Resolving the neural mechanism of core object recognition in space and time; A computational approach

Naser Sadeghnejad[1], Mehdi Ezoji[1*], Reza Ebrahimpour[2,3 *] and Sajjad Zabbah[3,4,5]

[1]Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

[2]Faculty of Computer Engineering, Shahid Rajaee Teacher Training University of Technology, Tehran, Iran

[3]School of Cognitive Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

[4]Wellcome Centre for Human Neuroimaging, University College London, London, UK.

[5]Max Planck UCL Centre for Computational Psychiatry and Aging Research, University College London, London, UK

Abstract

The underlying mechanism of object recognition- a fundamental brain ability- has been investigated in various studies. However, balancing between the speed and accuracy of recognition is less explored. Most of the computational models of object recognition are not potentially able to explain the recognition time and, thus, only focus on the recognition accuracy because of two reasons: lack of a temporal representation mechanism for sensory processing and using non-biological classifiers for decision-making processing. Here, we proposed a hierarchical temporal model of object recognition using a spiking deep neural network coupled to a biologically plausible decision-making model for explaining both recognition time and accuracy. We showed that the response dynamics of the proposed model can resemble those of the brain. Firstly, in an object recognition task, the model can mimic human's and monkey's recognition time as well as accuracy. Secondly, the model can replicate different speed-accuracy trade-off regimes as observed in the literature. More importantly, we demonstrated that temporal representation of different abstraction levels (superordinate, midlevel, and subordinate) in the proposed model matched the brain representation dynamics observed in previous studies. We conclude that the accumulation of spikes, generated by a hierarchical feedforward spiking structure, to reach abound can well explain not even the dynamics of making a decision, but also the representations dynamics for different abstraction levels.

Keywords: Temporal Object Recognition, Speed- accuracy Trade-off, Deep Spiking Convolutional Neural Network, Accumulation to Bound Model, Dynamical Representational Dissimilarity Matrix

1. Introduction

Object recognition is one of the main cognitive abilities in different species, especially mammals. Correctly recognizing an object in an efficient time is pivotal for proper interaction with the environment. Imagine how accurate and fast recognition of a predator is vital for an animal (Chittka et al., 2009). In order to understand the underlying mechanism of this cognitive ability, studies have suggested different computational models resampling humans' or monkeys' responses at both behavioral (choices) and neural levels (activities in the ventral visual pathway) (Riesenhuber and Poggio, 2000, Rajaei et al., 2019, Khaligh-Razavi and Kriegeskorte, 2014, Yamins and DiCarlo, 2016). Although there is strong evidence that both neural response (Dehaqani et al., 2016, Contini et al., 2017, Cichy et al., 2014, Isik et al., 2014) and behavioral choices (Mirzaei et al., 2013, Macé et al., 2009, Grill-Spector and Kanwisher, 2005) largely vary over time, the underlying mechanism is not well understood.

In addition to the temporal representation of sensory information, the dynamic of our behavior in an object recognition task is affected by another process in the brain called decision-making process. It is a process that entails reading out the sensory representations over time and balancing between the speed and the accuracy to maximize the expected gain (Okazawa et al., 2020, Heekeren et al., 2004, Zhan et al., 2019). Similar to the temporal representation of sensory information, the decision-making process and thus the speed-accuracy trade-off regime, has commonly been ignored in the computational model of object recognition. Most of these studies have investigated the process of object recognition while using a non-temporal linear readout mechanism to map the instantaneous population activity into the behavior (Chang and Tsao, 2017, Majaj et al., 2015).

The existing models which try to explain the process of object recognition in the brain have mostly different layers resembling different areas in the visual stream. In a hierarchical manner, these layers represent different features of the input stimuli from very simple (e.g., lines in specific angles) to very complex (e.g., faces) ones and increase the invariancy of the representations by pooling mechanisms between layers (Serre et al., 2007a, Kheradpisheh et al., 2016b, Riesenhuber and Poggio, 2000). In object recognition tasks, these hierarchical models, such as the HMAX model (Riesenhuber and Poggio, 1999), different extensions of the HMAX model (Farzmahdi et al., 2016, Zabbah et al., 2014, Rajaei et al., 2012, Ghodrati et al., 2012) and, more recently, deep convolutional neural network (DCNN)-based approaches (Cichy et al., 2016, Kriegeskorte, 2015), can categorize input images with the similar accuracy to that of humans or animals. Spiking versions of these models explain how neurons respond to specific features in the input stimulus via STDP learning rules (Kheradpisheh et al., 2018). These models are usually being evaluated by comparing their responses in different layers with neural or behavioral responses of humans or animals (see also (Cichy et al., 2016)). However, both models and evaluation methods suffer from two problems: (i) ignoring the dynamics of responses at both behavioral and neural levels. (ii) ignoring the decision-making process and thus the speed-accuracy trade-off regime. It worth mentioning that the underlying mechanism of response dynamic in the brain has recently been investigated in some studies (Kietzmann et al., 2019, Kar et al., 2019, Nayebi et al., 2022, Mirzaei et al., 2013). Kar et al. (Kar et al., 2019), showed that emerging to a solution in the brain for challenge images takes more time compared to control images. Brain behavior in presence of these images can better be explained by recurrent models compared to feedforward-only models. The necessity of recurrent connections for explaining the process of object recognition in the brain is also discussed in other studies (Kietzmann et al., 2019, Mirzaei et al., 2013, Nayebi et al., 2022, Spoerer et al., 2019). However, the underlying mechanism of behavioral and neural dynamics during feedforward processing is not yet investigated (see Kar et al. (Kar et al., 2019), Figure (5) for different dynamics during backward masking paradigm (Fahrenfort et al., 2007)).

In this study, in line with our recent work (Heidari Gorji et al., 2018), we proposed a hierarchical temporal model of object recognition using a spiking deep neural network connected to a biological plausible decision-making stage. The first stage of the model, which corresponds to the ventral stream in the brain, is comprised of several convolutional and pooling layers, in which neurons in each layer progressively learn frequent and salient patterns in the input image using STDP learning rules. Importantly, information between the layers is conveyed through spikes generated by trained neurons and, thus, the dynamic representation of the input stimulus is shaped. In a hierarchy, neurons are getting selective to more complex features from edges to faces for an instance. Spikes that are generated overtime in the first stage of the model are then transferred to the decision-making stage. There are accumulator units in this stage where the generated spikes are being accumulated toward a decision bound and, thus, shaping the accumulation to bound the mechanism, which is almost evident in the brain (Gold and Shadlen, 2007, Shadlen and Newsome, 2001).

Response dynamics of the proposed model can resemble that of the brain. Firstly, in an object recognition task, the model can mimic behavioral characteristics of humans' and monkeys' responses, not only in terms of the probability of correct response but also in terms of the reaction time. Secondly, similar to (Hanks et al., 2014), the speed-accuracy trade-off can be controlled with the decision bound of the proposed model. Finally and more importantly, the dynamics of neural representation for different abstraction levels (superordinate, midlevel and subordinate) in the proposed model match the brain representation dynamics observed in Dehaghani et al. (Dehaqani et al., 2016). As a result, the proposed temporal model can explain the dynamics of responses during object recognition in the brain from the very first layers of the temporal stream to behavioral responses.

2- Materials and methods

When an image is presented to the human visual system (HVS), various components of the image, such as prominent edges as well as coarse and fine information, are not processed simultaneously (Thorpe et al., 2001, Portelli et al., 2016). Considering these studies, we proposed a biologically plausible temporal model to explain the response dynamics of the brain during an object recognition task.

2-1- Structure of the proposed model

The proposed model consists of three main stages inspired by the object recognition system in the brain (Figure (1)). In the first stage, the input image is encoded into discrete spike trains in the temporal domain. Then, a deep spiking convolutional neural network (DSCNN) is used as a temporal feature extractor. Finally, accumulator units integrate evidence in support of possible choices, and the decision is made as soon as the accumulated evidence reaches a certain threshold (resembling the accumulation to bound model).

The active part of the cerebral cortex in HVS corresponding to each stage of the proposed model is highlighted in the upper part of Figure (1).

2-1-1 Temporal feeding

In this stage, we aim to model the behavior and characteristics of the ganglion cells and the lateral geniculate nucleus cells (Delorme et al., 2001, Kheradpisheh et al., 2016b, Masquelier and Thorpe, 2007). Thus, a difference of Gaussian (DoG) filter is used to precisely approximate the function of these sensitive cells according to (Kuffler, 1953).

Then, the output of this contrast detector filter is presented in the time-domain based on the amplitude of the detected contrast. Information from high contrast components in the image transfers sooner than the low-contrast ones. In other words, the higher the contrast amplitude, the faster the response would be. This rank-order coding, is efficient to replicate V1 like responses (Delorme et al., 2001). For simplicity, the contrast map is decomposed into discrete-slot temporal maps, in which each slot corresponds to a specific contrast slot and, thus, determines how fast the information will be transmitted.

2-1-2- Temporal feature representation (spiking convolutional neural network (SCNN))

The Spatio-temporal representation of the input image is then transferred to the next stage where a deep SCNN is used to simulate the neural responses of hierarchical layers in the ventral stream during the object recognition process. These layers progressively extract appropriate features for recognition tasks, from very simple features (i.e., edges or curves) to more complex ones (i.e., faces or cars) via an unsupervised learning algorithm (Kheradpisheh et al., 2018). Multiple spiking convolutional layers in the network structure form response selectivity to these features with different complexity levels via spike time-dependent plasticity (STDP) learning algorithm. A spiking pooling mechanism provides invariancy in neural responses to image local variations such as scale and displacement.

The following describes the convolution and pooling layers in the SCNN.

Spiking convolutional layer: A convolutional layer contains several kernels to detect similar visual properties in different locations. Each neuron receives spikes from neurons of previous layers within a specific receptive field (window). Spikes of all the neurons are generated based on the integrate-and-fire model. Based on this model, the membrane potential of a neuron is stated in terms of the synaptic inputs from presynaptic neurons. An action potential (spike) is generated when the membrane potential reaches a pre-determined threshold. In each time step t, the membrane potential of the i-th neuron is updated as follows:

$$V_i(t)=V\left(t-1\right)+\sum_j W_{j,i}\, S_j(t-1) \tag{1}$$

where Vi(t) is the membrane potential of the i-th convolutional neuron, Wij is the synaptic weight between the j-th presynaptic neuron and i-th postsynaptic neuron, and Sj is the spike train of the j-th presynaptic neuron. Sj (t - 1) is equal to 1 if the j-th presynaptic neuron generates a spike in time step t; it is set to 0 otherwise.

Vi controls the generation of action potential (spike) in the i-th neuron; if Vi exceeds a pre-determined threshold (V_{thr}), a spike is generated and Vi is reset to potential Vi = 0, i.e.:

$$V_i(t)=0 \quad , \quad S_i(t)=1 \qquad \text{if} \;\; V_i(t)\geq V_{thr} \tag{2}$$

There is also a lateral inhibition mechanism in all the convolutional layers. When one neuron fires in a particular position, it inhibits other neurons in that position i.e., their potentials are reset to zero. Neurons can generate as many spikes as they can. However, to increase the speed of the proposed model in the training phase, we limit the number of spikes to one spike (i.e., each neuron can only generate one spike).

Spiking pooling layers: Pooling layers which are interleaved between convolutional layers as shown in Figure (1) compress visual information and, thus, resemble what occurs in complex cells in the visual cortex(Serre et al., 2007b). Pooling neurons are of the integrate-and-fire type, with intra-synaptic weight and thresholds all set to one.

Finally, at the end of the temporal feature representation stage, there is a global aggregation layer where spikes generated from different locations of each kernel are being integrated; in each time step t, the output of this layer is calculated using Eq. (3):

$$O_i(t) = \sum_{j=1}^{m}\sum_{k=1}^{n} S_{j,k}^{i}(t) \tag{3}$$

In Eq. (3), Oi(t) is the output of the global aggregation layer corresponding to the i-th kernel, and m, n is the output size of the previous convolution layer. Also, $S_{j,k}^{i}(t)$ is the neuron spike corresponding to locations j and k of the output of the i-th filter.

STDP algorithm: As mentioned above, learning occurs only in convolutional layers. The learning process is layer by layer, in a way that the learning process in each layer starts whenever it finishes in the previous layer. When a new image is presented, neurons of each convolutional layer compete; those which fire earlier trigger the STDP algorithm to learn the input pattern. Weights in the STDP algorithm are updated as follows:

$$\Delta W_{ij} = \begin{cases} a^{+} W_{ij}(1-W_{ij}) & \text{if} \quad t_j - t_i \le 0 \\ a^{-} W_{ij}(1-W_{ij}) & \text{if} \quad t_j - t_i > 0 \end{cases} \tag{4}$$

where i and j refer to pre-and post-synaptic neurons, t_i and t_j are the corresponding spike times, respectively. ΔW_{ij} is the synaptic weight modification, and a^+ and a^- are two parameters specifying the learning rate. The sign, not the amount of time difference between the two spikes, affects the weight changes. The multiplicative term $W_{ij}(1-W_{ij})$ ensures the weights remain within [0,1] where W_{ij} is within [0,1]. Finally, Eq. (5) is used to calculate C_L as a criterion to stop the learning procedure (convergence) in the convolutional layers.

$$C_L = \sum_{f}\sum_{i} W_{f,i}(1-W_{f,i}) / n_w \tag{5}$$

where $W_{f,i}$ is the i-th synaptic weight of the f-th kernel and n_w is the total number of synaptic weights (independent of the features) in that layer. C_L tends to zero if each of the synaptic weights converges toward zero or one. Therefore, we stop the learning of the l-th convolutional layer whenever C_L is sufficiently close to zero.
The number of layers, size of the filters and learning parameters are set empirically.

2-1-3- Temporal decision making

After two stages of temporal coding, the input image will be represented in the spike trains of neurons in CNN. Neurons in the last layer on the CNN provide momentary evidence regarding the previously learned category-based patterns in the input image. Spikes of these neurons are transferred to the decision-making stage, in which a decision about the category of the input image is made via accumulating the bound model which is a well-known model of the decision-making process in the brain. In this stage, two accumulator units, corresponding to two categories of the input image, integrate the generated spikes in favor of each category. The decision will be made as soon as each of these accumulators reaches a pre-set threshold.

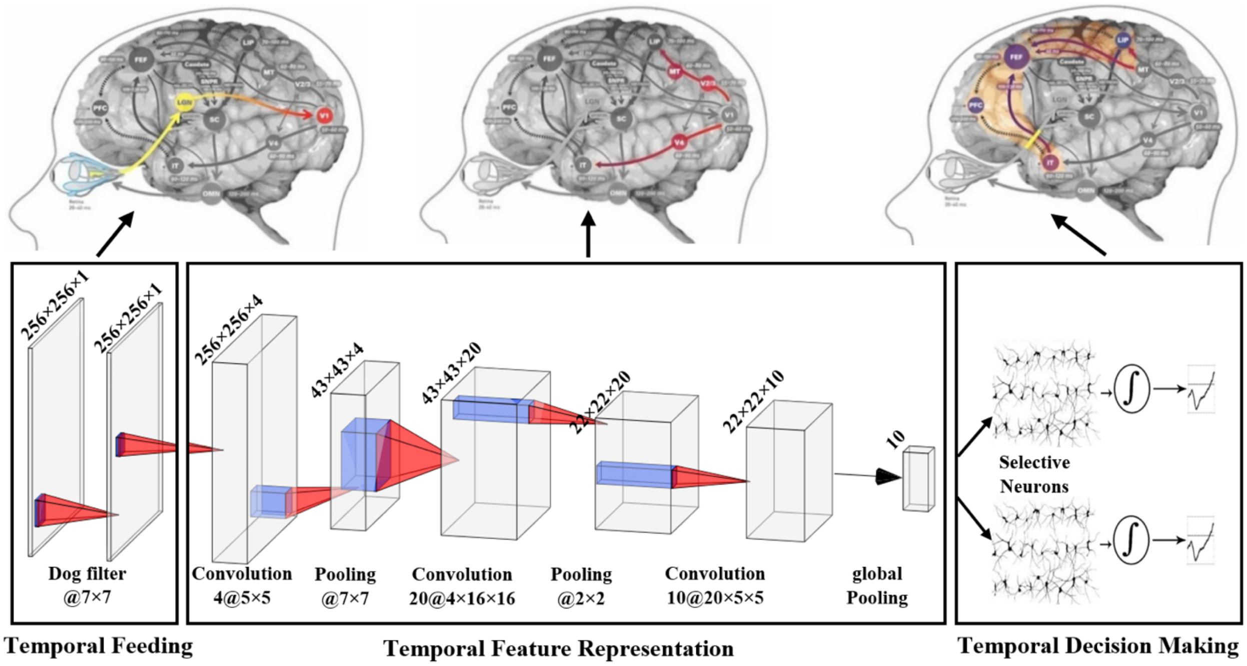

Figure (1). The general structure of the proposed model with three stages. In the first stage (temporal feeding), DOG filters are used to simulate the on-center off-center behavior of ganglion and geniculate cells. Then, the contrast values are being temporally codded, in a way that higher values will be sent to the next stage faster than lower values. The next stage (temporal feature representation) extracts more complex features using spiking neural networks which are simulating the ventral stream in the brain. There is another temporal coding in this stage. Patterns that are more similar to those previously learned ones will more potentiate the corresponding neurons and, thus, generate spikes sooner than the less similar patterns. Finally, spike trains will be accumulated in the temporal decision-making stage until reaching a specific level of threshold. The number and size of DOG, convolution and pooling filters are given as m@i×j×k at the bottom of the figure, where m shows the number of filters and i×j×k the size of each filter. Also at the top of the filters is the output size of each filter. (Of course, assuming an image with dimensions of 256 × 256 as input).

2-2- Implementing the proposed model

In the following, implementing various stages of the proposed structure, i.e., temporal coding of input images, feature representation, and decision making, is discussed.

2-2-1- Temporal feeding

In this paper, a DoG filter with the size of 7 × 7 pixels is constructed using Gaussian kernels with the standard deviations of 1 and 2. Higher values in DoG response show sharper edges and lower values are associated with smooth edges. Then, the output of the DoG filter is decomposed into 30 slot s based on the uniform multi-thresholding on the pixel values. Each of these 30 (considered as 30 different time points) slot s is applied as the input to the proposed model sequentially.

2-2-2- Temporal feature representation

In this stage, a spiking convolutional neural network with three convolution layers and three pooling layers is proposed. The initial weights of the convolution filters are randomly selected within [0, 1]. Using STDP, weights of convolution layers are being updated in an unsupervised fashion.

In the first experiment, the first, second and third layers of spiking convolution have 4, 20 and 10 kernels with the size of 5×5, 4×16×16 and 20×5×5, respectively. The size of the spiking pooling windows of the first and second layers is 7 × 7 and 2 × 2, respectively, with the size of strides 6 and 2. The third pooling layer performs a global pooling process that has 10 outputs due to the presence of 10 filters in the third layer of convolution.

Also, in the second experiment, the first, second and third layers of spiking convolution have 4, 50, and 30 neural maps (filters) with the dimensions of 5×5, 4×19×19, and 50×7×7, respectively. The size of the spiking pooling filters of the first and second layers is 7×7 and 4×4, respectively, with the size of strides 6 and 4. The third pooling layer performs a global pooling process that has 30 outputs due to the presence of 30 filters in the third layer of convolution.

It is important to note that we use simpler model for the first task, because this task is simpler than the second one. The simpler model in the representation stage speed-up the process. These models can temporally represent the input stimulus in neural firing rates and can accumulate represented information toward possible choices over time.

2-2-3- Temporal decision-making stage

For this stage, we use the accumulation to bound model. First, we empirically determine category-selective neurons from neurons in the last layer of the feature representation stage and call them category-selective neurons. This selection is based on the number of spikes generated by the neuron while presenting the images from a specific class. For example, car selective neurons are those that generate the greatest number of spikes when car images are shown and the fewest number of spikes in response to face images and similar for face-selective neurons. Then, the sum of the output of these neurons in each class is fed into two units of the decision model. These units accumulate the generated spikes in their inputs over time. A decision is made when the accumulated spikes reach a pr-set threshold. In the first experiment, we examine the behavior of the model for two different thresholds 20 and 30. Then, in another analysis we change the threshold in a wider range 0-50, and study the performance and reaction time of the decision.

3- Experiments and evaluation

In this paper, to evaluate the performance of the proposed model and its brain plausibility behavior, two different experiments are designed. In the first experiment, the decision of the model in response to images with different levels of noise is examined. In the second experiment, the neural representation in the model is compared with that in the inferior temporal (IT) cortex. We use logistic regression (Eq. (6)) and hyperbolic tangent function (Eq. (7)) as well-known psychometric and chronometric functions in behavioral studies to evaluate the probability of correct response and time of the response, respectively.

$$\mathrm{Logit}[P_{correct}]=\beta_0+\beta_1 C \qquad (6)$$

where Logit(P) is shorthand for log(p/1-p), β_i are regression coefficients and C stands for different stimulus strength. Fitting is by maximum likelihood under a binomial error model (i.e., a GLM).

$$\mathrm{RT}=\beta_0+\beta_1\frac{\tanh(\mathrm{C})}{\mathrm{C}} \tag{7}$$

where RT is the response time of the model and C stands for stimulus strength.
Also, we consider the regression models (Eq. (8) and Eq. (9)) to evaluate the number of spikes and the time of the peak response respectively.

$$\mathrm{NS}=\beta_0+\beta_1\mathrm{C} \tag{8}$$

$$\mathrm{TP}=\alpha_0+\alpha_1\mathrm{C} \tag{9}$$

where NS in Eq. (8) is the number of spikes at peak response, TP in Eq. (9) is the time of the peak response, β_i and α_i are regression coefficients and C stands for different stimulus strength.

In the following, using Eq. (10) we evaluate the effect of the decision threshold and different stimulus strengths on the reaction time.

$$\mathrm{RT}=\beta_0+\beta_1\mathrm{C}+\beta_2\theta \tag{10}$$

where RT is the response time of the model, β_i are regression coefficients, C stands for stimulus strength and θ is decision threshold.

We also use the "separability index" (SI) to quantify the discriminability of object categories in the neural population responses in both model and IT cortex.

Separability index: The separation of two categories of images in R^N can be defined as the ratio of the between- and the within-category scatter matrices (Duda et al., 1973).
Consider an M-category problem, in which there are n_i stimuli of category Ci, so that the mean vector of each category and the mean of the total data are given by:

$$\mu_i=\frac{1}{n_i}\sum_{j\in C_i}\vec{r}_{\mathrm{j}} \tag{11}$$

$$m=\frac{1}{n}\sum_{i=1}^{c}n_i\mu_i \tag{12}$$

where $\vec{r}_j$ is the jth and *n* is the number of total stimuli.

Then, the within-category scatter for each class and the whole data set is calculated by Eq. (13) and Eq. (14), respectively. The between-category scatter matrix is given by Eq. (15):

$$S_i=\sum_{j\in C_i}(\vec{r}_j-\mu_i)(\vec{r}_j-\mu_i)^T \tag{13}$$

$$S_w=\sum_{i\in M}S_i \tag{14}$$

$$S_B = \sum_{i \in M} n_i (\mu_i - m)(\mu_i - m)^T \tag{15}$$

Because we use the representation of images in neural space with N neurons, S_W and S_B have size N by N. Finally, the SI is calculated as follows:

$$SI = \frac{\|S_B\|}{\|S_W\|} \tag{16}$$

where ||S|| is the norm of S. In this paper, we use a spectral norm (or ||S||) (Horn and Johnson (McCarthy et al., 1990)).

3-1- Designed datasets

We use two different datasets in our experiments. In the first experiment, dataset-I is used for a two-category object classification problem; in the second experiment, dataset-II is employed to examine the temporal representation of classification levels in object recognition.

3-1-1- Dataset-I

The purpose of designing this dataset is to investigate the behavior of the model in two-class noisy decision space. In this dataset, face images are collected from the Caltech dataset (Griffin et al., 2007) and car images are collected from the Internet. The entire dataset consists of 350 face images and 350 car images with the size of 256 × 256 pixels. Some examples of these images are shown in Figure (2). Ten images from each set are selected randomly for testing. Using them, 120 different noisy images with different degrees of perceptual difficulty were formed via the proposed algorithm-I, as follows:

Figure (2) A few examples of Caltech and car face images (Griffin et al., 2007)

Algorithm-I

Input: Selected images from car and face classes in Caltech(Griffin et al., 2007)
Output: Images with different degrees of perceptual difficulty
1. Fourier transform is taken from all images.
2. The average amplitude of all the images is calculated.
3. The noise phase matrix is calculated as a product (1 - stimulus stiffness) in the interval (-π, π).
4. Hybrid phase matrix is calculated from the sum of the phase matrix of the face (car) image with the noise phase matrix.
5. Using the hybrid phase matrix and the mean amplitude matrix, the inverse Fourier transform is obtained.

Some examples of the obtained noisy images from algorithm-I are shown in Figure (3). The image with a strength of 40% is formed from 40% of the original image and 60% of the noise in phase. With 10 face images and 10 car images, 10,000 images with different strengths are formed, which are fed into the proposed model at 30 different temporal slot s. To test the network, several of these different stimuli with different strengths, which are 0, 20, 40, 60, 80 and 100%, respectively, are selected.

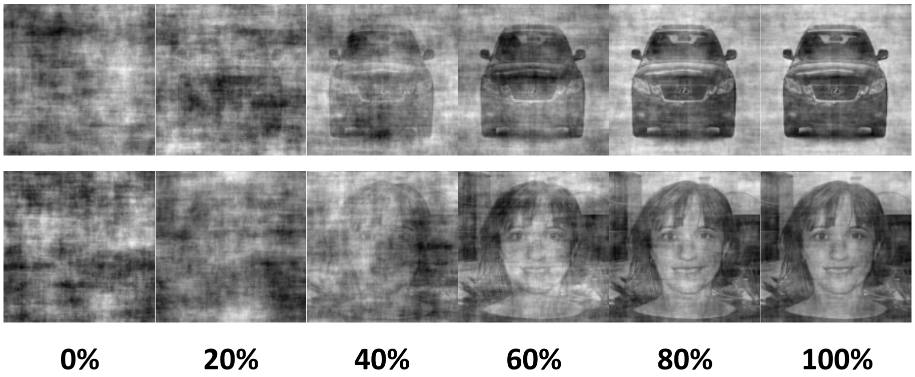

Figure (3) Several examples of visual stimuli with different strengths

3-1-2- Dataset-II

The second experiment aims to investigate the temporal representation of different abstract levels of categorization in a multi-class problem in the proposed model. Therefore, the corresponding dataset is selected to examine different levels of classification in object recognition, including superordinate level, midlevel and subordinate level classifications. For this purpose, eight different classes are selected as shown in Figure (4), which contains examples of images from each class. Inanimate classes include images of cars, airplanes, chairs, and tables. Animal classes also include images of monkeys, birds, dogs and human faces. All the images, except face images, are collected from the ImageNet dataset (Deng et al., 2009); face images are selected from the Caltech database (Griffin et al., 2007). In addition, face images are divided into two classes, i.e., male and female, as well as images of dogs which are divided into two classes (different breeds) to study the subordinate level classification.

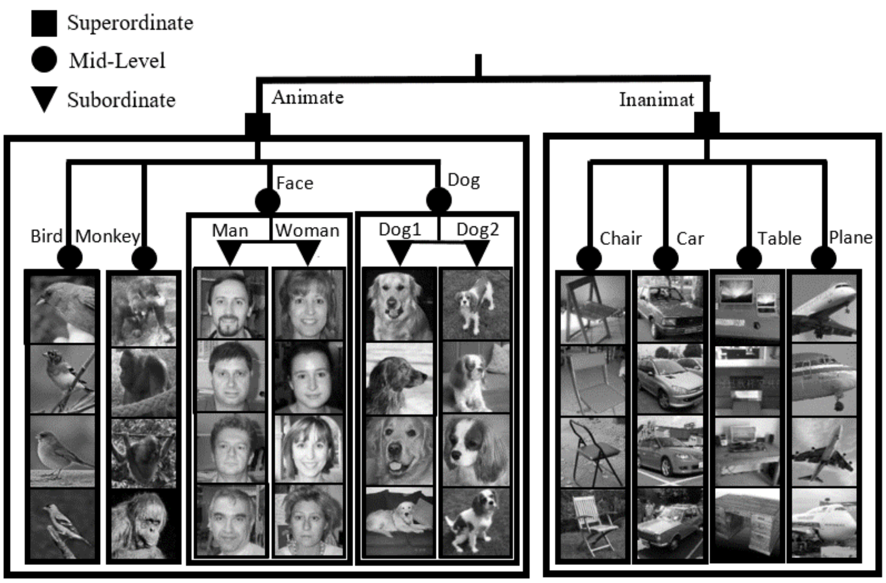

Figure (4) Hierarchical category structure of the dataset (2). Three category levels are defined: 1) superordinate level: animate vs. inanimate; 2) midlevel (basic level): face vs. body, bird vs. monkey, chair vs. table, car vs. plane; 3) subordinate level: man vs. woman, dog1 vs. dog2.

4- Results

The proposed model generates a temporal representation of the input stimulus in its feature representation stage. Once trained, neurons within this stage become selective to the features of the car and face in the input image. The presence of these features in the input stimulus causes more activation in the corresponding neurons. This will result in increasing the probability of generating spikes and stimulating next layers. Face (car) related spikes represented in the neural activity of the face-selective (or car-selective) neurons are then accumulated over time by an accumulator in the decision-making stage. Upon reaching a decision threshold, the model decides whether the input image is a face or a car. In the following, we first, show the firing pattern of the face and car selective neurons in response to either of these image categories. Then, we compare the speed and the accuracy of making a decision in the proposed model with what we expect from previous studies. Afterward, we study the role of decision bound in the model and compare it with its reported role in the brain. Finally, we focus on the temporal representation in the model and compare it with the representation in the inferotemporal cortex reported in previous electrophysiological studies.

4-1 The proposed model generates stimulus-difficulty and category-dependent spike trains.

Figure (5) shows spikes trains in the second stage of the model, replicating the behavior of the ventral stream in the brain. There are 20 neurons (Freiwald and Tsao, 2010) in the output of the second stage of the proposed model which represent neurons in the inferotemporal cortex (IT). Among them, 4 neurons for the face class and 4 neurons for the car class are selected which are called category-selective neurons (face-selective or car-selective neurons). Here, we use images with 20% strength in panels A and B and images with 80% strength in panels C and D. In panels A and C, the input images are face and, in panels B and D, the input images are car. As shown in this Figure, not only the number of spikes but also the timing of the spikes is being affected by the category and the strength of the input image. Neurons' responses reach their peak activity significantly faster, when the input image is their preferred one (face image for face neurons and car image for car neurons) compared to non-preferred input image (t-test, t=-18.9623, p-value=3.35e-60). In other words, face-selective neurons generate more and faster spikes in response to face images than the car-selective ones (panels A and C). It is similar for car-selective neurons in comparison with the face-selective ones when the input image is the car (panels B and D). In addition, both the number of spikes (Eq. (8), β_1=0.1905, p-value=3.46e-174) and the time of the peak response (Eq. (9), α_1=-0.0623, p-value=3.36e-32) are significantly modulated with the strength of the stimulus. These results are consistent with the previous findings (Afraz et al., 2006, Emadi and Esteky, 2013), which showed that stimulus strength and the category of the input image (preferred or non- preferred) affect the timing and the number of spikes in category selective neurons.

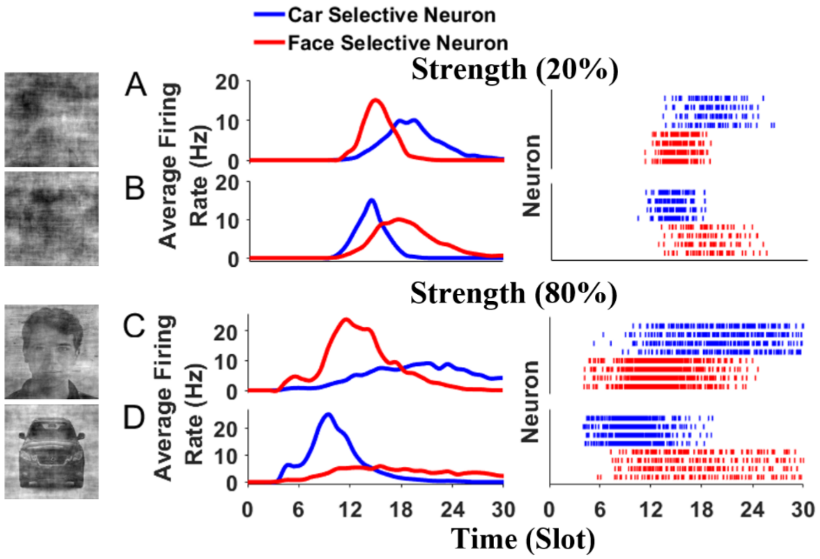

Figure (5): Raster plot and averaging firing rate of the car (blue) and face (red) selective neurons. Images with 20% strength in panels A and B and images with 80% strength in panels C and D are used. In panels A and C, we give the images of the face and, in panels B and D, we submit the images of the car to the model.

4-2 Consistent with behavioral studies, the model decides faster and more accurately as the signal-to-noise ratio increases in the input stimulus.

In order to investigate how well our proposed temporal object recognition model decides about the category of the input natural stimulus with different levels of difficulty, we use images in a dataset (1) with different noise levels (called stimulus strength). Then, we calculate the probability of correct decision as well as the average response time at each noise level. We also use logistic regression and hyperbolic tangent function as well-known psychometric and chronometric functions in behavioral studies (Shadlen et al., 2006) to investigate how well they can explain model's behavior.

Figure (6) panels A, B shows the recognition accuracy (performance) and response time of the model at 6 different noise levels in the input image (indicated on the x-axis) for different decision thresholds (indicated by different colors). The model, consistent with human behavior (Heidari-Gorji et al., 2021, Grill-Spector and Kanwisher, 2005) responds faster (Eq. (7), β_1=5.7477 p-value=7.49e-05) and more accurately (Eq. (6), β_1=0.0684, p-value=1.67e-07) as the stimulus signal to noise ratio increases. For example, when the threshold is 20 and the stimulus is the strongest one (red line, stimulus strength = 100%), the model can decide whether the stimulus is face or car with 100% accuracy (Figure (6) panel A red curve). This decision is made faster than decisions about weaker stimulus (Figure (6) panel B red curve), indicating that the input image provides high-frequency features which are very close to the previously learned features. In the weakest stimulus, the model spends more time observing the input image (receiving more time slots) and thus, makes its decision based on more information.

We also do same analysis on two well-known non-spiking deep neural networks, VGG-16 and ResNet-50, using a softmax on the last layer as a decision-making mechanism (Figure (6) panels E and F). In order to extract reaction time, like Spoerer et al. (Spoerer et al., 2019), we assume that the decision uncertainty reported by softmax represents reaction time. Models are pre-trained (see Spoerer et al. (Spoerer et al., 2019) for details) and then fine-tuned on the face-car dataset. It is important to note that these models (similar to our proposed model) are fed only with the strongest stimulus strength (100%) during fine-tuning phase. Although decision certainty/reaction time and the accuracy of non-spiking deep models differs as a function of signal to noise ratio of the input image, as shown in the Figure (6) panels E and F, they are more shaping a step function instead of a gradual changing from less-certain/slow to high-certain/fast classification.

4-3 The decision bound controls speed-accuracy regimes in a way that the brain does.

Logistic and hyperbolic tangent functions are widely used as psychometric and chronometric functions in behavioral studies (Spoerer et al., 2019, Bogacz et al., 2010). These functions can explain more than 99% of the variance of the accuracy (R2~0.99 for threshold=30, black curve in Figure (6) panel A) and more than 94% of the variance of the reaction time (R2~0.94 for threshold=30, black curve in Figure (6) panel B) of the proposed model. Decreasing the decision bound (threshold=20 Figure (6) panels A and B red curve) the variance of the accuracy explained by the behavioral psychometric function is still higher than 90% (R2~0.94 for threshold=20, red curve in Figure (6) panel A). However, explained reaction time variance by the behavioral chronometric function reduce to 73% (R2~0.73 for threshold=20, red curve in Figure (6) panel B). This reduction is mainly because of the reduction in the reaction time variance for low threshold. Overall, in all cases at least more than 73% of the variances generated by the model can be explained by behavioral psychometric and chronometric functions. This result provide support in favor of the capability of the model to generate reaction times and choices in a way that brain does. Moreover, changing the decision bound in the model affect both the reaction time (t-test, t=-10.88, p-value=1.12e-22) accuracy (t-test, t=-4.15, p-value=1.06e-04). Importantly, it differently affects the reaction time in different stimulus strengths (Eq. (10) (interaction between strength and threshold) β_2=0.4162, p-value=5.41e-11). The reaction time in weak stimuli is being more affected by the change in the decision bound compared to the strong stimuli. This effect is qualitatively consistent with what previous studies have shown (Figure (6) panels C and D). Figure (6) panels c and d adapted from (Hanks et al., 2014) shows how the behavior of a monkey in different speed-accuracy trade-off regimes alters during a perceptual decision-making task.

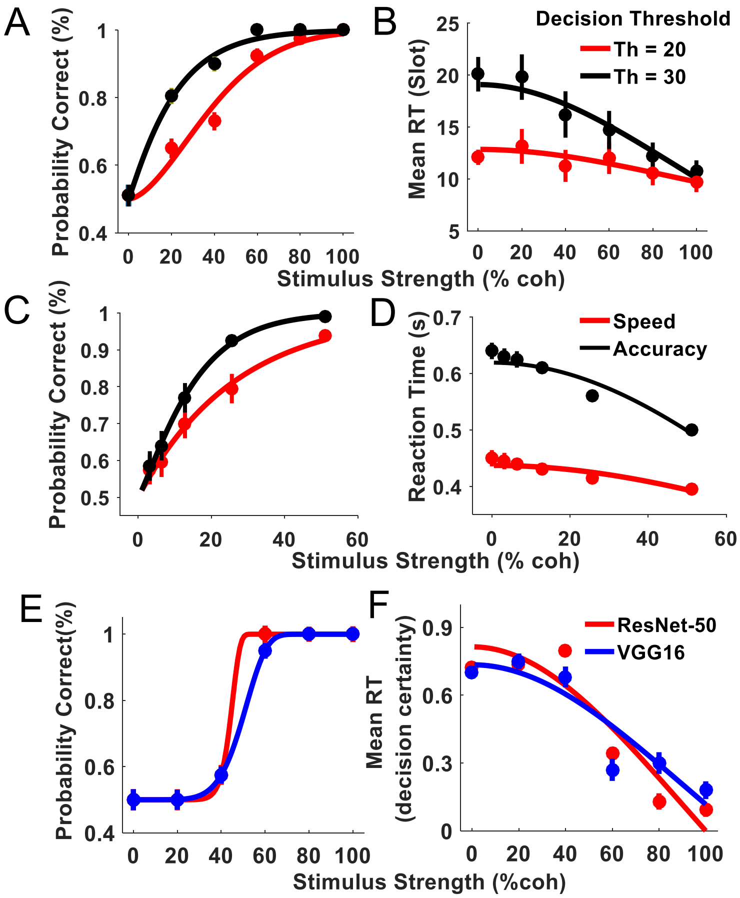

Figure (6) Performance and reaction time of monkeys, proposed model and non-spiking deep neural networks in the face-car categorization task. (A) demonstrates recognition accuracy (probability correct) as a function of stimulus difficulties for the six different thresholds. Lines are the fit of the logistic regression (see method) (B) displays the reaction time (number of slots) of the proposed model in terms of the different difficulties of the input image. Lines are the fit of a hyperbolic tangent function (see methods). As the difficulty level of the images decreases, the accuracy and recognition speed of the model increase. (C and D) are adapted from (Hanks et al., 2014) and show probability correct (C) and reaction time (D) in two different regimes of speed-accuracy trade-off during a reaction time task with a random dot motion stimulus. Data points in these panels are generated based on Figure (2) panel A in (Hanks et al., 2014). We use the same functions as in panels A and B for curve fitting. (E and F) performance (E) and reaction time (F) for two non-spiking deep neural networks (VGG16 blue , ResNet-50 red). Lines are fitted of same functions as A and B.

4-4 Spikes generated at different times in the model convey relevant, but not redundant, evidence about the input category.

In order to investigate whether generated spikes at different points at time convey informative evidence about the stimulus category, we examine the speed-accuracy trade-off in the model. As shown in Figure (7) panel A as the threshold increases, the accuracy of the model increases while the speed decreases. This shows the decision bound in the proposed model is truly playing its expected role (Wenzlaff et al., 2011, Van den Berg et al., 2016, Drugowitsch et al., 2012) and, more importantly, generated spikes in the second stage of the model are temporally informative and decorrelate in a way that accumulating more spikes results in improving the accuracy. Considering that this effect is consistent in all stimulus strength and given that the model in its decision-making stage accumulates spikes which are generated in the previous layer (the last layer of the second stage), we conclude these spikes at each time point (momentary evidence) not only contain information about the class of the input stimulus but also encode the level of uncertainty at that input.

In addition, increasing the threshold changes the decision time as well. As shown in Figure (7) panel B, the decision time of the model ranges from a very fast response (for low decision threshold) to a very slow response where the model should wait for the whole information in the input stimulus. Thus, even those spikes that are generated very late (which are about low-frequency features and not very similar to the patterns that the model learned) are important to improve the accuracy, independent of the stimulus strength.

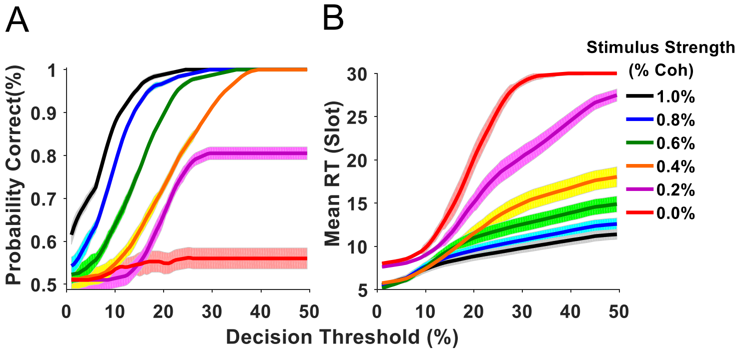

Figure (7) panel A shows the relationship between model performance and different decision bounds for different stimulus difficulties. Panel B shows the relationship between reaction time and different decision thresholds for different stimulus difficulties.

4-5 Temporal representation of semantic information in the model resembles neurophysiological findings.

The temporal representation in the model is based on a simple assumption that different timings of spikes are mainly due to the different contrast levels in the stimulus and the similarity of the extracted features with the pre-learned ones. However, one may speculate that these assumptions are enough for the model to generate the observed results on the stimulus with artificial noise levels. The second experiment aims to investigate whether the model can explain the temporal representation of the semantic information (not being manipulated by artificial noise) in the brain. Thus, we compare the temporal representation of different levels of categorization on natural images in the model with that of the brain (Dehaqani et al., 2016).

We use two different analyses to compare the temporal representation of semantic information in the model with that of the brain when both receive natural images as the input stimulus (this experiment is performed on the dataset (2)). As shown previously by Dehaghani et al. (Dehaqani et al., 2016), different levels of categorization such as superordinate, midlevel and subordinate are represented at different times in the brain.

In the first analysis, to better visualize the model outputs (like (Dehaqani et al., 2016), Figure (4) panel A), we represent the features in two dimensional space using PCA. Then, we employ these two dimensions as the coordinates to show each image in Figure (8) panel A (the surrounding scatter plots). This analysis demonstrates very detailed separation between images in different periods of time. Four circles surrounding the SI diagram stand for four time periods: 10, 16, 20 and 30. Colored squares (each color is specific to a class) are used when images are in one of the four animal classes (monkey, bird, dog and human face), and colored circles (each color is used for a class) are used when images are in one of the four non-animal classes (car, plane, chair and table). According to Figure (8) panel A, by the time point of 10, the images are very close to each other, and no separation takes place. The separation of squares and circles (superordinate level) continues until the time point 20; however, images in the midlevel category are again getting closer to each other after the time point 16.

The second analysis aims to investigate how the proposed model represents different images at different levels of categorization. To do so, the output of the model is separated based on the class of the corresponding input image. Then, the representation similarity between the classes is calculated in pairs (there are 8 image classes with total of 28 similarities extracted in pairs for midlevel, 16 for superordinate and 2 for subordinate). We employ the similarity index (SI) as (Dehaqani et al., 2016) to show the extent, to which the representation of different classes is separated from each other. The SI values are greater when there is more similarity within classes and less similarity between classes. The average SI at each level of abstraction is considered as the SI value for that level. As an example, in order to calculate the SI for superordinate level classification, we average all the SIs calculated for the superordinate level of each class (one of which is animate and another is inanimate). Similarly, the SI for midlevel classification is the average of all SIs in each midlevel class (the difference between subordinate level classes, such as men and women, is not considered here). Finally, for the subordinate level, the SI is averaged for men-women classes and for dog1-dog2 classes.

SI diagram in Figure (8) panel B shows the dynamics of separability index of features represented in the feature representation stage of the model which stands for inferotemporal (IT) cortex in the brain. In each time slot, using t-test, we showed whether the SI value is significant or not (p-value<0.0016 Bonferroni corrected). Peak latency is defined as the time that the SI exceeds 90% of its maximum value. Similar to what Dehaghani et al. (6) and Cichy et al. (Cichy et al., 2014) have reported, we observed that neurons in the last layer of the model represent mid-level categories (e.g., human faces) earlier than the superordinate level (e.g., animal) categories (Figure (8) panel C, p-value=0.0016). The model separates the midlevel with a higher SI than other categories (t-test, t=-3.1484, p-value=3.71e-07), which is again consistent with what is observed in IT neurons (Dehaqani et al., 2016). It is noticeable that the model has no special mechanism or training procedure for processing different levels of categorization.

Importantly, we show that the observed temporal pattern of the separability index (SI) doesn’t exist in the first layer’s representation (Figure (8) panel D). To do so, we calculate SI in a similar way for the output of the DOG filter where the contrast information is represented. As shown in Figure (8) panel D contrast representations cannot shape a significant non-zero separability index (p-value<0.0016 Bonferroni corrected).

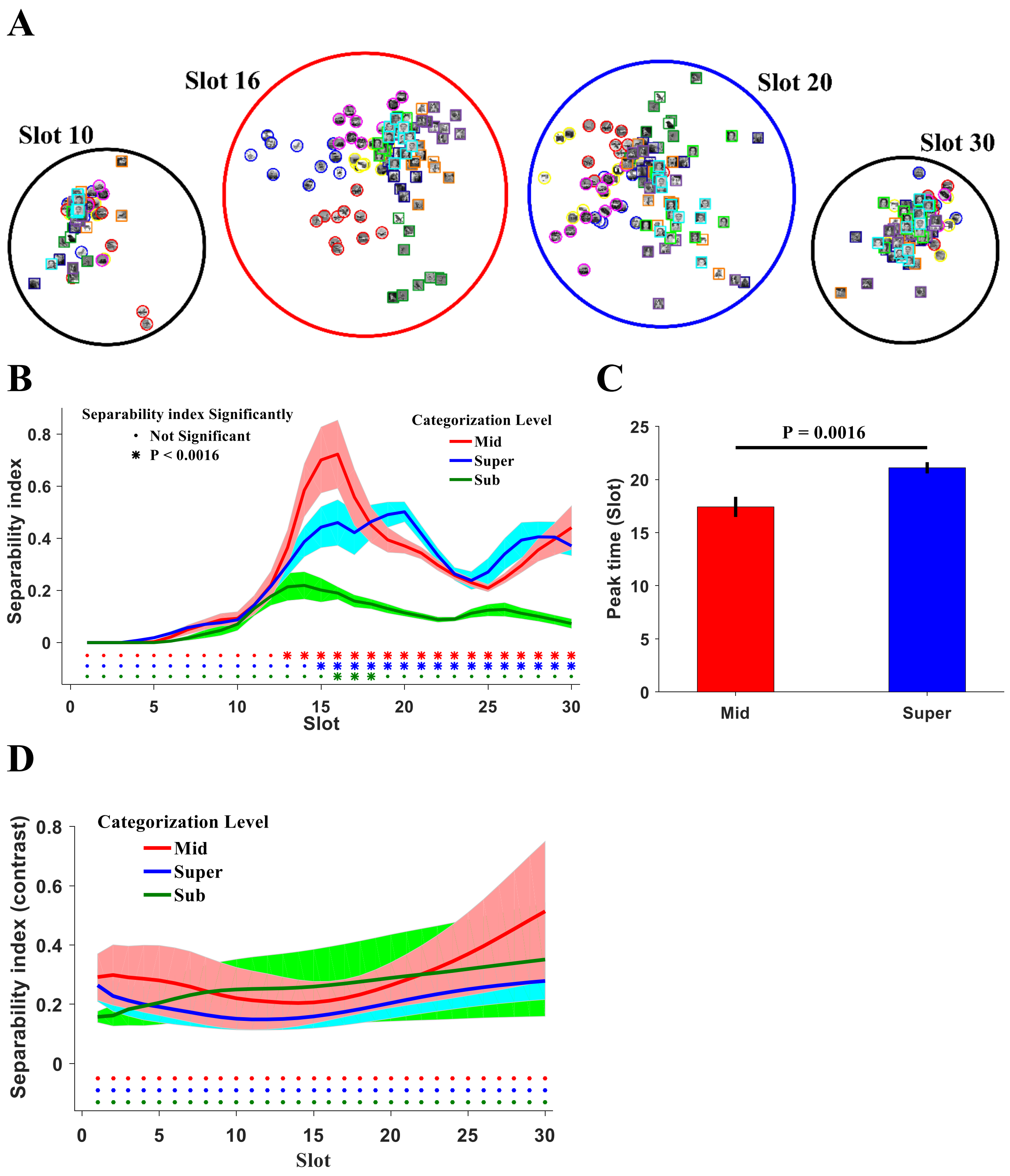

Figure (8) panel A representation of different categories in PCA (first two components) space in different time slots. Panel B Separability index of different levels of abstraction over time. Different symbols are used to show the significance level. Dots show non-significant SI value (compared to zero) stars show SI which are significantly differ from zero with (p-value<0.0016). Shaded areas show standard error. Panel C peak latencies of SI for mid and super ordinate level representation. Peak latency is defined as the time that the index exceeds 90% of its maximum value. Error bars are standard error. Panel D separability index for the output of the DOG filter where the contrast information is represented.

4-6 Dynamical representational dissimilarity matrix (DRDM)

Representational dissimilarity matrix (RDM) is widely used as a method to compare representation in the brain with that of the computational models (Kriegeskorte et al., 2008). However, as shown in previous studies (Dehaqani et al., 2016, Cichy et al., 2014, Kar and DiCarlo, 2021, Kar et al., 2019), the representation of input stimulus in the brain is not static. As the result of temporal coding in the brain, representations at different time points convey different information about the sensory stimulus. In order to make a reasonable comparison between models and the brain behavior in terms of representations, we suggest a DRDM, instead of a static RDM, in a way the dissimilarity matrix is calculated at different time points. Since the proposed model temporally encodes the input stimulus in its feature representation stage (second stage), we can compare DRDM with that of the brain reported in the literature. Each row (i) and column (j) of the matrix stand for a single image and each matrix stands for a single point at a time (k). To calculate this matrix, the correlation between the spike's vectors which are generated in the last layer of the deep spiking network at time k for images i and j is calculated. Then, the dissimilarity value is 1-correlation. This diagram illustrates the within-class similarities and between-class differences presented in Figure (9) panel A at different time points (16 and 20). Before time point 16, the matrix is very sparse and, thus, they are not plotted here. As illustrated in Figure (9) panel A, at time point 16, the classes are well separated. In this time step, blue squares are formed around the diameter, indicating the similarity of the features within the class, and values outside the original diameter show the degree of greater dissimilarity. There are also two larger squares at the top of the figure which belong to the class of human faces (containing male and female images related to the subordinate level of abstraction) and the dog image class (containing two different breeds of dogs related to the subordinate level of abstraction). The formation of these two large squares indicates that, at this time point, subordinate level separation has not yet occurred well. Also in time step 20, the within-class differences of animate and inanimate categories are being vanished, while between-class differences are being appeared. Therefore, as expected, the information represented in this time step provides better separation for the superordinate abstraction level. In Figure (9) panel B, there are three abstract true matrices of mid-level, super-level and sub-level. These matrices are formed in an ideal representation. In Figure (9) panel C, using the kendall τ_A calculation (Khaligh-Razavi and Kriegeskorte, 2014), the correlation between DRDM matrices and True matrices is shown which is similar to previous finding, slots 16 and 20 have the highest τ_A for mid-level and super-level respectively, indicating that representation in these time slots are more similar to the true representation matrices.

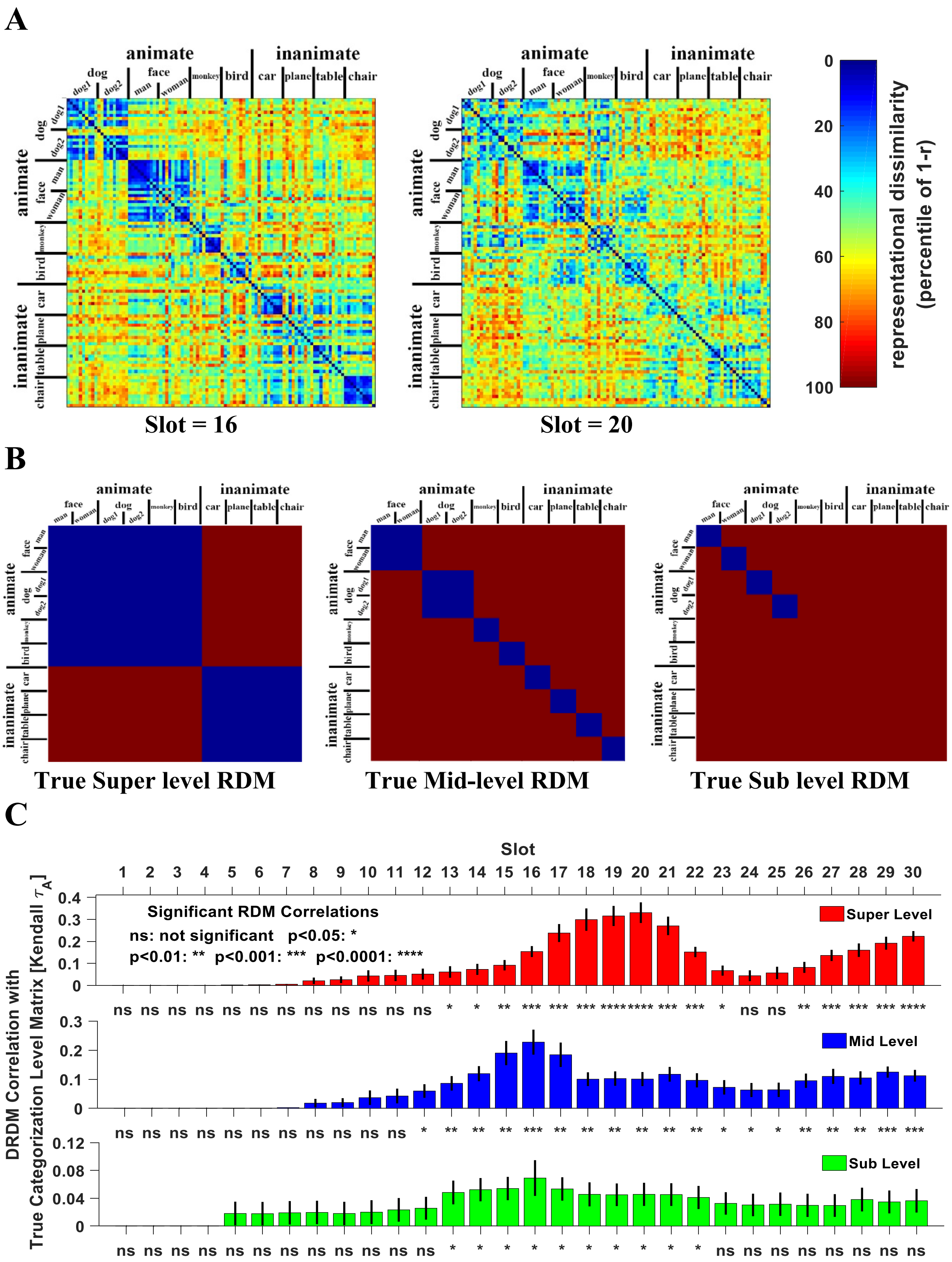
A
animate
inanimate
dog
dog1
dog2
face
man
woman
monkey
bird
car
plane
table
chair
Slot = 16
Slot = 20
0
20
40
60
80
100
representational dissimilarity
(percentile of 1-r)
B
True Super level RDM
True Mid-level RDM
True Sub level RDM
C
Slot
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
DRDM Correlation with
True Categorization Level Matrix [Kendall τ_A]
Significant RDM Correlations
ns: not significant p<0.05: *
p<0.01: ** p<0.001: *** p<0.0001: ****
Super Level
0.4
0.3
0.2
0.1
0
ns ns ns ns ns ns ns ns ns ns ns ns * * ** *** *** *** ******** *** *** * ns ns ** *** *** *** ****
Mid Level
0.3
0.2
0.1
0
ns ns ns ns ns ns ns ns ns ns ns * ** ** ** *** ** ** ** ** ** ** * * * ** ** ** *** ***
Sub Level
0.12
0.08
0.04
0
ns ns ns ns ns ns ns ns ns ns ns ns * * * * * * * * * * ns ns ns ns ns ns ns ns

Figure (9) panel A Dynamic Representational Dissimilarity Matrix (DRDM) for dataset (2) in 2-time steps. Panel B Three abstract true matrices of mid-level, super-level and sub-level. Panel C Correlation between DRDM matrices and True matrices using Kendall's τ_A. Error bars indicate the standard errors. Asterisks indicate significant RDM correlations (ns: not significant, p<0.05: *, p<0.01: **, p<0.001: ***, p<0.0001: ****).

5- Discussion

Object recognition and making a choice regarding the recognized object are important and vital abilities of the brain, either in humans (Aboudib et al., 2016, Andersen, 1995) or in much less developed creatures, like rodents (Bogacz et al., 2010, Dehaqani et al., 2016). The processes of object recognition and deciding about that in both representation (Delorme et al., 2010) and decision-making (Deng et al., 2009) parts take different times for different objects. For example, representation of ambiguous objects in the IT cortex (Farzmahdi et al., 2016) and the decisions to recognize them are slower and less accurate than those which are less ambiguous (Fukushima and Miyake, 1982, Ghodrati et al., 2014). Importantly, this speed and the accuracy can be adjusted via the decision process in the brain. Our brain, especially in those situations where the accurate choice is more rewarding than faster one, spends more time and collects more information in order to increase the probability of making a correct choice. On the other hand, in those situations when the speed is more rewarding (for example when we are doing a project very close to a deadline), we will make faster choices with the cost of being less accurate. Thus, this is not only the stimulus which determines our choice, but the speed-accuracy trade off regime of our decision process can also affect it. As a result, we believe that a reliable explanation for the underlying mechanisms of object recognition in the brain should be able to explain the choice considering the speed-accuracy trade off regimes as well as the temporal representation of input stimulus, otherwise we cannot distinguish between mechanisms that make similar choices but with different speeds.

In this work, we proposed a temporal feedforward model, explaining the process of object recognition in the brain in different speed-accuracy trade-off regimes. To this end, we consider not only the structure and the function of ventral stream as the brain areas which represent input stimulus, but also the lateral intraparietal cortex, which is accounted for making decision in different speed-accuracy trade-off regimes.

We proposed a spiking model for object recognition, in which the input image was represented in spike trains. Then generated spikes were accumulated toward decision bounds. The model consisted of three stages: neurons in the first stage, like V1 neurons in the occipital cortex, were selective to edges in different orientations. The timing of spikes generated in this layer depended on the contrast of the stimulus in the receptive field of the neuron. In the middle stage of the model, neurons were being selective to more complex shapes. In a deep structure spikes at the late layer of the model were generated in favor of different pre-defined categories in the input stimulus. Finally, in the last stage -the decision-making stage- these generated spikes were being accumulated in accumulator units until reaching a threshold. The model decided about the category of the input image as soon as any of the accumulators reached their threshold.

Multiple experiments showed that the proposed model can replicate what was expected from humans or animals while making decisions about an object. Firstly, we showed that the speed and the accuracy of the model in an abject recognition task with noisy stimuli followed a tangent hyperbolic chronometric and a sigmoid psychometric function, respectively in keeping with various studies of perceptual decision-making (57,58)(Heidari Gorji et al., 2018, Okazawa et al., 2020, Okazawa et al., 2021). We demonstrated that the decision bound in the decision-making stage of the model truly adjusted the speed and the accuracy of a decision, in a way that has been observed in behavioral studies (Hanks et al., 2014). The fact that the probability of correct choices increases if the decision stage of the model accumulates more information over time (higher decision bound) indicates that the represented temporal evidence in the second stage of the model is not redundant.

One may speculate that the above observations stem from the artificial noise we used in our images. Achieving a higher reaction time and lower performance for noisier stimulus might not be the result of a brain plausible temporal feature extractor, but the result of the way that we produce noise in the input image. In the next experiment, which is an important analysis in our study, we did not control the strength of the input evidence, instead, we used a set of different natural images in different categories (which is widely used in previous neurophysiological and behavioral studies) and compared the temporal representation of different abstract levels of categorization (superordinate, subordinate, and midlevel) with that in the brain. Results were consistent with those of the previous studies (Dehaqani et al., 2016, Cichy et al., 2014) in terms of the average peak and average strength of the decoding over time (Figure (8)). Moreover, the DRDM analysis represented that how single images within an abstract level of categorization were similarly represented over time. Thus, the second experiment, provided stronger evidence about the plausibility of temporal representation in the proposed model compared to previous models (Heidari Gorji et al., 2018, Kiani et al., 2013, Spoerer et al., 2019). Thus, we believe the model is a good candidate for delving deeper into the underlying mechanisms of core object recognition in the brain through computational modeling.

Bearing the structure of the model in mind, we may conclude that the representation dynamics in the brain in a rapid feedforward categorization task stems from three factors: (i) the contrast levels existing in the input image, (ii) the similarity of extracted features with those of pre-learned ones, and (iii) the decision bound which controls the speed and the accuracy trade-off. All these factors can affect reaction times when making a choice about a natural stimulus. Although these factors have been discussed separately in different studies, we showed that all should be considered to explain the object representation, the time, and the accuracy of object recognition in the brain.

According to (Macé et al., 2009, Dehaqani et al., 2016, Mack and Palmeri, 2015), classification at the midlevel was faster than the superordinate and subordinate levels because of the more within-class similarity and more between-class dissimilarity of features at the midlevel than the sub and superordinate levels. Results of our model in line with this hypothesis suggested that faster separation of the midlevel was not the effect of specific circuits or connections; instead, this is due to the contrast of different parts of the images and the more similarity between the learned and the presented features.

It is worth discussing that classic models of core object recognition in the brain neither have a mechanism for temporal representation (Riesenhuber and Poggio, 1999) nor for speed-accuracy trade-off. Recent studies give the role of temporal representation to recurrent connections (Spoerer et al., 2019, Kietzmann et al., 2019, Kar et al., 2019, Nayebi et al., 2022, Mirzaei et al., 2013) and leave the question of underlying mechanism of temporal dynamics in feedforward processing open. On the other hand, most of these models still suffer from the lack of a speed-accuracy trade-off mechanism and those which proposed such mechanism (Spoerer et al., 2019, Nayebi et al., 2022) are less plausible in terms of biological evidence (see below). In contrast, making a bridge between decision science and the science of object recognition, our model explains not only dynamics of making decisions about objects but also dynamics of representation of abstract information. The later one is less being discussed in previous studies. Finally, as shown by Wang 2002 (and later works in his lab), a neuronal implementation of the accumulation to bound model is a recurrent spiking network. Thus, in terms of the necessity of recurrent connections, our results are in line with (Spoerer et al., 2019, Kietzmann et al., 2019, Kar et al., 2019, Nayebi et al., 2022, Mirzaei et al., 2013). However, we are showing that these recurrent connections are not necessarily needed at the representation level but may be at the decision-making level.

Spoerer et al. (Spoerer et al., 2019), assume that reaction time is the time that the entropy of the model's posterior reaches a threshold. Thresholding on the posterior entropy as a mechanism of decision making is less evident in the literature compared to thresholding on the accumulated momentary evidence (which is used in our model). Although both may reflect a same concept, our proposed implementation of decision-making layer has much stronger biological supports (Shadlen and Kiani (Kiani et al., 2013)). Moreover, representation of reaction time in Spoerer et al. (Spoerer et al., 2019) model suffers from some minor issues which are not the case for our model: 1-Setting the threshold on the entropy (uncertainty) will result in a reverse relation between threshold level and decision accuracy/time; i.e. lower thresholds cause higher accuracy/reaction time. However, it has been shown that decision threshold has a direct relation with accuracy/reaction time; i.e. higher threshold results in higher accuracy/reaction time (Hanks et al. (Hanks et al., 2014)). 2- In order to calculate the entropy brain needs another neural mechanism which is not explained in the Spoerer et al. (Spoerer et al., 2019), however the neural implementation of the accumulation of momentary evidence in our model is well-known (Wang (Wang, 2002)). 3-Spoere et al. (Spoerer et al., 2019), assume that reaction time is the same as certainty, however although decision certainty is informed by reaction time, these two are not the same (Kiani et al. (Kiani et al., 2014)). In the bounded accumulation of momentary evidence model, which is used in our study, decision certainty can be represented differently from reaction time. In this model decision certainty represented by the distance of the looser accumulator from the decision bound while reaction time is the time that the winner accumulator spend to reach the decision bound.

We compared our spiking feedforward model with two non-spiking feedforward deep models (VGG-16 and ResNet-50 used in Spoerer et al. (Spoerer et al., 2019)) in shaping accuracy and reaction time as a function of signal to noise ratio (Figure (6) panels E and F). These models lack both temporal representation and speed-accuracy trade-off mechanisms. However, we can assume that decision certainty in these models can compensate for the lack of the first mechanism (i.e. temporal representation). In a way that, a more certain decision implies that stronger information was represented over time (Spoerer et al. (Spoerer et al., 2019)) and thus faster decision is made. Yet, the lack of the second mechanism (i.e. speed-accuracy trade-off) cannot be compensated because there is no potential mechanism in these models for changing speed-accuracy trade-off. As a result, we can expect that, represented certainty in these models can explain some variations of reaction times but not all. As shown in Figure (6), they are more shaping a step function instead of a gradual changing from less-certain/slow to high-certain/fast classification. This effect is in line with our expectation that these models can only explain a part of variance in reaction times.

Finally, we believe that the present work provides a new perspective for understanding the underlying mechanism of object recognition in the brain in three ways: 1- We proposed a feedforward computational model which provides a dynamical representation of a static input image. 2- The model makes decision based on a biologically plausible decision-making model, and thus can replicate neural and behavioral responses in different speed-accuracy trade-off regimes. 3- The proposed model explains temporal advantages of different abstract levels' representation in the brain.

We also conducted another analysis to compare the results of Grill-Spector and Kanwisher (Grill-Spector and Kanwisher, 2005) and Mack and Palmeri (Mack and Palmeri, 2015) with the behavior of the model in timing of the levels of the categorization. Grill-Spector and Kanwisher (Grill-Spector and Kanwisher, 2005) showed that a comparable performance on identification task (sub-ordinate level categorization) requires substantially more processing time compared to detection and categorization in the midlevel. Mack and Palmeri (Mack and Palmeri, 2015), in support of the previous separated works, systematically evaluated the effect of exposure time on the categorization level advantages. They showed that mid-level advantage appears only when the exposure time is less limited. We used the support vector machine (SVM) to classify the temporally represented information in different layers of the model in each time slot. The results are illustrated in Figure (10). As shown in this Figure the 3rd convolution layer, reaching to the performance of 80% for mid-level categorization takes around 10 time slots, while it takes 20 time slots for the sub-ordinate categorization to reach the same level of accuracy. This is similar to Grill-Spector and Kanwisher (Grill-Spector and Kanwisher, 2005). However, the model never shows an advantage of super-ordinate over mid-level in last representation layer (i.e. 3rd convolution layer), but this is to somehow evident in 2nd convolution layer where there are super-level advantages for lower exposure duration. Although making a strong conclusion in favor of these hypothesizes needs a systematic design of experiment for the model, these analyses show the potential of the proposed model to investigate different hypothesis in the field of object recognition and decision making especially those that are dealing with the recognition time and the speed-accuracy trade-off. It should be noted that in this experiment, we used images in dataset II (the train and the test set is similar to those we used for training and testing the proposed model). There are 8 image classes with a total of 28 pairs for midlevel, 16 pairs for superordinate and 2 pairs for subordinate. In each level of categorization, the average accuracy of these pairs is calculated as the performance in that level.

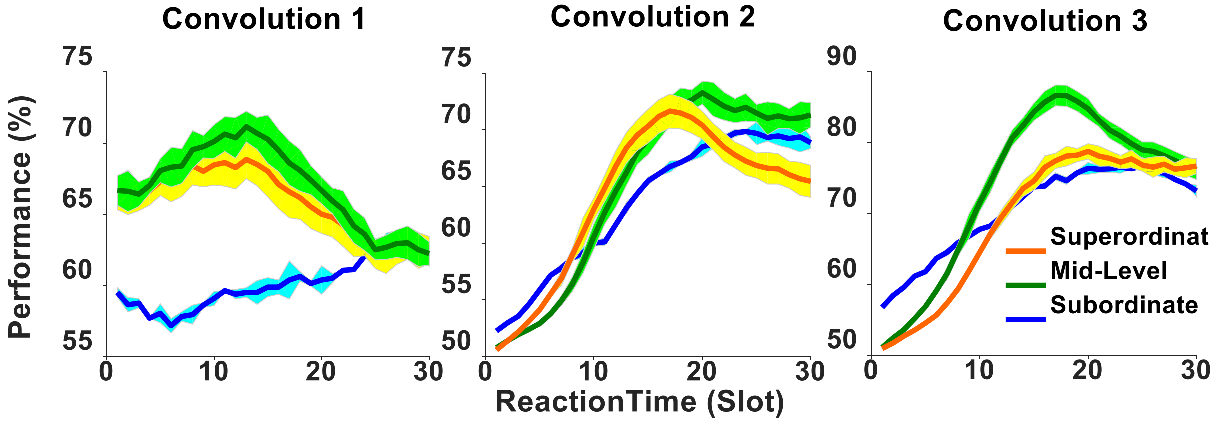

Figure (10). Performance of SVM classifier on different convolution layers of the model over time. The SVM was trained separately on each time slot and convolution layer to perform on superordinate, mid-level and subordinate categorization task. Each trained SVM is used to classify new images (test set) in the corresponding time slot and the convolution layer.

It is important to note that the existing computational models of core feedforward object recognition have usually ignored the brain's dynamics in the representation and decision-making parts (Gold and Shadlen, 2007, Hanks et al., 2014, He et al., 2016, Heekeren et al., 2004, Heidari-Gorji et al., 2015, Hubel and Wiesel, 1968, Kanwisher et al., 1997, Karimi-Rouzbahani et al., 2017, Kheradpisheh et al., 2016a), such that most of them are only capable of following either the accuracy (Gold and Shadlen, 2007, Hanks et al., 2014, He et al., 2016, Heidari-Gorji et al., 2015, Kiani et al., 2013, Hubel and Wiesel, 1968, Kheradpisheh et al., 2016a) or the reaction time (Kiani et al., 2013) of the recognition. The structure of the proposed model in its second stage was similar to the one proposed by Kheradpishe et al. (Kheradpisheh et al., 2018). However, there were some fundamental differences in our model borrowed from biological studies. Kheradpishe et al. (Kheradpisheh et al., 2018) used a maximization layer as the last layer of the feature representation stage, assuming that the strongest evidence was enough for deciding about the input object. However, using this layer, they eliminated other sources of information and only used the strongest one. On the contrary, in our proposed model, the generated spikes of all the neurons were being accumulated in the decision-making stage and, thus, all sources of information were being used for the categorization task. Moreover, in contrast to Kheradpishe et al. (Kheradpisheh et al., 2018), neurons in our proposed model were being reset after generating a spike and could fire again if their potential reached a threshold again. More importantly, in order to shape a temporal tuning curve for each neuron, we decreased the threshold of neurons after being trained. Therefore, the number of spikes generated by a neuron depends on the similarity of the input with its preferred feature.

Limitations of the model are understood by considering to the model structure. i) The model is a feedforward spiking implementation of the visual system. Due to lack of any feedback path, it is not expected to mimic the visual system in variety of object recognition problems. ii) In addition to the trainable parameters of the SNN, there are several parameters (such as number of time slot, number of filters and decision bound) which should be set to achieve the best performance. Nevertheless, these parameters improve the degree of the freedom of the model. iii) Due to the learning strategy, the proposed model can't be applied to online application especially on 3D or 4D images.

Code availability. The custom code for data analysis and models is available upon request from the corresponding author.

Acknowledgement

The authors acknowledge the funding support of Babol Noshirvani University of Technology through Grant program No. BNUT/389079/ 1400-1.

References

ABOUDIB, A., GRIPON, V. & COPPIN, G. 2016. A biologically inspired framework for visual information processing and an application on modeling bottom-up visual attention. *Cognitive computation,* 8**,** 1007-1026.

AFRAZ, S.-R., KIANI, R. & ESTEKY, H. 2006. Microstimulation of inferotemporal cortex influences face categorization. *Nature,* 442**,** 692-695.

ANDERSEN, R. A. 1995. Encoding of intention and spatial location in the posterior parietal cortex. *Cerebral Cortex,* 5**,** 457-469.

BOGACZ, R., WAGENMAKERS, E.-J., FORSTMANN, B. U. & NIEUWENHUIS, S. 2010. The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences,* 33**,** 10-16.

CHANG, L. & TSAO, D. Y. 2017. The code for facial identity in the primate brain. *Cell,* 169**,** 1013-1028. e14.

CHITTKA, L., SKORUPSKI, P. & RAINE, N. E. 2009. Speed–accuracy tradeoffs in animal decision making. *Trends in ecology & evolution,* 24**,** 400-407.

CICHY, R. M., KHOSLA, A., PANTAZIS, D., TORRALBA, A. & OLIVA, A. 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports,* 6**,** 1-13.

CICHY, R. M., PANTAZIS, D. & OLIVA, A. 2014. Resolving human object recognition in space and time. *Nature neuroscience,* 17**,** 455.

CONTINI, E. W., WARDLE, S. G. & CARLSON, T. A. 2017. Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia,* 105**,** 165-176.

DEHAQANI, M.-R. A., VAHABIE, A.-H., KIANI, R., AHMADABADI, M. N., ARAABI, B. N. & ESTEKY, H. 2016. Temporal dynamics of visual category representation in the macaque inferior temporal cortex. *Journal of neurophysiology,* 116**,** 587-601.

DELORME, A., PERRINET, L. & THORPE, S. J. 2001. Networks of integrate-and-fire neurons using Rank Order Coding B: Spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomputing,* 38**,** 539-545.

DELORME, A., RICHARD, G. & FABRE-THORPE, M. 2010. Key visual features for rapid categorization of animals in natural scenes. *Frontiers in psychology,* 1**,** 21.

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. & FEI-FEI, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition, 2009. Ieee, 248-255.

DRUGOWITSCH, J., MORENO-BOTE, R., CHURCHLAND, A. K., SHADLEN, M. N. & POUGET, A. 2012. The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience,* 32**,** 3612-3628.

DUDA, R. O., HART, P. E. & STORK, D. G. 1973. *Pattern classification and scene analysis*, Wiley New York.

EMADI, N. & ESTEKY, H. 2013. Neural representation of ambiguous visual objects in the inferior temporal cortex. *PloS one,* 8**,** e76856.

FAHRENFORT, J. J., SCHOLTE, H. S. & LAMME, V. A. 2007. Masking disrupts reentrant processing in human visual cortex. *Journal of cognitive neuroscience,* 19**,** 1488-1497.

FARZMAHDI, A., RAJAEI, K., GHODRATI, M., EBRAHIMPOUR, R. & KHALIGH-RAZAVI, S.-M. 2016. A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Scientific reports,* 6**,** 1-17.

FREIWALD, W. A. & TSAO, D. Y. 2010. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science,* 330**,** 845-851.

FUKUSHIMA, K. & MIYAKE, S. 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. *Competition and cooperation in neural nets.* Springer.

GHODRATI, M., FARZMAHDI, A., RAJAEI, K., EBRAHIMPOUR, R. & KHALIGH-RAZAVI, S.-M. 2014. Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in computational neuroscience,* 8**,** 74.

GHODRATI, M., KHALIGH-RAZAVI, S.-M., EBRAHIMPOUR, R., RAJAEI, K. & POOYAN, M. 2012. How can selection of biologically inspired features improve the performance of a robust object recognition model? *PloS one,* 7**,** e32357.

GOLD, J. I. & SHADLEN, M. N. 2007. The neural basis of decision making. *Annual review of neuroscience,* 30.

GRIFFIN, G., HOLUB, A. & PERONA, P. 2007. Caltech-256 object category dataset.

GRILL-SPECTOR, K. & KANWISHER, N. 2005. Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science,* 16**,** 152-160.

HANKS, T., KIANI, R. & SHADLEN, M. N. 2014. A neural mechanism of speed-accuracy tradeoff in macaque area LIP. *Elife,* 3**,** e02260.

HE, K., ZHANG, X., REN, S. & SUN, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 770-778.

HEEKEREN, H. R., MARRETT, S., BANDETTINI, P. A. & UNGERLEIDER, L. G. 2004. A general mechanism for perceptual decision-making in the human brain. *Nature,* 431**,** 859-862.

HEIDARI-GORJI, H., EBRAHIMPOUR, R. & ZABBAH, S. 2021. A temporal hierarchical feedforward model explains both the time and the accuracy of object recognition. *Scientific reports,* 11**,** 1-12.

HEIDARI-GORJI, H., ZABBAH, S., AKHAVAN, M., BAGHERI, N. & EBRAHIMPOUR, R. STDP based HAMX behavior in response to homogeneous and heterogeneous categories. Bernstein Conference. Germany, 2015.

HEIDARI GORJI, H., ZABBAH, S. & EBRAHIMPOUR, R. 2018. A temporal neural network model for object recognition using a biologically plausible decision making layer. *arXiv e-prints***,** arXiv: 1806.09334.

HUBEL, D. H. & WIESEL, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology,* 195**,** 215-243.

ISIK, L., MEYERS, E. M., LEIBO, J. Z. & POGGIO, T. 2014. The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology,* 111**,** 91-102.

KANWISHER, N., MCDERMOTT, J. & CHUN, M. M. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience,* 17**,** 4302-4311.

KAR, K. & DICARLO, J. J. 2021. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron,* 109**,** 164-176. e5.

KAR, K., KUBILIUS, J., SCHMIDT, K., ISSA, E. B. & DICARLO, J. J. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience,* 22**,** 974-983.

KARIMI-ROUZBAHANI, H., BAGHERI, N. & EBRAHIMPOUR, R. 2017. Average activity, but not variability, is the dominant factor in the representation of object categories in the brain. *Neuroscience,* 346**,** 14-28.

KHALIGH-RAZAVI, S.-M. & KRIEGESKORTE, N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology,* 10**,** e1003915.

KHERADPISHEH, S. R., GANJTABESH, M. & MASQUELIER, T. 2016a. Bio-inspired unsupervised learning of visual features leads to robust invariant object recognition. *Neurocomputing,* 205**,** 382-392.

KHERADPISHEH, S. R., GANJTABESH, M., THORPE, S. J. & MASQUELIER, T. 2018. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Networks,* 99**,** 56-67.

KHERADPISHEH, S. R., GHODRATI, M., GANJTABESH, M. & MASQUELIER, T. 2016b. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports,* 6**,** 1-24.

KIANI, R., CHURCHLAND, A. K. & SHADLEN, M. N. 2013. Integration of direction cues is invariant to the temporal gap between them. *Journal of Neuroscience,* 33**,** 16483-16489.

KIANI, R., CORTHELL, L. & SHADLEN, M. N. 2014. Choice certainty is informed by both evidence and decision time. *Neuron,* 84**,** 1329-1342.

KIETZMANN, T. C., SPOERER, C. J., SöRENSEN, L. K., CICHY, R. M., HAUK, O. & KRIEGESKORTE, N. 2019. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences,* 116**,** 21854-21863.

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.

KRIEGESKORTE, N., MUR, M. & BANDETTINI, P. A. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience,* 2**,** 4.

KUFFLER, S. W. 1953. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology,* 16**,** 37-68.

MACE, M. J.-M., JOUBERT, O. R., NESPOULOUS, J.-L. & FABRE-THORPE, M. 2009. The time-course of visual categorizations: you spot the animal faster than the bird. *PloS one,* 4**,** e5927.

MACK, M. L. & PALMERI, T. J. 2015. The dynamics of categorization: Unraveling rapid categorization. *Journal of Experimental Psychology: General,* 144**,** 551.

MAJAJ, N. J., HONG, H., SOLOMON, E. A. & DICARLO, J. J. 2015. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience,* 35**,** 13402-13418.

MASQUELIER, T. & THORPE, S. J. 2007. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput Biol,* 3**,** e31.

MCCARTHY, T. V., HEALY, J., HEFFRON, J. J., LEHANE, M., DEUFEL, T., LEHMANN-HORN, F., FARRALL, M. & JOHNSON, K. 1990. Localization of the malignant hyperthermia susceptibility locus to human chromosome 19ql2–13.2. *Nature,* 343**,** 562-564.

MIRZAEI, A., KHALIGH-RAZAVI, S.-M., GHODRATI, M., ZABBAH, S. & EBRAHIMPOUR, R. 2013. Predicting the human reaction time based on natural image statistics in a rapid categorization task. *Vision research,* 81**,** 36-44.

NAYEBI, A., SAGASTUY-BRENA, J., BEAR, D. M., KAR, K., KUBILIUS, J., GANGULI, S., SUSSILLO, D., DICARLO, J. J. & YAMINS, D. L. 2022. Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition. *bioRxiv*, 2021.02. 17.431717.

OKAZAWA, G., HATCH, C. E., MANCOO, A., MACHENS, C. K. & KIANI, R. 2021. The geometry of the representation of decision variable and stimulus difficulty in the parietal cortex. *bioRxiv*, 2021.01. 04.425244.

OKAZAWA, G., SHA, L. & KIANI, R. 2020. Linear integration of sensory evidence over space and time underlies face categorization. *bioRxiv*.

PORTELLI, G., BARRETT, J. M., HILGEN, G., MASQUELIER, T., MACCIONE, A., DI MARCO, S., BERDONDINI, L., KORNPROBST, P. & SERNAGOR, E. 2016. Rank order coding: a retinal information decoding strategy revealed by large-scale multielectrode array retinal recordings. *Eneuro,* 3.

RAJAEI, K., KHALIGH-RAZAVI, S.-M., GHODRATI, M., EBRAHIMPOUR, R. & ABADI, M. E. S. A. 2012. A stable biologically motivated learning mechanism for visual feature extraction to handle facial categorization. *PLOS one,* 7**,** e38478.

RAJAEI, K., MOHSENZADEH, Y., EBRAHIMPOUR, R. & KHALIGH-RAZAVI, S.-M. 2019. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS computational biology,* 15**,** e1007001.

RIESENHUBER, M. & POGGIO, T. 1999. Hierarchical models of object recognition in cortex. *Nature neuroscience,* 2**,** 1019-1025.

RIESENHUBER, M. & POGGIO, T. 2000. Models of object recognition. *Nature neuroscience,* 3**,** 1199-1204.

SERRE, T., OLIVA, A. & POGGIO, T. 2007a. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences,* 104**,** 6424-6429.

SERRE, T., WOLF, L., BILESCHI, S., RIESENHUBER, M. & POGGIO, T. 2007b. Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence,* 29**,** 411-426.

SHADLEN, M. N., HANKS, T. D., CHURCHLAND, A. K., KIANI, R. & YANG, T. 2006. The speed and accuracy of a simple perceptual decision: a mathematical primer. *Bayesian brain: Probabilistic approaches to neural coding***,** 209-37.

SHADLEN, M. N. & NEWSOME, W. T. 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of neurophysiology,* 86**,** 1916-1936.

SPOERER, C. J., KIETZMANN, T. C. & KRIEGESKORTE, N. 2019. Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. *BioRxiv***,** 677237.

THORPE, S., DELORME, A. & VAN RULLEN, R. 2001. Spike-based strategies for rapid processing. *Neural networks,* 14**,** 715-725.

VAN DEN BERG, R., ZYLBERBERG, A., KIANI, R., SHADLEN, M. N. & WOLPERT, D. M. 2016. Confidence is the bridge between multi-stage decisions. *Current Biology,* 26**,** 3157-3168.

WANG, X.-J. 2002. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron,* 36**,** 955-968.

WENZLAFF, H., BAUER, M., MAESS, B. & HEEKEREN, H. R. 2011. Neural characterization of the speed–accuracy tradeoff in a perceptual decision-making task. *Journal of Neuroscience,* 31**,** 1254-1266.

YAMINS, D. L. & DICARLO, J. J. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience,* 19**,** 356-365.

ZABBAH, S., RAJAEI, K., MIRZAEI, A., EBRAHIMPOUR, R. & KHALIGH-RAZAVI, S.-M. 2014. The impact of the lateral geniculate nucleus and corticogeniculate interactions on efficient coding and higher-order visual object processing. *Vision research,* 101**,** 82-93.

ZHAN, J., INCE, R. A., VAN RIJSBERGEN, N. & SCHYNS, P. G. 2019. Dynamic construction of reduced representations in the brain for perceptual decision behavior. *Current Biology,* 29**,** 319-326. e4.