Front. Eng. Manag. 2021, 8(4): 572–581
https://doi.org/10.1007/s42524-021-0169-x

RESEARCH ARTICLE

Dongli DUAN, Xixi WU, Shubin SI

Novel interpretable mechanism of neural networks based on network decoupling method

Abstract The lack of interpretability of the neural network algorithm has become the bottleneck of its wide application. We propose a general mathematical framework, which couples the complex structure of the system with the nonlinear activation function to explore the decoupled dimension reduction method of high-dimensional system and reveal the calculation mechanism of the neural network. We apply our framework to some network models and a real system of the whole neuron map of Caenorhabditis elegans. Result shows that a simple linear mapping relationship exists between network structure and network behavior in the neural network with high-dimensional and nonlinear characteristics. Our simulation and theoretical results fully demonstrate this interesting phenomenon. Our new interpretation mechanism provides not only the potential mathematical calculation principle of neural network but also an effective way to accurately match and predict human brain or animal activities, which can further expand and enrich the interpretable mechanism of artificial neural network in the future.

Keywords neural networks, interpretability, dynamical behavior, network decouple

Received March 29, 2021; accepted June 9, 2021

Dongli DUAN, Xixi WU
School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710311, China

Shubin SI (✉)
Ministry of Industry and Information Technology Key Laboratory of Industrial Engineering and Intelligent Manufacturing, Northwestern Polytechnical University, Xi'an 710072, China; School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China
E-mail: sisb@nwpu.edu.cn

This work was supported by the National Natural Science Foundation of China (Grant Nos. 72071153, 71631001, and 71771186), the Natural Science Foundation of Shaanxi Province (Project No. 2020JM-486), and the Fund of the Key Laboratory of Equipment Integrated Support Technology (Project No. 6142003190102).

1 Introduction

1.1 Background

Neural network is a complex system. This network is nonlinear because it is formed by a large number of processing units, which are neurons or similar to neurons. The network has a nonlinear parallel structure to complete the information processing function by simulating the way of the neurons in human and animal brain processing and memorizing information. Hence, the mechanism within this work is proposed on the basis of the research results of modern neuroscience. The operation mode of the neural network includes two types, namely, feedforward and feedback. The feedforward network adopts a hierarchical network structure and realizes the nonlinear mapping from the state space of the input layer to the state space of the output layer. This network is widely used in pattern classification and feature extraction. The feedback network takes the form of an interconnected network structure, where the node is not only an input but also a calculation unit and an output, which is used for optimization calculation and associative memory. The artificial neural network is abstracted from the human brain network; however, it is quite different from the real human brain. Some big blind spots still persist in the academical field about the operating principle of the human brain. Therefore, the artificial neural network simulates each neuron node through a number of computers to form an array by calculating the mathematical functions to allocate the weight of each computer and achieve the effect of parallel computing for greatly enhancing the processing capacity of the computer system.

However, the neural network is considered to be a "black box". The reason for this problem is not just the difficulty in understanding the neural network features and its decision logic at the semantic level, which may be caused by its super large scale. Another important reason is the lack of mathematical tools to diagnose and evaluate the feature expression ability of the network (for example, to

explain the amount of knowledge modeled by the depth model, its generalization ability, and convergence speed) and to explain the information processing characteristics of different neural network models.

In recent years, the artificial intelligence technology based on big data and deep learning has made great strides in many application fields, such as image recognition, speech recognition, and natural language processing. However, the opacity and incomprehensibility of the neural networks lead them to slow down their theoretical development and practical application. The performance of the intelligent system significantly decreases with the changes in face environment, incomplete information, interference, or false information. This situation may cause catastrophic consequences at this time if users follow the results of the system unconditionally. In the military, medical, financial, transportation, and other high-risk decision-making field, users often need the intelligent system to provide the basic information to make the final evaluation and judgment on the decision of the intelligent system, such as rejecting, accepting, or adjusting the results. Therefore, the main goals of the future development of deep learning are to understand and interpret the neural network. These initiatives will change the human understanding of intelligence from "only knowing what it is" to "knowing why it is", which enable the system to move from perception to cognition, thus promoting the development of deep network theory and application.

1.2 Literature review

The feature of black box problem for neural network has attracted extensive research interest from academia, industry, and even military scholars, which enable the interpretable mechanism to be one of the difficult points and hot spots in current artificial intelligence research. At present, the interpretation of neural network mainly focuses on different network structures, including convolutional neural networks (CNN), recurrent neural networks (RNN), and general adverse networks (GAN). The methods can be classified as visual interpretation, text interpretation, and multimodal interpretation.

CNN is composed of one or more convolution layer, pooling layer, and top full connection layer. This network is widely used in image recognition, speech recognition, semantic recognition, automatic driving, and other systems. Based on the hidden layer representation method, Zhou et al. (2019) output or reconstruct the feature map learned by the filter to the original image with a Network Dissection approach, which is suitable for network structure from the convolutional feature graph to global average pooling layer and to softmax layer. The basic principle of sensitivity analytical method is to remove or change the value of components (such as pixels or semantic segmentation area in the image) and observe the change of decision results. If the decision results constantly change, then it corresponds to important components (Bach et al., 2015; Samek et al., 2017). Based on the analysis of linear models, the PatternNet and PatternAttribution methods proposed by Kindermans et al. (2017) improve the theory of linear classification model, in which only one step of back propagation calculation is needed. Decision tree is another method to imitate deep neural network. Zharov et al. (2018) extract the network by a series of hierarchical gradient enhanced decision trees and encode the neuron activation sequence with the index of leaf nodes. Wu et al. (2017) use the decision tree model to analyze the local behavior of the deep neural network and train a network that is easier to understand through the tree normalization method. The text interpretation method is used to generate natural language description in the form of text for the model decision results to explain the basis of decision, such as the characteristics contained in the target. Hendricks et al. (2016) propose a loss function based on sampling and reinforcement learning, which learn to generate sentences with global attributes. A multimodal interpretation method generates not only the text interpretation for model decision-making but also the corresponding visual evidence area on the image. PJ-X model uses attention mechanism to construct a model for generating visual interpretation of decision-making results and produces natural language interpretation of text description for the output results, which is verified on visual problem and activity recognition tasks (Park et al., 2018). In summary, many challenges still persist in understanding and interpreting CNN in the complex multi-layer nonlinear network structure even though CNN has high-precision prediction performance. On the one hand, the semantics of hidden neurons in CNN is difficult to determine and limit. On the other hand, the decision-making process is hard to comprehensively interpret from visual semantics and text.

RNN is a type of neural network used to process sequence data. The common RNN networks include long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), bidirectional LSTM (BLSTM) (Schuster and Paliwal, 1997), and gated recurrent unit (GRU) (Cho et al., 2014), which are widely used in semantic analysis, emotion analysis, image text annotation, and language translation. Kádár et al. (2017) prove that RNN learning carries semantic information of lexical categories and grammatical functions. LSTMV is a visual analytical tool for RNNs that focuses on understanding the dynamic process of the hidden states (Strobelt et al., 2018). Li et al. (2015) find important words based on the significance of gradient and uses traditional methods and simple strategies to evaluate the contribution of neurons for semantic construction. Tang et al. (2017) propose to use visual storage vectors to understand the behavior of LSTM and GRU in speech recognition tasks. RNNV visualizes the hidden state based on the expected response of RNN to the input. When a negative/positive input is present, a large

number of negative/positive hidden cell clusters will be activated.

GAN is a generative model (Goodfellow et al., 2014) that is composed of a generator and a discriminator. This model is trained by the way of confrontation learning. The purpose is to estimate the potential distribution of data samples and generate new data samples. In recent years, various GAN models have been proposed, such as conditional GAN (CGAN) (Mirza and Osindero, 2014), deep convolutional GAN (DCGAN) (Radford et al., 2015), InfoGAN (Chen et al., 2016), StackGAN (Zhang et al., 2017), Wasserstein GAN (WGAN) (Arjovsky et al., 2017), sequence GAN (SeqGAN) (Yu et al., 2016), and plug & play GAN (PPGAN) (Nguyen et al., 2017), to improve the model structure. GAN has broad application prospects in various fields, such as image and visual computing, speech and language processing, information security, and chess games. Radford et al. (2015) use the method of manipulating potential vectors and observing the manner to which the results change accordingly to understand and explain the process of GAN. Zhang et al. (2017) use the knowledge map to explain concepts, entities, and word bags of neural networks.

In conclusion, the above research rarely focuses on the correlation between nodes of the neural network, especially the dynamic characteristics of neuron nodes. We compared the current common visual interpretation and text interpretation methods and found that it is more complete and profound to couple the activation function and network structure together and explore the computing architecture of various neural networks. This mechanism is a more universal framework to shed light on the interpretable mechanism of neural networks.

1.3 Motivation

The forms of the activation function are no longer limited to linear functions to enhance the generalized performance of neural networks. Some complex nonlinear functions, such as sigmoid, tanh, softmax, and relu, are taken to greatly improve the representation ability. Meanwhile, the interpretability of the algorithm becomes worse with the increase in model complexity. From the perspective of network science, we reveal the dynamical behavior of neural networks with the coupling of network structure and activation function, which can help explore the general calculation principle of the neural network in a universal framework. Such data can overcome the inherent shortcomings of the above interpretation methods, which attempts to revise and explain the black box calculation method of the neural network from the perspective of phenomenon and result.

The rest of this paper is organized as follows. Section 2 describes the activation function of neural networks and its decoupling method. Section 3 introduces the interpretable mechanism of neural networks. Section 4 compares the simulation results and the theoretical solution of neural network with scale-free network model and then explores the correlation of the structure and the behavior of Caenorhabditis elegans (hereinafter referred to as C. elegans). Section 5 provides the conclusions and future work of this paper.

2 Activation function of neural networks and its decoupling method

Positive and negative nonlinear regulation relationships exist between each node or synapse in the artificial neural network (ANN) or human brain neural network. Taking ANN for example, if the activation function is not used, then the output of each layer would be a linear function of the input of the upper layer. This notion means that the output is a linear combination of the inputs regardless of the number of layers the neural network has. The hyperbolic tangent function, a typical nonlinear function, gradually replaces the sigmoid function as the standard activation function, because it has many good characteristics for deep neural networks, such as it is completely differentiable and antisymmetric. Some smoother variants of this function can be used to solve the problem of slow learning and/or gradient disappearance. This work takes hyperbolic tangent excitation function as an example. We provide a neural network decoupling method with a general network structure. The two regulatory relationships of excitation and inhibition between trillions of neuron synapses in the human or animal brain would allow the individuals to show a much more complex nonlinear phenomenon of affine transformation. The activation function can be used to describe the coupling mechanism among synapses as follows:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = I - \frac{x_i}{R} + \frac{J_1}{2}\sum_{j=1}^{N}\hat{A}_{ij}(1+\tanh(n(x_j-a))) + \frac{J_2}{2}\sum_{j=1}^{N}\overline{A}_{ij}(1-\tanh(n(x_j-a))), \quad (1)$$

where x_i represents the activity of neuron i $(i = 1, 2, ...)$; (i, j) is a connectome linked neuron i with j; a is the firing threshold; n describes the slope of the sigmoid function; I is the basal activity of the neuron; R represents the inverse of the death rate of the neuron; J_1 and J_2 denote the excitation and inhibition strength between neurons, respectively; and $\tanh(x)$ is a hyperbolic tangent function, which can be explained as follows:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{2}{1+e^{-2x}} - 1. \quad (2)$$

The third and the fourth terms of Eq. (1) represent the exciting and inhibitory relationships between neurons. **A** is

the adjacency matrix of a network that describes the interactions between pair of components,

$$\mathbf{A} = \hat{\mathbf{A}} + \overline{\mathbf{A}}, \tag{3}$$

where $\hat{\mathbf{A}}$ and $\overline{\mathbf{A}}$ represent the active and suppressive adjacency matrix of the network, respectively, which are separated from the adjacency matrix $\mathbf{A}$. Variables $\hat{\mathbf{A}}$ and $\overline{\mathbf{A}}$ are nonnegative matrixes that capture the interactions between neurons. The fraction of the numbers of nonzero elements in $\hat{\mathbf{A}}$ and $\overline{\mathbf{A}}$ are $1-p=\frac{|\hat{\mathbf{A}}|}{|\mathbf{A}|}$ and $p=\frac{|\overline{\mathbf{A}}|}{|\mathbf{A}|}$, respectively, where p is the fraction of inhibiting links in the network. We denote $s_i^{\text{in}}=\sum_{j=1}^{N}A_{ij}$, $s_j^{\text{out}}=\sum_{i=1}^{N}A_{ij}$, $s_i^{-\text{in}}=\sum_{j=1}^{N}\overline{A}_{ij}$, $s_j^{-\text{out}}=\sum_{i=1}^{N}\overline{A}_{ij}$, $s_i^{+\text{in}}=\sum_{j=1}^{N}\hat{A}_{ij}$, and $s_j^{+\text{out}}=\sum_{i=1}^{N}\hat{A}_{ij}$. Then, we obtain the indegree and outdegree sequences of the coupled network for $\mathbf{A}$ as $\mathbf{s}^{\text{in}}=(s_1^{\text{in}}, s_2^{\text{in}}, ..., s_N^{\text{in}})$ and $\mathbf{s}^{\text{out}}=(s_1^{\text{out}}, s_2^{\text{out}}, ..., s_N^{\text{out}})$, the indegree and outdegree sequences of the suppressive network for $\overline{\mathbf{A}}$ as $\mathbf{s}^{-\text{in}}$ and $\mathbf{s}^{-\text{out}}$, the indegree and outdegree sequences of the active network for $\hat{\mathbf{A}}$ as $\mathbf{s}^{+\text{in}}$ and $\mathbf{s}^{+\text{out}}$, respectively. The equations of (1)–(3) represent the dynamical models when x_i is given the corresponding actual meaning.

When the system of Eq. (1) tends to be in a relatively stable state, which means that the activities of the N individuals in the network have a constant value, the N-dimensional nonlinear rate equations all equal to zero. However, calculating the stable state of the system requires the numerical or analytical computation of the N-dimensional nonlinear rate equations, which could be difficult or almost impossible, especially when the system is large-scale. We assume that the influence of the neuron j on i is equivalent to the effect of the average network activity on i. Specifically, we assume that the activity of each neuron in the system is uniformly distributed. Thus, the system of Eq. (1) can be decoupled as follows:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=I-\frac{x_i}{R}+\frac{J_1 s_i^{+}}{1+e^{(-2n(\langle x\rangle-a))}}+\frac{J_2 s_i^{-}e^{(-2n(\langle x\rangle-a))}}{1+e^{(-2n(\langle x\rangle-a))}}, \tag{4}$$

where s_i^{+}/s_i^{-} is the number of exciting/inhibiting links of neuron i, and $\langle x\rangle$ is the average activity of the neural network. Then we denote that

$$f(x_i, s_i^{+}, s_i^{-})$$

$$=I-\frac{x_i}{R}+\frac{J_1 s_i^{+}}{1+e^{(-2n(\langle x\rangle-a))}}+\frac{J_2 s_i^{-}e^{(-2n(\langle x\rangle-a))}}{1+e^{(-2n(\langle x\rangle-a))}}. \tag{5}$$

When the system reaches a stable state, the following condition should be satisfied to ensure its linear stability

$$f(x_i, s_i^{+}, s_i^{-})=0. \tag{6}$$

From Eq. (6), the steady activity of neuron i is

$$x_i=\frac{J_1 s_i^{+}R}{1+e^{(-2n(\langle x\rangle-a))}}+\frac{J_2 s_i^{-}Re^{(-2n(\langle x\rangle-a))}}{1+e^{(-2n(\langle x\rangle-a))}}+IR. \tag{7}$$

We have decoupled the complex system of Eq. (1) into N independent subsystems by using Eq. (7).

3 Interpretable mechanism of neural networks

We can recouple Eq. (1) into a 1D equation to obtain the average activity of the system $\langle x\rangle$ and calculate the expectations of both sides of Eq. (7). Accordingly, we can further obtain the relationship between the average behavior and the structure of the system as follows:

$$\langle x\rangle=\frac{J_1(1-p)\langle s\rangle R}{1+e^{(-2n(\langle x\rangle-a))}}+\frac{J_2 p\langle s\rangle Re^{(-2n(\langle x\rangle-a))}}{1+e^{(-2n(\langle x\rangle-a))}}+IR, \tag{8}$$

where $\langle s\rangle$ is the average degree of the neuron network, and we have the average degree of $\hat{\mathbf{A}}$ is $\langle s^{+}\rangle=(1-p)\langle s\rangle$, and the average degree of $\overline{\mathbf{A}}$ is $\langle s^{-}\rangle=p\langle s\rangle$. Hence, $\langle x\rangle$ is obtained by solving Eq. (8) with $\langle s\rangle$ and p. In particular, $n\to\infty$, $e^{(-2n(\langle x\rangle-a))}$ takes only a discrete value of zero, which means that the neuron is inactive or firing at the maximum rate, corresponding to the logic approximation. Hence, the first and second terms in Eq. (8) can be approximated as follows:

$$\lim_{n\to\infty}\frac{J_1(1-p)\langle s\rangle R}{1+e^{(-2n(\langle x\rangle-a))}}=J_1(1-p)\langle s\rangle R, \tag{9}$$

$$\lim_{n\to\infty}\frac{J_2 p\langle s\rangle Re^{(-2n(\langle x\rangle-a))}}{1+e^{(-2n(\langle x\rangle-a))}}=0. \tag{10}$$

Therefore, the average activity of the neurons in the system can be approximated by substituting Eqs. (9) and (10) into Eq. (8):

$$\langle x\rangle\approx J_1(1-p)\langle s\rangle R+IR. \tag{11}$$

According to Eq. (11), from the perspective of inhibition, the average activity of neurons is only determined by the proportion of inhibiting interactions in the network, and the effect of the inhibition intensity J_2 can be ignored. In addition, the lower (L) and upper (H) bounds of the average activity of neurons can be obtained as follows:

$$\langle x\rangle_{\text{L}}=\lim_{p\to 1}\langle x\rangle=IR, \tag{12}$$

$$\langle x\rangle_{\text{H}}=\lim_{p\to 0}\langle x\rangle=J_1\langle s\rangle R+IR. \tag{13}$$

In particular, an implicit condition that $\langle x\rangle>a$ exists in the approximating process of Eqs. (9) and (10). According to

Eq. (11), we have

$$IR > a. \tag{14}$$

Equation (11) can be separated as follows to clearly distinguish the dynamic behavior and structure of the neural network:

$$\left\{ \langle x \rangle_{\mathrm{J}} = \frac{\langle x \rangle}{J_1 R} - \frac{I}{J_1} \right\} = \{(1-p)\langle s \rangle = \langle s^{+} \rangle\}, \tag{15}$$

where $\langle x \rangle_{\mathrm{J}}$ is a linear combination of the average neuron activity and kinetic parameters denoting the average performance of the system, and $\langle s^{+} \rangle$ is the average positive connectivity of the network structure. Equation (15) can help predict the system performance through the structural parameters. Even if the excitation function is nonlinear, our results show that the behavior of the system is linear with the system structure.

We can conclude that the average activity of synapses is determined by the proportion of inhibitory links in the neural network, but it has nothing to do with the inhibition intensity between synapses. The upper bound of the average activity is positively correlated with the average density of the network. Meanwhile, the lower bound is independent of the network structure. Our framework can be used to effectively separate the dynamics and structure of neural networks for exploring the system behavior. The result showed that the weighted activity of neural networks is equal to the average links of positive excitation of each neuron in the network.

4 Case studies for network models and real neural networks

4.1 Predicting the system behavior of network models

In this work, we test our mathematical framework for neural dynamics on the network models and real neural networks. The simulation results are based on ode45 function with MATLAB for Eq. (1). We draw the conclusions by comparing the numerical results and theoretical solutions.

We take scale-free network as the neural structure model. While fixing $I = R = 1$ in the system, Fig. 1(a) shows that the theoretical solution of our framework is consistent with the simulation results of the scale-free network. We compared the impact of the excitation strength J_1 and inhibition strength J_2 on system behavior. The result shows that the average activity of the neural network has nothing to do with the inhibition intensity in the system; it is only determined by the inhibition proportion. We draw the theoretical solutions of the average activity of neural network in Fig. 1(b) to examine the contribution of the basal activity I and inverse of death rate R to system behavior. The system shows a lower activity with the increase in the fraction of inhibition links. Meanwhile, the product IR determines the lower bound of the system behavior.

The above simulation results in Fig. 1 show that when the structural and dynamic parameters of the neural network are determined, the steady-state behavior of the neural system is also defined as a certain value, which is in a fixed range. Specifically, the behavior of the system has clear upper and lower bounds when the activating and inhibiting behaviors are mixed together. We draw the upper bound $\langle x \rangle_{\mathrm{H}}$ and lower bound $\langle x \rangle_{\mathrm{L}}$ of the average activity of neurons in Fig. 2 to obtain these critical values for the system.

4.2 Critical threshold of the neural dynamics

From the perspective of system resilience, the system will remain resilient while $\langle x \rangle > \langle x \rangle_{\mathrm{L}}$, as illustrated in the phase diagram of Fig. 3(a). The lower bound $\langle x \rangle_{\mathrm{L}}$ is determined

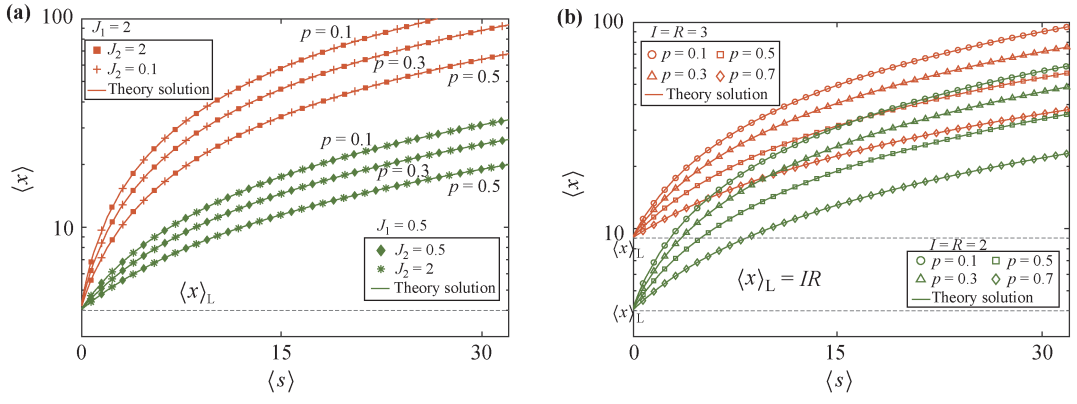

Fig. 1 Testing the model approximations for neural dynamics on scale-free networks with $N = 1000$ and $\langle s \rangle = 8$ (the line is the theoretical solution of Eq. (11), and the various symbols represent the simulation results). (a) Comparison of the impact of the excitation strength J_1 and inhibition strength J_2 on system behavior (here, we set $I = R = 1$, and the inhibition links are randomly selected from the adjacency matrix). (b) Comparison of the impact of the basal activity I and inverse of death rate R on system behavior.

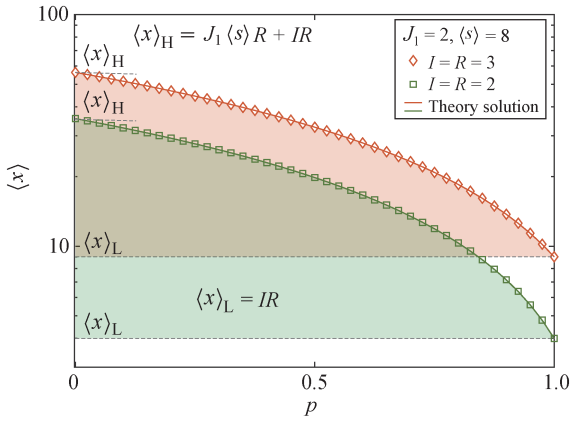

Fig. 2 Upper and lower bounds of the average activity of neurons (the colored solid lines represent the theoretical solutions of $\langle x\rangle$, the square and diamond represent the simulation results of $\langle x\rangle$, and the dotted lines represent the lower bound $\langle x\rangle_{\rm L}$ and the upper bound $\langle x\rangle_{\rm H}$).

by the product IR (Fig. 3(b)), which is independent of the system structure and the coupling strength between neurons. However, the upper bound of the system behavior is much more complicated, which depends on not only the system's dynamic parameters (I and R) but also the structural density $\langle s\rangle$ and the excitation strength J_1 (Fig. 3(c)). Hence, the contour map of the upper bound $\langle x\rangle_{\rm H}$ in Fig. 3(d) is not exactly symmetric compared with the case of the lower bound in Fig. 3(b).

4.3 Linear phenomenon for the nonlinear dynamics

Equation (8) or (11) can well describe the macrobehavior of the system. Our above experiments also test the results of this theoretical framework. In addition to the accuracy of the theoretical results, we also believe that a more concise linear relationship exists between the macrodynamic behavior and the structure of the system coupled by the nonlinear excitation function and complex network structure. We decompose the structural and dynamic parameters of Eq. (11) as follows: The fraction of the excitation links $(1-p)$ is multiplied with the average connection density of the network $\langle s\rangle$ to describe the structure of the network, which is denoted as $\langle s^+\rangle$; and the

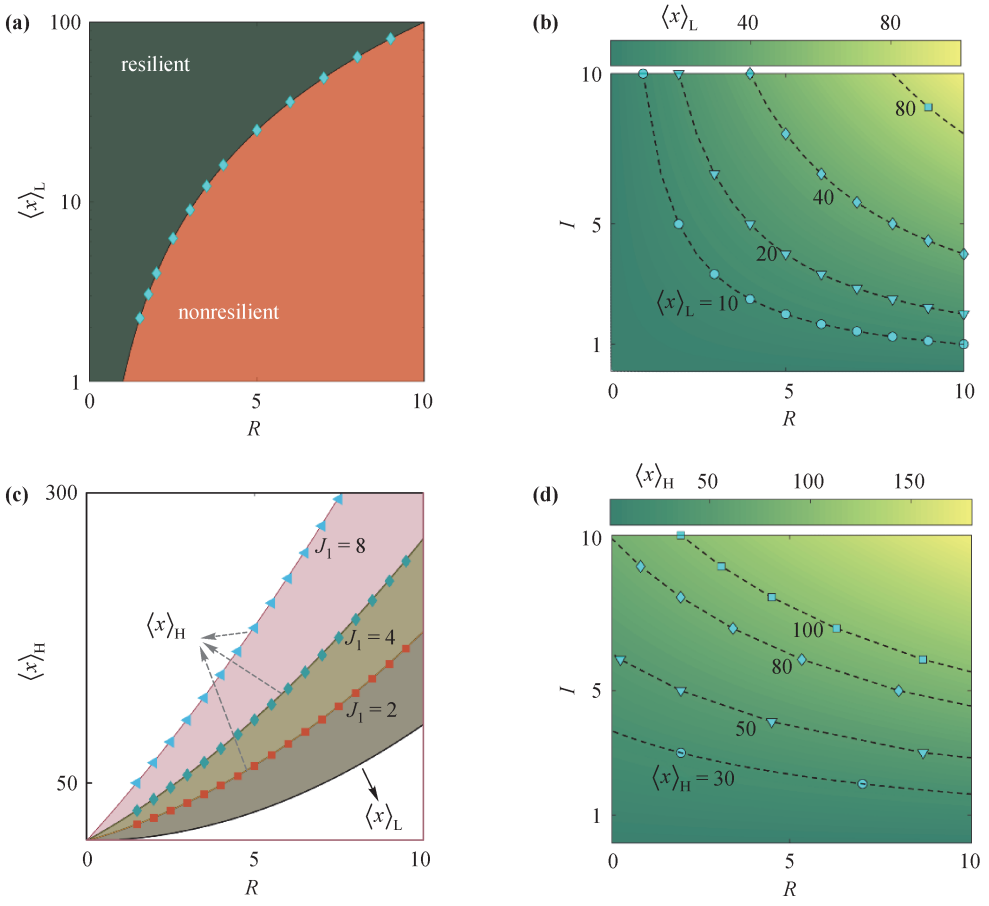

Fig. 3 Contour maps of the upper and lower bounds of the average activity. (a) Phase diagram of the system behavior with $\langle x\rangle_{\rm L}$, which is independent of the network structure and other dynamical parameters. (b) Contour map of the lower bound $\langle x\rangle_{\rm L}$. (c) Phase diagram of the system behavior with $\langle x\rangle_{\rm H}$, which is coupled as $J_1\langle s\rangle R + IR$ (here we set $\langle s\rangle = 4$). (d) Contour map of the upper bound $\langle x\rangle_{\rm H}$.

average activity of the network $\langle x \rangle$, the basal activity I, the inverse of death rate R, and the excitation strength J_1 between neurons are combined as a weighted average activity, which is recorded as $\langle x \rangle_J$ to describe the macrodynamic behavior. The result show that $\langle x \rangle_J = \langle s^+ \rangle$. Specifically, we obtain the linear characteristics of the macrobehavior of this nonlinear activation function with its neural structure. The simulation results and theoretical solution are shown in Fig. 4.

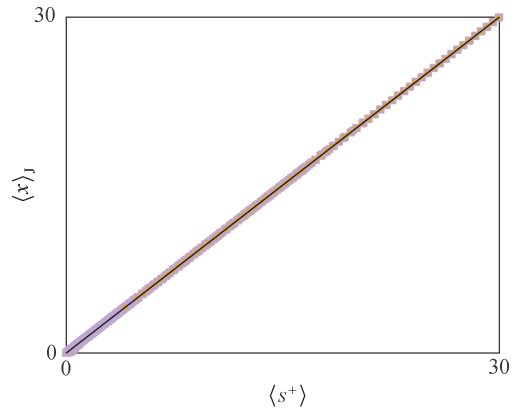

Fig. 4 Linear correlation between the macroscopic behavior and the structure of the nonlinear neural networks (the black solid line is drawn from Eq. (11), and the squares are the simulation results for 1000 networks).

4.4 Application to neural networks for C. elegans

The neural structure of C. elegans is simple — the adult hermaphrodite has only 302 neurons, and the adult male has only 383 neurons. Nevertheless, the system is perfect, and the behavioral characteristics, such as movement, foraging, excretion, mating, and temperature tropism, are also rich. Meanwhile, the system has behaviors like learning, memory, and sleep. This model is ideal to study the whole neural map. The study of the neural network is of great significance for further understanding the structure and function of the nervous system of higher animals and exploring the manner by which the brain structure of animals and humans determines the function. Bumbarger et al. (2013) compare and analyze the pharyngeal nervous system composed of 20 neurons in C. elegans and Pacific nematode with PageRank and centrality as indicators. The results show that the neuronal wiring structure led to different predator–prey behavioral patterns in the two species. Cook et al. (2019) provide a complete map of the neural connections between the two sexes of C. elegans from sensory input to the end organ output of the whole animal, which is an important milestone in the field of connectomes. This map provides basic data for further comprehensive exploration of the interaction mechanism of C. elegans at the cellular level.

We applied our approach to the whole-animal connectomes of both C. elegans sexes. The result shows significant differences between the male and the hermaphrodite within the evolution, as shown in Fig. 5. The sensory neurons and inter-neurons of adult hermaphrodite C. elegans were more advanced, while the activity of the sex specific cells and inter-neurons of adult male accounted for approximately 70% of the total activity.

We can investigate the behavioral differences between adult male and adult hermaphrodite of C. elegans from the steady-state behavior of the neuron map. In this work, we simply compare the behavioral differences between electrical signal and chemical detection (Fig. 6). This work further shows that our framework can theoretically explain the macrobehavior of the complex neural system.

Our method can not only quantify the behavioral differences between the two C. elegans sexes (Fig. 6) but also accurately predict the steady-state behavior of each neuron. According to the link relationship of neurons, the behavior value of each neuron in steady-state must be predicted. If we can predict the activity value of each neuron, then we can correlate the structure of neurons with various behaviors, to precisely explore the mechanism of neuron action at the cellular level. In our framework, the behavior of the node in the steady-state can be calculated by Eq. (7). Furthermore, we need to first determine the macrobehavior of the system and the number of links of a single neuron to infer the steady-state value of a single neuron. According to the previous analysis, this type of nonlinear neural network has linear behavior characteristics, which can be inferred by Eq. (15). Therefore, we only need to determine the number of links of neurons to predict the steady-state activity of a single neuron. With regard to the four types of neuron link data provided in the literature (Cook et al., 2019), we perform the numerical simulation in time dimensions and compare the convergence of each neuron activity.

The result showed that the theoretical solution of the steady-state activity of neurons in our framework is accurate and consistent with the simulation results. In Fig. 7, the line represents the simulation process with ode45 function for the nonlinear neural networks, while the dots are the steady-state of each neuron with our method.

5 Conclusions and future work

Whether the neural network is an artificial neural network or human brain, it is still a black box problem in the existing research and application, which greatly limits its application scenarios. However, the current interpretable research of the neural network focuses on the result-oriented interpretation mechanism. In this study, we mainly consider the coupling method of the nonlinear activation function and its high-dimensional complex network structure. We establish a mathematical framework

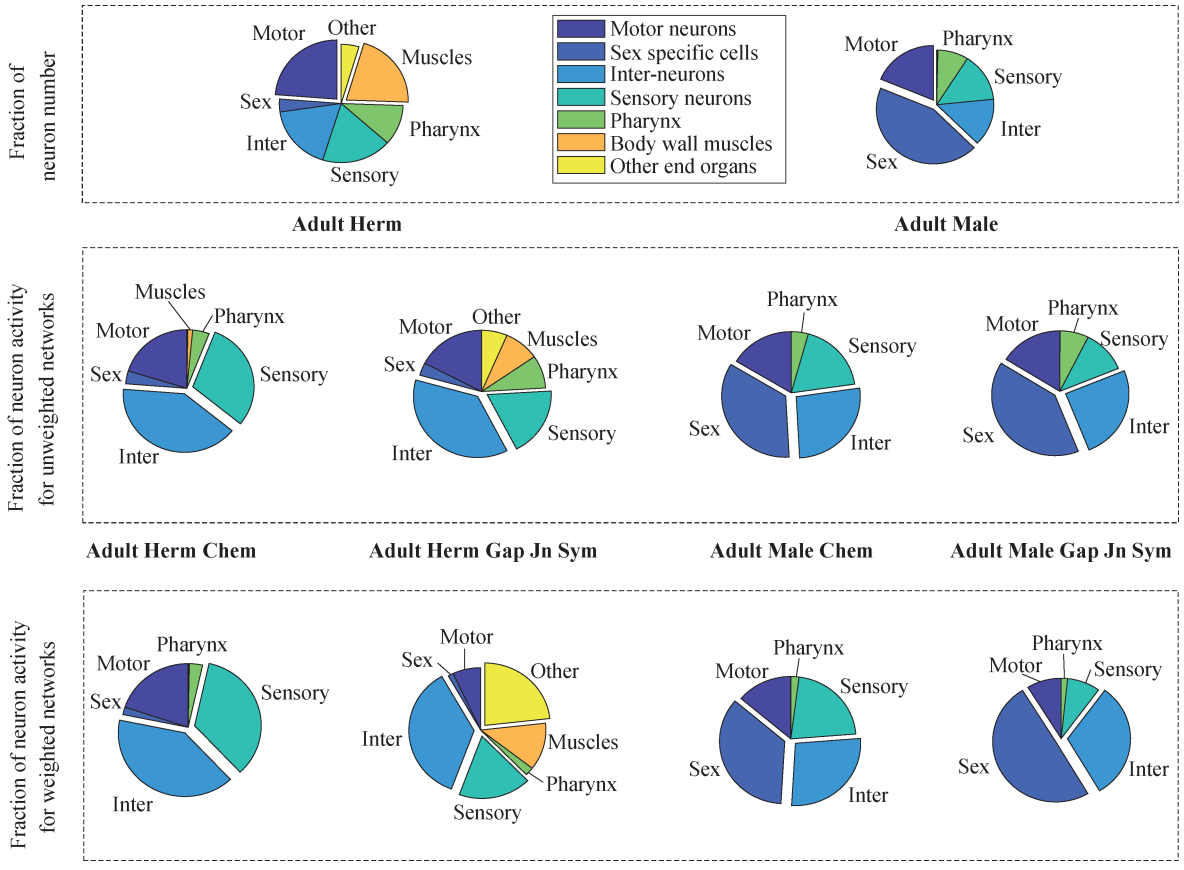

Fig. 5 Comparison of the structure and dynamical behaviors of C. elegans with the whole-animal connectomes ("Chem" and "Gap Jn Sym" mean the chemical and the gap junction connectivity network for the neuron systems, respectively). Top: the pie charts show the number of all types of neurons in the whole neuron map for adult hermaphroditism and adult male of C. elegans. Middle: the connection data of the neuron map are incorporated into the neuron dynamic model to calculate the activity distribution of different types of neurons (here, the system is unweighted network). Bottom: the weight of links is considered when calculating the activity of neurons.

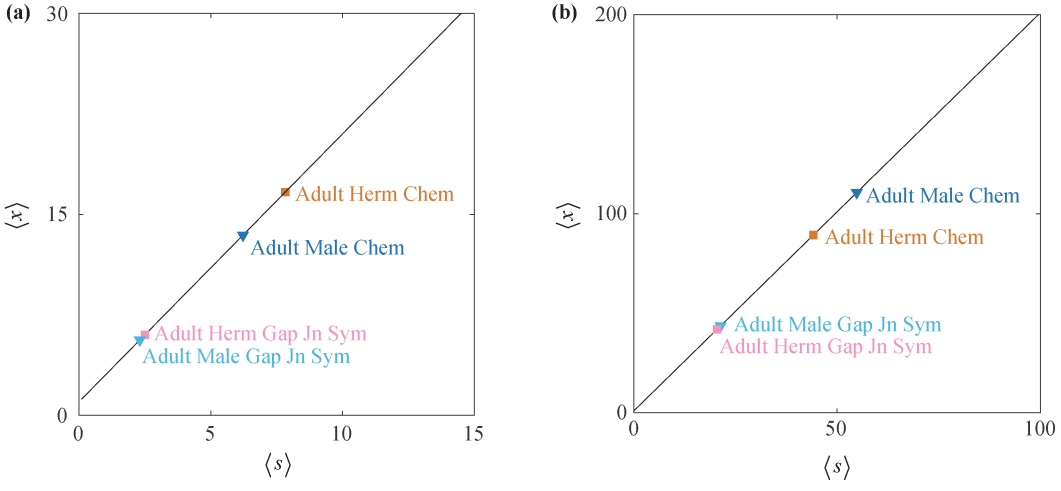

Fig. 6 Quantification of the behavioral differences between the two C. elegans sexes with our framework. (a) Simulation and theoretical solution for unweighted neuron networks. (b) Simulation and theoretical solution for weighted neuron networks.

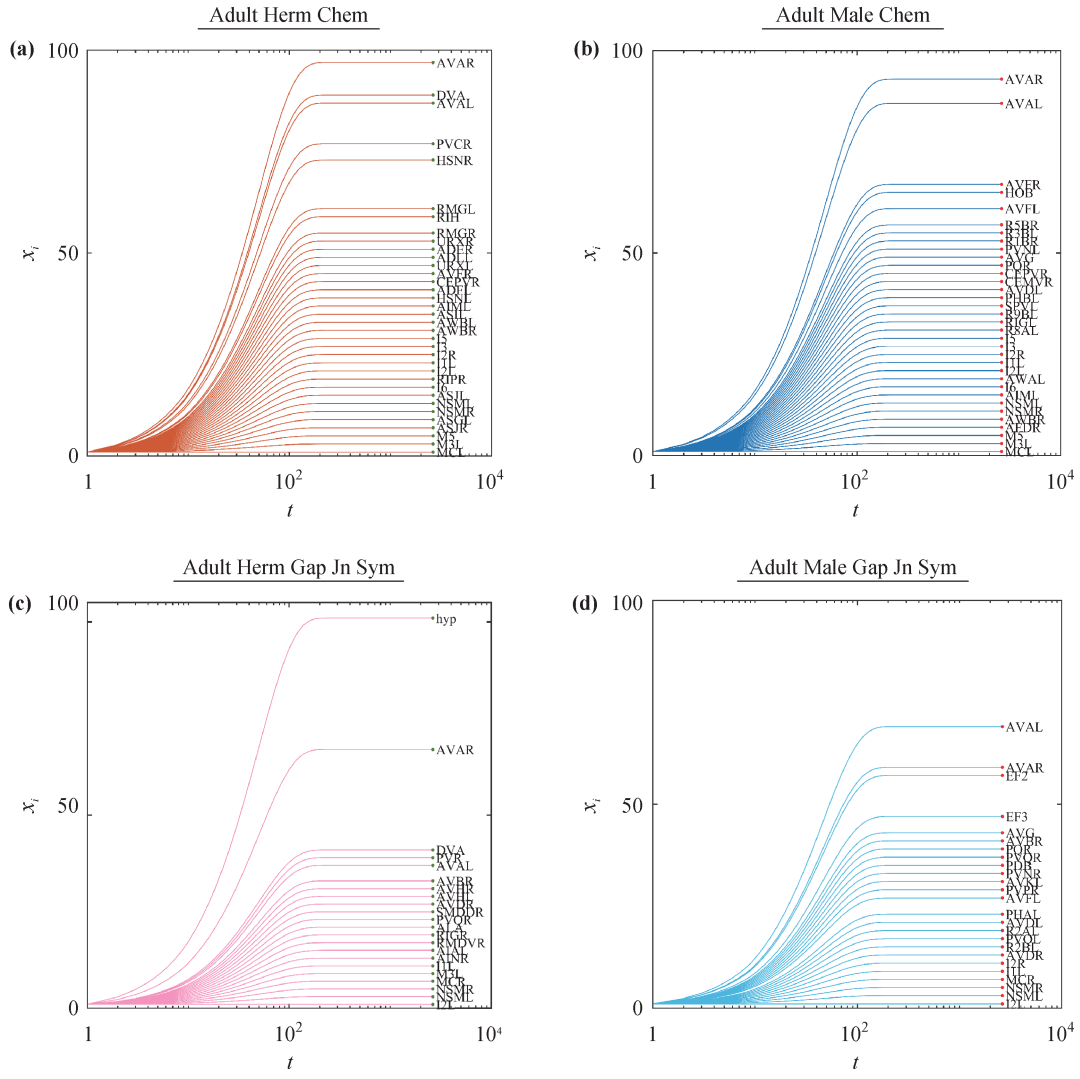

Fig. 7 Predicting the steady-state of each neuron with our framework.

for the highly nonlinear coupling system with a network decoupling approach to reveal the calculating mechanism of the neural network. A simple linear mapping relationship exists between the structure and its dynamical behavior with the highly nonlinear model. Our numerical results and theoretical solutions fully confirm this interesting conclusion. Our mathematical framework provides not only the potential mathematical calculation principle of the neural network but also an effective way to accurately match and predict human or animal brain activities, which can further expand and enrich the interpretable mechanism of the artificial neural network in future.

Our findings have yet to be validated and utilized in the artificial neural network, which can be further studied. In the future, we will offer specific quantitative methods to provide analytical interpretation methods and models for deep learning and machine learning in engineering according to the hierarchical structure and activation function of various types of deep neural networks. We believe that our study of the interpretable mechanism based on the network decoupling method is an important step in the integrative analysis of the coupling mechanism between structure and dynamics toward the understanding of the overall neural network performance. The gained insights can be utilized to extract meaningful information for its application to various artificial neural networks.

References

Arjovsky M, Chintala S, Bottou L (2017). Wasserstein generative adversarial networks. In: 34th International Conference on Machine Learning. Sydney, 214–223

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

Bumbarger D J, Riebesell M, Rödelsperger C, Sommer R J (2013). System-wide rewiring underlies behavioral differences in predatory

and bacterial-feeding nematodes. Cell, 152(1–2): 109–119

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: 30th Conference on Neural Information Processing System (NIPS). Barcelona, 2180–2188

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, 1724–1734

Cook S J, Jarrell T A, Brittin C A, Wang Y, Bloniarz A E, Yakovlev M A, Nguyen K C Q, Tang L T, Bayer E A, Duerr J S, Bülow H E, Hobert O, Hall D H, Emmons S W (2019). Whole-animal connectomes of both Caenorhabditis elegans sexes. Nature, 571 (7763): 63–71

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014). Generative adversarial nets. In: 27th International Conference on Neural Information Processing Systems (NIPS). Montreal, Quebec, 2672–2680

Hendricks L A, Akata Z, Rohrbach M, Donahue J, Schiele B, Darrell T (2016). Generating visual explanations. In: European Conference on Computer Vision. Amsterdam, 3–19

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

Kádár Á, Chrupała G, Alishahi A (2017). Representation of linguistic form and function in recurrent neural networks. Computational Linguistics, 43(4): 761–780

Kindermans P J, Schütt K T, Alber M, Müller K R, Erhan D, Kim B, Dähne S (2017). Learning how to explain neural networks: PatternNet and PatternAttribution. arXiv preprint, arXiv: 1705.05598

Li J, Chen X, Hovy E, Jurafsky D (2015). Visualizing and understanding neural models in NLP. In: 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, 681–691

Mirza M, Osindero S (2014). Conditional generative adversarial nets. arXiv preprint, arXiv: 1411.1784

Nguyen A, Clune J, Bengio Y, Dosovitskiy A, Yosinski J (2017). Plug & Play Generative Networks: Conditional iterative generation of images in latent space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, 3510–3520

Park D H, Hendricks L A, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, 8779–8788

Radford A, Metz L, Chintala S (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint, arXiv: 1511.06434

Samek W, Binder A, Montavon G, Lapuschkin S, Müller K R (2017). Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11): 2660–2673

Schuster M, Paliwal K K (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11): 2673–2681

Strobelt H, Gehrmann S, Pfister H, Rush A M (2018). LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Transactions on Visualization and Computer Graphics, 24(1): 667–676

Tang Z, Shi Y, Wang D, Feng Y, Zhang S (2017). Memory visualization for gated recurrent neural networks in speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, 2736–2740

Wu M, Hughes M C, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F (2017). Beyond sparsity: Tree regularization of deep models for interpretability. In: 32nd AAAI Conference on Artificial Intelligence. New Orleans, LA, 1670–1678

Yu L, Zhang W, Wang J, Yu Y (2016). SeqGAN: Sequence generative adversarial nets with policy gradient. In: 31st AAAI Conference on Artificial Intelligence. San Francisco, CA, 2852–2858

Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas D N (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE International Conference on Computer Vision (ICCV). Venice, 5908–5916

Zhang Y, Xiao Y, Hwang S W, Wang H, Wang X S, Wang W (2017). Entity suggestion with conceptual explanation. In: 26th International Joint Conference on Artificial Intelligence (IJCAI). Melbourne, 4244–4250

Zharov Y, Korzhenkov D, Shvechikov P, Tuzhilin A (2018). YASENN: Explaining neural networks via partitioning activation sequences. arXiv preprint, arXiv: 1811.02783

Zhou B, Bau D, Oliva A, Torralba A (2019). Interpreting deep visual representations via network dissection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2131–2145