IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Learning "What" and "Where": An Interpretable Neural Encoding Model

Haibao Wang
Research Center for Brain-Inspired Intelligence
National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Beijing, China
haibaow@hotmail.com

Lijie Huang
Research Center for Brain-Inspired Intelligence, CASIA
University of Chinese Academy of Sciences
Beijing, China
lijie.huang@ia.ac.cn

Changde Du
Research Center for Brain-Inspired Intelligence
National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Beijing, China
duchangde@gmail.com

Huiguang He*
Research Center for Brain-Inspired Intelligence
National Laboratory of Pattern Recognition, CASIA
University of Chinese Academy of Sciences
Center for Excellence in Brain Science and Intelligence Technology, CAS
Beijing, China
huiguang.he@ia.ac.cn

***Abstract*—Neural encoding modeling aims to reveal how brain processes perceived information by establishing a quantitative relationship between stimuli and evoked brain activities. In the field of visual neuroscience, many studies have been dedicated to building the neural encoding model for primary visual cortex and demonstrate that the population receptive field (pRF) models can be used to explain how neurons in primary visual cortex work. However, these models rely on either the inflexible prior assumptions imposed on the spatial characteristics of pRF or the clumsy parameter estimation methods which requiring too much manual adjustment. Suffering from these issues, current methods yield dissatisfactory performance on mimicking brain activity. In this paper, we address the problems under a novel "what and where" neural encoding framework. Basing on deep neural network (DNN) and the separability of the spatial ("where") and visual feature ("what") dimensions, the proposed method is not only powerful in extracting nonlinear features from images, but also rich in interpretability. Owing to two forms of regularization: sparsity and smoothness, receptive fields are estimated automatically for each voxel without prior assumptions on shape, which gets rid of the shortcomings of previous methods. Extensive empirical evaluations on publicly available fMRI dataset show that the proposed method has superior performance gains over several existing methods.**

***Index Terms*—neural encoding, deep network, "what" and "where", Laplacian, population receptive field**

I. Introduction

Vision is one of the most important channels for human to perceive and understand the world. How the human brain processes visual information, especially how the various brain regions encode the visual information, has always been an important open question in neuroscience and artificial intelligence researches. Towards this goal, numerous studies [1]–[4] have attempted to build neural information encoding models, and find that the information processing mechanism of neurons in visual cortex follows the retinotopic mapping rule. Specifically, in the visual system, the 2-dimensional plane of the retina is mapped multiple times onto the surface of visual cortex [5], and a strong coherence between the blood oxygen level-dependent signal arising in a voxel and stimulus locations in the visual field is revealed. Approaches to building encoding models for visual areas make use of visual features including, but not limited to, Gabor wavelet pyramid model [6], luminance contrast models [7]–[11], the motion energy model [12], semantic models [13], deepnet models [14], and other machine learning methods [15], [16]. The estimated characteristics of visual cortex derived from these models can be roughly divided into two categories: "what" and "where". "where" characterizes the spatial characteristic of neuron populations in visual cortex, i.e. the location and extent of pooling over visual features, while "what" characterizes the feature selection property of neuron populations in visual cortex.

Generally, "where" centers around classical receptive fields. In the population receptive field (pRF) model [7], the visual feature is a binary map of the pixels occupied by a high-contrast stimulus (e.g., bar, ring, wedge). For each voxel, the model is constructed by an isotropic Gaussian area, the pRF, that pooled the visual feature map within a spatially localized area. This seminal approach quantitatively measured population receptive field properties in human visual areas for the first time and was extended to other pRF models [7]–[11]. "What" focuses on what feature is meaningful and feature tuning functions. In the semantic model [13], several object category features (e.g. the presence of an animal, or a car.) are encoded as the vector of binary variables. Then, every object category was assigned a tuning parameter for each voxel to construct the model. Similarly, different visual features are further studied in subsequent models [6], [12], [15]–[17].

Recently, based on powerful representation ability of deep

978-1-7281-2009-6/$31.00 paper N-19793.pdf

neural networks (DNNs), a new approach [14] to encoding visual features, the feature-weighted receptive field (FWRF), was proposed. The key idea of the FWRF is that activity in each voxel encodes visual features in a spatially localized region across multiple feature maps. Through the DNN, visual stimulus was transformed into feature maps. FWRF further regressed all feature maps onto brain activities simultaneously, which yielded state-of-the-art prediction accuracy. However, while previous work demonstrated promising results of processing in visual cortex, voxel-wise encoding models still lack adequate examination and require plenty of efforts to improve.

There are two main challenges getting in the way of development of effective models. On one hand, conventional approaches [7], [8] are endowed with inflexible prior assumptions on the spatial characteristics of population receptive field (pRF), which limit the representation performance and effectiveness of models to a large extent. For example, in the population receptive model [7], it assumed that the pRF has an isotropic Gaussian topography while the potentially suppressive surround is neglected. There have been subsequent models [8], [10], [11], [14] which have adopted the same principles with different pRF topographies. In general, any assumption about receptive field structure puts a prior constraint on the ability to extract the pRF topography of the model. Specifically, inaccurate assumptions about pRF topography may lead to erroneous estimation of pRF characteristics, resulting in an ineffective model. Hence, it is meaningful to propose a new approach that can extract pRF topography in an unbiased manner.

On the other hand, previous approaches [7], [8], [10], [11], [14], [16] to obtaining pRF are based on grid search, which set search parameters according to experience and are prone to pRF center mislocalization and size miscalculation. In the population receptive models [7], [8], [14], the pRF topography parameters were set to certain parameters, which can be obtained by minimizing the residual between the observed fMRI signal and predicted signal. In this case, according to different shape parameters (i.e. center and radius), these models will inevitably generate large quantities of candidate pRF. That is, the grid fitting requires searching over quite large model-parameter spaces, which were assessed with respect to the explained variance to find the best parameters. Consequently, their encoding performances depend on the amount and parameter interval of candidate pRF, which are set artificially. Obviously, more often than not, it is not optimal and requires lots of manual effort. It would therefore be significant to obtain the pRF automatically in a more reasonable way.

Existing methods are prone to suffer from one or both of these issues and yield dissatisfaction. Attaching great importance to these bottlenecks, we proposed a "what" and "where" neural encoding (WWNE) network architecture (Fig. 1). On the basis of the convolutional neural network with a special pooling layer, WWNE enables us to take full advantage of classical population receptive field (pRF) without prior assumptions on shape, maintaining interpretability and powerful representation ability. Furthermore, owing to the regularization: sparsity and smoothness, our WWNE gets rid of the previous methods of searching best pRF shape from candidate pRF, and estimates the pRF for each voxel automatically. In addition, voxels in the same brain area tend to perform similar computations at different positions in the visual field [18]. In consideration of this, we construct our WWNE model with the voxel-shared module and voxel-specific module, which is beneficial to suppressing noise and improving the prediction performance.

To summarize, the main contributions of WWNE are outlined as follows:

- WWNE is a novel end-to-end "what" and "where" neural encoding model, performing feature learning from scratch without inflexible prior assumptions on the spatial characteristics of pRF.
- The estimation of pRF is essential to neural encoding models. Unlike selecting optimal shape parameters via grid search, WWNE adopted L1 regularization and Laplacian smoothing to estimate the pRF for each voxel automatically.
- In consideration of the computational similarities between voxels, WWNE model is divided into voxel-shared module and voxel-specific module, remarkably reducing the parameters and complexity of the model.
- Extensive empirical evaluations on the publicly available fMRI dataset show that our WWNE outperforms other baselines to achieve the state-of-the-art performance.

II. Problem definition

In this section, we introduce the notation and problem definition of this paper.

A. Notation

Boldface lowercase letters like $\mathbf{w}$ are used to denote vectors, while boldface uppercase letters like $\mathbf{W}$ denote matrices. The (i,j)th element of $\mathbf{W}$ is denoted as W_{ij}, the ith row of $\mathbf{W}$ and the jth column of $\mathbf{W}$ are denoted as $\mathbf{W}_{i*}$ and $\mathbf{W}_{*j}$, respectively. $\mathbf{W}^T$ is the transpose of $\mathbf{W}$. Specially, $\otimes$ and $*$ denote the element-wise product and convolution operation, respectively. $||\cdot||_1$ and $||\cdot||_F$ denote the L_1 norm ("entriwise" norm) and the Frobenius norm of a matrix, respectively.

B. Neural encoding model

Assume the training set consists of paired observations from two distinct modalities $(\mathbf{X}, \mathbf{Y})$, denoted by $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)$, where N is the training set size, $\mathbf{x}_i \in \mathbb{R}^{D_1}$ and $\mathbf{y}_i \in \mathbb{R}^{D_2}$ for $i = 1, \ldots, N$. Here $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^{D_1 \times N}$ denotes the visual image modality, where $\mathbf{x}_i$ can be the raw pixels or hand-crafted features of image i. Similarly, $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{D_2 \times N}$ denotes fMRI activity modality, where $\mathbf{y}_i$ is the fMRI activity pattern related to image i.

Given the training information $(\mathbf{X}, \mathbf{Y})$, neural encoding models aim to learn a quantitative model that describes how evoked brain activities $(\mathbf{Y})$ respond to the stimuli $(\mathbf{X})$. Specifically, for particular voxel j, the response to natural image i is y_{ij}, the bijective function from images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N \in$

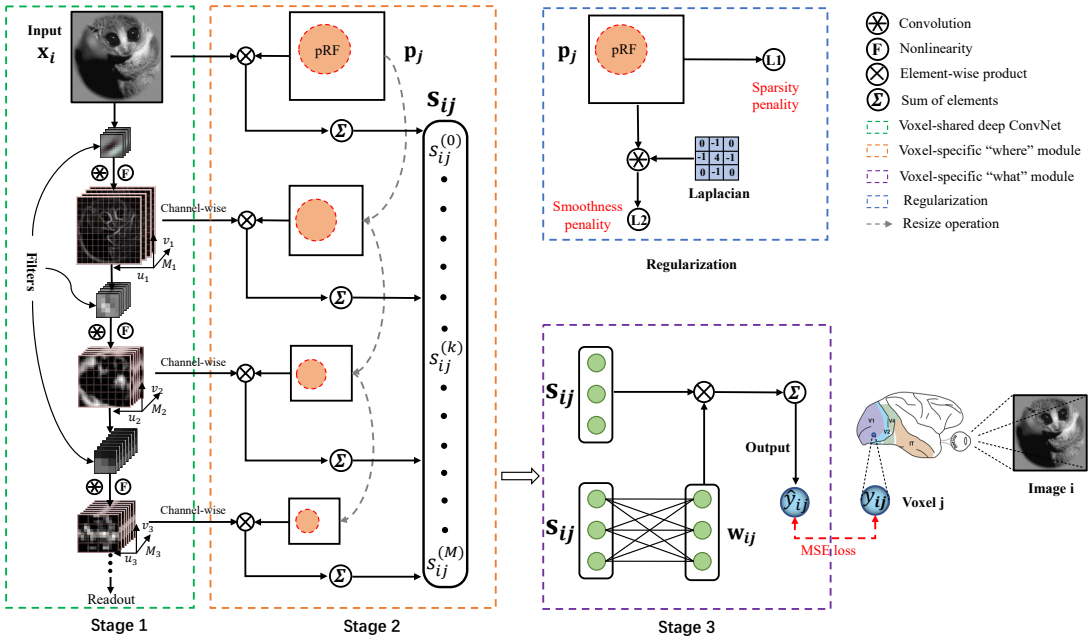

Fig. 1. The schematic illustration of WWNE. The proposed method predicts the brain activity $\hat{y}_{ij}$ in the voxel j, in response to visual stimulus $\mathbf{x}_i$ (i.e. image i). It consists of a voxel-shared feature representation network and two voxel-specific modules ("what" and "where"). In stage 1, image $\mathbf{x}_i$ is input into the feature representation network and is transformed into one and more feature maps (feature maps in three convolutional layers, denoted by $\{\mathbf{x}_i^{(k)} : k \in M_1\}$, $\{\mathbf{x}_i^{(k)} : k \in M_2\}$ and $\{\mathbf{x}_i^{(k)} : k \in M_3\}$, are shown in the figure, where $M_1 + M_2 + M_3 = M$, M_m is the number of feature maps in mth convolutional layer and M is the total number of feature maps). In stage 2, a special pooling field (i.e. pRF) is applied to learn "where" to attend in visual information processing. The pRF is initialized as a spatial mask $\mathbf{p}_j$, and is regularized by L1 penalty and L2 penalty on the Laplacian of the pRF. The output of pRF filtering operation for each feature map is denoted as $\mathbf{s}_{ij} = \{s_{ij}{}^{(k)}\}_{k=0}^{M}$, which is then weighted by a feature weight $\mathbf{w}_{ij}$ in stage 3. Here, the feature weight $\mathbf{w}_{ij}$ is produced by a fully connected layer. Finally, the weighted outputs are summed to produce the predicted brain activity $\hat{y}_{ij}$. It is worth noting that the object function of WWNE consists of three parts: sparsity penalty, smoothness penalty and MSE loss.

$\mathbb{R}^{D_1 \times N}$ to voxel responses $\mathbf{y}_{*j} = \{y_{ij}\}_{i=1}^{N} \in \mathbb{R}^{1 \times N}$ needs modeling by neural encoding models.

III. Model Architecture

In this section, we present the proposed method in detail. To ease understanding, our full model is described in parts. First of all, the feature representation network (stage 1) is described in Sec. III-A. Our proposed "where" module (stage 2) and "what" module (stage 3) is then described in Sec. III-B and Sec. III-C, respectively. Finally, Sec. III-D shows optimization for the WWNE model.

A. Feature representation network

Recently, deep learning with neural networks [19], [20] has been widely used to perform feature learning from scratch with promising performance, which sparks interest in using deep learning methods for understanding information processing in visual cortex [21]–[24].

The internal representations used by DNNs provide a natural and compelling set of hypotheses about the visual features encoded by activity in real brains [14]. For this reason, our feature representation network is a simple CNN model adopted from AlexNet [20]. There are five layers used in this paper, which are all convolutional layers.

Please note that the main goal of this paper is to introduce the idea that learning "what" and "where" with deep neural networks. However, how to design different feature representation networks is not the focus of this paper. Other feature representation networks might be more effective to perform feature learning for our WWNE, which remains to be further studied.

To perform feature learning from visual images, we input each image $\mathbf{x}_i$ into the feature representation network, then obtain related features $\mathbf{f}_i$ from a stack of feature maps, where $\mathbf{f}_i = \{\mathbf{x}_i^{(k)}\}_{k=0}^{M}$ denotes the set of feature maps, $\mathbf{x}_i^{(k)}$ is the kth feature map and M is the number of feature maps. Formally, a feature map is a matrix function, the kth feature map $\mathbf{F}^{(k)}(\mathbf{x}_i)$ outputs a matrix $\mathbf{x}_i^{(k)}$, where elements are feature map pixels. Specially, $\mathbf{F}_i^{(0)}$, $\mathbf{F}_i^{(1)}$, and $\mathbf{F}_i^{(2)}$ are the identical mapping, i.e., the first three feature maps are the raw image $\mathbf{x}_i$.

A question arises whether multiple feature maps included in WWNE are meaningful. Actually, we do not know up what features can better explain activity in the visual cortex under many circumstances. In this way, the full set of feature maps $\mathbf{f}_i$ is able to contain enough feature maps to capture the breadth of reasonable hypotheses about what is encoded in the visual cortex. Based on enough feature maps, we can use training samples (paired observations) to infer which features are more important for explaining the activity in the voxel. Each feature map will be assigned an associated feature weight, which indicates the importance of the feature map for predicting the

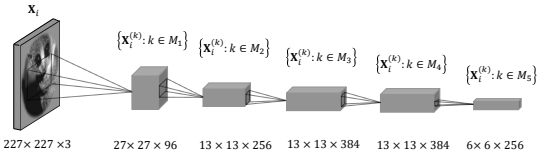

Fig. 2. Configuration of the feature maps.

activity of each voxel. In this sense, all the feature map weights are "what" parameters of the WWNE, which is introduced detailedly in "what" module. Similarly, "where" parameters are introduced in "where" module, the feature map pixels are focused on selectively, since concentrating on importance locations and suppressing unnecessary ones are conducive to improving the encoding performance.

B. "Where" module

The key idea of WWNE is the separability of "what" and "Where", which contributes to explanation of the activity in the voxel. To achieve this, we sequentially apply a special pooling layer ("Where") and a full-connected layer ("what"), so that the proposed method can learn "where" and "what" to attend in visual information processing separately. In "where" module, the visual feature maps are pooled within a spatially localized area, which is the population receptive field (pRF).

From the feature representation network, we obtain the features $\mathbf{f}_i$ from a stack of feature maps, where $\mathbf{f}_i = \{\mathbf{x}_i^{(k)}\}_{k=0}^{M}$ and $M = \sum_{m=0}^{5} M_m$, M_m is the number of feature maps in the mth convolutional layer. The size of the raw image i is $227 \times 227 \times 3$, which is denoted as $\{\mathbf{x}_i^{(k)} : k \in M_0\}$. Similarly, corresponding to $\{\mathbf{x}_i^{(k)} : k \in M_m\}, m = 1, ..., 5$, the size of feature maps for each convolutional layer is $27 \times 27 \times 96$, $13 \times 13 \times 256$, $13 \times 13 \times 384$, $13 \times 13 \times 384$, $6 \times 6 \times 256$ respectively, which are showed in Fig. 2.

On the basis of feature maps, we apply a special pooling layer to encoding each voxel's pRF location. Initially, For particular voxel j, the pRF is denoted as $\mathbf{p}_j$ and its size is 227×227 , which is same as each channel size of the raw image i. In consideration of different sizes of feature maps, pRF size was adapted to the each channel size for different convolutional layers by the means of reshaping. Taking the first convolutional layer as an example, the corresponding pRF size is 27×27. In this way, there are six sizes of the pRF denoted as $\{\mathbf{p}_j^{(k)} : k \in M_m\}, m = 0, ..., 5$, which have one-to-one correspondence to the channel size of the raw image i and five convolutional layers. As a direct way to combine feature maps with pRF, element-wise product is more natural and general than specially designed operations. Hence, we further obtain the element-wise product vector $\mathbf{s}_{ij} = \{{s_{ij}}^{(k)}\}_{k=0}^{M}$, where $s_{ij}^{(k)}$ denotes the output of pRF filtering operation for the kth feature map, which is computed as:

$$s_{ij}^{(k)} = \sum_{*} (\mathbf{x}_i^{(k)} \otimes \mathbf{p}_j^{(k)}) \tag{1}$$

where $\sum_{*}$ denotes the sum of elements of a matrix.

Note that if there is not any other operations applied to the pRF, in fact, it is just an ordinary mask. In order to make full use of advantages of deep learning and classical pRF, our WWNE model adopted two forms of regularization: sparsity and smoothness. Specifically, for particular voxel j, since we expect its pRF to be highly sparse, the pRF was regularized by L1 penalty with strength λ_s:

$$\mathcal{L}_{sparsity} = ||\mathbf{p}_j||_1 \tag{2}$$

To ensure WWNE model can optimize an area as effectively as possible, we use an L2 penalty on the Laplacian of the pRF with the strength λ_l:

$$\mathcal{L}_{laplace} = ||\mathbf{p}_j * \mathbf{L}||_F \ , \quad \mathbf{L} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{3}$$

In this way, sparsity and smoothness made the ordinary mask transform into the pRF we need.

Generally speaking, the original intention of neural encoding models is to account for the responses of different visual processing stages and reveal the information processing mechanism of neurons in visual cortex. As an interpretable method, pRF estimation is beneficial to explaining the evoked brain activities. In our model, the pooling layer focuses on "Where" is an efficient extent of visual features, which is complementary to convolutional layers and achieved by pRF. In order to overcome the drawbacks brought by the inflexible prior assumptions or the clumsy parameter estimation methods, the regularization: sparsity and smoothness was adopted in the pooling layer. It yielded explicit pRF automatically and encoded features within a contiguous region of the visual field, enhancing the interpretability of our WWNE model.

C. "What" module

Through the "Where" module, features $\mathbf{f}_i$ from a stack of feature maps turn into the output vector $\mathbf{s}_{ij} = \{{s_{ij}}^{(k)}\}_{k=0}^{M}$, where $s_{ij}^{(k)}$ means the integrated information of the kth feature map. Here, i indicates the image i and j indicates the voxel j. One natural way to cope with the output vector is to regress feature maps in each layer onto brain activity independently, which is proved highly effective [22], [24]. However, it reduces the model scale at the expense of model expressiveness.

Actually, in neural encoding models, it is a reasonable way that each of different feature maps is assigned an appropriate weight, which indicates the importance of the feature map for explaining the activity of voxels. It is certain that important feature maps are corresponding to larger weights while unimportant feature maps are corresponding to smaller weights. The weights for different feature maps can be learned from training data directly by the means of appropriate optimization algorithm. In our proposed method, the "What" module plays the same role and focus on "What" is meaningful given an input image.

To obtain the appropriate weights, we use a simple fully connected layer (activation function is the identity) to capture the relationship between feature maps, as these feature maps

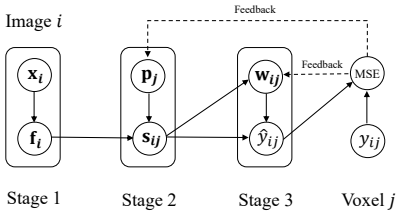

Fig. 3. Transformation of the mathematical representation (MSE loss).

are closely related. More specially, the feature maps in the latter layer are further abstraction of those in the former layer to a certain extent. The fully connected layer contained M units, which is the same number as feature maps. Without loss of generality, let $f(\mathbf{s}_{ij};\theta_j) \in \mathbb{R}^M$ denote the the output of fully connected layer for image i and voxel j. Here, θ_j is the network parameter of fully connected layer. Furthermore, the weights of different feature maps can be obtained as:

$$\mathbf{w}_{ij} = f(\mathbf{s}_{ij};\theta_j) \tag{4}$$

Finally, to predict the response of the voxel j to the natural image i, the wights $\mathbf{w}_{ij}$ is multiplied by the output from "Where" module. Formally:

$$\begin{aligned}\widetilde{y}_{ij} &= \mathbf{s}_{ij} \bullet \mathbf{w}_{ij} \\ &= \sum_{k=0}^{M} s_{ij}^{(k)} f(\mathbf{s}_{ij};\theta_j)^{(k)}\end{aligned} \tag{5}$$

where $\bullet$ is the vector inner product and k is the kth component.

Strictly speaking, there is tend to be an additional voxel-wise bias b_j. Hence, the prediction $\widetilde{y}_{ij}$ is modified as:

$$\hat{y}_{ij} = \widetilde{y}_{ij} + b_j \tag{6}$$

Here, we obtained the prediction $\hat{y}_{ij}$ and mean-squared error (Fig. 3) can be formulated as:

$$\mathcal{L}_{mse} = \frac{1}{B}\sum_{i}(y_{ij} - \hat{y}_{ij})^2 \tag{7}$$

where y_{ij} is the measured activity of voxel j in response to image i, $\hat{y}_{ij}$ is the predicted activity of WWNE model, B indicates the minibatch size.

D. Optimization for the WWNE model

The objective function of WWNE for particular voxel j is defined as follows:

$$\mathscr{T} = \mathcal{L}_{mse} + \lambda_s \mathcal{L}_{sparsity} + \lambda_l \mathcal{L}_{laplace} \tag{8}$$

where λ_l and λ_s are hyper-parameters. Note that WWNE model is a voxel-wise neural encoding model, without loss of generality, the objective function for any other voxel is formulated in the same way.

The first term $\mathcal{L}_{mse}$ is the MSE loss. Intuitively, minimizing this loss, which is equivalent to making generated values approximate true values, can result in more accurate predictions.

It is obvious that the L1 regularization term $\mathcal{L}_{sparsity}$ plays an sparse role. The pRF is supposed to be such a highly sparse area that optimizing the second term contributes to the generation of pRF.

The third term $\mathcal{L}_{laplace}$ is the L2 penalty on the Laplacian of the pRF, which is used to make the pRF smooth. More Specifically, by minimizing this term, the pixels in the particular area of pRF tend to be numerically consistent. Therefore, this constraint ensured that pRF encodes features within a contiguous region, strengthening the interpretability of WWNE model.

When optimizing the WWNE model, the goal is to infer pRF area $\mathbf{p}_j$ and weights $\mathbf{w}_{ij}$ of feature maps (i.e. the network parameter θ_j of fully connected layer) that lead to more accurate predictions of the voxel's response to any image. More specifically, expressed as an explicit function of $\mathbf{p}_j$, θ_j, the objective function can be minimized by Adam Optimizer [25] in our WWNE model.

IV. Experiments

In this section, we present extensive experimental results on the publicly available fMRI dataset to demonstrate the effectiveness of the proposed neural encoding model. Specifically, we compare our WWNE model with the following algorithms, each of which is either a traditional or a deep architecture:

- **Compressive spatial summation model** (CSS): The CSS model takes a contrast image (i.e. an image representing the location of contrast in the visual field), computes a weighted sum of the contrast image using an isotropic 2D Gaussian, and applies a static power-law nonlinearity [10]. The nonlinearity is typically compressive, hence the name of the model.
- **Second – order contrast model** (SOC): The SOC model starts with a grayscale image (luminance values), applies Gabor filters as a way of computing local contrast, computes second-order contrast within the spatial extent of an isotropic 2D Gaussian, and applies a static power-law nonlinearity [16]. Whereas the CSS model explains only how the location and size of the stimulus relate to the response, the SOC model is more general, explaining how an arbitrary grayscale image relates to the response.
- **Feature – weighted receptive field** (FWRF): A latest neural encoding method based on convolutional neural network and multivariate linear regression [14]. Initially, FWRF model starts with a natural image, obtains feature maps in a pre-trained convolutional neural network, and computes a weighted sum within the spatial extent of an 2D Gaussian receptive field. Afterwards, it regresses all the feature maps onto brain activity simultaneously and outputs accurate predictions, which outperforms other comparable encoding models to achieve the state-of-the-art performance.

A. Experimental testbed and setup

Data description. Experiments are conducted on the public fMRI dataset vim-1 from kay et al. and naselaris et al., which is described in detail in [6]. In summary, functional BOLD activity was measured in the occipital lobe with 4T INOVA MR scanner (Varian, Inc.) at a spatial resolution of 2mm×2mm×2.5mm and a temporal resolution of 1 Hz. During the acquisition, subjects viewed sequences of $20^o \times 20^o$ greyscale natural photographs while fixating on a central white square. Photographs were presented for 1s with a delay of 3s between successive photographs.

The data is partitioned into distinct training and testing sets. The training set consists of estimated voxel activity in response to 1750 photographs while the testing set consists of estimated voxel activity in response to 120 photographs, which is available online at https://crcns.org/data-sets/vc/vim-1. We used fMRI data from visual area V1, V2, V3, V4, LO, V3a, V3b of subject 1 (S1) for the analysis. Specially, the original training set is further partitioned into distinct training set and validation set in our experiments, separate validation set consists of 20% of the original training set. Note that all comparing algorithms were trained on the training set and evaluated on the testing set. The details of the vim-1 data set used in our experiments had been summarized in Table I.

TABLE I
THE DETAILS OF THE DATA SET USED IN OUR EXPERIMENTS.

ROIs	#Instances	#Pixels	#Voxels	#Training	#Validation	#Testing
V1	1870	227×227	1294	1400	350	120
V2	1870	227×227	2083	1400	350	120
V3	1870	227×227	1790	1400	350	120
V4	1870	227×227	1535	1400	350	120
LO	1870	227×227	928	1400	350	120
V3a	1870	227×227	484	1400	350	120
V3b	1870	227×227	314	1400	350	120

Voxel selection. Voxel selection is a crucial component to fMRI brain encoding, as plenty of voxels may not respond to the visual stimulus in reality. A common approach is to choose those voxels that are maximally correlated with the visual images during training. Voxels to which the model provided better predictability (encoding performance) is chosen, which caters to our intuition that the voxels better predicted with the visual images are those to be analyzed in neural encoding models. The goodness-of-fit between model predictions and measured voxel activities was quantified using the coefficient of determination (R^2) which indicates the percentage of variance that is explained by the model. Initially, the R^2 of each voxel on validation data is computed in our experiments, then voxels with positive R^2 were selected for further analyses.

Model fitting. The AlexNet [20] architecture pre-trained on ImageNet dataset is exploited to initial the five convolutional layers, and other parameters of WWNE model are randomly initialized. The hyper-parameters of the proposed WWNE were set to $(\lambda_s, \lambda_l) = (1, 1)$, while validation was carried out on the validation set to choose better regularization parameters. In our experiments, the minibatch size B is set to 20, and the Adam Optimizer [25] with an initial learning rate of 0.0001 with early stopping is adopted. Specifically, we monitored the validation loss every iteration of totally 1000 iterations and early stopped when the validation loss had not decreased. For fair comparison, model parameters of other methods had also been tuned carefully.

B. Performance evaluation

Visualization. The pRF shapes and contribution of each convolution layer produced by the proposed WWNE are visualized in Fig. 4, which are the typically effective voxels from V1, V2, V3, V4, LO. Note that the visualization of V3a and V3b is not presented in Fig. 4, while it is conducted in our experiments as well. Due to the limitation of fMRI data, each of the ROIs whose voxel number is either nearly 1000 or above 1000 was chosen for visualization, as it contains adequate voxels for further analyses.

It is easy to find that the pRF (circled by red line, showed in Fig. 4A) are smooth and unique for the particular voxels. Although the pRF may not be regular shape for all voxels, the main shapes can be clearly distinguished. Actually, On account of regularization, the pRF in our proposed method is able to capture the reasonable distribution of the training data and is estimated with higher generalization capability. In particular, for arbitrary visual image, the pRF can be optimized automatically in this way. In addition, the analysis of the WWNE feature weights is demonstrated in the form of the contribution to prediction accuracy each layer made. Specially, there are two forms of contributions: one is based on single layer (L contribution) while the other is based on single feature map (F contribution). L contribution represents the overall contribution (the sum contribution of feature maps) of each layer, F contribution represents the average contribution of feature maps in each layer. With the analysis on the contribution in Fig. 4B, there are subtle differences can be outlined. it is obvious that F contribution focus more on conv3 and conv4 than L contribution. We attribute this phenomenon to the fact that the conv3 and conv4 consist of more feature maps than other layers, which made greater real contribution to prediction accuracy practically. In fact, F contribution indicates the importance of feature maps from each layer and is more significant. The gradual decline of F contribution for later visual areas might be explained by the fact that primary visual cortex tend to focus on fundamental features rather than abstract features.

Evaluation. To evaluate the encoding performance quantitatively, we used several standard similarity metrics, including **mean squared error** (**MSE**), **Pearson's correlation coefficient** (**PCC**), and **fraction of explainable variance** (**FEV**, i.e. R^2). Note that MSE is not highly indicative of predictions, while PCC and FEV can address the shortcomings by taking variable texture and Goodness of fit into account. In addition, we also performed the **statistical significance test (SST)** of WWNE model prediction. Specifically, for each voxel, the

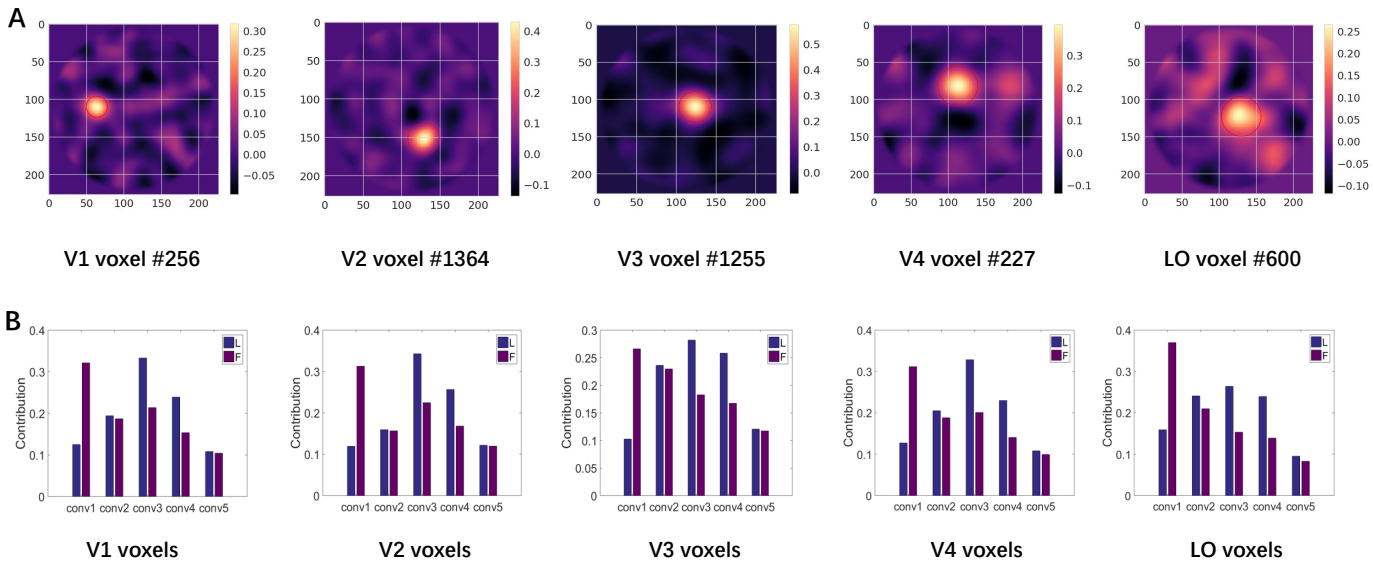

Fig. 4. Visualization of "What" and "Where". (A) panels show the typical voxel from distinct ROIs. Each pRF of the particular voxel is circled by red line. (B) panels demonstrate the contribution of each convolution layer. L means the contribution based on single layer while F means the contribution based on single feature map.

Pearson's correlation coefficient between the model prediction and measured response above 0.27 was significant $p < 0.001$ relative to its null hypothesis distribution. Performance comparisons and the count of significant voxels were listed in Table II. Several observations can be drawn as follows.

First and foremost, by comparing WWNE against the other algorithms, it is easy to find that WWNE performs considerably better in most brain ROIs. In particular, the MSE values and FEV values of WWNE remarkably surpass the baseline algorithms.

Second, by examining WWNE against CSS, which is a nonlinear model for visual images, we can find that WWNE always outperform CSS. The encouraging result shows that the WWNE with a DNN model for visual images is more able to extract nonlinear features from visual images, which may contribute to revealing the visual information processing in the primary visual cortex. Aside from it, the performance of SOC is moderate for all ROIs. We attribute this to the fact that it is a two-stage method, which is hard to obtain the global optimal solution of model parameters.

Third, WWNE shows obvious better performance than FWRF, especially in MSE and FEV. In spite of the same feature representation network, it is maybe caused by the fact that FWRF is endowed with the two-dimensional Gaussian assumption and searches optimal pRF form large quantities of candidate pRF. In another word, the amount and parameter interval of candidate pRF which are set artificially may not result in the optimal solution, especially for the high dimensionality of limited fMRI data instances. Accordingly, Owing to the regularization, WWNE made full use of training data and estimated optimal pRF automatically, which is more desirable.

Last but not least, statistical significance test of WWNE model prediction on ROIs verifies that the proposed model is significant for most selected voxels. Specially, it is worth noting that all the presented methods obtain good performance in early visual processing stages (V1,V2 and V3) while little effect on higher visual cortex (V4 and LO). Nevertheless, our WWNE model is still superior to other methods, which makes progress in higher visual cortex.

In summary, the substantial superior performance of WWNE verifies that estimating pRF automatically without prior assumptions imposed on the spatial characteristics is beneficial to enhancing the encoding performance.

V. Conclusion and Future Work

We have proposed a novel end-to-end "what" and "where" architecture to tackle the current bottlenecks faced by neural encoding models, which performs feature learning from scratch without prior assumptions imposed on the spatial characteristics of pRF. Owing to the regularization: sparsity and smoothness, WWNE get rid of shortcomings of clumsy parameter estimation methods which requiring too much manual adjustment. The pRF for each voxel can be estimated automatically in our method, which maintains interpretability as well. Moreover, in consideration of computational similarities between voxels, WWNE is partitioned into voxel-shared module and voxel-specific module, remarkably reducing the parameters and complexity of the model. Empirical evaluations show that the proposed method outperforms other baselines to achieve the state-of-the-art performance.

Two challenging and promising directions can be considered in the future. First, considering the attention mechanism in our framework, we can explore the relationship between feature maps from a new perspective and compress the network to force all computations to be performed in the pooling layer.

TABLE II
PERFORMANCE OF SEVERAL NEURAL ENCODING MODELS ON THE TEST SET. THE BEST PERFORMANCE ON EACH DATASET WAS HIGHLIGHTED.

Evaluation	Algorithms	V1	V2	V3	V4	LO	V3a	V3b
MSE	CSS [10]	0.2827	0.2684	0.2349	0.2308	0.2403	0.2268	0.2293
	SOC [16]	0.2815	0.2688	0.2355	0.2306	0.2400	0.2256	0.2272
	FWRF [14]	0.3070	1.2577	0.2969	1.4354	0.5210	0.2592	0.2273
	WWNE	**0.2389**	**0.2441**	**0.2270**	**0.2259**	**0.2354**	**0.2253**	**0.2268**
PCC	CSS [10]	0.0855	0.0983	0.0846	0.0812	0.0705	0.0702	0.0677
	SOC [16]	0.1442	0.1100	0.1116	0.1152	0.1418	0.0019	0.0414
	FWRF [14]	0.5045	**0.4908**	0.3922	**0.3338**	0.2350	0.1672	0.1653
	WWNE	**0.5089**	0.4581	**0.4050**	0.2869	**0.2572**	**0.1732**	**0.1657**
FEV	CSS [10]	0.0201	0.0138	0.0151	0.0246	0.0194	0.0052	0.0122
	SOC [16]	0.0302	0.0261	0.0320	0.0299	0.0264	0.0007	0.0012
	FWRF [14]	0.1845	-0.1513	0.06996	0.0349	-1.5323	0.0009	-1.2794
	WWNE	**0.2306**	**0.1711**	**0.1041**	**0.0598**	**0.0507**	**0.0085**	**0.0205**
SST	WWNE	**390**	**441**	**171**	**109**	**60**	**4**	**5**

Second, in consideration of each subject's fMRI measurements, multi-subject neural encoding can be further explored.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No. 91520202), CAS Scientific Equipment Development Project under Grant YJKYYQ20170050, Youth Innovation Promotion Association CAS and Strategic Priority Research Program of CAS, Beijing Municipal Science&Technology Commission (Z181100008918010).

REFERENCES

[1] S. A. Engel, D. E. Rumelhart, B. A. Wandell, A. T. Lee, G. H. Glover, E.-J. Chichilnisky, and M. N. Shadlen, "fMRI of human visual cortex." *Nature*, 1994.

[2] M. I. Sereno, A. Dale, J. Reppas, K. Kwong, J. Belliveau, T. Brady, B. Rosen, and R. Tootell, "Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging," *Science*, vol. 268, no. 5212, pp. 889–893, 1995.

[3] R. F. Dougherty, V. M. Koch, A. A. Brewer, B. Fischer, J. Modersitzki, and B. A. Wandell, "Visual field representations and locations of visual areas V1/2/3 in human visual cortex," *Journal of vision*, vol. 3, no. 10, pp. 1–1, 2003.

[4] E. A. DeYoe, G. J. Carman, P. Bandettini, S. Glickman, J. Wieser, R. Cox, D. Miller, and J. Neitz, "Mapping striate and extrastriate visual areas in human cerebral cortex," *Proceedings of the National Academy of Sciences*, vol. 93, no. 6, pp. 2382–2386, 1996.

[5] G. M. Holmes, "Ferrier lecture-the organization of the visual cortex in man," *Proceedings of the Royal Society of London. Series B-Biological Sciences*, vol. 132, no. 869, pp. 348–361, 1945.

[6] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, "Identifying natural images from human brain activity," *Nature*, vol. 452, no. 7185, p. 352, 2008.

[7] S. O. Dumoulin and B. A. Wandell, "Population receptive field estimates in human visual cortex." *Neuroimage*, vol. 39, no. 2, pp. 647–660, 2008.

[8] W. Zuiderbaan, B. M. Harvey, and S. O. Dumoulin, "Modeling center–surround configurations in population receptive fields using fMRI," *Journal of vision*, vol. 12, no. 3, pp. 10–10, 2012.

[9] S. Lee, A. Papanikolaou, N. K. Logothetis, S. M. Smirnakis, and G. A. Keliris, "A new method for estimating population receptive field topography in visual cortex." *Neuroimage*, vol. 81, no. 11, pp. 144–157, 2013.

[10] K. N. Kay, J. Winawer, A. Mezer, and B. A. Wandell, "Compressive spatial summation in human visual cortex," *Journal of neurophysiology*, vol. 110, no. 2, pp. 481–494, 2013.

[11] P. Zeidman, E. H. Silson, D. S. Schwarzkopf, C. I. Baker, and W. Penny, "Bayesian population receptive field modelling," *NeuroImage*, vol. 180, pp. 173–187, 2018.

[12] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[13] T. Naselaris, D. E. Stansbury, and J. L. Gallant, "Cortical representation of animate and inanimate objects in complex natural scenes," *Journal of Physiology - Paris*, vol. 106, no. 5-6, pp. 239–249, 2012.

[14] G. St-Yves and T. Naselaris, "The feature-weighted receptive field: an interpretable encoding model for complex feature spaces," *NeuroImage*, vol. 180, pp. 188–202, 2018.

[15] C. Wang, "Variational Bayesian approach to canonical correlation analysis," *IEEE Transactions on Neural Networks*, vol. 18, no. 3, pp. 905–910, 2007.

[16] K. N. Kay, J. Winawer, A. Rokem, A. Mezer, and B. A. Wandell, "A two-stage cascade model of bold responses in human visual cortex," *PLoS computational biology*, vol. 9, no. 5, p. e1003079, 2013.

[17] C. Du, C. Du, L. Huang, and H. He, "Reconstructing perceived images from human brain activities with bayesian deep multi-view learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[18] D. Klindt, A. S. Ecker, T. Euler, and M. Bethge, "Neural system identification for large populations separating "what" and "where"," in *Advances in Neural Information Processing Systems*, 2017, pp. 3506–3516.

[19] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[21] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[22] M. Eickenberg, G. Varoquaux, B. Thirion, and A. Gramfort, "Convolutional network layers map the function of the human visual cortex," *ERCIM NEWS*, no. 108, pp. 12–13, 2017.

[23] N. Kriegeskorte, "Deep neural networks: a new framework for modeling biological vision and brain information processing," *Annual review of vision science*, vol. 1, pp. 417–446, 2015.

[24] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.