A computational model of familiarity detection for natural pictures, abstract images, and random patterns: Combination of deep learning and anti-Hebbian training

Yakov Kazanovich[1], Roman Borisyuk[1,2]

[1]Institute of Mathematical Problems of Biology, the Branch of M.V. Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Russia

[2]University of Exeter, College of Engineering, Mathematics and Physical Sciences Exeter, UK

ABSTRACT

We present a neural network model for familiarity recognition of different types of images in the perirhinal cortex (*the FaRe model*). The model is designed as a two-stage system. At the first stage, the parameters of an image are extracted by a pretrained deep learning convolutional neural network. At the second stage, a two-layer feed forward neural network with anti-Hebbian learning is used to make the decision about the familiarity of the image. FaRe model simulations demonstrate high capacity of familiarity recognition memory for natural pictures and low capacity for both abstract images and random patterns. These findings are in agreement with psychological experiments.

Keywords: Recognition memory, Familiarity recognition, Deep learning, Anti-Hebbian rule, Memorization

1. Introduction

Recognition memory is a form of declarative memory. It is subdivided into two types [Mandler, 1980; Yonelinas, 2002]: *recollection*, when an object or event is recalled together with the context, and *familiarity detection (recognition)*, when the judgment about knowing an object or event is made, but any context information is absent. Psychological experiments show that the time required for recollection is usually greater than for familiarity recognition. Thus, one can assume that familiarity recognition and recollection are two different processes that follow one after the other being dissociated at least at the behavioral level.

Some aspects of familiarity recognition by humans are paradoxical in comparison to other forms of declarative memory. While the capacity of the short-term memory is usually estimated

as 7±2 elementary objects (chunks) (Miller, 1956), the capacity of memory for familiarity recognition seems to be unlimited (Standing, 1970; Standing, 1973). In Standing's experiments the memory for familiarity was created by presenting a large number of stimuli. Each stimulus was presented only once for 5 s. The accuracy achieved under these conditions for 10000 pictures was about 85%. It is surprising that the capacity of memory for familiarity recognition in the Standing's experiments is so radically greater than the capacity of the short-term memory.

In the experimental and theoretical literature, there is no generally accepted view on whether the components of the recognition memory are two distinct cognitive processes or different stages of the same process. Some researchers support the *single-process theory* (SPT) that states that recollection is simply a more detailed version of familiarity recognition. The recognition memory is considered as a continuum where the strength of memory traces varies from weak to strong (Squire et al., 2007; Medina, 2008; Rutishauseret et al., 2008; Wixted & Squire, 2011; Slotnick, 2013). Other researchers argue that the *dual-process theory* (DPT) is in better agreement with experimental evidence. They refer to fMRI data and the results of investigations of patients with different forms of amnesia.

A review of theoretical arguments in favor of DPT can be found in the paper (Diana et al., 2006). According to DPT, there are two regions in the brain where the processes of the recognition memory are implemented, namely, the hippocampus and the perirhinal cortex, which are responsible for recollection and familiarity detection, respectively (Aggleton et al., 2005; Eichenbaum et al., 2007; Bowles et al., 2010; Montaldi & Mayers, 2010). Some studies reveal that such a dichotomy is not exact: the hippocampus participates in both types of recognition memory (Wais et al., 2006; Merkow et al., 2015). The situation is even more complicated since the lesion of the entorhinal cortex leads to the deficit of familiarity detection but not recollection (Brandt et al., 2016). The difference between the structures that are responsible for recollection and familiarity recognition extends beyond the medial temporal cortex, in particular to the prefrontal cortex (Scalici et al., 2017).

The recognition memory is evidently a complex phenomenon (Kafkas & Montaldi, 2018; Bastin et al., 2019). The neural network models (Norman & O'Reilly, 2003; Norman, 2010) were proposed to separate the hippocampal and neocortical contributions to the recognition memory. The main idea is that the neocortex assigns similar representations to similar stimuli, while the hippocampus assigns distinct patterns (separated representations) to stimuli, regardless of their similarity. The memory capacity for familiarity detection in these models is lower than in Standing’s experiments (Standing, 1973).

The relatively independent functioning of the perirhinal cortex during familiarity recognition proclaimed by the DPT is helpful for modeling familiarity recognition as an isolated process. Based on this idea, several computer models have been developed that correspond (more or less) to the Standing's results and neurophysiological data on the activity in the perirhinal cortex during familiarity recognition (Bogacz & Brown, 2003; Androulidakis et al., 2008).

The models (Bogacz et al., 2001; Bogacz & Brown, 2002; Greve et al., 2010; Sacramento & Wichert, 2012) operate with artificial binary patterns. Their functioning is based on the energy function in Hopfield networks (Hopfield, 1982) or on a biologically plausible analogue of this function. The absolute value of this function is greater for known (learned) patterns than for novel patterns. This fact is used as a discrimination criterion.

Some models are designed as two-layer networks with the feedforward flow of signals. Anti-Hebbian learning (Brown & Xiang, 1998; Bogacz & Brown, 2003) or the Info-max algorithm (Androulidakis et al., 2008, Lulham et al., 2011) were used for memorization since they weaken the connections between the layers and decrease the activity in the output layer (the perirhinal cortex) in response to familiar stimuli, which corresponds to neurophysiological evidence.

It has been shown by analytical methods and computer simulations (Bogacz, Brown, 2002; Budilova et al., 2009, Cortes et al., 2010) that the capacity of memory for familiarity recognition provided by some models is of the order n^2, where n is the number of units in the neural network. Note that the standard Hebbian learning in a Hopfield network gives the capacity of the order n (Amit, 1989). This is in agreement with the theoretical analysis (Frolov & Murav'ev, 1993) of informational characteristics of neural networks. These results can be considered as a mathematical explanation of the phenomenal memory capacity in familiarity recognition.

The models that simulated Standing's experiments (Standing, 1970; Standing, 1973) used input binary patterns (Androulidakis et al., 2008). This was reasonable since one of the purposes of modeling was to compare information characteristics of different models for various parameter values. Analytical estimation of the memory capacity and error rates was only possible under significant simplification of real experimental conditions.

In this work, we try to fulfill the gap between formal binary patterns and natural stimuli investigating how a model for familiarity recognition would function if it received the features of natural pictures as its inputs. We would also like to know whether modeling can demonstrate a difference in familiarity recognition between natural pictures and artificial random or abstract patterns. This difference was observed in the experiments of Bellhouse-King and Standing (2007). It has been shown that the performance of subjects in the familiarity recognition task is much better for natural than for abstract pictures.

Most of familiarity recognition models operate with formal input patterns. The paper (Ji-An et al., 2019) is an exception. It uses real pictures of human faces as the input. The process of familiarity detection is divided into two stages. At the first stage, a picture is processed by a deep learning neural network that extracts the parameters of the picture. At the second stage, these parameters are used to determine the familiarity of the picture. The network for the first stage was pretrained on a large set of photographs of faces. The network for the second stage learned input pictures via a special type of synaptic plasticity between the feature extraction module and memory module. Each synapse is described by a set of hidden variables that determine the value of the connection strength. It is shown that the memory capacity for this model is of the order n^2.

Our familiarity recognition model (*the FaRe model*) is similar to the one in (Ji-An et al., 2019) and also includes two stages (Fig. 1). The first stage reproduces information processing before the perirhinal cortex and provides input patterns for further processing at the second stage in the perirhinal cortex. At the first stage, the parameters of an image are extracted by a pretrained deep learning convolutional neural network. At the second stage, we use a two-layer feed forward neural network for familiarity recognition. Our recognition network is inspired by the models with the anti-Hebbian learning rule (Bogacz & Brown, 2002; Bogacz & Brown 2003, Androulidakis et al., 2008). Model simulations show high memory capacity for familiarity recognition when natural pictures are used as stimuli and low capacity when abstract images or random patterns are presented. This is in agreement with the results of psychological experiments.

2. Methods

2.1. The input and feature extraction

The perirhinal cortex receives highly processed visual information. It is not known what kind of features describe the image at the input to this brain structure. It is reasonable to assume that these features are formed by brain structures related to the early development of the visual system of humans being adapted to the visual information in real surrounding. As a substitute of this experience, we use the features that are formed in a high layer of a pretrained deep learning neural network. Thus, the feature extraction module (Fig. 1) of the FaRe model is based on the large, deep convolutional neural network (CNN) called AlexNet (Krizhevsky et al., 2012). This network has been pretrained to classify images into 1000 classes. The training set included 1.3 million of high-resolution natural images (photographs).

AlexNet consists of 25 layers. First nineteen layers combine convolutional, pooling and normalization modules to analyze the image and to define the features at different complexity levels. The 20th fully connected layer (fc7) represents the result of the multi-stage feature extraction procedure. Layers from 21st to 25th receive the features from the fc7 layer to classify the image by a back-propagation based algorithm with a 1000-way classifier. It has been proved that such image representation at the fc7 layer can be used to successfully train another classifier, e.g., the support vector machine, to classify natural pictures (see MATLAB R2020b: Feature Extraction Using AlexNet).

We used the output of the 20th AlexNet layer as the vector of features representing the image. These features can be considered as a compressed representation of the image. Extracted image features are fed to the input of the memory module (Fig. 1) for training and recognition.

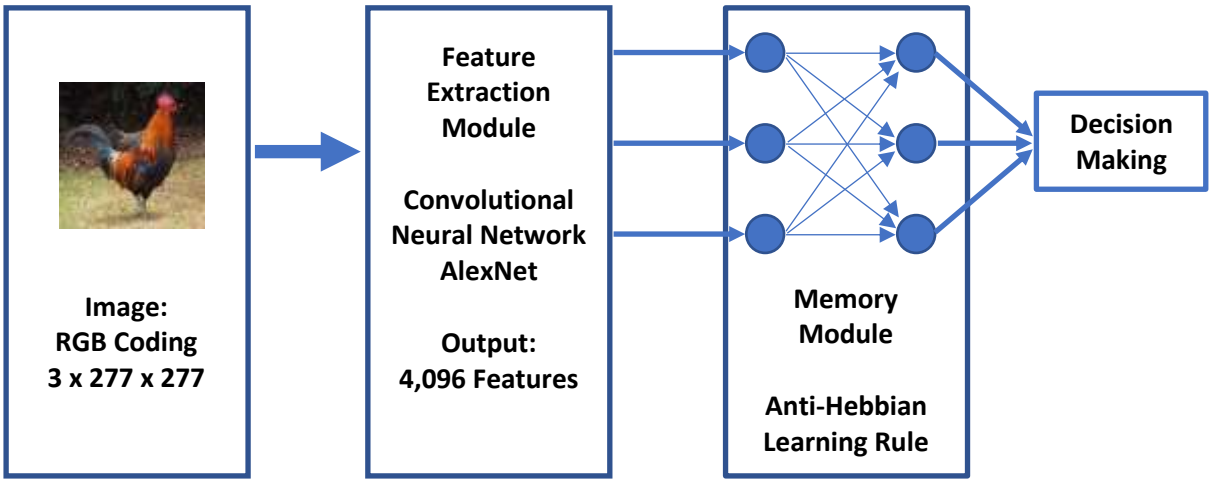

Fig. 1. The diagram of information flow of the FaRe model.

Before image processing, the RGB representation of each picture was normalized to the size 3x227x227, where 227x227 is the size of the image and 3 is the number of layers corresponding to RGB colors. Each image was processed by AlexNet implemented in the MATLAB software to define the corresponding vector of 4096 features.

The performance of the model was tested on three types (groups) of images (Fig. 2).

The **first group** is a collection of about 30,000 natural pictures (photographs) (Fig. 2a) selected from the CALTEX256 database (Griffin et al., 2007; CALTEX256 internet link).

The **second group** contains abstract images (Fig. 2b) similar to those used by Bellhouse-King and Standing (2007) (in their paper these images were called diverse abstract stimuli). An abstract image consists of rectangles, ovals, ellipses, and circles of different size and color. The

colors of the background, of all objects in the image, and of their edges were selected randomly from the RGB spectrum. The details of abstract image generation are presented in Appendix A.

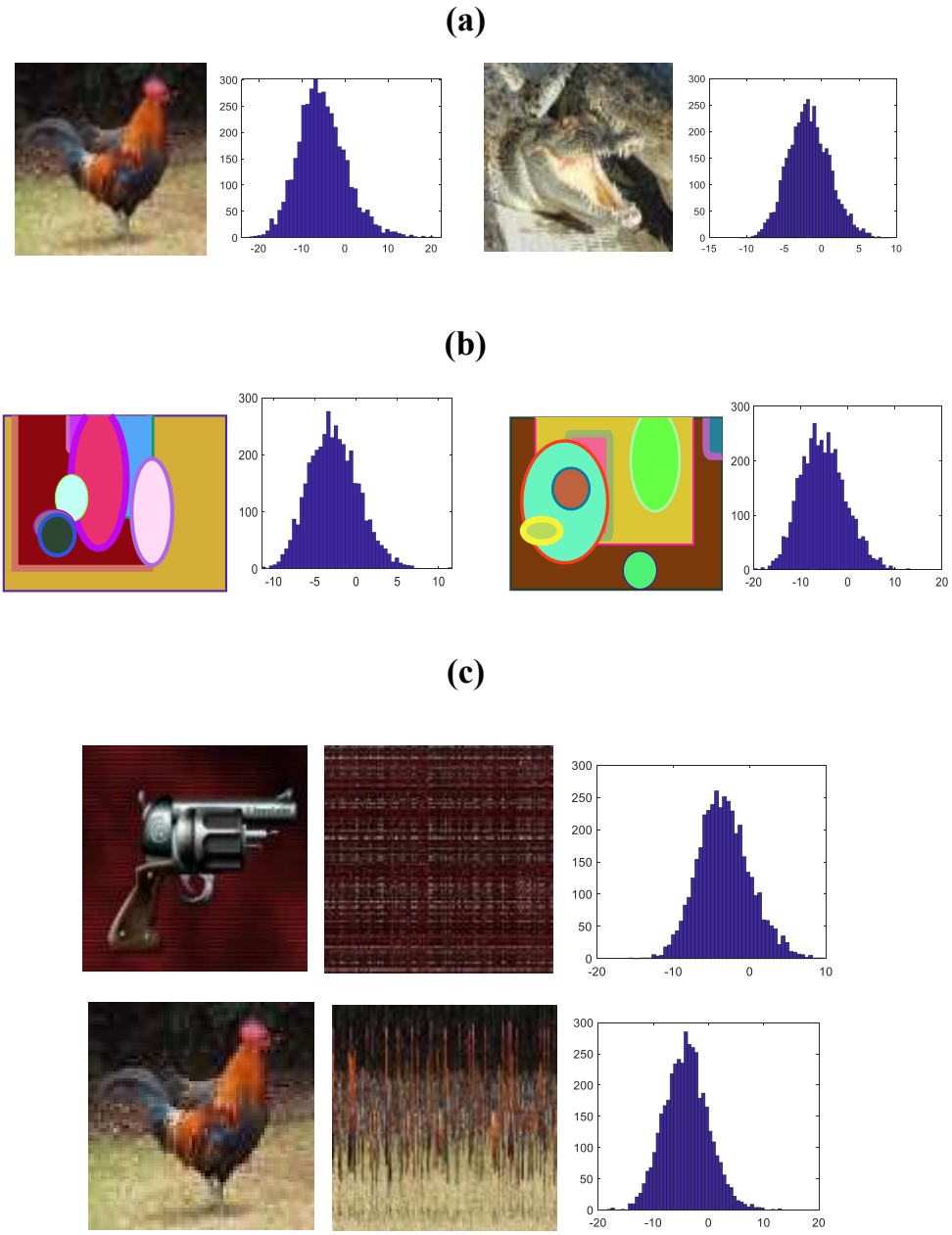

Fig. 2. Examples of images used in simulations.
(a) Natural pictures and the histograms of their feature values.
(b) Abstract images and the histograms of their feature values.
(c) Left: natural pictures; center-top: random pattern of type 1; center-bottom: random pattern of type 2; right: histograms of feature values of random patterns.

The **third group** contains random patterns (Fig. 2c) obtained from the photographs of the first group by a random permutation of pixels. We used two procedures to transform natural pictures into random patterns. According to the first procedure, the random pattern was obtained

from a natural picture by randomly permuting rows first and then columns of pixels in the picture (random patterns of *type 1*, see the top row in Fig. 2c). In the second procedure, we restricted permutations to randomly rearranging columns of pixels only (random patterns of *type 2*, see the bottom row in Fig. 2c).

In addition to images, Fig. 2 shows the histograms of feature values obtained at the output of the feature module. The shapes of distributions for different types of images are rather similar. All histograms are bell-shaped, resembling the Gaussian distribution, but they differ from the Gaussian distribution because they are not symmetrical about the mean. The shapes of distributions may give a hope that random patterns with the Gaussian distribution of feature values can be used as a suitable substitution of image features in testing the capacity of the memory module. Our computations show that this is not the case. For each image we normalized the vector of 4096 features (real numbers) associated with the image and generated by the AlexNet. For the normalized vector of features the mean is zero and the standard deviation is 1. We use the normalized vector as an input to the memory module.

2.2. The memory module for familiarity recognition

The memory (familiarity recognition) module (Fig. 1) is a feedforward two-layer neural network with the anti-Hebbian learning rule which is adapted to the inputs of real numbers instead of binary values used in models by Bogacz & Brown (2003) and Androulidakis et al., (2008). The choice of the anti-Hebbian neural network is caused by the following reasons. First, such networks reproduce the experimental evidence about the decrease of the activity in the perirhinal cortex for familiar stimuli. Second, these models use a simple biologically plausible type of learning. Third, it has been shown that the memory with the anti-Hebbian learning rule works well in reproducing the Standing's data for formal (binary) input patters (Androulidakis et al., 2008).

Our FaRe model includes the following specifications:

- The input layer contains $n = 4096$ neurons.
- The output layer contains $m = 4096$ novelty detection neurons.
- The layers are coupled by feedforward all-to-all connections.
- Initial values of connection strengths are randomly distributed in the interval (-1, 1).
- The connection strengths are modified during the learning stage.

The activity of neurons in the output layer is determined by the formula:

$$h_j = \sum_{i=1}^{n} w_{ij} x_i, \quad j = 1, \ldots, m, , \quad i = 1, \ldots, n, \tag{1}$$

where x_i are the components of the input X, w_{ij} are connection strengths between neurons of the input and output layers.

The output layer works in the regime of m/2-winners: only half of neurons with the highest activity levels are considered to be in active state, other neurons are in the rest state and they do not participate in the modification of connection strengths. This rule has earlier been used in modeling familiarity recognition using anti-Hebbian learning (Bogacz & Brown, 2002; Bogacz & Brown 2003, Androulidakis et al., 2008).

During the learning stage, the connection strengths between active neurons are modified in such a way that the average activity of neurons in the output layer decreases (Brown et al., 1987; Li et al., 1993; Brown & Xiang, 1998; Bogacz & Brown, 2002; Bogacz & Brown 2003). This can be achieved by the anti-Hebbian learning rule,

$$w_{ij} \to w_{ij} - \eta\ x_i, \tag{2}$$

where $\eta > 0$ denotes the learning rate. This modification of connection strengths is implemented for each pattern X of the learning set.

The average output activity during the testing stage for a pair of images *(X, Z)* is computed as:

$$d(X) = \frac{1}{m}\left(\sum_{j\in M_1}\sum_{i=1}^{n} w_{ij}x_i - \sum_{j\in M_2}\sum_{i=1}^{n} w_{ij}x_i\right),$$
$$d(Z) = \frac{1}{m}\left(\sum_{j\in M_1}\sum_{i=1}^{n} w_{ij}z_i - \sum_{j\in M_2}\sum_{i=1}^{n} w_{ij}z_i\right), \tag{3}$$

where M_1 and M_2 are the sets of m/2 "winners" and "losers" in the output layer, respectively, X is a known and Z is an unknown image.

It is assumed that a familiar image elicits a lower activity in the output layer than an unknown image, so the correct decision about the familiarity of X is made if $d(X) < d(Z)$, otherwise an error in familiarity recognition is registered.

Simulations of the FaRe model are organized similar to psychological experiments (Standing, 1970; Standing, 1973; Bellhouse-King & Standing, 2007). First, we randomly collect from the whole pool of input images two sets of the same size N: the set L of target (learning) items and the set T of unknown (distractor) items. The patterns from L are learnt one by one via a proper adaptation of connection strengths between the input and output layers of the memory module using the anti-Hebbian type rule (2).

Then comes the testing stage. During this stage the model receives a pair of items (X_k, Z_k) (k = 1, …, N), where the first item belongs to L and the second one is selected from T. Running the test through all N pairs and making the decision about familiarity according to (3), we compute the whole number of errors and the probability of errors.

Remark. In the study by Androulidakis et al. (2008), input vectors are binary and the connection strengths are normalized after learning each pattern so that for each neuron the mean of incoming connection strengths is 0 and the Euclidian length of the vector of weights is 1. In the FaRe model, input vectors to the memory module are real numbers and our simulations have shown that such normalization increases the probability of errors especially in the case of large sets of natural pictures. Thus, the normalization of connection strengths was excluded.

In our simulations of the FaRe model, we studied how the probability of errors depends on the number of tested pairs N, the model parameters, and the type of images. We show that the memory capacity for familiarity recognition significantly depends on the type of images. It can be high in the case of natural pictures, but the memory capacity is drastically reduced if either abstract or random patterns are memorized. This finding is in good agreement with psychological data (Bellhouse-King & Standing, 2007).

3. Simulation results

The performance of the anti-Hebbian model of familiarity recognition was analyzed in (Bogacz & Brown, 2002; Bogacz & Brown, 2003; Androulidakis et al., 2008). In particular, it has been shown that for random binary input patterns the probability of errors in familiarity recognition only weakly depends on the learning set size (see Fig. 7 in (Androulidakis et al., 2008)). We performed similar computations of the probability of errors for patterns with normally (the mean is 0, the standard deviation is 1) and independently distributed components as the inputs to the memory module. The errors were averaged through 20 runs of the module. The results of our computations are presented in Table 1 together with experimental data extracted from the paper (Standing, 1973).

Table 1. The probability of errors in experiments and simulations.

The number of stimuli N	The probability of errors			
	Experimental data (normal pictures)	Simulation results		
		$\eta = 0.0003$	$\eta = 0.0004$	$\eta = 0.0005$
20	0.01	0.17	0.08	0.045
40	0.045	0.15	0.11	0.05
100	0.05	0.12	0.1	0.02

200	0.085	0.18	0.12	0.02
400	0.14	0.14	0.1	0.03
1000	0.115	0.11	0.05	0.02
4000	0.189	0.17	0.05	0.02
10000	0.17	0.11	0.05	0.02

In contrast to the Standing's data and in agreement with the results of Androulidakis et al. (2008), the probability of errors is nearly stable when the size of the learning set increases. At the same time, there is significant sensitivity to the value of the learning rate η. In a reasonable range of η values the probability of errors decreases with the increase of η.

Let us go to the assessment of the model performance where the features of natural (real) pictures are used as the input information for the FaRe model. The learning rates in simulation experiments were equal to $\eta = 0.01$ or $\eta = 0.02$. The considerations behind this choice are explained in Appendix B. The probability of errors was computed through 100 runs of the model and random selection of picture sets for learning and testing stages from the whole database for each run. After each run is completed, we compute the probability of errors during this run. These probabilities are averaged to find the mean and standard deviation. The number of learned pictures N varied as in the Standing's experiments.

Besides the probability of errors, we follow Standing's approach to estimate the number of items N_{ret} retained in memory after learning. This is computed according to the formula

$$N_{ret} = N\,(1 - 2\,P_{er}), \tag{4}$$

where N is the number of learned items, P_{er} is the probability of errors.

The results of FaRe model simulations are shown in Fig. 3. The number of retained objects N_{ret} vs the number of images N is shown in Figs 3a,c in $\log_{10}$ scale. The averaged probabilities of errors vs the number of images N is shown in Figs 3b,d. Note, the values of N used in our simulations coincide with the values of N in Standing's experiments (Standing, 1973).

In contrast to the simulation results presented in Table 1, the probability of errors increases when the size of the learning set increases. In this respect our results are in agreement with experimental data. However, the experimental data are somewhat underestimated for $N < 400$ and overestimated for $N > 1000$, especially in the case $\eta = 0.02$, so the behaviour of the experimental and simulated curves in Figs 3b,d are different. Radical increase of the number of neurons m in the output layer of the memory module does not significantly affect our results. The corresponding arguments are presented in Appendix C.

The results of FaRe model computations are shown in Fig. 4. Probabilities of familiarity recognition errors for random patterns of type 1 are shown in Figs. 4a,b and for abstract images in Figs. 4c,d. Colored graphs represent the mean probabilities of errors and vertical bars show the standard deviations. For the reference, the black graphs (the same graphs as in Figs. 3b,d) represent the Standing's experimental data (Standing, 1973).

Comparing Figs. 3 and 4, it can be seen that the performance of the model in the case of random patterns and abstract images is radically poorer than in the case of natural pictures. Even dozens of images present difficulties for familiarity recognition.

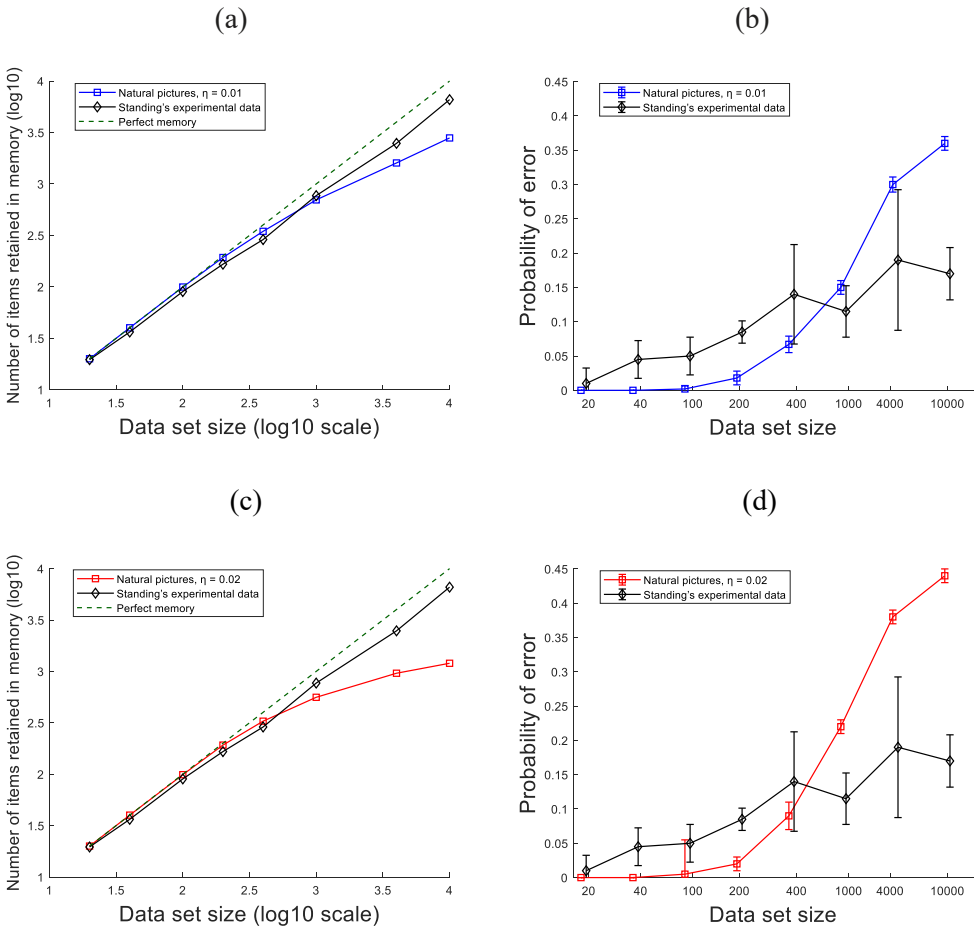

Fig. 3. The results of simulations of familiarity recognition for natural pictures.
(a), (c) The number of items retained in memory as a function of the number of presented items during learning in $\log_{10}$ scale.
(b), (d) Averaged probabilities of errors during testing. Black lines correspond to experimental data (Standing, 1973).
Blue lines in **(a)** and **(b)** correspond to computations with $\eta = 0.01$. Red lines in **(c)** and **(d)** correspond to computations with $\eta = 0.02$. Green dashed lines in **(a)** and **(c)** correspond to perfect recognition memory, where all the presented images are retained. Vertical bars in **(b)** and **(d)** show standard deviations.

Fig. 5 allows us to compare the results of simulations with the experimental data of Bellhouse-King and Standing for diverse abstract stimuli (Bellhouse-King & Standing, 2007). The authors of this paper experimented with three types of images which they called concrete, regular abstract, and diverse abstract. Concrete pictures were the pictures of everyday objects and scenes similar to those which we used in our simulations with natural pictures. Regular abstract images were photographs of snowflakes. Diverse abstract stimuli were meaningless images constructed using a combination of geometrical shapes of different colour, size, and orientation. They are similar to our abstract images (Fig. 2b). The size of training and testing sets was $N = 30$.

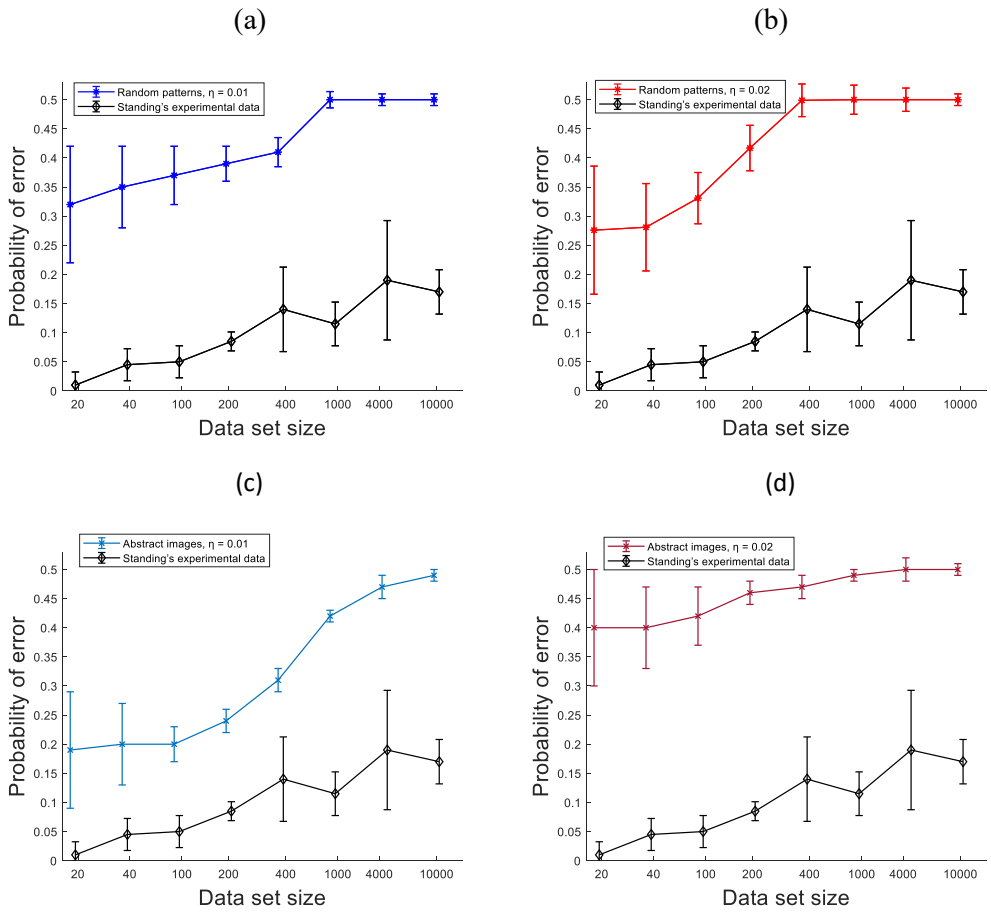

Fig. 4. The probability of familiarity recognition errors for random patterns of type 1 and abstract images.
(a-b) Computations for random patterns of type 1 with $\eta = 0.01$ **(a)** and $\eta = 0.02$ **(b)**.
(c-d) Computations for abstract images with $\eta = 0.01$ **(c)** and $\eta = 0.02$ **(d)**.
Colored lines show the results of computations. Black graphs represent psychological experimental data for natural pictures (Standing, 1973)). Vertical bars show standard deviations.

Fig. 5 shows the average probability of errors in the Bellhouse-King and Standing experiments by green and magenta circles (see the part of the figure on the left from the vertical

dotted line). Green circles correspond to the experiments with the stimuli presented to the subject in the order Diverse Abstract – Regular Abstract – Concrete. Magenta circles correspond to the experiments with the stimuli presented in the reverse order: Concrete - Regular Abstract - Diverse Abstract. Vertical bars of the corresponding colour show standard deviations.

Besides experimental data, Fig. 5 shows the results of FaRe model simulations with random patterns of types 1 and 2 and abstract images (see the part of the figure on the right from the vertical dotted line). There was no necessity for us to exactly follow the conditions of psychological experiments in the way of stimuli presentation since they have been adjusted for convenience of subjects. Note that the number of items retained in memory for both orders of presentation were qualitatively similar. Thus, in the simulations we presented only one type of stimuli (either random patterns of a particular type or abstract images).

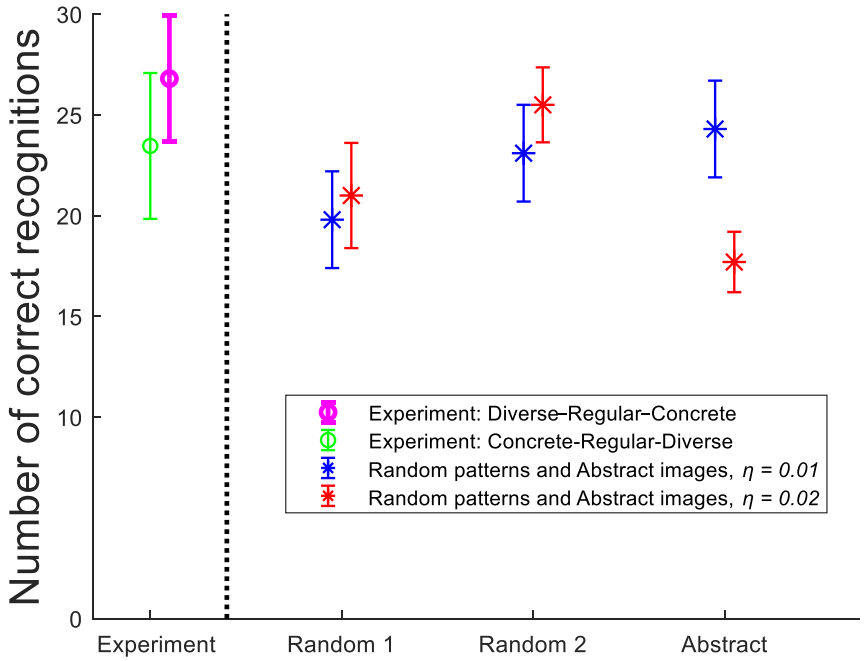

Fig. 5. The average number of items retained in memory of subjects (Experiment) and the FaRe model (random and abstract images).
The learning set size is $N = 30$. Circles and stars show the mean values, vertical bars show standard deviations. Blue and red bars correspond to computations with $\eta = 0.01$ and $\eta = 0.02$, respectively.
Experiment: Experimental data of Bellhouse-King and Standing (2007) for diverse abstract stimuli. The green bar shows the experimental results when the order of presentation of stimuli was Diverse Abstract – Regular Abstract – Concrete. The magenta bar corresponds to the order of presentation Concrete - Regular Abstract - Diverse Abstract.
Random 1: Simulation results for random patterns of type 1.
Random 2: Simulation results for random patterns of type 2.
Abstract: Simulation results for abstract images.

As in the case of natural pictures, the average probabilities of errors and their standard deviations were computed through 100 runs of the model and random selection of items of a definite type for each run. The number of stimuli used in training and testing stages of simulations was as in the psychological experiments, $N = 30$. The results of simulations are shown in the right part of Fig. 5. Blue and red stars show the probabilities of errors under the parameter values $\eta = 0.01$ and $\eta = 0.02$, respectively, and vertical bars show standard deviations.

Our simulations are in a good agreement with psychological data of Bellhouse-King and Standing (2007). Simulations with abstract images and with random patterns show a relatively high number of errors and low number of items retained in memory in comparison with natural pictures. The random patterns of type 1 are more difficult to recognize as familiar than random patterns of type 2. This is in agreement with intuitive expectations. The number of retained items for $\eta = 0.01$ is slightly lower than for $\eta = 0.02$ for both types of random patterns. For abstract images the results of familiarity recognition are much better for $\eta = 0.02$ than for $\eta = 0.01$. For $\eta = 0.01$ the number of retained abstract images is similar to the experimental data with diverse abstract stimuli.

4. Discussion

Experimental evidence shows that visual recognition memory, in particular familiarity recognition, is more reliable and works faster if it operates with natural meaningful pictures in comparison to abstract, artificial meaningless images (Bellhouse-King & Standing, 2007; Boucher et al. 2016). This is usually explained by the dual-coding theory, according to which natural pictures are processed and memorized not in purely visual, but also in verbal categories. Not denying this concept, we would like to draw attention to another possible source of this phenomenon. We assume that the experience during the early stages of brain development (the childhood) forms the visual system in such a way that it allows for efficient coding everyday images, but it is inefficient in coding abstract, artificial images.

The results of our modeling support this assumption. The deep learning neural network trained by natural pictures is inefficient in coding abstract images or random patterns, which is reflected in poor capability of the memory module in distinguishing between known and unknown images. Taking into account the results of (Bogacz & Brown, 2003) on the correlated patterns, one may suggest that low efficiency of familiarity recognition of random patterns and abstract images is due to higher correlations between their features. The results presented in Appendix D do not support this hypothesis. Averaged characteristics of correlations are similar for all types of images.

It seems that some deeper dependencies between particular pairs of features are responsible for the radical difference in the performance of the model for images of different types.

Thus, we conclude that successful solution of the familiarity recognition task significantly depends on the pretraining, when the subject develops an efficient system of feature extraction for the analysis and representation of pictures. This feature system is not universal. For special types of images, e.g. random patterns and abstract images, it results in poor familiarity recognition.

If the system of features is optimized by preliminary training, a relatively simple neural network for familiarity recognition can memorize and correctly recall thousands of natural pictures with the probabilities of errors similar to those observed in psychological experiments. The same model demonstrates that in familiarity recognition of random patterns and abstract images the probability of errors is as high as in psychological experiments for diverse abstract stimuli when the number of memorized images is about three dozen only.

Our computations show that stimulation of the FaRe model by artificial random patterns can be misleading in assessing how well the model reproduces experimental data. For random Gaussian patterns the FaRe model failed to demonstrate the increase of the probability of errors with the increase of the training set size observed in psychological experiments. Nevertheless, such increase was demonstrated by the model in the case of natural pictures. The distribution of feature values in the latter case has a bell-shape form similar to the Gaussian distribution that we used in generating artificial random stimuli, but internal dependency between the features of natural pictures cannot be formulated in simple stochastic terms. This explains the difference in simulation results.

Earlier computer experiments with random stimuli have shown that the anti-Hebbian learning, being biologically plausible in the perirhinal cortex upon familiarity recognition, provides the memory capacity of the order n^2, where n is the size of the network (Bogacz & Brown, 2002; Bogacz & Brown, 2003). Our simulations show that the high memory capacity of the anti-Hebbian model is maintained under the stimulation by the features of natural pictures. Note that the high capacity is reached without complex synapses that were used in the paper (Ji-An et al., 2019). Moreover, the paper (Ji-An et al., 2019) was devoted to the study of the capacity of their familiarity recognition model without any reference to psychophysical data. In contrast, our main results are directed to the reproduction of experimental data of (Standing, 1973; Bellhouse-King & Standing, 2007).

Our computations show that the learning rate η is a critical parameter for anti-Hebbian model functioning. We have chosen the range of the values of this parameter that would allow us to minimize the probability of errors of familiarity recognition. The model is too abstract to justify

our choice by neurobiological observations. We can only hope that in real biological systems the optimization of features important for survival is achieved in the process of evolution.

Qualitatively reflecting some aspects of experimental data, the anti-Hebbian network and its learning algorithm are far from being perfect in simulating the form of experimental error curves. Also the model was not able to reach subject's results for large learning sets. We have done numerous computations in an attempt to reduce the probability of errors for large N, varying η and radically increasing the number of neurons in the output layer. Unfortunately, this did not lead to positive results. Evidently, the model is too simple to properly reflect complex process of human decision-making during familiarity recognition. To improve model performance, more detailed structure of the perirhinal cortex and its interaction with other brain structures such as the parahippocampal cortex and hippocampus should be taken into account. It is important to emphasize that the models should be tested on the features of real images, both natural and artificial.

Appendix A. Generation of abstract images

Each abstract image contained a large rectangle, two ovals (medium and small) and five ellipses (large, medium, and three small). All objects in the image were generated by a MATLAB program *rectangle* with various parameter values. The intervals of the parameters for the background and for objects are summarized in Table A1. The parameters (except curvature) were randomly selected from the specified intervals and rounded to the nearest integer. The colors of the background, of all objects in the image, and of their edges were selected randomly from the RGB spectrum.

Table A1. Ranges of parameters for generation of abstract images.

Objects	Position x	Position y	Size x	Size y	Curvature x	Curvature y	Edge width
Background	0	0	30	10	0	0	(0, 5)
Large oval	(1, 4)	(1, 3)	(12, 25)	(7, 12)	0	0	(0, 12)
Medium oval	(1, 30)	(1, 7)	(5, 10)	(3, 9)	0.1	0.2	(0, 12)
Small oval	(1, 30)	(1, 10)	(2, 5)	(2, 4)	0.5	0.3	(0, 12)
Large ellipse	(1, 11)	(1, 5)	(7, 13)	(5, 10)	1	1	(0, 12)
Medium ellipse	(15, 19)	(0, 8)	(4, 7)	(4, 8)	1	1	(0, 12)
Small ellipse 1	(0, 30)	(0, 10)	(3, 5)	(1, 4)	1	1	(0, 12)
Small	(1, 11)	(1, 4)	(5, 6)	(1, 2)	1	1	(0, 12)

ellipse 2							
Small ellipse 3	(1, 9)	(1, 5)	(4, 6)	(1, 3)	1	1	(0, 12)

Appendix B. Selection of the learning rate

The error rates produced by the model critically depend on the learning rate η. In the paper (Androulidakis et al., 2008) a cost function was used to provide the best fit between simulation results and Standing's experimental data. In our simulations our main aim was to explain the difference in recognition memory for both natural and artificial images, so we used a simpler approach to the selection of η that is based on minimization of the mean probability of errors. The mean probability of errors was computed for N=1000 natural pictures and for η varying in the range (0, 0.03) (Fig. B1). Averaging was done through 100 runs of the model. The minimal value was obtained for η = 0.01. The results of simulations presented in Figs. 3-5 allows one to compare the functioning of the model for the values η = 0.01 and η = 0.02 which are expected to be somewhere near the optimum.

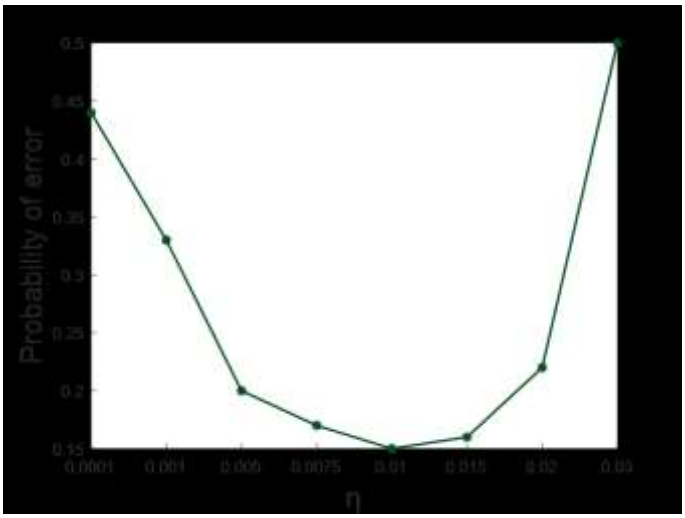

Fig. B1. The average probability of errors of familiarity recognition for natural pictures as a function of the learning rate.

Appendix C. The influence of the size of the memory module on the probability of error

The experience obtained with the familiarity recognition model FamE based on the energy in the Hopfield network (Androulidakis et al., 2008) may lead to the suggestion that restricted recognition memory capacity is a result of low size of the output layer of the memory module. To test this hypothesis, we varied the number of neurons m in the output layer and computed the average probability of errors. Averaging was done through 100 runs of the model. The results of computations are presented in Fig. C1. The graphs in the figure clearly show that though the

probability of errors gradually decreases with increasing m, even the radical increase of m does not significantly improve the performance of the model.

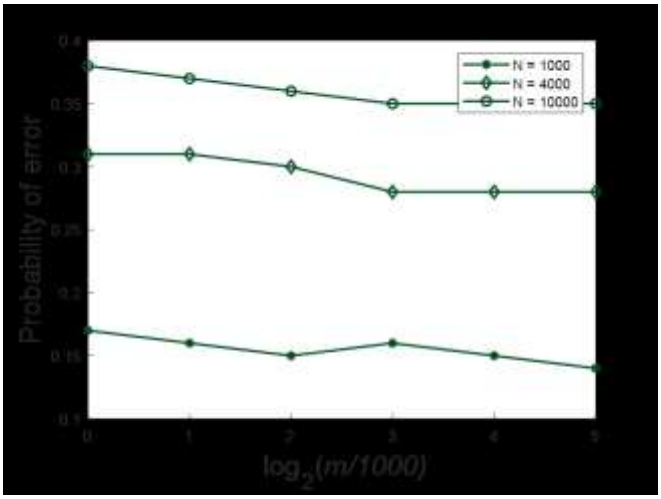

Fig. C1. The average probability of errors of familiarity recognition for natural pictures as a function of m for m = 1000, 2000, 4000, 16000, 32000, η = 0.01.

Appendix D. Dependences between the features of images

The paper (Bogacz & Brown, 2003) shows that in the case of random binary input patterns the correlations between the components of the patterns increases the number of familiarity recognition errors. One may think that such correlations are the source of the increased number of errors for random and abstract images. To test this hypothesis, we computed various parameters that describe mutual dependences between the features of images that were used in our simulations.

The computations were organized in the following way. First, we selected a large enough set of images of a particular kind. The size of each set is shown in the second column of Table D1. Each image is described by 4096 features. Then, the correlations between all 4096 x 4096 pairs of features (except of self-correlations) were computed. Averaging these values, we obtain the mean correlations which are shown in the third column (Mean 1) of Table D1 and in brackets we report the standard deviation. It is seen that there is no significant difference of this parameter for different types of images.

In principle, low values of the mean correlations can be a result of mutual compensation of positive and negative correlations. To eliminate this effect, we computed the means of the absolute values of correlations (Mean 2). They are presented in the fourth column of Table D1. The last two columns of Table D1 show minimal and maximal values of the absolute values of correlations.

The data in Table D1 allow us to make the following conclusion: the values in each of the columns 3-6 of Table D1 are too similar to explain the difference in the results of familiarity recognition. However, the correlation between some pairs of features is very strong. Averaging hides a complex structure of mutual dependences between the image features which is the source of radically different memory capacity for different image types.

Table D1. Correlation of features

Type of images	The number of images	Correlations			
		Mean 1 (std)[a]	Mean 2 (std)[b]	Min[e]	Max[d]
Natural pictures	29,700	0.11 (0.14)	0.15 (0.11)	8.6e-08	0.86
Random patterns of type 1	19,720	0.13 (0.26)	0.25 (0.17)	5.7e-08	0.93
Random patterns of type 2	19,720	0.15 (0.24)	0.23 (0.16)	5.8e-0.8	0.89
Abstract images	21,000	0.12 (0.19)	0.18 (0.13)	6.2e-0.9	0.81

[a] Mean 1 (std) – the mean and standard deviation of correlation values (in brackets);
[b] Mean 2 (std) – the mean and standard deviation of correlation absolute values (in brackets);
[c] Min - minimum of correlation absolute values;
[d] Max - maximum of correlation absolute values.

References

Aggleton, J.P., Vann, S.D., Denby C., et al. (2005). Sparing of the familiarity component of recognition memory in a patient with hippocampal pathology. Neuropsychologia, 43, 1810–1823.

Amit, D.J. (1989). Modeling brain function. Cambridge, Cambridge University Press.

Androulidakis, Z., Lulham. A., Bogacz, R., and Brown, M.W. (2008). Computational models can replicate the capacity of human recognition memory. Network: Comput. Neural Syst., 19, 161–182.

Bastin, C., Besson, G., Simon, J., Delhaye, E., Geurten, M., Willems, S., and Salmon, E. (2019). An integrative memory model of recollection and familiarity to understand memory deficits. Behav. Brain Sci., 42, 1-66.

Bellhouse-King M.W. and Standing L.G. (2007). Recognition memory for concrete, regular abstract, and diverse abstract pictures. Perceptual and Motor Skills, 104, 758-762.

Bogacz, R and Brown, M.W. (2002). The restricted influence of the sparseness of coding on the capacity of the familiarity discrimination networks. Network: Comput. Neural Syst., 13, 457–485.

Bogacz, R. and Brown, M.W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. Hippocampus, 13, 494–524.

Bogacz, R, Brown, M.W., and Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. J. Comput. Neurosci., 10, 5-23.

Boucher, O., Chouinard-Leclaire, C., Muckle, G., Westerlund, A., Burden, M.J., Jacobson, S.W., and Jacobson, J.L. (2016). An ERP study of recognition memory for concrete and abstract pictures in school-aged children. Int. J. Psychophysiol., 106, 106-114.

Bowles, B, Crupi, C., Pigott, S., et al. (2010). Double dissociation of selective recollection and familiarity impairments following two different surgical treatments for temporal-lobe epilepsy. Neuropsychologia, 48, 2640–2647.

Brandt, K.R., Eysenck, M.W., Nielsen, M.K., and von Oertzen, T.J. (2016). Selective lesion to the entorhinal cortex leads to an impairment in familiarity but not recollection. Brain and Cognition, 104, 82–92.

Brown, M.W., Wilson, F.A.W., and Riches, I.P. (1987). Neuronal evidence that inferotemporal cortex is more important than hippocampus in certain processes underlying recognition memory. Brain Res., 409, 158 –162.

Brown, M.W. and Xiang, J.Z. (1998). Recognition memory, Neuronal substrates of the judgement of prior occurrence. Progr. Neurobiol., 55, 149–189.

Budilova, E.V., Karpenko, M.P., Kachalova, L.M., and Terekhin A.T. (2009). Familiarity recognition and recollection: A neural network model. Biophysics, 54, 500–507.

CALTEX256 - database of images: https://www.kaggle.com/jessicali9530/caltech256.

Cortes, J.M., Greve, A, Barrett, A., and van Rossum M.C.W. (2010). Dynamics and robustness of familiarity memory. Neural Comput., 22, 448-466.

Diana, R.A., Reder, L.M., Arndt, J., and Park, H. (2006). Models of recognition, A review of arguments in favor of a dual-process account. Psychon. Bull. Rev., 13, 1–21.

Eichenbaum, H., Yonelinas, A.R., and Ranganath, C. (2007). The medial temporal lobe and recognition memory. Annu. Rev. Neurosci., 30, 123–152.

Frolov, A.A. and Murav'ev, I.P. (1993). Information characteristics of neural networks capable of associative learning based on Hebbian plasticity. Network: Comput. Neural Syst., 4, 495-536.

Greve, A., Donaldson, D.I., and van Rossum, M.C.W. (2010). A Single-Trace Dual-Process Model of Episodic Memory: A Novel Computational Account of Familiarity and Recollection. Hippocampus, 20, 235–251.

Griffin, G., Holub, A.D., and Perona, P. (2007). Caltech-256. Object category dataset. Caltech Technical Report. *ImageNet.* http://www.image-net.org

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. (USA), 79, 2554–2558.

Ji-An, L., Stefanini, F., Benna, M.K., and Fusi, S. (2019). Face familiarity detection with complex synapses. bioRxiv, doi: http://dx.doi.org/10.1101/854059.

Kafkas, A. and Montaldi, D. (2018). How do memory systems detect and respond to novelty? Neurosci. Lett., 680, 60–68.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, Proceedings of the 25th International Conference on Neural Information Processing Systems, v.1: pp. 1097–1105.

Li, L., Miller, E.K., and Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex, J. Neurophysiol., 69, 1918 –1929.

Lulham, A., Bogacz, R., Vogt, S., and Brown, M.W. (2011). An infomax algorithm can perform both familiarity discrimination and feature extraction in a single network. Neural Comput., 23, 909–926.

Mandler*,* G. *(*1980*).* Recognizing, The judgment of previous occurrence. Psychol. Rev., 87, 252–271.

MATLAB R2019b. Deep Learning Toolbox: Extract Image Features Using Pretrained Network.

Medina, J.J. (2008). The biology of recognition memory. Psychiatric Times., 13, 13-16.

Merkow, M.W., Burke, J.F., and Kahanac, M.J. (2015).The human hippocampus contributes to both the recollection and familiarity components of recognition memory. Proc. Natl. Acad. Sci. (USA), 112, 14378–14383

Miller, G.A. (1956). The magical number seven, plus or minus two. Some limits on our capacity for processing information. Psychol. Rev., 63, 81-97.

Montaldi, D. and Mayers, A.R. (2010). The role of recollection and familiarity in the functional differentiation of the medial temporal lobes. Hippocampus, 20, 1291–1314.

Norman, K.A. (2010). How hippocampus and cortex contribute to recognition memory: revisiting the complementary learning systems model. Hippocampus, 20, 1217-1227.

Norman, K.A. and O'Reilly, R.C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. Psychol. Rev., 110, 611-646.

Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2008). Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. Proc. Natl. Acad. Sci. (USA), 105, 329–334.

Sacramento, O. and Wichert, A. (2012). Binary Willshaw learning yields high synaptic capacity for long-term familiarity memory. Biol. Cybern., 106,123-133.

Scalici, F. Caltagirone, C., and Carlesimo, G.A. (2017). The contribution of different prefrontal cortex regions to recollection and familiarity, a review of fMRI data. Neurosci. Biobehav. Rev., 83, 240–251.

Slotnick, S.D. (2013). The nature of recollection in behavior and the brain. NeuroReport, 24, 663-670.

Squire, L.R., Wixted, J.T., and Clark, R.E. (2007). Recognition memory and the medial temporal lobe, a new perspective. Nature Rev. Neurosci., 8, 872-883.

Standing, L. (1970). Perception and memory for pictures, Single trial learning of 2500 visual stimuli. Plychol. Sci., 19, 73-74.

Standing, L. (1973). Learning 10,000 pictures. Quat. J. Exp. Psychol., 25, 207-222.

Wais, P.E., Wixted, J.T., Hopkins, R.O., and Squire, L.R. (2006). The hippocampus supports both the recollection and the familiarity components of recognition memory. Neuron, 49, 459-466.

Wixted, J.T. and Squire, L.R. (2011). The medial temporal lobe and the attributes of memory. Trends Cogn. Sci., 15, 210-217.

Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. J. Memory and Language, 46, 441-517.