A Convolutional Neural Network Interpretable Framework for Human Ventral Visual Pathway Representation

Mufan Xue[1], Xinyu Wu[1], Jinlong Li[1], Xuesong Li[2], Guoyuan Yang[1,3]*

[1]Advanced Research Institute of Multidisciplinary Sciences, Beijing Institute of Technology, Beijing 100081, China
[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[3]School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China
yanggy@bit.edu.cn

Abstract

Recently, convolutional neural networks (CNNs) have become the best quantitative encoding models for capturing neural activity and hierarchical structure in the ventral visual pathway. However, the weak interpretability of these black-box models hinders their ability to reveal visual representational encoding mechanisms. Here, we propose a convolutional neural network interpretable framework (CNN-IF) aimed at providing a transparent interpretable encoding model for the ventral visual pathway. First, we adapt the feature-weighted receptive field framework to train two high-performing ventral visual pathway encoding models using large-scale functional Magnetic Resonance Imaging (fMRI) in both goal-driven and data-driven approaches. We find that network layer-wise predictions align with the functional hierarchy of the ventral visual pathway. Then, we correspond feature units to voxel units in the brain and successfully quantify the alignment between voxel responses and visual concepts. Finally, we conduct Network Dissection along the ventral visual pathway including the fusiform face area (FFA), and discover variations related to the visual concept of 'person'. Our results demonstrate the CNN-IF provides a new perspective for understanding encoding mechanisms in the human ventral visual pathway, and the combination of ante-hoc interpretable structure and post-hoc interpretable approaches can achieve fine-grained voxel-wise correspondence between model and brain. The source code is available at: https://github.com/BIT-YangLab/CNN-IF.

Introduction

The ventral visual pathway is a remarkable feat of nature, capable of processing complex visual stimuli with predominant efficiency and accuracy. Recently, CNNs have emerged as optimal quantitative encoding models for capturing neural activities and hierarchical structures in the ventral visual pathway (Yamins et al. 2014; Kriegeskorte 2015; Schrimpf et al. 2018; Cadena et al. 2019; Schrimpf et al. 2020; Storrs et al. 2021). These models provide a hierarchical correspondence to the early visual cortex (V1-V4) and inferior temporal (IT) (Khosla et al. 2022): early CNN layers predict V1 best, while intermediate and late layers predict V4 and IT best (Yamins et al. 2014; Cichy et al. 2016; Güçlü and van Gerven 2015). Comparable strategies have proven successful in understanding the human auditory cortex (Kell et al. 2018) and motor cortex (Sussillo et al. 2015), highlighting the universality of CNN encoding models (Zhuang et al. 2021; Konkle and Alvarez 2022). Nonetheless, the current understanding of the fundamental mechanisms within the brain and model systems remains incomplete. Unraveling the nature of representational transformations and computations in the ventral visual pathway has long been a vital aim in neuroscience. Importantly, the use of computational models enables the simulation of visual hierarchical processing and facilitates the exploration of hypotheses that may not be readily accessible in the human brain (Beguš, Zhou, and Zhao 2023).

To date, CNNs have been predominantly regarded as black-box models, posing a significant challenge in terms of interpretability. The inherent inability to delve into the internal workings of these models impedes our progress in comprehending the fundamental mechanisms by which they encode visual representations (Ribeiro et al. 2022). Despite ongoing efforts from researchers to enhance the interpretability of CNNs, such as utilizing Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017) and LRP (Layer-wise Relevance Propagation) (Bach et al. 2015), striking a balance between accuracy and interpretability remains a major obstacle. With the emergence of large-scale brain imaging datasets (Chang et al. 2019; Allen et al. 2022), both goal-driven and data-driven approaches have the potential to provide advanced encoding models for the ventral visual pathway (Qiao et al. 2021; Gu et al. 2022). This creates a further appetite for model interpretability. The goal-driven approach is to characterize voxel responses through the feature space trained on high-level tasks, and the data-driven approach is to directly train the model with fMRI data to characterize voxel responses (Cadena et al. 2019; Xiao et al. 2022). It is essential to note that biological visual learning is a process of differentiation, wherein the learning involves discerning differences in existing visual features present in visual inputs rather than constructing new features for each new category (Konkle and Alvarez 2022). This highlights the need to strike a balance between predictive performance and interpretability. Overemphasizing the predictive performance of CNN models may lead to high accuracy in voxel response prediction but often at the cost of understanding

*Corresponding author: Guoyuan Yang

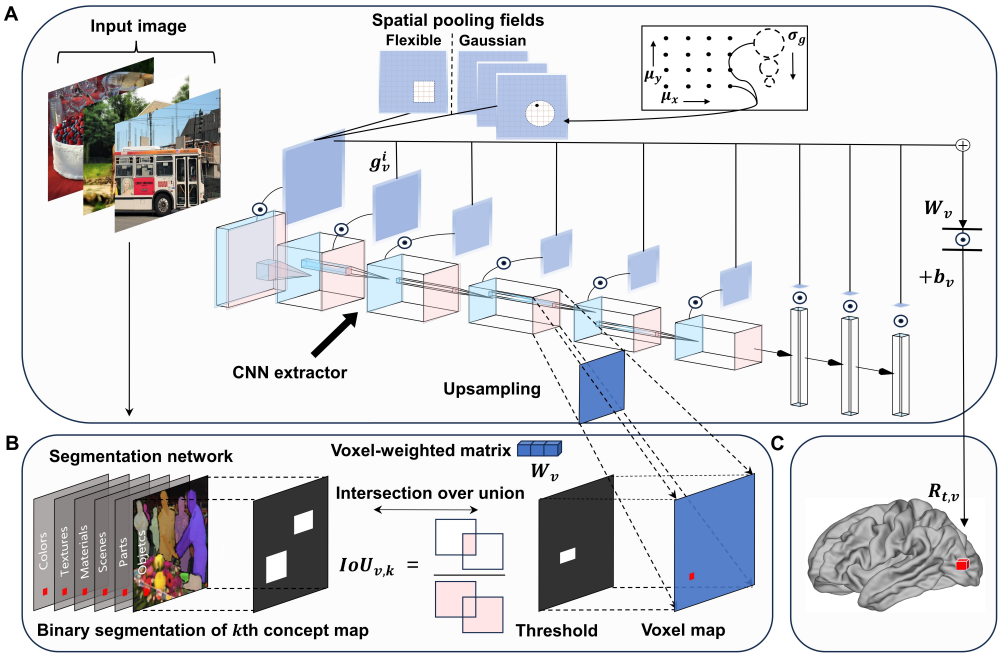

Figure 1: The architecture of CNN-IF. (A) All voxels within the ROI are characterized by a shared CNN extractor. To effectively predict voxel responses for input images, spatial pooling fields, and voxel-weighted matrices are specifically allocated to each voxel. The input images utilized in this study are obtained from the NSD experiment, leveraging the COCO dataset, comprising a total of 73,000 images. (B) The input images are segmented by a binary segmentation network, yielding up to six concept maps per image. The IoU (Intersection over Union) is then calculated between the concept maps and the thresholded voxel maps. (C) For each voxel, the Pearson correlation between the predictive responses and the measured responses is computed on a testing dataset and then corrected by the noise ceiling of that voxel.

their underlying neural processing mechanisms (Cole et al. 2017; Kohoutová et al. 2020). Therefore, it is of utmost importance to give more attention to both goal-driven and data-driven approaches in terms of interpretability.

Here, we propose CNN-IF which is used for the interpretation of CNN models in the human ventral visual pathway. First, we adapt the feature-weighted receptive filed (fwrf) model (St-Yves and Naselaris 2018) to establish an interpretable component for our encoding models. Then, we train encoding models on a large-scale fMRI dataset named the Natural Scenes Dataset (NSD) (Allen et al. 2022). Next, we utilize the voxel-weighted matrix derived from the parameters of the fwrf to correspond feature units to voxel units in the brain and successfully quantify the alignment between voxel responses with visual concepts. Finally, we conduct Network Dissection (Bau et al. 2017, 2020) along the visual ventral pathway and achieve similar results to a previous study (Khosla and Wehbe 2022) on the fusiform face area (FFA) (Kanwisher, McDermott, and Chun 1997), the extrastriate body area (EBA) (Downing et al. 2001), the visual word form area (VWFA) (Cohen et al. 2000), and the retrosplenial cortex (RSC) (Dilks et al. 2013). We demonstrate that our framework achieves fine-grained hierarchical alignment between the model and the brain. Overall, our main contributions are as follows:

- The CNN-IF can provide transparent interpretable encoding models for the ventral visual pathway.
- The CNN layer-wise predictions align with the functional hierarchy of the ventral visual pathway.
- We captured fine-grained hierarchical alignment between voxel units and a set of visual concepts.
- We discovered variations related to the visual concept of 'person' in V1-V2-V3-hV4-FFA brain regions.
- We visualized the spatial pooling field and the activation map to explain this sensational finding.

Related Work

Human Ventral Visual Pathway

The human ventral visual pathway is a major neural pathway in the brain that is responsible for object recognition and visual perception. It is located primarily in the ventral (lower) region of the brain, specifically the temporal lobe. The visual process of object recognition along the ventral visual pathway mainly includes four stages retinal imaging, feature extraction, feature combination, and finally object recognition. Among these brain functional regions, V1 is responsible for capturing information such as edges and curvatures as well as detecting simple features such as shape, color, and position (Hubel and Wiesel 1962; Cadena et al. 2019). V2 which receives most of the input from V1 to extract more complex local features has similar selectivity to direction and spatial scale (Levitt, Kiper, and Movshon 1994; Coggan et al. 2017) and shows stronger selectivity and tolerance to visual texture

(Ziemba et al. 2016). Finally, the high-level regions, including FFA, EBA, VWFA, and RSC, complete the final coding semantic category to achieve object recognition tasks such as faces, scenes, etc (Khosla et al. 2022).

Network Dissection

Network dissection is a technique used in computer vision to understand the inner workings of deep neural networks (Bau et al. 2017). It involves analyzing the intermediate layers of a neural network model to identify and interpret the function of individual units or groups of units. The goal of network dissection is to uncover the semantics and meanings learned by the network for specific tasks. It helps researchers gain insights into what the network has learned and how it has encoded and represented information. Network dissection has been used to analyze and interpret various deep network architectures, such as CNNs for image classification and generative adversarial networks (GANs) for image synthesis (Bau et al. 2020). Network dissection is also used to interpret the tumor segmentation results (Natekar, Kori, and Krishnamurthi 2020). In the human visual system, network dissection is applied to the last layer of their encoding model to demonstrate strong selectivity and functional specialization of the high-level visual areas (Khosla and Wehbe 2022; Sarch et al. 2023), but they don't discover the encoding processing for this selectivity and specialization.

Methods

CNN-IF

The architecture of CNN-IF is shown in Figure 1. We adopt the fwrf model to separate the "where" parameter, which indicates the location of feature pooling, from the "what" parameter, which fine-tunes the feature weights, establishing interpretable components called spatial pooling fields for each voxel of in human brain (Fig. 1A). The size of these spatial pooling fields matches the size of the feature map in each layer of the model. For the goal-driven model, a single isotropic 2D Gaussian pooling field is selected from a predefined set and applied to all feature maps. In contrast, for the data-driven model, an independent and flexible pooling field is applied to each layer of feature maps. Feature maps are grouped based on the model layers and then multiplied pixel-wise by the corresponding spatial pooling field g_v^i that determines the region of visual space that drives voxel response. The weighted pixel values in each feature map are then weighted by the voxel-weighted matrix W_v to yield predictive voxel responses. The CNN extractors remain identical for all voxels across the encoding models, while the spatial pooling fields are optimized and vary across voxels. After training, feature maps of each convolutional layer in the extractor are multiplied by the voxel-weighted matrix to obtain voxel maps. We then perform network dissection with these voxel maps. This allows us to quantify the alignment between a set of semantic concepts with all voxels in a specific region of interest (ROI) (Fig. 1B). Our procedure closely follows the work of (Bau et al. 2017).

Dataset

All encoding models were trained on the NSD[1] (Allen et al. 2022), which consists of individual high-density sampling fMRI data obtained from 8 participants (6 females, aged 19-32 years). During 30-40 sessions of 7T MRI (whole-brain gradient-echo EPI, 1.8 mm isotropic voxels, and 1.6 s TR), each participant viewed 9,000-10,000 different colored natural scenes, with each scene repeated 2-3 times. A special set of 1,000 images was shared across subjects, while the rest were mutually exclusive. The trained model is validated on these 1000 shared pictures to obtain validation accuracy. The images that subjects viewed (3 s on and 1 s off) were from the Microsoft Common Objects in Context (COCO) database (Lin et al. 2014) with a square crop resized to 8.4 × 8.4°.

Regions of Interests

We focus on modeling responses within 8 ROIs in the ventral visual in the study. Four ROIs belonging to the retinotopic early visual cortex, namely, V1, V2, V3, and hV4 are defined using a population receptive field (pRF) localizer scan session, and four higher-level visual ROIs, namely FFA, EBA, RSC, VWFA are manually drawn based on the results of the functional localizer (fLoc) experiment after a liberal thresholding procedure (Allen et al. 2022). We use the cortical flap map resent eight selected ROIs on the ventral visual pathway (Fig. 2C). To better demonstrate the generalization of CNN-IF, we register the ROI of each subject to a common anatomical space (MNI152), and the model predictions are presented in the same manner.

Model Architecture

We employ AlexNet (Krizhevsky, Sutskever, and Hinton 2012) and GNet (Allen et al. 2022; St-Yves et al. 2023) to predict voxel responses. These models possess intricate brain-inspired architectures and provide biologically plausible interpretations, enabling the effective capture of hierarchical representations of visuals in the human brain. AlexNet has previously been shown to deliver state-of-the-art performance in visual response modeling (Güçlü and van Gerven 2015; St-Yves and Naselaris 2018). GNet is a data-driven encoding model that has been shown to train models from scratch and accurately predict voxel responses for V1-V2-V3-hV4. Both AlexNet and GNet consist of a CNN feature extractor and an interpretable fwrf model used to predict the voxel response.

The CNNs utilized in the AlexNet and GNet models are constructed by hierarchically composing functions that process an input image denoted as t:

$$f_l(t) = f_{l-1}(t) \cdot \xi_l$$

where ξ_l is a CNN extractor that operates at layer l on the output of $f_{l-1}(t)$. $f_l(t)$ is the output of layer l and is fed into the next layer as input. The encoding models leverage the intermediate representations $f_l(t)$, which are feature maps

[1] http://naturalscenesdataset.org

with pixels donated by $[f_l(t)]_{k,i,j}$, where (i, j) is the location of the pixel in the kth feature map. The predictive response of voxel v to the input t is expressed by the following formula:

$$\tilde{R}_{t,v,l} = b_v + \sum_k W_{k,v} \cdot \sigma_{k,v,l}(t)$$

where $W_{k,v}$ is the feature weight for voxel v and feature k. The summation $\sum_k W_{k,v}$ of voxel v indicates the voxel-weighted matrix by summing up the weights of all features in encoding models. b_v is a bias item for voxel v.

$$\sigma_{k,v,l}(t) = \sum_{i,j} [f_l(t)]_{k,i,j} \cdot g^l_{v,i,j}$$

where $g^l_{v,i,j}$ indicates the spatial pooling field of voxel v in CNN-IF to reduce the spatial dimensions of feature maps while preserving important features. Important features are located by pixel (i, j). The spatial pooling field of each voxel in different layers is initialized with the same parameters.

Model Training and Testing

We aim to maximize the utilization of data from all eight subjects for model training, a rigorous evaluation resulted in excluding data from subjects 4, 6, 7, and 8 due to significantly lower signal-to-noise ratios, particularly in the higher-level visual brain areas focused on in this study. Finally, we selected data from subjects 1, 2, 3, and 5 for both training and testing. The dataset consists of a total of 37,000 natural scene images, with 1,000 images shared by all subjects, and each subject contributing 9,000 unique images. The model is tested on the 1,000 shared images, while the remaining 36,000 images were split into a training set (90%) and a validation set (10%). To fully exploit the advantages of the NSD dataset, we jointly optimized our CNN extractor using data from the four subjects. Specifically, for the goal-driven encoding model, the CNN extractor parameters were pre-trained based on object classification in the ImageNet database (Deng et al. 2009). As for the data-driven encoding model, the CNN extractor parameters, spatial pooling fields, and feature weights were all optimized using stochastic gradient descent and an L_2-norm weighted loss function:

$$Loss(\tilde{R}, R) = \sum_{t \in Batch} \sum_s \sum_l \sum_t (\tilde{R}_{t,s,v,l} - R_{t,s,v})^2$$

where t, s denote the image t presented to subject s, $\tilde{R}_{t,s,v,l}$ denotes the predictive response of model layer l for stimulu t received by voxel v of subject s. $R_{t,s,v}$ denotes the measured response of voxel v of subject s to stimulu t.

We quantified the predictive accuracy of the model by calculating the Pearson correlation coefficient between the predictive responses of each voxel and the measured response and then compared the predictive accuracy with the noise ceiling (Fig. 2B). We employed the ADAM optimizer (Kingma and Ba 2014) with a learning rate of $1e$-3, β_1=0.9, β_2=0.999, 50 epochs, and a batch size of 50 for training. Additionally, in order to promote stability during the training process, parameter updates were alternated between feature extractors, spatial pooling fields, and feature weights.

Experiments

In the following experiment, we first trained interpretable encoding models AlexNet and GNet in goal-driven and data-driven approaches on the NSD dataset. We found differences in training methods and predictive performance between the two models. Then, we evaluated the predictive accuracy of each layer of the model and found that network layer-wise predictions align with the functional hierarchy of the ventral visual pathway. Next, we used the voxel-weighted matrix to correspond feature units to voxel units and successfully quantified the alignment between voxel responses with a set of visual concepts by network dissection. Finally, we performed network dissection along the ventral visual pathway and visualized spatial pooling fields and activation maps, explaining variations related to the visual concept of 'person' in V1-V2-V3-hV4-FFA ROIs.

Interpretable Encoding Models for the Ventral Visual Pathway

To validate the effectiveness of our CNN-IF, we carefully selected eight ROIs (V1, V2, V3, hV4, FFA, EBA, RSC, VWFA) along the ventral visual pathway with hierarchical relationships (Fig. 2C). Voxel responses corresponding to these ROIs were extracted from the NSD dataset and utilized for training. To assess the generalizability of the CNN-IF, we employed both goal-driven and data-driven approaches to train the AlexNet and GNet encoding models. In the data-driven approach, we further partitioned the model by initializing the CNN extractor with the identical parameters utilized in our 'goal-driven-pretrained encoding model', which we referred to as the 'data-driven-pretrained encoding model'. Additionally, we conducted training from scratch with the random initialization, denoting this particular variant as the 'data-driven-unpretrained encoding model'. Detailed information about the construction and training of the models can be found in the method section.

When only a small amount of data is available, we found that the goal-driven encoding model exhibits significantly higher predictive accuracy compared to the data-driven encoding model. However, as we further increase the amount of data, the predictive accuracy of the models eventually levels off. The difference in predictive accuracy between the two approaches narrows. This suggests that data-driven approaches demonstrate impressive performance improvements when there is a large amount of available data, approaching the performance of the goal-driven encoding model, particularly evident on GNet (Fig. 2A, 2E). Providing the model with pretraining parameters does indeed improve the predictive accuracy, which is more pronounced in the case of AlexNet. To further understand the performance of the GNet model, we compared its predictive performance with the noise ceiling estimate (Fig. 2B). Throughout the voxels, the predictive accuracy is closely related to the noise ceiling, indicating that voxel differences in predictive accuracy simply reflect differences in signal-to-noise ratio (SNR). Additionally, the predictive accuracy approaches but does not reach the noise ceiling. Next, the cortical flapmap reveals voxel-wise predictive performance (Fig.

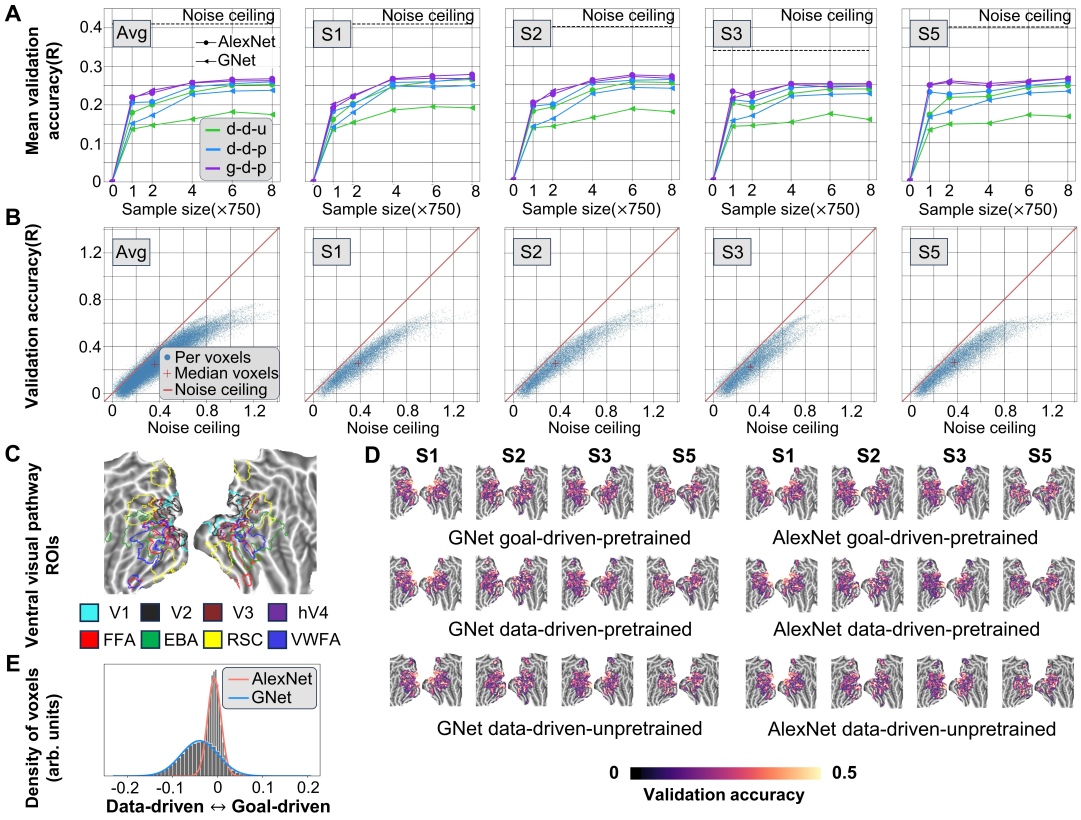

Figure 2: Prediction of voxel responses in the ventral visual pathway. (A) The results of the change in the predictive accuracy of the training results of six models with the amount of training data. The validation accuracy is estimated as the Pearson correlation coefficient between measured voxel responses and predictive responses on the testing dataset. (B) The distribution of prediction accuracy per voxel relative to the corresponding noise ceiling shows voxel differences in predictive accuracy simply reflect differences in SNR. (C) Illustration of ROIs of the ventral visual pathway for encoding models. (D) The cortical flat map demonstrates the achieved predictive accuracy of our models across all voxels in the eight ROIs, revealing high accuracy across extensive regions within these ROIs. (E) The distribution of voxel-wise differences in predictive accuracy between goal-driven and data-driven approaches shows that pretrained parameters contribute to an increase in predictive accuracy. d-d-u, data-driven-unpretrained; d-d-p, data-driven-pretrained; g-d-p, goal-driven-pretrained.

2D). The predictive performance of the early visual cortex is higher than the predictive accuracy of the floc ROIs. This is due to the higher SNR in the primary visual cortex.

Fine-grained Hierarchical Alignment between Model and the Ventral Visual Pathway

Layer-wise hierarchy To evaluate the alignment between network layer-wise predictions and the functional hierarchy of the ventral visual pathway, we divided the eight ROIs into the hierarchy of early visual cortex (V1, V2, V3, hV4) and floc ROIs (FFA, EBA, RSC, VWFA) (Fig. 3A). Then, we quantified the correlation between predictive responses of all goal-driven encoding model layers of AlexNet and GNet with measured responses of the ventral visual pathway hierarchy to obtain the AlexNet hierarchy (Fig. 3B) and GNet hierarchy (Fig. 3C). The results of the correlation of data-driven encoding models are similar to goal-driven encoding models, which are provided in the Appendix[2]. Results from both models consistently demonstrate a hierarchical alignment between the model and the brain. Specifically, the early layers of the model exhibit the strongest predictive performance for the early visual cortex, whereas the intermediate and late layers of the model exhibit the strongest predictive performance for the floc ROIs.

Fine-grained voxel-wise hierarchy To further obtain fine-grained hierarchical alignment, we first used the voxel-weighted matrix to correspond feature units to voxel units, quantifying the alignment between voxel responses and a set of visual concepts. Then, we performed network dissection for each of the four floc ROIs (FFA, EBA, RSC, VWFA). The alignment between each concept map and individual voxel map is quantified by the Intersection over Union (IoU), which is computed on Broden, a broadly and densely labeled dataset (Bau et al. 2017). The units with semantics are given labels across a range of objects, parts, scenes, and materials. We chose an IoU threshold of 0.04 and a voxel map activation threshold of 0.01 to quantify the

[2]https://github.com/BIT-YangLab/CNN-IF

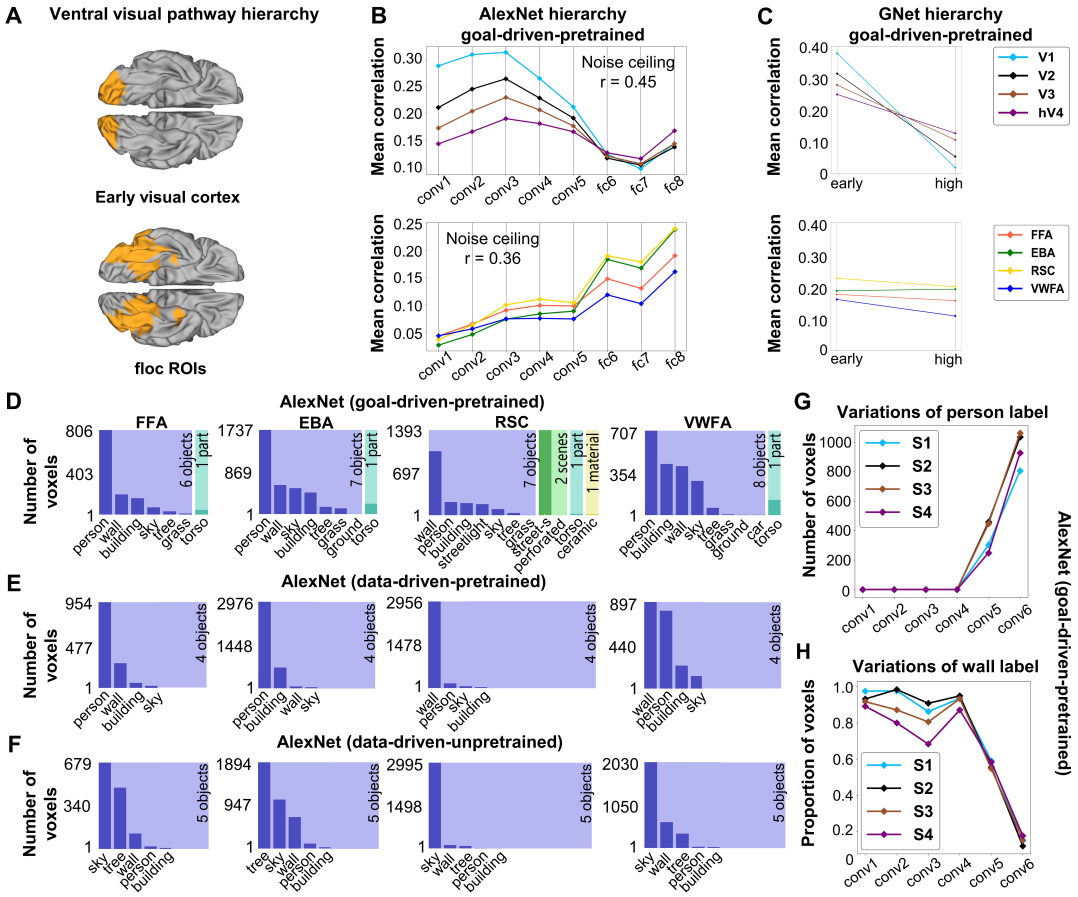

Figure 3: Alignment between model and brain. (A) Visualization of the hierarchical structure of the ventral visual pathway, proceeding from posterior to anterior regions (including the early visual cortex and floc ROIs). (B) and (C) show the correlation between predictive voxel responses and measured voxel responses for each ROI from all goal-driven encoding model layers. The results were averaged across four subjects. (D) The number of voxels that are detected on AlexNet (goal-driven-pretrained). (E) The number of voxels that are detected on AlexNet (data-driven-pretrained). (F) The number of voxels that are detected on AlexNet (data-driven-unpretrained). Floc ROIs are aligned with the last convolutional layer of AlexNet for network dissection. The number of voxels reflects the highest alignment (IoU $>$ 0.04) of all voxels within a certain ROI with a set of visual concepts. (G) Variations in the number of voxels aligned with 'person' are detected by different model layers. (H) Variations in the proportion of voxels aligned with the 'wall' are detected by different model layers.

interpretability of a layer. Specifically, if $IoU_{v,k}$ calculated by voxel map for voxel v and concept map k for concept c exceed 0.4, we consider voxel v to have successfully detected representations encoded by the model feature space for the input image. To quantify the interpretability of a layer, we only select the concept with the highest IoU score per voxel and count the number of unique concepts aligned with voxel units. FFA, EBA, and VWFA all show selectivity for the concept of 'person' (Fig. 3D), but FFA is more concentrated on the face (Fig. 4A), EBA is more concentrated on the body parts (Fig. 4B), and VWFA is not only highly selective for 'word' but also for 'person'. This is due to the local anatomical overlap of FFA, EBA, and VWFA. Due to the lack of a Word-related label in Broden, this concept was detected as 'building', which can be seen in the activation map visualized in Fig. 4D. RSC shows selectivity for the concepts of 'wall' and 'street-s' (Fig. 3D), which can be seen in the activation map visualized in Fig. 4C. We achieved similar results to a previous study by (Khosla and Wehbe 2022). Additionally, we found that the AlexNet (goal-driven-pretrained) model (Fig. 3D) detected more concepts than the AlexNet (data-driven-pretrained) model (Fig. 3E), and this alignment was more difficult to capture in the AlexNet (data-driven unpretrained) model (Fig. 3F). The results of the GNet network dissection are in the Appendix[2].

Variations of the 'person' concept in V1-V2-V3-hV4-FFA ROIs Our model successfully simulated the face recognition function of human high-level visual areas. To further explore the variations of this process along the ventral visual pathway, we performed network dissection along the ventral visual pathway (V1-V2-V3-hV4-FFA ROIs) to seek variations related to the visual concept of 'person' that occur at the functional hierarchy learned by AlexNet (goal-

Figure 4: Activation map of AlexNet (goal-driven-pretrained). For each region of floc ROIs, we took the top five activation maps with the highest activation from the top five voxels with the highest IoU score (top 1% quantile level). (A) Voxels in FFA are aligned with 'person'. (B) Voxels in EBA are aligned with 'person' but more focused on the body. (C) Voxels in RSC are aligned with 'wall' and 'street-s'. (D) Voxels in VWFA are aligned with 'person' and 'building' (the concept of 'word' is hidden in the label of 'building'). (E) The activation of two images at each layer of AlexNet (goal-driven-pretrained) and they are aligned with 'person' only from conv5 and conv6.

driven-pretrained) models. Our results show that the brain encodes the visual concept of 'person' along the ventral visual pathway and ultimately forms the representation of 'person' in the FFA in an unprecedented way (Fig. 3G). We also counted the proportion of voxels aligned with the 'wall' concept in the corresponding ROI in this hierarchy (Fig. 3H). According to the concept map of the 'wall', it contains some basic semantic information such as color and text, which is easier to detect in the early layer of the model. As expected, when a high-level concept contains more low-level semantic information, this problem may more easily occur. The inability of early layers to detect the 'person' label may be because these voxels are involved in encoding other concepts.

Although V1-hV4 (corresponding to conv1-conv4 of the model) didn't merge unique voxel units aligned with the 'person' concept, these ROI voxels are still involved in encoding representation that contains the semantic of 'person' when we paid attention to two images that contain 'person' (Fig. 4E). Conv1 and conv2 activate only the regions around the person, indicating a focus on some basic semantic information such as color and texture. Since V3, although voxels corresponding to 'person' has not been detected at this time, the activation map of conv4 shows that voxels in V3 and hV4 are sensitive to the circle (the face is also round). By comparing conv5 and conv6 (where voxel is starting to appear that can detect 'person') with other earlier layers, we can find that conv5 and conv6 have a larger whole of active areas, which can include all the information related to a 'person', while earlier layers are still a few scattered activated areas. This is the reason why we find that representations about the 'person' are ultimately formed at FFA rather than at the early visual cortex.

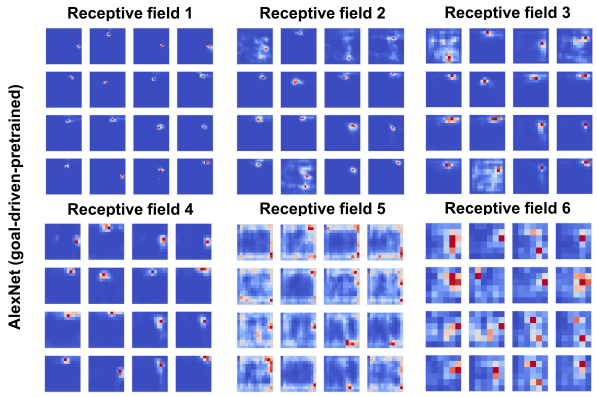

Figure 5: Visualization of the spatial pooling field of AlexNet (goal-driven-pretrained). We fit the corresponding spatial pooling field for each voxel in each layer of the model to visualize this interpretable component.

Finally, we exhibited the variations of interpretable spatial pooling fields of the AlexNet (goal-driven-pretrained) model. The parameters of spatial pooling fields reflect where the corresponding model layer should focus the most when predicting voxel responses (Fig. 5). All receptive fields are initialized in the same way. As the model layer deepens, the activation region of the receptive field expands, however, the size of the activated area decreases. This observation suggests that the spatial pooling field prioritizes local pivotal information while fitting voxel responses.

Conclusion

We propose the CNN-IF which provides an interpretable encoding model for the human ventral visual pathway and well balances the predictive performance and interpretability of the model. By exploiting the interpretable hierarchical fwrf model of two high-performing encoding models, including AlexNet and GNet, we discover that the network layer-wise predictions align with the functional hierarchy of the ventral visual pathway. Using network dissection, we quantify the alignment between voxel responses and a set of visual concepts. The results show variations related to the visual concept of 'person' in the high-level visual area corresponding to higher layers of the model. Finally, we exhibit the spatial pooling field and the activation map to explain this sensational finding. We demonstrate that CNN-IF provides a new perspective on the interpretability of the CNN model for understanding encoding mechanisms in the human ventral visual pathway and achieving fine-grained hierarchical alignment between the model and the ventral visual pathway.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grants number 82302175, 62071049 and 62336002); the National Science and Technology Innovation 2030 Program (grant number 2021ZD0200500); the Beijing Municipal Science and Technology Commission (grants number Z171100000117012 and Z181100001518003); the Beijing Municipal Natural Science Foundation Project (grant number 4222018); the China Postdoctoral Science Foundation (grant number 2021M700015).

References

Allen, E. J.; St-Yves, G.; Wu, Y.; Breedlove, J. L.; Prince, J. S.; Dowdle, L. T.; Nau, M.; Caron, B.; Pestilli, F.; Charest, I.; et al. 2022. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1): 116–126.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

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

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48): 30071–30078.

Beguš, G.; Zhou, A.; and Zhao, T. C. 2023. Encoding of speech in convolutional layers and the brain stem based on language experience. *Scientific Reports*, 13(1): 6480.

Cadena, S. A.; Denfield, G. H.; Walker, E. Y.; Gatys, L. A.; Tolias, A. S.; Bethge, M.; and Ecker, A. S. 2019. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS computational biology*, 15(4): e1006897.

Chang, N.; Pyles, J. A.; Marcus, A.; Gupta, A.; Tarr, M. J.; and Aminoff, E. M. 2019. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Scientific data*, 6(1): 49.

Cichy, R. M.; Khosla, A.; Pantazis, D.; Torralba, A.; and Oliva, A. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1): 27755.

Coggan, D. D.; Allen, L. A.; Farrar, O. R.; Gouws, A. D.; Morland, A. B.; Baker, D. H.; and Andrews, T. J. 2017. Differences in selectivity to natural images in early visual areas (V1–V3). *Scientific Reports*, 7(1): 2444.

Cohen, L.; Dehaene, S.; Naccache, L.; Lehéricy, S.; Dehaene-Lambertz, G.; Hénaff, M.-A.; and Michel, F. 2000. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2): 291–307.

Cole, J. H.; Poudel, R. P.; Tsagkrasoulis, D.; Caan, M. W.; Steves, C.; Spector, T. D.; and Montana, G. 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163: 115–124.

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

Dilks, D. D.; Julian, J. B.; Paunov, A. M.; and Kanwisher, N. 2013. The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, 33(4): 1331–1336.

Downing, P. E.; Jiang, Y.; Shuman, M.; and Kanwisher, N. 2001. A cortical area selective for visual processing of the human body. *Science*, 293(5539): 2470–2473.

Gu, Z.; Jamison, K.; Sabuncu, M.; and Kuceyeski, A. 2022. Personalized visual encoding model construction with small data. *Communications Biology*, 5(1): 1382.

Güçlü, U.; and van Gerven, M. A. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27): 10005–10014.

Hubel, D. H.; and Wiesel, T. N. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106.

Kanwisher, N.; McDermott, J.; and Chun, M. M. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311.

Kell, A. J.; Yamins, D. L.; Shook, E. N.; Norman-Haignere, S. V.; and McDermott, J. H. 2018. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3): 630–644.

Khosla, M.; Jamison, K.; Kuceyeski, A.; and Sabuncu, M. 2022. Characterizing the ventral visual stream with response-optimized neural encoding models. *Advances in Neural Information Processing Systems*, 35: 9389–9402.

Khosla, M.; and Wehbe, L. 2022. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, 2022–03.

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

Kohoutová, L.; Heo, J.; Cha, S.; Lee, S.; Moon, T.; Wager, T. D.; and Woo, C.-W. 2020. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature protocols*, 15(4): 1399–1435.

Konkle, T.; and Alvarez, G. A. 2022. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1): 491.

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

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

Levitt, J. B.; Kiper, D. C.; and Movshon, J. A. 1994. Receptive fields and functional architecture of macaque V2. *Journal of neurophysiology*, 71(6): 2517–2542.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Natekar, P.; Kori, A.; and Krishnamurthi, G. 2020. Demystifying brain tumor segmentation networks: interpretability and uncertainty analysis. *Frontiers in computational neuroscience*, 14: 6.

Qiao, K.; Zhang, C.; Chen, J.; Wang, L.; Tong, L.; and Yan, B. 2021. Effective and efficient roi-wise visual encoding using an end-to-end cnn regression model and selective optimization. In *Human Brain and Artificial Intelligence: Second International Workshop, HBAI 2020, Held in Conjunction with IJCAI-PRICAI 2020, Yokohama, Japan, January 7, 2021, Revised Selected Papers 2*, 72–86. Springer.

Ribeiro, F. L.; Bollmann, S.; Cunnington, R.; and Puckett, A. M. 2022. An explainability framework for cortical surface-based deep learning. *arXiv preprint arXiv:2203.08312*.

Sarch, G. H.; Tarr, M. J.; Fragkiadaki, K.; and Wehbe, L. 2023. Brain Dissection: fMRI-trained networks reveal spatial selectivity in the processing of natural images. *bioRxiv*, 2023–05.

Schrimpf, M.; Kubilius, J.; Hong, H.; Majaj, N. J.; Rajalingham, R.; Issa, E. B.; Kar, K.; Bashivan, P.; Prescott-Roy, J.; Geiger, F.; et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007.

Schrimpf, M.; Kubilius, J.; Lee, M. J.; Murty, N. A. R.; Ajemian, R.; and DiCarlo, J. J. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3): 413–423.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

St-Yves, G.; Allen, E. J.; Wu, Y.; Kay, K.; and Naselaris, T. 2023. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nature communications*, 14(1): 3329.

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

Storrs, K. R.; Kietzmann, T. C.; Walther, A.; Mehrer, J.; and Kriegeskorte, N. 2021. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10): 2044–2064.

Sussillo, D.; Churchland, M. M.; Kaufman, M. T.; and Shenoy, K. V. 2015. A neural network that finds a naturalistic solution for the production of muscle activity. *Nature neuroscience*, 18(7): 1025–1033.

Xiao, W.; Li, J.; Zhang, C.; Wang, L.; Chen, P.; Yu, Z.; Tong, L.; and Yan, B. 2022. High-Level visual encoding model framework with hierarchical ventral stream-optimized neural networks. *Brain Sciences*, 12(8): 1101.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and 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.

Zhuang, C.; Yan, S.; Nayebi, A.; Schrimpf, M.; Frank, M. C.; DiCarlo, J. J.; and Yamins, D. L. 2021. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3): e2014196118.

Ziemba, C. M.; Freeman, J.; Movshon, J. A.; and Simoncelli, E. P. 2016. Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences*, 113(22): E3140–E3149.