., You, S., & Neumann, U. (2018). Learning to prune filters in convolutional neural networks. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 709–718.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, *148*(3), 574.

Huettel, S. A., Song, A. W., McCarthy, G., et al. (2004). *Functional magnetic resonance imaging* (Vol. 1). Sinauer Associates Sunderland.

Huntenburg, J. M., Bazin, P.-L., & Margulies, D. S. (2018). Large-scale gradients in human cortical organization. *Trends in cognitive sciences*, *22*(1), 21–31.

Jiang, M., Yang, S., Yan, J., Zhang, S., Liu, H., Zhao, L., Dai, H., Lv, J., Zhang, T., Liu, T., et al. (2020). Exploring functional difference between gyri and sulci via region-specific 1d convolutional neural networks. *International Workshop on Machine Learning in Medical Imaging*, 250–259.

Jiang, X., Li, X., Lv, J., Zhang, T., Zhang, S., Guo, L., & Liu, T. (2015). Sparse representation of hcp grayordinate data reveals novel functional architecture of cerebral cortex. *Human brain mapping*, *36*(12), 5301–5319.

Jiang, X., Li, X., Lv, J., Zhao, S., Zhang, S., Zhang, W., Zhang, T., Han, J., Guo, L., & Liu, T. (2016). Temporal dynamics assessment of spatial overlap pattern of functional brain networks reveals novel functional architecture of cerebral cortex. *IEEE Transactions on Biomedical Engineering*, *65*(6), 1183–1192.

Jiang, X., Zhang, T., Zhang, S., Kendrick, K. M., & Liu, T. (2021). Fundamental functional differences between gyri and sulci: Implications for brain function, cognition, and behavior. *Psychoradiology*, *1*(1), 23–41.

Jiang, X., Zhao, L., Liu, H., Guo, L., Kendrick, K. M., & Liu, T. (2018). A cortical folding pattern-guided model of intrinsic functional brain networks in emotion processing. *Frontiers in neuroscience*, *12*, 575.

Kanwisher, N., & Wojciulik, E. (2000). Visual attention: Insights from brain imaging. *Nature reviews neuroscience*, *1*(2), 91–100.

Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M. H., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., & Moradi, M. (2021). Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*, *8*, 1–18.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, *23*(1), 315–341.

Kather, J. N., Halama, N., & Marx, A. (2018). 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, *5281*.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, *53*(8), 5455–5516.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS computational biology*, *4*(11), e1000209.

Kim, J., Sangjun, O., Kim, Y., & Lee, M. (2016). Convolutional neural network with biologically inspired retinal structure. *Procedia Computer Science*, *88*, 145–154.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kok, E. M., & Jarodzka, H. (2017). Before your very eyes: The value and limitations of eye tracking in medical education. *Medical Education*, *51*, 114–122.

Kostoska, O., Mitikj, S., Jovanovski, P., & Kocarev, L. (2020). Core-periphery structure in sectoral international trade networks: A new approach to an old theory. *PloS one*, *15*(4), e0229547.

Kourtzi, Z., & Huberle, E. (2005). Spatiotemporal characteristics of form analysis in the human visual cortex revealed by rapid event-related fmri adaptation. *Neuroimage*, *28*(2), 440–452.

Krishnan, K. S., & Krishnan, K. S. (2021). Vision transformer based covid-19 detection using chest x-rays. *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, 644–648.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

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

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention Perception & Psychophysics*, *72*(5), 1205–1217.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, *242*, 396–402.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, L., & Talwalkar, A. (2020). Random search and reproducibility for neural architecture search. *Uncertainty in artificial intelligence*, 367–377.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., & Chen, M. (2014). Medical image classification with convolutional neural network. *2014 13th international conference on control automation robotics & vision (ICARCV)*, 844–848.

Li, Q., Dong, Q., Ge, F., Qiang, N., Wu, X., & Liu, T. (2021). Simultaneous spatial-temporal decomposition for connectome-scale brain networks by deep sparse recurrent auto-encoder. *Brain Imaging and Behavior*, *15*(5), 2646–2660.

Li, Q., Dong, Q., Ge, F., Qiang, N., Zhao, Y., Wang, H., Huang, H., Wu, X., & Liu, T. (2019). Simultaneous spatial-temporal decomposition of connectome-scale brain networks by deep sparse recurrent auto-encoders. *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, 579–591.

Li, Q., Wu, X., & Liu, T. (2021). Differentiable neural architecture search for optimal spatial/temporal brain function network decomposition. *Medical Image Analysis*, *69*, 101974.

Li, Q., Zhang, W., Zhao, L., Wu, X., & Liu, T. (2021). Evolutional neural architecture search for optimization of spatiotemporal brain network decomposition. *IEEE Transactions on Biomedical Engineering*.

Li, X., Chen, H., Zhang, T., Yu, X., Jiang, X., Li, K., Li, L., Razavi, M. J., Wang, X., Hu, X., et al. (2017). Commonly preserved and species-specific gyral folding patterns across primate brains. *Brain structure and function*, *222*(5), 2127–2141.

Li, Y., Zhang, K., Cao, J., Timofte, R., & Van Gool, L. (2021). Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*.

Liberman, L., & Menell, J. H. (2002). Breast imaging reporting and data system (bi-rads). *Radiologic Clinics of North America*, *40*, 409–430.

Lindquist, K. A., & Barrett, L. F. (2012). A functional architecture of the human brain: Emerging insights from the science of emotion. *Trends in cognitive sciences*, *16*(11), 533–540.

Liu, H., & Heynderickx, I. (2011). Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE transactions on Circuits and Systems for Video Technology*, *21*(7), 971–982.

Liu, H., Simonyan, K., & Yang, Y. (2018). Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Liu, H., Jiang, X., Zhang, T., Ren, Y., Hu, X., Guo, L., Han, J., & Liu, T. (2017). Elucidating functional differences between cortical gyri and sulci via sparse representation hcp grayordinate fmri data. *Brain research*, *1672*, 81–90.

Liu, H., Zhang, M., Hu, X., Ren, Y., Zhang, S., Han, J., Guo, L., & Liu, T. (2017). Fmri data classification based on hybrid temporal and spatial sparse representation. *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 957–960.

Liu, H., Zhang, S., Jiang, X., Zhang, T., Huang, H., Ge, F., Zhao, L., Li, X., Hu, X., Han, J., et al. (2019). The cerebral cortex is bisectionally segregated into two fundamentally different functional units of gyri and sulci. *Cerebral Cortex*, *29*(10), 4238–4252.

Liu, N., Han, J., Zhang, D., Wen, S., & Liu, T. (2015). Predicting eye fixations using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 362–370.

Liu, S., Ge, F., Zhao, L., Wang, T., Ni, D., & Liu, T. (2022). Nas-optimized topology-preserving transfer learning for differentiating cortical folding patterns. *Medical Image Analysis*, *77*, 102316.

Liu, S., Lv, J., Hou, Y., Shoemaker, T., Dong, Q., Li, K., & Liu, T. (2016). What makes a good movie trailer? interpretation from simultaneous eeg and eyetracker recording. *Proceedings of the 24th ACM international conference on Multimedia*, 82–86.

Liu, T., Hu, X., Li, X., Chen, M., Han, J., & Guo, L. (2014). Merging neuroimaging and multimedia: Methods, opportunities, and challenges. *IEEE Transactions on Human-Machine Systems*, *44*(2), 270–280.

Liu, T., Shen, D., & Davatzikos, C. (2004). Deformable registration of cortical structures via hybrid volumetric and surface warping. *NeuroImage*, *22*(4), 1790–1801.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021a). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Llinares-Benadero, C., & Borrell, V. (2019). Deconstructing cortical folding: Genetic, cellular and mechanical determinants. *Nature Reviews Neuroscience*, *20*(3), 161–176.

Logothetis, N. K. (2008). What we can do and what we cannot do with fmri. *Nature*, *453*(7197), 869–878.

Luo, F., Li, B., Wan, X.-F., & Scheuermann, R. H. (2009). Core and periphery structures in protein interaction networks. *BMC bioinformatics*, *10*(4), 1–11.

Luo, J.-H., Zhang, H., Zhou, H.-Y., Xie, C.-W., Wu, J., & Lin, W. (2018). Thinet: Pruning cnn filters for a thinner net. *IEEE transactions on pattern analysis and machine intelligence*, *41*(10), 2525–2538.

Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., & Heng, P.-A. (2021). Rethinking annotation granularity for overcoming deep shortcut learning: A retrospective study on chest radiographs. *arXiv preprint arXiv:2104.10553*.

Luo, W., Li, Y., Urtasun, R., & Zemel, R. (2016). Understanding the effective receptive field in deep convolutional neural networks. *Advances in neural information processing systems*, *29*.

Luo, X., Wei, L., Wen, L., Yang, J., Xie, L., Xu, Z., & Tian, Q. (2021). Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, *34*, 13073–13085.

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., Zhang, S., Hu, X., Han, J., Huang, H., et al. (2015). Sparse representation of whole-brain fmri signals for identification of functional networks. *Medical image analysis*, *20*(1), 112–134.

Lv, J., Jiang, X., Li, X., Zhu, D., Zhang, S., Zhao, S., Chen, H., Zhang, T., Hu, X., Han, J., et al. (2014). Holistic atlases of functional networks and interactions reveal reciprocal organizational architecture of cortical function. *IEEE Transactions on Biomedical Engineering*, *62*(4), 1120–1131.

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. *Proceedings of the European conference on computer vision (ECCV)*, 116–131.

Mahapatra, D., Poellinger, A., & Reyes, M. (2022). Interpretability-guided inductive bias for deep learning based medical image classification and segmentation. *Medical Image Analysis*, 102551.

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, *120*(3), 233–255.

Mall, S., Brennan, P. C., & Mello-Thoms, C. (2018). Modeling visual search behavior of breast radiologists using a deep convolution neural network. *Journal of medical imaging*, *5*, 035502–035502.

Mall, S., Krupinski, E., & Mello-Thoms, C. (2019). Missed cancer and visual search of mammograms: What feature-based machine-learning can tell us that deep-convolution learning cannot. *Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment*, *10952*, 281–287.

McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Minderer, M., Bachem, O., Houlsby, N., & Tschannen, M. (2020). Automatic shortcut removal for self-supervised representation learning. *International Conference on Machine Learning*, 6927–6937.

Mondal, A. K., Bhattacharjee, A., Singla, P., & Prathosh, A. (2021). Xvitcos: Explainable vision transformer based covid-19 screening using radiography. *IEEE Journal of Translational Engineering in Health and Medicine*, *10*, 1–10.

Monti, M. M. (2011). Statistical analysis of fmri time-series: A critical review of the glm approach. *Frontiers in human neuroscience*, *5*, 28.

Moon, C. H., Fukuda, M., & Kim, S.-G. (2013). Spatiotemporal characteristics and vascular sources of neural-specific and-nonspecific fmri signals at submillimeter columnar resolution. *Neuroimage*, *64*, 91–103.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.

Moreira, I., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, *19*, 236–248.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. *Academic radiology*, *19*(2), 236–248.

Mu, J., & Andreas, J. (2020). Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, *33*, 17153–17163.

Mwangi, B., Tian, T. S., & Soares, J. C. (2014). A review of feature reduction techniques in neuroimaging. *Neuroinformatics*, *12*(2), 229–244.

Nauta, M., Walsh, R., Dubowski, A., & Seifert, C. (2022). Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. *Diagnostics*, *12*(1), 40.

Nichols, E. A., Kao, Y.-C., Verfaellie, M., & Gabrieli, J. D. (2006). Working memory and long-term memory for faces: Evidence from fmri and global amnesia for involvement of the medial temporal lobes. *Hippocampus*, *16*(7), 604–616.

Nie, J., Guo, L., Li, K., Wang, Y., Chen, G., Li, L., Chen, H., Deng, F., Jiang, X., Zhang, T., et al. (2012). Axonal fiber terminations concentrate on gyri. *Cerebral cortex*, *22*(12), 2831–2839.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, *42*(1), 188–204.

Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, *1*(10), e3.

Parhizi, B., Daliri, M. R., & Behroozi, M. (2018). Decoding the different states of visual attention using functional and effective connectivity features in fmri data. *Cognitive neurodynamics*, *12*(2), 157–170.

Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., Moon, S., Lim, J.-K., & Ye, J. C. (2021). Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. *arXiv preprint arXiv:2103.07055*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018). Efficient neural architecture search via parameters sharing. *International conference on machine learning*, 4095–4104.

Qiang, N., Dong, Q., Liang, H., Ge, B., Zhang, S., Sun, Y., Zhang, C., Zhang, W., Gao, J., & Liu, T. (2021). Modeling and augmenting of fmri data using deep recurrent variational auto-encoder. *Journal of neural engineering*, *18*(4), 0460b6.

Qiang, N., Dong, Q., Sun, Y., Ge, B., & Liu, T. (2020). Deep variational autoencoder for modeling functional brain networks and adhd identification. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 554–557.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10428–10436.

Real, E., Aggarwal, A., Huang, Y., & Le, Q. V. (2019). Regularized evolution for image classifier architecture search. *Proceedings of the aaai conference on artificial intelligence*, *33*(01), 4780–4789.

Ren, Y., Nguyen, V. T., Guo, L., & Guo, C. C. (2017). Inter-subject functional correlation reveal a hierarchical organization of extrinsic and intrinsic systems in the brain. *Scientific reports*, *7*(1), 1–12.

Ren, Y., Xu, S., Tao, Z., Song, L., & He, X. (2021). Hierarchical spatio-temporal modeling of naturalistic functional magnetic resonance imaging signals via two-stage deep belief network with neural architecture search. *Frontiers in Neuroscience*, *15*, 794955.

Richman, D. P., Stewart, R. M., Hutchinson, J., & Caviness Jr, V. S. (1975). Mechanical model of brain convolutional development: Pathologic and experimental data suggest a model based on differential growth within the cerebral cortex. *Science*, *189*(4196), 18–21.

Robinson, E. C., Jbabdi, S., Glasser, M. F., Andersson, J., Burgess, G. C., Harms, M. P., Smith, S. M., Van Essen, D. C., & Jenkinson, M. (2014). Msm: A new flexible framework for multimodal surface matching. *Neuroimage*, *100*, 414–426.

Rombach, M. P., Porter, M. A., Fowler, J. H., & Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied mathematics*, *74*(1), 167–190.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211–252.

Saab, K., Hooper, S. M., Sohoni, N. S., Parmar, J., Pogatchnik, B. P., Wu, S., Dunnmon, J., Zhang, H., Rubin, D. L., & Ré, C. (2021). Observational supervision for medical image classification using gaze data. *Medical Image Computing and Computer-Assisted Intervention*.

Sabatinelli, D., Frank, D., Wanger, T., Dhamala, M., Adhikari, B., & Li, X. (2014). The timing and directional connectivity of human frontoparietal and ventral visual attention networks in emotional scene perception. *Neuroscience*, *277*, 229–238.

Salazar, R., Dotson, N., Bressler, S., & Gray, C. (2012). Content-specific fronto-parietal synchronization during visual working memory. *Science*, *338*(6110), 1097–1100.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Satpute, A. B., & Lindquist, K. A. (2019). The default mode network's role in discrete emotion. *Trends in cognitive sciences*, *23*(10), 851–864.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2022). Transformers in medical imaging: A survey. *arXiv preprint arXiv:2201.09873*.

Sheliga, B. M., Riggio, L., & Rizzolatti, G. (1994). Orienting of attention and eye movements. *Experimental brain research*, *98*(3), 507–522.

Shen, D., & Davatzikos, C. (2002). Hammer: Hierarchical attribute matching mechanism for elastic registration. *IEEE transactions on medical imaging*, *21*(11), 1421–1439.

Shen, X., & Lam, W. (2021). Towards domain-generalizable paraphrase identification by avoiding the shortcut learning. *Proceedings of the Interna-*

tional Conference on Recent Advances in Natural Language Processing (RANLP 2021), 1318–1325.

Shi, F., Yap, P.-T., Fan, Y., Gilmore, J. H., Lin, W., & Shen, D. (2010). Construction of multi-region-multi-reference atlases for neonatal brain mri segmentation. *Neuroimage*, *51*(2), 684–693.

Shome, D., Kar, T., Mohanty, S. N., Tiwari, P., Muhammad, K., AlTameem, A., Zhang, Y., & Saudagar, A. K. J. (2021). Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, *18*(21), 11086.

SIIM-ACR pneumothorax segmentation [[online] Available: https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation]. (2020).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences*, *106*(31), 13040–13045.

Sood, E., Tannert, S., Frassinelli, D., Bulling, A., & Vu, N. T. (2020). Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*.

Swensson, R. G. (1980). A two-stage detection model applied to skilled visual search by radiologists. *Attention Perception & Psychophysics*, *27*, 11–16.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., & Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2820–2828.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, 6105–6114.

Tang, Z., Liu, X., Li, Y., Yap, P.-T., & Shen, D. (2020). Multi-atlas brain parcellation using squeeze-and-excitation fully convolutional networks. *IEEE Transactions on Image Processing*, *29*, 6864–6872.

Toga, A. W., & Thompson, P. M. (2001). The role of image registration in brain mapping. *Image and vision computing*, *19*(1-2), 3–24.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning*, 10347–10357.

Touvron, H., Cord, M., Sablayrolles, A., Synnaeve, G., & Jégou, H. (2021). Going deeper with image transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 32–42.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The wu-minn human connectome project: An overview. *Neuroimage*, *80*, 62–79.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, H., Zhao, S., Dong, Q., Cui, Y., Chen, Y., Han, J., Xie, L., & Liu, T. (2018). Recognizing brain states using deep sparse recurrent neural network. *IEEE transactions on medical imaging*, *38*(4), 1058–1068.

Wang, J., Ren, Y., Hu, X., Nguyen, V. T., Guo, L., Han, J., & Guo, C. C. (2017). Test–retest reliability of functional connectivity networks during naturalistic fmri paradigms. *Human brain mapping*, *38*(4), 2226–2241.

Wang, Q., Zhao, S., He, Z., Zhang, S., Jiang, X., Zhang, T., Liu, T., Liu, C., & Han, J. (2022). Modeling functional difference between gyri and sulci within intrinsic connectivity networks. *Cerebral Cortex*.

Wang, S., Ouyang, X., Liu, T., Wang, Q., & Shen, D. (2022). Follow my eye: Using gaze to supervise computer-aided diagnosis. *IEEE Transactions on Medical Imaging*.

Wang, W., & Shen, J. (2017). Deep visual attention prediction. *IEEE Transactions on Image Processing*, *27*(5), 2368–2378.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, *13*(4), 600–612.

Wang, Z., Li, C., & Wang, X. (2021). Convolutional neural network pruning with structural redundancy reduction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14913–14922.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *nature*, *393*(6684), 440–442.

Welker, W. (1990). Why does cerebral cortex fissure and fold? *Cerebral cortex*, 3–136.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, *14*(6), 1370–1386.

Wu, G., Jia, H., Wang, Q., & Shen, D. (2011). Sharpmean: Groupwise registration guided by sharp mean image and tree-based registration. *NeuroImage*, *56*(4), 1968–1981.

Xiao, K., Engstrom, L., Ilyas, A., & Madry, A. (2021). Noise or signal: The role of image backgrounds in object recognition. *International Conference on Learning Representations*.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Xie, S., Kirillov, A., Girshick, R., & He, K. (2019). Exploring randomly wired neural networks for image recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1284–1293.

Xie, X., Bratec, S. M., Schmid, G., Meng, C., Doll, A., Wohlschläger, A., Finke, K., Förstl, H., Zimmer, C., Pekrun, R., et al. (2016). How do you make me feel better? social cognitive emotion regulation and the default mode network. *NeuroImage*, *134*, 270–280.

Xu, G., Knutsen, A. K., Dikranian, K., Kroenke, C. D., Bayly, P. V., & Taber, L. A. (2010). Axons pull on the brain, but tension does not drive cortical folding. *Journal of biomechanical engineering*, *132*(7).

Xu, Z., & Niethammer, M. (2019). Deepatlas: Joint semi-supervised learning of image registration and segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 420–429.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yang, S., Zhao, Z., Cui, H., Zhang, T., Zhao, L., He, Z., Liu, H., Guo, L., Liu, T., Becker, B., et al. (2019). Temporal variability of cortical gyral-

sulcal resting state functional activity correlates with fluid intelligence. *Frontiers in neural circuits*, *13*, 36.

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, *8*(8), 665–670.

You, J., Leskovec, J., He, K., & Xie, S. (2020). Graph structure of neural networks. *International Conference on Machine Learning*, 10881–10891.

Yu, X., Zhang, L., Dai, H., Zhao, L., Lyu, Y., Wu, Z., Liu, T., & Zhu, D. (2023). Gyri vs. sulci: Disentangling brain core-periphery functional networks via twin-transformer. *arXiv preprint arXiv:2302.00146*.

Yu, X., Zhang, L., Zhao, L., Lyu, Y., Liu, T., & Zhu, D. (2022). Disentangling spatial-temporal functional brain networks via twin-transformers. *arXiv preprint arXiv:2204.09225*.

Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., & Wu, W. (2021). Incorporating convolution designs into visual transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 579–588.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, *15*(11), e1002683.

Zhang, S., Liu, H., Huang, H., Zhao, Y., Jiang, X., Bowers, B., Guo, L., Hu, X., Sanchez, M., & Liu, T. (2018). Deep learning models unveiled functional difference between cortical gyri and sulci. *IEEE Transactions on Biomedical Engineering*, *66*(5), 1297–1308.

Zhang, T., Chen, H., Guo, L., Li, K., Li, L., Zhang, S., Shen, D., Hu, X., & Liu, T. (2014). Characterization of u-shape streamline fibers: Methods and applications. *Medical image analysis*, *18*(5), 795–807.

Zhang, T., Chen, H., Razavi, M. J., Li, Y., Ge, F., Guo, L., Wang, X., & Liu, T. (2018). Exploring 3-hinge gyral folding patterns among hcp q3 868 human subjects. *Human brain mapping*, *39*(10), 4134–4149.

Zhang, T., Zhu, D., Jiang, X., Zhang, S., Kou, Z., Guo, L., & Liu, T. (2016). Group-wise consistent cortical parcellation based on connectional profiles. *Medical image analysis*, *32*, 32–45.

Zhang, W., Zhao, L., Li, Q., Zhao, S., Dong, Q., Jiang, X., Zhang, T., & Liu, T. (2019). Identify hierarchical structures from task-based fmri data via hybrid spatiotemporal neural architecture search net. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 745–753.

Zhang, Z., Zhang, H., Zhao, L., Chen, T., & Pfister, T. (2021). Aggregating nested transformers. *arXiv preprint arXiv:2105.12723*.

Zhao, J., Tomasi, D., Wiers, C. E., Shokri-Kojori, E., Demiral, Ş. B., Zhang, Y., Volkow, N. D., & Wang, G.-J. (2017). Correlation between traits of emotion-based impulsivity and intrinsic default-mode network activity. *Neural plasticity*, *2017*.

Zhao, L., Dai, H., Jiang, X., Zhang, T., Zhu, D., & Liu, T. (2021). Exploring the functional difference of gyri/sulci via hierarchical interpretable autoencoder. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 701–709.

Zhao, L., Dai, H., Wu, Z., Xiao, Z., Zhang, L., Liu, D. W., Hu, X., Jiang, X., Li, S., Zhu, D., et al. (2022). Coupling visual semantics of artificial neural networks and human brain function via synchronized activations. *arXiv preprint arXiv:2206.10821*.

Zhao, L., Liu, H., Jiang, X., Zhao, S., He, Z., Liu, T., Guo, L., & Zhang, T. (2019). A task performance-guided model of functional networks identification. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1590–1593.

Zhao, L., Wu, Z., Dai, H., Liu, Z., Zhang, T., Zhu, D., & Liu, T. (2022). Embedding human brain function via transformer. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 366–375.

Zhao, L., Zhang, T., Guo, L., Liu, T., & Jiang, X. (2021). Gyral-sulcal contrast in intrinsic functional brain networks across task performances. *Brain Imaging and Behavior*, *15*(3), 1483–1498.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921–2929.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *40*(6), 1452–1464.

Zhu, D., Zhang, D., Faraco, C., Li, K., Deng, F., Chen, H., Jiang, X., Guo, L., Miller, L. S., & Liu, T. (2011). Discovering dense and consistent landmarks in the brain. *Biennial International Conference on Information Processing in Medical Imaging*, 97–110.

Zilles, K., Armstrong, E., Schleicher, A., & Kretschmann, H.-J. (1988). The human pattern of gyrification in the cerebral cortex. *Anatomy and embryology*, *179*(2), 173–179.

Zilles, K., Palomero-Gallagher, N., & Amunts, K. (2013). Development of cortical folding during evolution and ontogeny. *Trends in neurosciences*, *36*(5), 275–284.

Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.