Optimization of computational complexity of an artificial neural network

Nikolay Vershkov[1‡†], **Viktor Kuchukov**[1‡‡], **Natalia Kuchukova**[1††], **Nikolay Kucherov**[1†] **and Egor Shiriaev**[1‡]

[1] North-Caucasus Center for Mathematical Research, North-Caucasus Federal University,1, Pushkin Street, 355017, Stavropol, Russia

E-mail: [‡†]vernick61@yandex.ru [‡‡]vkuchukov@ncfu.ru, [††]nkuchukova@ncfu.ru, [†]nkucherov@ncfu.ru, [‡]ea_or@list.ru

Abstract. The article deals with the modelling of Artificial Neural Networks as an information transmission system to optimize their computational complexity. The analysis of existing theoretical approaches to optimizing the structure and training of neural networks is carried out. In the process of constructing the model, the well-known problem of isolating a deterministic signal on the background of noise and adapting it to solving the problem of assigning an input implementation to a certain cluster is considered. A layer of neurons is considered as an information transformer with a kernel for solving a certain class of problems: orthogonal transformation, matched filtering, and nonlinear transformation for recognizing the input implementation with a given accuracy. Based on the analysis of the proposed model, it is concluded that it is possible to reduce the number of neurons in the layers of neural network and to reduce the number of features for training the classifier.

1. Introduction

The necessity to represent functions of n variables in the form of a superposition of functions of smaller number variables arose in connection with the development of the neural networks theory and practice. The appearance of neural networks is associated with an article by McCulloch et al. [1], which describes a mathematical model of a neuron and a neural network. It has been proven that both Boolean functions and finite state machines can be represented by neural networks. Later, a serious mathematical analysis of perceptron revealed limitations on the area of their applicability. Later, the restrictions were weakened by replacing the threshold activation functions of neurons with sigmoid ones. The theoretical basis of Artificial Neural Networks (ANNs) was the Kolmogorov-Arnold theorem proved as a result of scientific discussion [2, 3], which showed the possibility of representing a continuous real function of n variables $f(x_1, x_2, \ldots, x_n)$ in the form of a superposition of functions of a smaller number variables. A significant theoretical development of ANN was the proof of the Hecht-Nielsen theorem [4], which showed the possibility of approximating a function of several variables with a given accuracy of ANN with one hidden layer in a non-constructive form. Interest in Deep Networks stems from the limitations of the perceptron. The use of multilayer networks was initially limited

by the complexity of their training. Due to the ideas of Hinton's team, training multilayer ANNs became possible [5]. Multilayer networks enabled solving problems of classification, extrapolation, feature extraction, etc. in conditions of high uncertainty, i. e. with a sufficiently small volume of the training sample, obtain satisfactory results. Thus, the modern theory of ANN is based on a vector (geometric) approach [2, 3, 4, 5, 6, 7].

An interesting approach to the problem of constructing image recognition systems of optimal architecture was proposed by Rao et al. [8]. They proposed to present the pattern recognition system functionally in the form of two blocks: feature selection and a trained classifier. The selection of features is carried out using orthogonal transformations of the input signal. In order to increase the efficiency of the system, a method for decreasing the dimension of the feature vector for training the classifier is proposed. The ANN input receives a sequence of values $\{X_i\} = (x_1^i, x_2^i, \ldots, x_n^i)$, which can be represented as discrete samples of some continuous function $x(t)$. ANN output is a sequence $\{Y_i\} = (y_1^i, y_2^i, \ldots, y_m^i)$, which can also represent discrete samples of the function y(t). Here i is the number of the sample, m and n are the dimensions of the output and input samples of the sequences, respectively. This approach makes it possible to study the output sequence of the ANN not only as a geometric interpretation of the input samples, but also to consider the information interaction of the layers of complex (deep) ANNs using the mathematical apparatus of the information transmission theory.

We developed the ideas of Ahmed and Rao, widely applying the methods of decoding and separating signals against the background of noise, used in the theory of information transmission [9]. However, the use of the discrete Fourier transform (DFT) did not allow obtaining a significant gain in reducing the computational complexity of the ANN [9]. Despite the significant improvement in the algorithms of the ANN operation and the reduction of the training time by tens and hundreds of times, a number of problems remain in this area that require theoretical comprehension. First of all, these include a significant computational load and, as a result, a significant training time.

The solution of the above problem using existing theoretical approaches is difficult, and the proposed ANN model in the form of an information transmission system will allow studying the interaction of layers and significantly reduce the complexity of training.

Within the framework of this article, the following restrictions apply. We did not seek to consider all the diversity of the ANN architecture, but limited themselves to feed-forward networks. This article did not set the task of a complete study of the ANN and obtaining practically significant results, but considered the problem of constructing a mathematical model of the ANN based on the processes of parallel digital processing of input information, presented as a random process containing a deterministic signal that must be attributed to a certain class. We suggest that such an approach will significantly reduce the costs of training ANNs and, thereby, increase the efficiency of their application.

2. Feature selection as an orthogonal transformation of the input vector

Studying the problem of feature selection, we relied on the provisions formulated by Rao et al. [8]. ANN input vector $\{X_i\} = (x_1^i, x_2^i, \ldots, x_n^i)$ is considered as discretization of the continuous signal $x(t)$ considering the provisions of the Kotelnikov theorem (better known abroad as the Nyquist-Shannon theorem). The output signal $y(t)$ can also be represented by discrete values $\{Y_i\} = (y_1^i, y_2^i, , y_m^i)$. Investigating the ANN, we proceed from the classical scheme of the information transmission system using wideband signals [10, 11]. The transmission of broadband (complex) signals is characterized by the shape of a time-frequency matrix [11]. In [6] distinguish 3 types of time-frequency matrices: parallel, serial and serial-parallel. A signal with a serial-parallel matrix is fed to the input of the ANN.

Let the function $x(t)$ be a complex signal consisting of i variants, each of which is encoded with a sequence of n symbols. At the output, we get the function $y(t)$, consisting of i variants, each

of the length m. Considering the principles of complex signals analysis, it is more convenient to represent them in a generalized spectral form, i.e. each version of the signal can be represented in the form [5, 10, 11]:

$$\begin{cases} x_r(t) = \sum_{k=k_{r_1}}^{k_{r_2}} a_{k_r}\phi_k(t) \\ y_l(t) = \sum_{k=k_{l_1}}^{k_{l_2}} a_{k_l}\phi_k(t) \end{cases}, t \in [0,T] \tag{1}$$

There $T = n\Delta t_x = m\Delta t_y$, and the expansion coefficients

$$\begin{cases} a_{kr} = \frac{1}{\int_0^T \phi_k^2(t)dt}\int_0^T x_r(t)\phi_k(t)dt \\ a_{kl} = \frac{1}{\int_0^T \phi_k^2(t)dt}\int_0^T y_l(t)\phi_k(t)dt \end{cases} \tag{2}$$

The coordinate functions $\phi_k(t)$ satisfy the orthogonality condition [11]. Representation (1) makes it possible to understand the procedure for processing complex signals not only in the time domain, but in the time-frequency domain, as it happens with the classical ANN design methods. Using orthogonal transformations, the features (spectrum) of the input signal will be selected, which, after processing, can be used to train the classifier. Thus, the ANN in the form of an information transmission system (ITS) performs the transformation $y(t_i) = F(x(t_i))$, where $x(t_i) = \sum_{i=1}^n x_i(i\Delta t_x)$ and $y(t_i) = \sum_{i=1}^n y_i(i\Delta t_y)$ and the functional of F is the subject of this article.

The classical approach (McCulloch-Pitts model [1]) considers a mathematical model of a neuron in the form $y_{k,l} = f(\sum_{i=1}^n w_i^{k,l}x_i^{k,l})$, where k, l are the number of the layer and the number of the neuron in the layer, $y_{k,l}$ is the output of the neuron, $x_i^{k,l}$ are the inputs of the neuron, $w_i^{k,l}$ are the weights (synapses) of input signals, f is the output function of the neuron, which may or may not be linear. The transformation of a signal in a neuron can be considered in the traditional sense as an algebraic sum of products of input signals and weights (adaptive adder), and in the sense of expressions (1),(2). Indeed, if the set of weights of the k-th neuron of the i-th layer $\{w_k^i\}_{k=0,1,2,\ldots}$ is a discrete representation of the i-th orthogonal function $\phi_i(t)$, then the output will be

$$a_i = \sum_{k=0}^{n} x(k\Delta t)\phi_i(k\Delta t) \tag{3}$$

Applying to expression (3) a normalizing coefficient of the form $k = \frac{1}{\int_0^T \phi_i^2(t)dt}$, denoting

$$w_i(k\Delta t) = \frac{1}{\int_0^T \phi_i^2(t)dt}\phi_i(k\Delta t) \tag{4}$$

and summing over the period, we arrive at expression (2).

Let us consider the application of the orthogonal transform based on the widely used Fourier transform [10, 11]. However, in some cases, instead of trigonometric functions, some others are more appropriate as a core, such as Lagger, Legendre, Hermite, Walsh, Chebyshev, Hadamard, etc. Since the input and output signals are presented in discrete form, we use a kind of Fourier transform for discrete signals, discrete Fourier transform (DFT).

Thus, choosing the weights $w_i(k\Delta t)$ in accordance with (4), at the output of the i-th neuron, we obtain the value of one of the spectral components a_i. A layer of neurons with selected weights $w_i(k\Delta t)$ for spectral components with serial numbers $i = 0, 1, 2, 3, \ldots$ gives the spectrum of the input signal with a given accuracy as a result of the transformation.

Let us consider in more detail the process of orthogonal transformation of the DFT based on the McCulloch-Pitts model [1]. The DFT is based on the well-known expression for the

continuous Fourier transform $X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\infty t}dt$ Passing from the continuous form to the discrete one, we replace the integration with the summation and, introducing a restriction on the signal spectrum width, we obtain the expression $X(m) = \sum_{i=0}^{n-1} x(i)e^{-j2\pi im/n}$. A similar expression can be implemented using complex numbers, but, more conveniently, it can be reduced to the form

$$\begin{array}{l} X(m) = \sum_{i=0}^{n-1} x(i)\left(\cos\left(\frac{2\pi im}{n}\right) - j\sin\left(\frac{2\pi im}{n}\right)\right) = \\ = \sum_{i=0}^{n-1} x(i)\cos\left(\frac{2\pi im}{n}\right) - j\sum_{i=0}^{n-1} x(i)\sin\left(\frac{2\pi im}{n}\right) \end{array} \quad (5)$$

using Euler's identity $e^{-j\phi} = \cos(\phi) - j\sin(\phi)$.

Here $X(m)$ is the m-th harmonic of the DFT, m is the index in the frequency domain, $x(i)$ is a sequence of input samples of size n. In this case, the number of neurons in the layer must be at least $2n$ to convert the real (cosine) and imaginary (sine) parts of the complex number, which is in agreement with the Hecht-Nielsen theorem [8]. It requires the number of neurons in the first hidden layer at least $2n + 1$. If a different orthogonal transformation is used as a core function, then the substitution of weights is performed based on the transformation used.

In addition to DFT, in the practice of digital signal processing, the discrete cosine transform (DCT) is widely used [8]. DCT is used to compress images in MPEG and JPEG formats. DCT is closely related to the Fourier transform and is a homomorphism of its vector space. Since DCT operates with real numbers, it does not require $2n$ neurons in a layer, n is enough, which reduces the number of neurons in the first hidden layer by 2 times compared to DFT. To solve the problem of reducing the computational load on the ANN, we use DCT.

3. Implementation of the classifier based on the McCulloch-Pitts model

For successful classification of the i-th implementation of the input value, it is necessary to perform the operation of determining the degree of similarity of the implementation with the values that determine the classification levels. The classical approach for training the classifier is based on adaptive algorithms that minimize the value of the working function [12, 13] As a working function for training ANN with a teacher, the mean square of the neural layer error is widely used. Newton's and steepest descent methods are used as algorithms, as well as their variations [13].

Since the proposed model widely uses digital signal processing methods, the implementation of the classifier is based on the criteria of optimal filtering. A similar approach was considered in [9], where the correlation function of the ANN output signal and the training sequence was proposed as a working function. Let each realization of the input value be a deterministic signal, which is the average value of all realizations from the training sample belonging to a certain class mixed with additive noise, i.e. $X_i = Z_j + N$. Here X_i is the input implementation belonging to the j-th class. $Z_j = \sum_{i=0}^{p-1} X_i/p$ is the number of input implementations belonging to the j-th class. N is a random process, which has a Gaussian distribution. In the theory of information transfer, the ratio of the maximum value of the input implementation to the standard deviation of the noise is used as a characteristic that determines the ratio of the implementation to a certain class [10]:

$$q = \frac{\max(X_i)}{\sigma_{out}} \quad (6)$$

The optimal filter tends to maximize the value (6), i.e. $q \to \max$. Since after passing through the first hidden layer, the input implementation is presented in the form of a spectrum, $X_{iout}(\omega) = X_i(\omega)H(\omega)$, where $H(\omega)$ is the frequency response of the output layer. Omitting intermediate calculations [10], we get:

$$q_{opt} = \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|X(\omega)|^2}{N(\omega)} d\omega} \quad (7)$$

Based on the Bunyakovsky-Schwarz inequality and expression (7), we obtain the complex frequency response of the optimal circuit:

$$H(\omega) = k\frac{X^*(\omega)}{N(\omega)}e^{-j\omega t} \tag{8}$$

Based on expression (8), the amplitude-frequency characteristic of the optimal filter will be the amplitude-frequency characteristic of the input implementation with an accuracy up to the coefficient k:

$$H(\omega) = k\,|X(\omega)| \tag{9}$$

Expression (9) allows us to define the ANN output layer as an $m \times n$ matrix, i. e. the number of matrix rows is determined by the number of classes, columns – by the dimension of the input implementation, and the values of the matrix rows contain the averaged values of the input implementations related to each class. In accordance with the McCulloch-Pitts model, the value at the output of the j-th neuron of the output layer is defined as

$$Y_j = \sum_{i=1}^{n} X_i H_{ij} \tag{10}$$

In expression (10), H_{ij} is the j-th row of the output layer's matrix of weights, which is the mathematical expectation of the implementations X_i of the training sample belonging to the j-th class. The physical meaning of expression (10) means the correlation function of the input implementation with the representation of the class, defined in the form of the output layer weights.

Based on the above reasoning and the obtained expression (10), an ANN with one hidden layer was built based on the PyTorch library [14]. The experiment was carried out using the MNIST database [11]. The ANN layer weights were filled in as follows: the first hidden layer implements DCT, its dimension is equal to the dimension of the input implementation n, and the output layer performs the function of an optimal receiver based on the correlator. Without the use of nonlinear functions of the layers, the ANN gives the recognition accuracy on test data of 72%. In order to assess the correctness of the choice of the mathematical expectation criterion for the output layer, we will train the classifier, i.e. ANN output layer using the gradient method. To do this, we artificially prohibit changing the weights of all layers, except for the last one. The result of training the classifier is shown in figure 1. Due to the training, the recognition reliability was increased to 90%, which indicates that the weights in expression (10), represented by the mathematical expectation of class realizations, are not an ideal criterion for the optimal receiver. The use of a nonlinear layer, which is a ReLu function, increases the recognition capabilities of test data from the MNIST database up to 95%.

4. Minimization of the feature vector in training the classifier

The next step, in accordance with [8], is to minimize the feature vector (in our case, the spectrum of the input signal) in such a way that, with insignificant losses in the information content of the feature vector, reduce its dimension. To do this, Ahmed and Rao propose to conduct an analysis of variance of the feature vector in order to identify features with minimum variance, i. e. not informative, and exclude them from the number of analyzed ones. To assess the degree of influence of harmonics with low dispersion, we use an ANN with one hidden layer and the ReLu function. The result of the successive removal of uninformative harmonics is shown in figure 2.

It is clearly seen from the presented graph that the removal of more than 300 of 784 harmonics has no effect at all on the recognition accuracy of test data. Removing also 200 leads to a decrease in the recognition quality by 1%. That is, leaving approximately 250 features of 784, we can

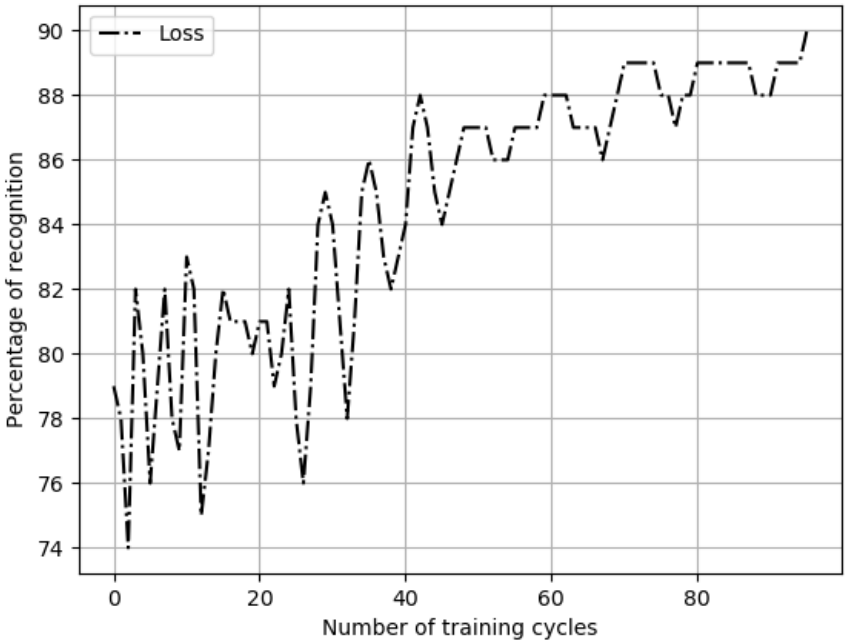

Figure 1. Classifier training result.

recognize the test data with 94% reliability. Thus, the first hidden layer of the ANN can be reduced by approximately 70% with a decrease in reliability of no more than 1%. The weight matrix of the output layer also will be reduced by 70%. Therefore, the gain from the application of the proposed model in comparison with the standard one [2, 3, 4] is approximately 84%.

5. Conclusion

Despite the significant gain in ANN performance, the proposed model has a number of disadvantages. These include the low reliability of test data recognition: 95% versus 98% or more for ANNs implemented on the basis of the traditional model. This can be explained by the fact that the simplest ANNs with one hidden layer have been simulated so far. In addition, the use of the averaged value as a measure of the similarity with the implementation requires further study, since further training is needed, and therefore computational costs. We hope that further research in the proposed direction will reduce the recognition error and expand the range of application of the proposed model.

Acknowledgements This work was supported by a grant from the Russian Science Foundation Grant No. 19-71-10033.

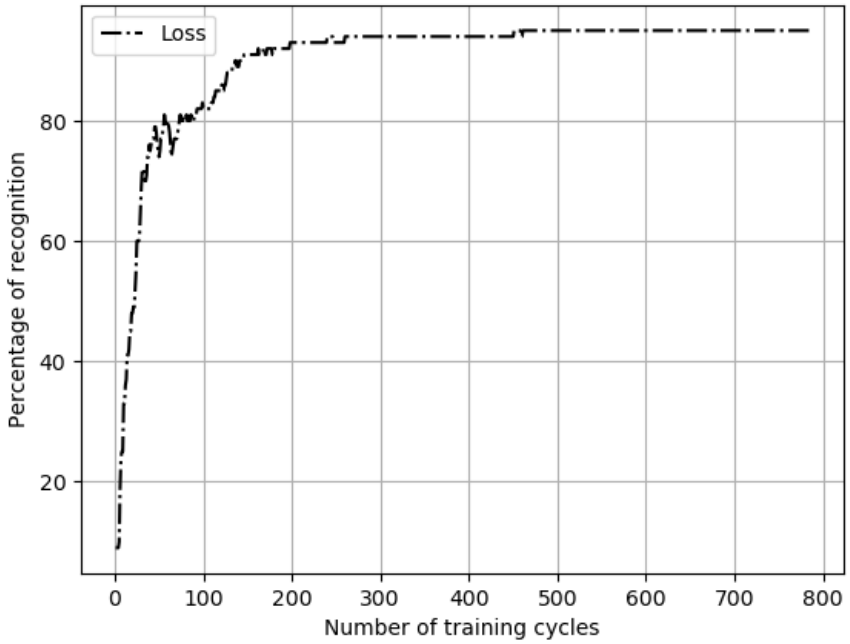

Figure 2. The result of removing harmonics from the spectrum.

References

[1] McCulloch W S and Pitts W 1943 *The bulletin of mathematical biophysics* **5** 115–133

[2] Kolmogorov A N 1957 On the representation of continuous functions of several variables in the form of superpositions of continuous functions of one variable and addition *Doklady Akademii nauk* vol 114 (Rossijskaya akademiya nauk) pp 953–956

[3] Arnol'd V I 1958 *Mat. Prosveshchenie* **3** 41–61

[4] Hecht-Nielsen R 1988 *IEEE spectrum* **25** 36–41

[5] Hinton G E 2007 *Trends in cognitive sciences* **11** 428–434

[6] Alexandrovich S A 1983

[7] Klod S 1963 *Works on information theory and cybernetics* (Foreign Literature Publishing House)

[8] Rao K and Ahmed N 1976 Orthogonal transforms for digital signal processing *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing* vol 1 (IEEE) pp 136–140

[9] Vershkov N A, Kuchukov V A, Kuchukova N N and Babenko M 2020 The wave model of artificial neural network *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)* (IEEE) pp 542–547

[10] Vasilievich S A 1967 *Information theory and its application to automatic control problems* (Publishing House "Science" Head edition of physical and mathematical literature)

[11] Deng L 2012 *IEEE Signal Processing Magazine* **29** 141–142

[12] Hebb D O 1949 *A Wiley Book in Clinical Psychology* **62** 78

[13] Widrow B 1959 *Part* **4** 74–85

[14] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L *et al.* 2019 *arXiv preprint arXiv:1912.01703*