Analysis of Neural Networks for Image Classification

Nikolay Vershkov[1(✉)], Mikhail Babenko[1,2], Viktor Kuchukov[3], Nataliya Kuchukova[1], and Nikolay Kucherov[1]

[1] North-Caucasus Federal University, Stavropol, Russia
nvershkov@ncfu.ru
[2] Institute for System Programming of the Russian Academy of Sciences, Moscow, Russia
[3] North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

Abstract. The article explores the option of using information theory's mathematical tools to model artificial neural networks. The two primary network architectures for image recognition, classification, and clustering are the feedforward network and convolutional networks. The study investigates the use of orthogonal transformations to enhance the effectiveness of neural networks and wavelet transforms in convolutional networks. The research proposes practical applications based on the theoretical findings.

Keywords: artificial neural networks · orthogonal transformations · wavelets · feature vector · convolution · correlation

1 Introduction

The construction of artificial neural networks (ANN) is based on the organization and operation principles of their biological equivalents [1]. The research on ANN has its roots in the theory of brain functioning, which was established in 1943 by W. McCulloch and W. Pitts. Their work is widely regarded as a significant contribution to this field [2]. The theory of ANN has undergone substantial development over the past 80 years, including advances in architecture and learning methods. However, it is crucial to note that the development of ANN is primarily intuitive and algorithmic rather than mathematical. Many ANN architectures have been borrowed from the biological realm [3]. The mathematical depiction of ANN received a significant advancement through the works of Kolmogorov-Arnold [4,5] and Hecht-Nielsen [6]. These works are regarded as notable milestones in this field.

The use of information theory in the study of ANN has been relatively uncommon. Claude Shannon's groundbreaking work in information theory [7] established the basis for measuring and optimizing information transmission through communication channels, including the role of coding redundancy in improving error detection and correction. Since ANNs are essentially systems that process

A. Alikhanov et al. (Eds.): APAMCS 2022, LNNS 702, pp. 258–269, 2023.
https://doi.org/10.1007/978-3-031-34127-4_25

information, researchers have applied the mathematical tools of information theory to investigate self-organization models [3].

The authors of this article conducted a series of studies [8–10] to examine the information processes in feedforward ANNs. These studies have shown potential benefits, including reduced redundancy, energy consumption, and training time, when considering the information processing characteristics of neural networks. The mathematical model of a neuron's ability to perform various transformations in the ANN layers enables us to analyze the input information processing methods from an information theory standpoint. The conventional approach, known as the McCulloch-Peets model [2], regards the mathematical model of a neuron as follows:

$$y_{k,l} = f\left(\sum_{i=1}^{n} w_i^{k,l} x_i^{k,l}\right) \tag{1}$$

where k and l are the number of layer and neuron in the layer, respectively, $y_{k,l}$ is the output of the neuron, $x_i^{k,l}$ signifies the inputs of the neuron, $w_i^{k,l}$ symbolizes the weights (synapses) of the input signals, and f is the neuron output function, which can be linear or not. Several linear transformations in information theory possess a comparable structure, such as orthogonal transformations, convolution, correlation, filtering in the frequency domain, among others. Previous research [8–10] addressed problems such as the optimal loss function, non-linear neuron characteristics, and neural network volume optimization. The goal of this article is to examine neural networks for image processing from an information theory perspective and establish general principles for building ANNs to solve specific problems. The research is entirely theoretical, and the article does not aim to experimentally validate the authors' propositions using mathematical tools of information theory.

2 Materials and Methods

2.1 The Wave Model of Feedforward ANN

According to previous studies [8], the information model of a feedforward ANN involves a multidimensional input vector $X_i = \left\{x_1^i, x_2^i, \ldots, x_n^i\right\}$, which can be discretized in time and level values of some input function $x(t)$. This input value X_i is processed by each neuron in each layer of the ANN according to Eq. (1), resulting in discrete output values $Y_i = \left\{y_1^i, y_2^i, \ldots, y_m^i\right\}$. The Kotelnikov theorem, also known as the Nyquist criterion, is used to discretize the functions $x(t)$ and $y(t)$ in the information model of a feedforward ANN. It should be noted that the set $\{X_i\}_{i=1,2,\ldots,n}$ is not complete, which means that some input values may not be included in the training alphabet of the ANN. This is different from the decoding process in an information channel, where the alphabet of transmitted discrete messages is finite and predefined, as described by Shenon [7]. Additionally, the weight values in all neurons of the ANN are assumed to be randomly assigned before the learning process begins. When training the ANN with a teacher, the output function $y(t)$ is completely known.

ANNs are composed of an input layer that can handle X_i, an output layer with a capacity of Y_i, and one or more hidden layers. Depending on the application, ANNs can perform different tasks like image classification, clustering, and function approximation. To better understand the function being studied, the operation of the network will be analyzed in the application domain that is most appropriate.

The output layer is a critical component in ANNs for tasks such as classification or clustering. Its purpose is to assign input signals to their corresponding classes or clusters, similar to how a received signal in communication systems is observed to determine the transmitted signal [12]. However, just like in communication systems, ANNs can also experience interference, which depends on the set of input information used for classification or clustering rather than the communication channel. To mathematically describe the ANN, a transition probability $p\left[x\left(t\right)|y\left(t\right)\right]$ is used, which represents the probability of converting a received realization into the correct class or cluster. A model using additive white Gaussian noise, similar to communication theory, can be applied to the data [13]. This model is suitable when there is a large number of data in the set, such as in the MNIST database [14], which contains 60,000 records. The transition probability decreases exponentially with the square of the Euclidean distance $d^2\left(x,y\right)$ between the obtained value of X_i and the ideal representation of class Y_i given by:

$$p\left[x\left(t\right)|y\left(t\right)\right] = k\exp\left(-\frac{1}{N_0}d^2\left(x,y\right)\right), \tag{2}$$

where k is a coefficient independent of $x(t)$ and $y(t)$, N_0 is the spectral density of noise, and

$$d^2\left(x,y\right) = \int_0^T \left[x\left(t\right) - y\left(t\right)\right]^2 dt. \tag{3}$$

In some problems involving approximation and prediction, it is assumed that the signals $x(t)$ and $y(t)$ have the same period, but in the specific problem classes, they are treated as separate. For instance, in image classification problems like those found in the MNIST database, the input vector comprises 784 pixel values, and there are ten image classes. To solve these problems, it's necessary to establish a clear mapping $Y_i \leftrightarrow \widetilde{X}_i$, where an observation X_i is compared to an "ideal representation" of class $\widetilde{X}_i$, and if they are similar, it is inferred that observation X_i belongs to class Y_i. This process is expressed mathematically in the following equation:

$$\left(X_i \in Y_i\right) = \min_j d^2\left(X_i, \widetilde{X}_j\right). \tag{4}$$

Opening the parentheses in Eq. (3) and replacing the representation y with $\widetilde{x}\left(t\right)$, we obtain:

$$d^2\left(x,\widetilde{x}\right) = \int_0^T x(t)^2 dt - 2\int_0^T x\left(t\right)\widetilde{x}\left(t\right)dt + \int_0^T \widetilde{x}(t)^2 dt = \|x\|^2 - 2z + \|\widetilde{x}\|^2. \tag{5}$$

The Eq. (5) involves the energy of the input realization and the cluster representation $\widetilde{x}$ denoted by $|x|^2$ and $|\widetilde{x}|^2$, respectively. These values are constants when the input signal is normalized. The term z in the equation represents the correlation between the input realization x and the cluster representation $\widetilde{x}$ and can be calculated as follows:

$$z = \int_0^T x\,(t)\,\widetilde{x}\,(t)\,dt. \tag{6}$$

This quantity is often referred to as the mutual energy of the two signals. Taking Eqs. (5) and (6) into account, we can represent Eq. (4) as follows:

$$(X_i \in Y_i) = \max_j z_j^i \tag{7}$$

The correlation between input signal X_i and the j-th cluster representation $\widetilde{X}j$ is denoted by z_j^i. To prevent signal distortion, it is necessary to normalize the cluster representations as shown in Eq. (5)). When dealing with input signals of different lengths, the process is referred to as volume packing, where the average energy $\bar{E} = \frac{1}{n}\sum i = 1^n E_i = const$ is constant. If all input signals have the same length and their endpoints are on a spherical surface, it is called spherical packing.

Let's revisit Eq. (1), which forms the foundation of all neuron operations. If the weights of the output layer, denoted by $W^{k,l}$, are randomly assigned, then the vector $W^{k,l}$ acts as a multiplying interference, causing an increase in the packing volume. However, during the learning process, the weights become meaningful values determined by Eq. (7) by computing the error function and converting it into the gradient vector $W^{k,l}$. This operation is called "matched filtering" in information theory, and as the ANN's output layer is optimized during learning, it takes on the form of a matched filter. According to information theory, the condition for achieving the maximum response from a device with an impulse response is given by [15]:

$$h\,(t) = kx(-t). \tag{8}$$

n order to determine the weights of a neuron for a particular class Y_i, it is necessary for them to have a Hilbert-conjugate relationship with the ideal representation of class $\widetilde{X}_i$. This implies that if the weights are set in each neuron of the output layer based on expression (8) for each class and the function $\max_i Y_i$ is used as the output layer's function, a matched filter with a dimension of m can be obtained. However, there is an issue with this proposed solution. However, there is a certain issue with this proposed solution. The correlation integral (6) can be represented in both the time and frequency formats:

$$z_i = \int x\,(t)\,\widetilde{x}_i\,(t-\tau)\,d\tau = X\,(j\omega)\,\widetilde{X}_i\,(j\omega)\,. \tag{9}$$

Equation (1) is not suitable for calculating the correlation function in the time domain. This is because if the signals X and $\widetilde{X}$ are decomposed into an

orthogonal basis, such as the Fourier basis, all products with non-coinciding indices are set to zero, resulting in expression (1). However, if the orthogonality condition is not met, using Eq. (1) will produce correlation values (9) that contain errors. This can lead to an increase in classification errors and results that deviate from the expected outcomes.

Equation (9) indicates that the most favorable outcome could be achieved if the inputs to the output layer are orthogonal vectors. To accomplish this, a group of orthogonal functions, denoted as $\{u_n(t)\} = \{u_1(t), u_2(t), \ldots, u_n(t)\}$, is utilized. These functions fulfill the criteria (10) for each pair, and they are utilized to determine the conversion coefficients.

$$\int_0^T u_i(t)\, u_j(t)\, dt = \begin{cases} a, \forall i = j \\ 0, \forall i \neq j \end{cases} \tag{10}$$

The conversion coefficients are not difficult to determine as

$$c_j = \frac{1}{a} \int_0^T x(t)\, u_j(t)\, dt,\ j = 1, 2, \ldots, m. \tag{11}$$

The original Eq. (11) was used to transform continuous images represented by $x(t)$ to the discrete space of clusters. In the context of digital image processing, the integral in Eq. (11) was substituted with a sum.

$$c_j = \frac{1}{a} \sum_{k=0}^{n-1} x_k u_j^k. \tag{12}$$

The article by Ahmed et al. [16] provides an extensive discussion of various types of orthogonal transformations that can be used for pattern recognition. These transformations are linear and establish a one-to-one correspondence between the input vector X and the output vector of coefficients C, resulting in an n-dimensional output vector. Comparing Eqs. (12) and (1), it becomes evident that they are identical. In other words, if we substitute weights w_j^k for u_j^k, the ANN layer can represent an orthogonal transformation, and the output of the layer will have values $\{c_j\}$. By representing vector $\widetilde{X}$ as an orthogonal transformation $\widetilde{C}_x$, we obtain expression (9) in the following form:

$$Z_i = \sum_{j=0}^{n-1} X_i \widetilde{X}_i = \sum_{j=0}^{n-1} x_j \widetilde{x}_j. \tag{13}$$

Therefore, using an orthogonal transformation allows for the implementation of a feedforward ANN-based pattern recognition system.

In the study of the wave model of ANN [8–10], it was noted that both the standard and wave models had similar classification errors during the learning process, but they took different amounts of time to achieve this. This is because the standard learning algorithm, which primarily relies on the gradient method and error backpropagation, modifies the weights from the last layer to the first

(error backpropagation). Consequently, the decomposition functions in the first layer are selected based on the classification errors in the last layer. The key characteristic of the gradient used in ANN training is that it determines the direction in which a function $f(x)$ increases the most.

$$\nabla f\,(x) = \frac{df}{dx_1}e_1 + \frac{df}{dx_2}e_2 + \ldots + \frac{df}{dx_n}e_n. \tag{14}$$

The error function is represented by the vector $E = (e_1, e_2, \ldots, e_n)$, and the direction in which the function $f(x)$ does not increase is indicated by the opposite of the gradient. Using this information, the algorithm calculates the correction vector for the weights of the last layer and the previous layer based on the respective errors. The algorithm selects the decomposition functions of the first hidden layer, which become complex due to the nonlinearity of the neuron transfer function. This complexity was predicted by V.I. Arnold in [5].

The above examples indicate that using orthogonal transformations in artificial neural networks can enhance information processing. Such transformations allow for operations like correlation and convolution to be performed in appropriate planes, and the multiplication of elements with non-coincident indices in different planes is automatically excluded due to the orthogonal properties. Consequently, the use of orthogonal transformations can greatly reduce the computational burden required for image processing tasks in neural networks.

2.2 Wave Model of Convolutional ANN

Classification or clustering tasks are better suited for convolutional neural networks (CNN) than feedforward neural networks. CNN were proposed by Ian Lekun in 1988 and are known for their efficiency. They consist of one or more convolutional layers that use a small-sized kernel for the convolution operation. This operation reduces the size of the image, which is particularly beneficial for color images. A 3-dimensional kernel is used in this case to produce a single image on the layer output instead of 3. Typically, the convolutional layer is the first layer in the ANN structure and may be followed by pooling (subsampling) operations. However, we will not discuss this aspect in detail here. The output of the convolution operation is a feature map that can be classified using the last layer of the feedforward ANN.

Since the convolution integral is similar to the correlation integral (9), the advantages of using orthogonal transformations discussed in the previous section also apply to the convolution operation. Therefore, using an orthogonal transformation to represent the input signal and kernel can improve the efficiency and simplicity of the convolution calculation. Consequently, it is reasonable to use a layer that performs orthogonal transformations as the first layer in a typical CNN.

Linear transformations are commonly used in signal processing for information theory. Among them, subband encoding, which is a linear transformation, has several advantageous properties that are relevant to ANN theory. There are

two types of encoders based on linear transformation: transformation encoders and subband encoders [17]. The Fourier transform, which decomposes a signal into sinusoidal components, is an example of the first type, while the discrete cosine transform (DCT) and the Karhunen-Loève theorem are examples of the second type. These transformations are computed by convolving a finite-length signal with a set of basis functions, resulting in a set of coefficients that can be further processed. Most of these transformations are applied to non-overlapping signal blocks, and efficient computational algorithms have been developed for many of them [17].

Subband encoding applies several bandpass filters to the signal and then thins out the result by decimation. Each resulting signal carries information about a specific spectral component of the original signal on a particular spatial or temporal scale. There are several crucial properties to consider when encoding images using this method [17], including:

- scale and orientation;
- spatial localization;
- orthogonality;
- fast calculation algorithms.

In subband coding, orthogonality is not usually emphasized in communication theory. Instead, orthogonal transformations are used to decorrelate signal samples. While Fourier bases have good frequency localization, they lack spatial localization, which is not a problem when encoding a signal described by a Gaussian process. However, certain image features cannot be accurately represented by this model and require bases that are spatially localized. Filter blocks that are local and in space provide better decorrelation on average. The correlation between pixels decreases exponentially with distance, as shown by the equation:

$$R_l = e^{-\omega_0|\delta|}, \tag{15}$$

where δ is the distance variable. The corresponding spectral power density is

$$\Phi_l(\omega) = \frac{2\omega_0}{\omega_0^2 + (2\pi\omega)^2}. \tag{16}$$

To obtain smooth segments of the spectrum, it is necessary to accurately divide the spectrum at lower frequencies and approximately divide it at higher frequencies, as revealed by the Eq. (16). This process will generate subbands that exhibit white noise characteristics, with the variance directly proportional to the power spectrum within that range.

The Fourier transform is known to have a drawback in that it necessitates all of the time-related data of a signal in order to produce a single conversion coefficient. This leads to the time peak of the signal spreading throughout the frequency domain of the Fourier transform. To address this issue, the windowed Fourier transform is frequently utilized.

$$\Phi_x(\omega,\ b) = \int x(t)\, e^{-j\omega t} w(t-b)\, dt. \tag{17}$$

In this particular case, the transformation characterization involves a time window of the form $w(t-b)$. As a result, the transformation becomes time-dependent, generating a time-frequency matrix of the signal as described in [18]. By selecting the Gaussian function as the window, the inverse transformation can also be conducted using the same function.

The fixed size of the window in Eq. (17) is a major drawback, as it cannot be adapted to suit the features of the image. A wavelet transform can be used instead of the Fourier transform to overcome this limitation. The wavelet transform has the form:

$$\psi_{a,b}(t) = a^{-\frac{1}{2}} \psi\left(\frac{t-b}{a}\right). \tag{18}$$

It is evident that the basic wavelet functions are real and located at different positions in proximity to the x-axis. These wavelets are defined for a brief time interval, which is shorter than the signal period. The fundamental functions can be seen as rescaled and time-shifted versions of one another, according to Eq. (18), where b and a denote the time position and scaling factor, respectively. The direct wavelet transform can be mathematically formulated as:

$$\Phi_x(a,b) = a^{-\frac{1}{2}} \int x(t)\, \psi\left(\frac{t-b}{a}\right) dt. \tag{19}$$

The convolutional layer of a CNN is responsible for computing the convolution of the input signal block X with a core J of size $s \times s$, i.e.

$$C_{i,j} = \sum_{k=0}^{s-1} \sum_{l=0}^{s-1} X_{i+k,j+l}\ J_{k,l}. \tag{20}$$

Through the conversion of Eq. (19) for discretized signals and functions and comparing it with (20), the fundamental wavelet transform function can be depicted as the essential component of a convolutional layer. This implies that utilizing multiple fundamental functions is equivalent to applying several filters with distinct kernel sizes. Consequently, it is feasible to choose adaptable parameters for the window that accommodate the signal, enabling greater flexibility in the convolutional layer of the CNN.

The use of wavelet transforms in ANNs is not a novel concept, as it has been investigated in prior research [20]. Nonetheless, a more recent approach entails using the wavelet transform as the foundation of the convolutional layer in the initial layer of a feedforward CNN, as presented in [21]. This method is more attractive since the convolutional layer can function with several kernels simultaneously, making it possible to obtain multiple approximations within a single layer.

In communication theory, a signal can be expressed as a series of successive approximations, which can be advantageous for signal analysis. For instance, in image transmission, an initial rough version of an image can be transmitted and subsequently refined in sequence, facilitating rapid viewing of numerous images

from a database. A similar method can be employed for image recognition. If an image cannot be classified into a specific category based on the coarsest approximation, there is no need to compare it in a more precise approximation. This technique is referred to as multiscale analysis.

Multiplescale analysis involves describing the space $L^2(R)$ using hierarchical nested subspaces V_m, that do not overlap, and their union results in the limit $L^2(R)$, i.e. $\ldots \cup V_2 \cup V_1 \cup V_0 \cup V_{-1} V_{-2} \cup \ldots$, $\bigcap_{m \in Z} V_m = \{0\}$, $\bigcup_{m \in Z} V_m = L^2(R)$. These subspaces have the property that any function $f(x)$ belonging to V_m will have a compressed version that belongs to V_{m-1}, i.e. $f(x) \in V_m \Leftrightarrow f(2x) \in V_{m-1}$. Additionally, there exists a function $\varphi(x) \in V_0$, whose shifted versions $\varphi_{0,m}(x) = \varphi(x - m)$ form an orthonormalized basis of space V_0. The functions $\varphi_{n,m}(x) = 2^{-\frac{m}{2}} \varphi(2^{-m}x - n)$ form an orthonormal basis of space V_m. These basis functions are called scaling functions as they create scaled versions of functions in $L^2(R)$ [17]. Thus, a function $f(x)$ in $L^2(R)$ can be represented by its set of successive approximations $f_m(x)$ in V_m.

Therefore, it is possible to perform image analysis at various resolution or scale levels by selecting the value of m, which is known as the scale factor or level of analysis. A higher value of m results in a coarser approximation of the image, lacking in details, but allowing for identification of broader generalizations. Decreasing the scaling coefficient enables identification of finer details. In essence, $f_m(x)$ is an orthogonal projection of $f(x)$ onto V_m [17], i.e.

$$f_m(x) = \sum_n \langle \varphi_{m,n}(x), f(x) \rangle \varphi_{m,n}(x) = \sum_n c_{m,n} \varphi_{m,n}(x). \tag{21}$$

Without delving into the specifics of wavelet analysis at present, it is worth mentioning that any function $f(x)$ within the space $L^2(R)$ can be expressed as a combination of orthogonal projections. When analyzing the function up to a specific scale factor m, the function $f(x)$ can be represented as the addition of its crude approximation and various details. The Haar wavelet family, for example, offers such functionalities [18].

When employing subband transforms, the potential for constructing filter banks must be taken into account, which involve filtering followed by downsampling [17,19]. In a two-band filter bank, the low-frequency component provides a crude estimation of the signal without capturing intricate details, while the high-frequency component contains finer details. Depending on the particular processing objective, an ANN can utilize the low-frequency approximation to emphasize broad and smooth features, or the high-frequency component to emphasize specific details.

Utilizing wavelets as the kernel of a CNN enables the extraction and enhancement of the necessary image features. While this approach is not new in information processing and transmission theory, it is being utilized to establish an information model for CNNs. This technique not only advances our comprehension of the process of feature map generation but also simplifies the development of a lifting scheme for information processing in a multi-layer CNN.

3 Results

Using orthogonal transformations can be advantageous when working with images, irrespective of the ANN architecture employed. For instance, in feedforward ANNs, the use of orthogonal transformations can improve the efficiency of the final layer where image classification or clustering is performed. Orthogonalizing the data can enhance the accuracy of computing the correlation integral for the classified signal and ideal class representation.

Convolutional neural networks (CNNs) employ feedforward networks in their last layer, similar to traditional feedforward ANNs, which is essential for feature map classification. To enhance the efficiency of the last layer in CNNs, orthogonal transformations are utilized, as in feedforward ANNs. However, when analyzing image details, the Fourier transform (or similar ones) does not offer significant benefits. Therefore, wavelet transforms are more promising as they have localization in both frequency and time, unlike the window Fourier transform. Wavelets can also function as orthogonal transformations and enable the creation of filter banks for general and detailed image analysis based on specific criteria. This approach not only allows for general image classification, as in the case of the MNIST database, but also enables complex image classification based on specific details.

To confirm the effectiveness of the approach described above, experimental validation is necessary. The next step is to explore the wavelet transforms currently available for CNNs and their implementation in convolutional layers. It is essential to ensure that the feature maps are sufficiently detailed to enable efficient processing in subsequent layers.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Re- search within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The study was financially supported by the Russian Foundation for Basic Research within the framework of the scientific project No. 20-37-51004 “Effective intelligent data management system in edge, fog and cloud computing with adjustable fault tolerance and security” and Russian Federation President Grant SP-3186.2022.5.

References

1. Kruglov, V.V., Borisov, V.V.: Artificial neural networks. Theory and practice (Iskusstvennye nejronnye seti. Teoriya i praktika). — M.: Goryachaya liniya — Telekom (2002). (in Russian)
2. McCulloch, W., Pitts, W.: A logical calculus of ideas immament to nervous activity (Logicheskoe ischislenie idej, otnosyashchihsya k nervnoj aktivnosti). — Avtomaty. — M.: Izd. inostr. lit. (1956). (in Russian)
3. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall (1999)

4. Kolmogorov, A.N.: On the Representation of Continuous Functions of Several Variables as Superpositions of Continuous Functions of One Variable and Addition (O predstavlenii nepreryvnyh funkcij neskol'kih peremennyh v vide superpozicij nepreryvnyh funkcij odnogo peremennogo i slozheniya). - Dokl. AN SSSR, 1957, T. 114, vol. 5, pp. 953–956 (1957). (in Russian)
5. Arnol'd, V.I.: On the Representation of Functions of Several Variables as a Superposition of Functions of Fewer Variables (O predstavlenii funkcij neskol'kih peremennyh v vide superpozicii funkcij men'shego chisla peremennyh). Mat. Prosveshchenie **3**, 41–61 (1958). (in Russian)
6. Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley (1989)
7. Shannon, K.: Works on information theory and cybernetics. (Raboty po teorii informacii i kibernetike). Izd-vo inostrannoj literatury (1963). (in Russian)
8. Vershkov, N.A., Kuchukov, V.A., Kuchukova, N.N., Babenko, M.: The wave model of artificial neural network. In: Proc. IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus: Moscow, St. Petersburg 2020, pp. 542–547 (2020)
9. Vershkov N.A., Babenko M.G., Kuchukov V.A., Kuchukova N.N. Advanced supervised learning in multi-layer perceptrons to the recognition tasks based on correlation indicator. //Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 1, pp. 33–46 (2021)
10. Vershkov, N.A., Kuchukov, V.A., Kuchukova, N.N. The Theoretical Approach to the Search for a Global Extremum in the Training of Neural Networks. Trudy ISP RAN/Proc. ISP RAS, vol 31, issue 2, pp. 41–52 (2019). https://doi.org/10.15514/ISPRAS-2019-31(2)-4
11. Kotel'nikov, V.A.: Theory of Potential Noise Immunity. (Teoriya potencial'noj pomekhoustojchivosti). - Radio i svyaz' (1956). (in Russian)
12. Harkevich, A.A.: Selected Works. Vol. 3. Information Theory. Pattern Recognition. (Izbrannye trudy. T.3. Teoriya informacii. Opoznanie obrazov.) – M.; Nauka (1972). (in Russian)
13. Ipatov, V.: Broadband systems and code division of signals. Principles and Applications. (Shirokopolosnye sistemy i kodovoe razdelenie signalov. Principy i prilozheniya.) - M.: Tekhnosfera (2007). (in Russian)
14. Qiao, Yu.: THE MNIST DATABASE of handwritten digits (2007). Accessed 04.08.2021
15. Cook, C., Bernfeld, M.: Radar signals. Theory and application. (Radiolokacionnye signaly. Teoriya i primenenie.) - Sovetskoe Radio, Moscow (1971). (in Russian)
16. Ahmed, N., Rao, K.R.: Orthogonal transformations in digital signal processing (Ortogonal'nye preobrazovaniya pri obrabotke cifrovyh signalov): Per. s angl./Pod red. I.B. Fomenko. - M.: Svyaz', (1980). (in Russian)
17. Vorob'ev, V.I., Gribunin, V.G.: Theory and practice of wavelet transforms. (Teoriya i praktika vejvlet-preobrazovaniya.) - S.-Peterburg: Voennyj universitet svyazi (1999). (in Russian)
18. Sikarev, A.A., Lebedev, O.N.: Microelectronic devices for forming and processing complex signals. (Mikroelektronnye ustrojstva formirovaniya i obrabotki slozhnyh signalov.) - M.: Izd-vo <<Radio i svyaz'>> 1983. (in Russian)
19. Haar, A.: Zur theorie der orthogonalen funktionensysteme. Georg-August-Universitat, Gottingen (1909)
20. Genchaj, R., Sel'chuk, F., Uitcher, B.: Introduction to wavelets and other filtering techniques in finance and economics. (Vvedenie v vejvlety i drugie metody fil'tracii v finansah i ekonomike.) - Academic Press (2001). (in Russian)

21. Alexandridisa, A.K., Zapranisb, A.D.: Wavelet neural networks: a practical guide. Neural Networks, vol. 42, pp. 1–27 (2013)
22. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification (2016). arXiv preprint arXiv:1603.06995