UDC 519.71

DOI: 10.15587/1729-4061.2024.318554

# OPTIMIZATION OF IMAGE COMPRESSION USING ARTIFICIAL NEURAL NETWORKS

**Oleksandr Lytvyn***
**Nadiya Kolos**
*Corresponding author*
PhD*
E-mail: nadiya.kolos@lnu.edu.ua
*Department of Discrete Analysis
and Intelligent System
Ivan Franko National University of Lviv
Universytetska str., 1, Lviv, Ukraine, 79000

***The object of research is artificial neural networks of adaptive resonance theory (ART). ART neural networks are classified by matching input data to one of the existing classes, provided that the data is sufficiently similar to the Class prototypes. Continuous and discrete adaptive resonance theory networks ART-1 and ART-2 work effectively in recognition systems, especially in conditions of high uncertainty, when it is necessary to identify a large number of different images.***

***The main problem that was solved in this study was to optimize the process of image compression using artificial neural networks, because image compression is widely used in many scientific and technical fields and becomes especially relevant when transmitting over narrow-band communication channels. A way to overcome these difficulties may be to select basic data for reconstruction from an open data set (Modified National Institute of Standards and Technology) – Fashion-MNIST. There are still unresolved issues related to the fact that lossy compression algorithms with increasing compression ratio usually generate artifacts that are clearly visible to the human eye.***

***A compression algorithm based on neural networks is described, which establishes a correspondence between the input and output spaces consisting of elements of the codebook and neurons. The proposed method uses a different approach (First Order), rather than a simple difference coding scheme (zero order), where the new code is calculated by subtracting the previous encoded block. The peak signal-to-noise ratio of PSNR and the root-mean-square error (MSE) of these algorithms is 24.7 DB with a compression ratio of 25.22.***

***The main area of practical use of the results obtained is improved image compression for processing large – volume video and photo materials without significant loss of quality***

***Keywords: image compression, image processing, neural network, compression method, compression algorithm***

*Received 04.10.2024*
*Received in revised form 27.11.2024*
*Accepted 11.12.2024*
*Published 25.12.2024*

***How to Cite**: Lytvyn, O., Kolos, N. (2024). Optimization of image compression using artificial neural networks. Eastern-European Journal of Enterprise Technologies, 6 (2 (132)), 23–35. https://doi.org/10.15587/1729-4061.2024.318554*

## 1. Introduction

The constant growth of the amount of processed information, in particular visual information, creates the need for its compact storage. Traditional approaches that use prefix or arithmetic coding require knowledge of the statistics of the appearance of elements. The use of statistical coding methods that take into account the conditional probabilities of the appearance of symbols in different contexts requires significant computational resources. The problem with the combinatorial approach is that the image has both width and height, which makes it difficult to determine repeated substrings [1].

Image quality can be described both from a technical point of view and objectively, indicating deviations from an ideal or reference image. It can also relate to subjective perception or prediction, for example, the image of a person's appearance. The appearance of noise affects the quality of the image, and this effect depends on how much the noise interferes with the information that the viewer is trying to obtain from the image. Visual information processing includes several stages, such as capture, enhancement, compression and transmission. After processing, some of the information embedded in the image characteristics may be distorted. Therefore, the image quality assessment should be carried out taking into account human perception.

There are various methods and metrics for objectively assessing image quality, which are divided into two main groups, depending on the availability of a reference image. The main categories are as follows:

1) full-reference (FR) methods: these methods compare the test image with a reference image, which is considered to be an image of ideal quality. For example, to assess the quality of JPEG image compression, the original and compressed images can be compared;

2) no-reference (NR) method: does not require a reference image and is aimed only at assessing the quality of the test image itself [2].

One of the most common and simplest metrics of full-reference assessment is the mean square error (MSE), which is calculated as the average of the squared differences between the pixel intensities of the reference and distorted images. Based on this, the peak signal-to-noise ratio (PSNR) is also calculated. Metrics such as MSE and PSNR are widely used due to their simplicity, physical interpretation, and convenience for mathematical implementation during optimization. However, sometimes they do not fully correspond to the visual perception of image quality and may be unnormalized in the context of presenting the results. However, the presence of sampling operations in multimedia data allows for effective lossy compression of information. In some cases, such as in the representation of still images, this compression is justified. The JPEG standard includes both types of compression, but does not always meet modern requirements. Artificial neural networks are promising for lossless compression, for example, in statistical coding methods for estimating the probability of occurrence of symbols. They can also be used for lossy compression, in particular in the JPEG 2000 format, which is based on the wavelet transform. In the case of lossy compression, the use of artificial neural networks to implement vector quantization is particularly effective. That is why the use of artificial neural networks (ANNs) is promising for the development of new methods of image compression.

There are 3 main types of artificial neural networks capable of image compression: the Kohonen network and its variations, neural networks with associative memory, for example, the Hopfield network, and autoencoders.

The Kohonen network refers to unsupervised learning methods. The network itself and its variations are often used to compress images with loss of quality. It allows to allocate similar fragments of data into classes. The class number usually takes up much less space in memory than the class kernel. If to transfer to the recipient all the class kernels and class numbers encoding each fragment of data, the data can be restored. In this case, losses are inevitable if the number of classes is less than the number of different data fragments.

"Data compression" for Hopfield networks is a side effect, since their main purpose is to restore the original image from noisy or damaged input data.

Autoencoders are feed-forward artificial neural networks that learn to reconstruct the input signal at the output of the network. They are characterized by the presence of a hidden layer, usually of lower dimension. In the general case, it is a code that describes the data entering the network input and is used to reconstruct it in the decoder.

## 2. Literature review and problem statement

In [3], the results of research on the image compression module based on neural network autoencoders are presented. Approaches to image reproduction using neural network convolution and inverse convolution layers are analyzed. It is shown that data compression in the form of a neural network module based on the structure of autoencoders has the most optimal learning time, compression level and obtains a sufficiently clear image reconstruction. A new approach to data compression in the form of a neural network module based on the structure of autoencoders is proposed, which has the most optimal learning time, compression level and obtains a sufficiently clear image reconstruction. However, issues related to atypical behavior during the increase in layers in the structure of the autoencoder, which do not lead to an increase in the quality of image reproduction, remain unresolved. The reason for this may be objective difficulties associated with a significant loss of image quality when increasing the layers, which makes the relevant studies impractical. An option to overcome the corresponding difficulties may be the selection of basic data for reconstruction from the open Fashion-MNIST data set, which will allow for simplified testing of neural network structures, the process of their training and obtaining results.

In [4], the results of research on image compression and protection using neural networks are presented. It is shown that lossy compression algorithms, when the compression ratio increases, usually generate artifacts that are clearly visible to the human eye. However, the issues related to the fact that lossy compression algorithms, when the compression ratio increases, usually generate artifacts that are clearly visible to the human eye remain unresolved. An option to overcome the corresponding difficulties may be the combination of two well-known algorithms: Kohonen artificial neural network and Grossberg star (Fig. 1, 2).

At the same time, properties appear that none of them have separately. The used algorithm for compressing image data using an artificial neural network can be attributed to the class of lossy compression, which allows to obtain images that are much smaller in size than the original with quite significant deformations. This, in turn, allows to significantly speed up data transmission over communication channels and, if intercepted by an attacker, make it impossible to restore the original without using this algorithm, thus protecting them. Also, images compressed using this algorithm are convenient to use in steganography, because it will be difficult to determine the fact of transmitting hidden information.

In [5], the results of a comparative study of image compression based on compression measurement are presented. For this purpose, an artificial Hopfield neural network was developed to obtain stereo images as a cost minimization function (Fig. 3, 4). This cost function is minimal when the system is in an equilibrium or stable state.

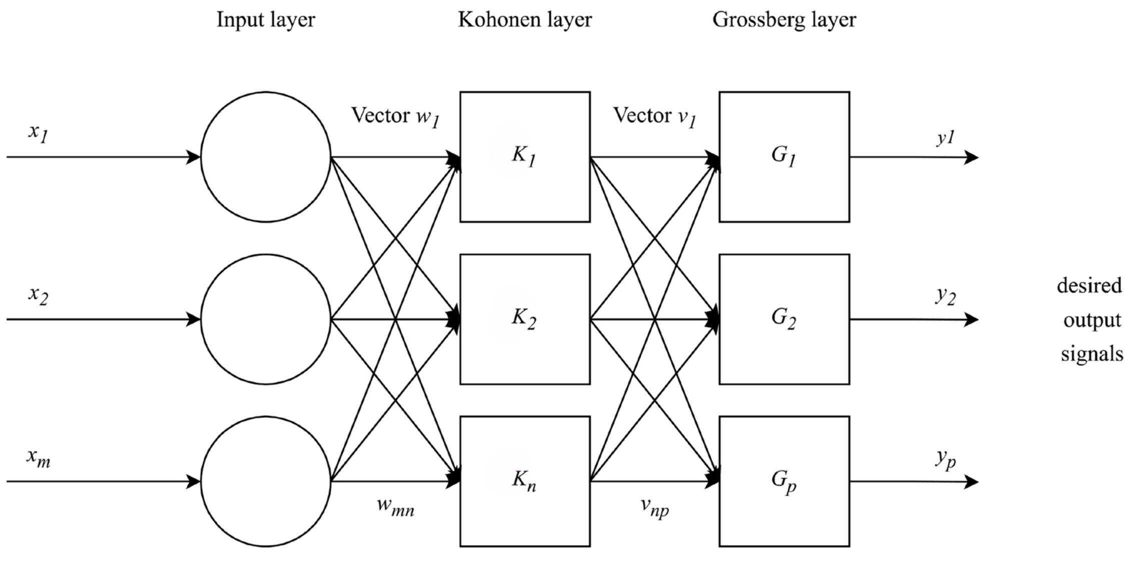

Fig. 1. Schematic view of the Kohonen network [4]

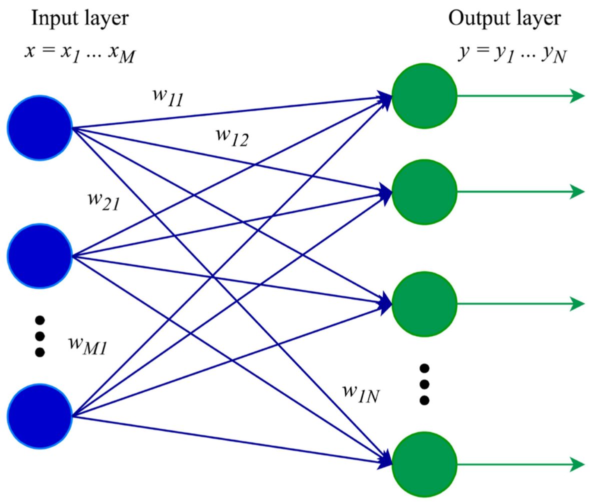

Fig. 2. Image of a backpropagation network without feedback [4]

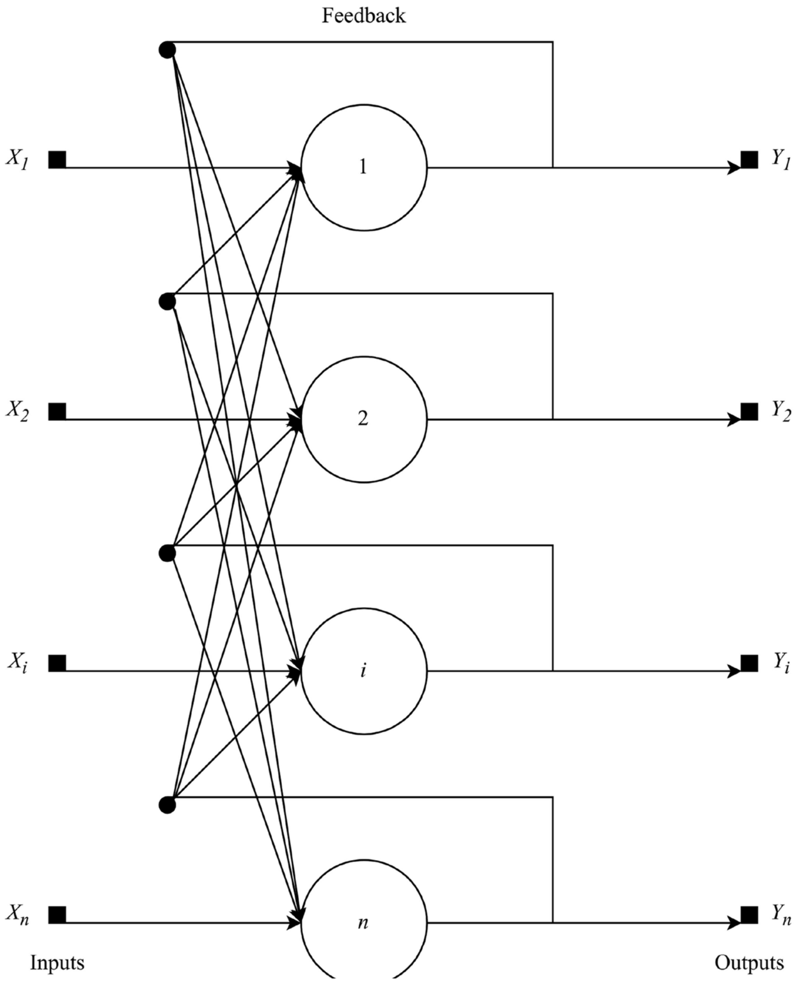

Fig. 3. Schematic representation of the structural diagram of an artificial Hopfield neural network [5]

For better image optimization, a Hamming network is proposed as an extension of the Hopfield network and which implements a classifier based on the smallest error for binary input vectors, where the error is determined by the Hamming distance (Fig. 5, 6). This distance is defined as the number of bits that differ between two corresponding fixed-length input vectors. In the training mode, the input vectors are distributed into categories for which the distance between the sample input vectors and the running input vector is minimal. The desired training image is input to the input layer, and the value of the desired class to which the vector belongs is output from the output layer. The output contains only the value of the class to which the input vector belongs. The recursive nature of the Hopfield layer provides a means of correcting all connection weights.

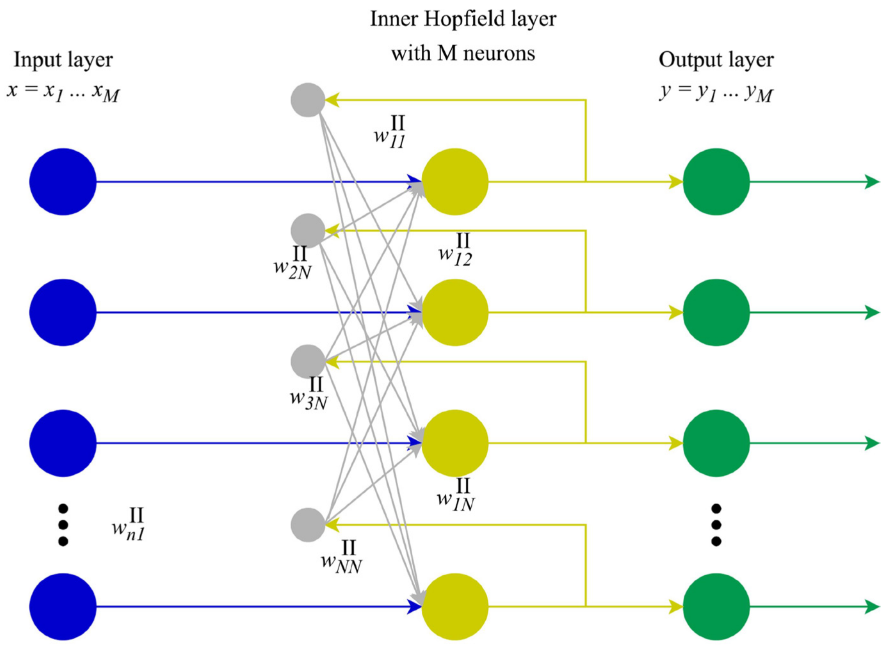

Fig. 4. Hard threshold activation function [5]

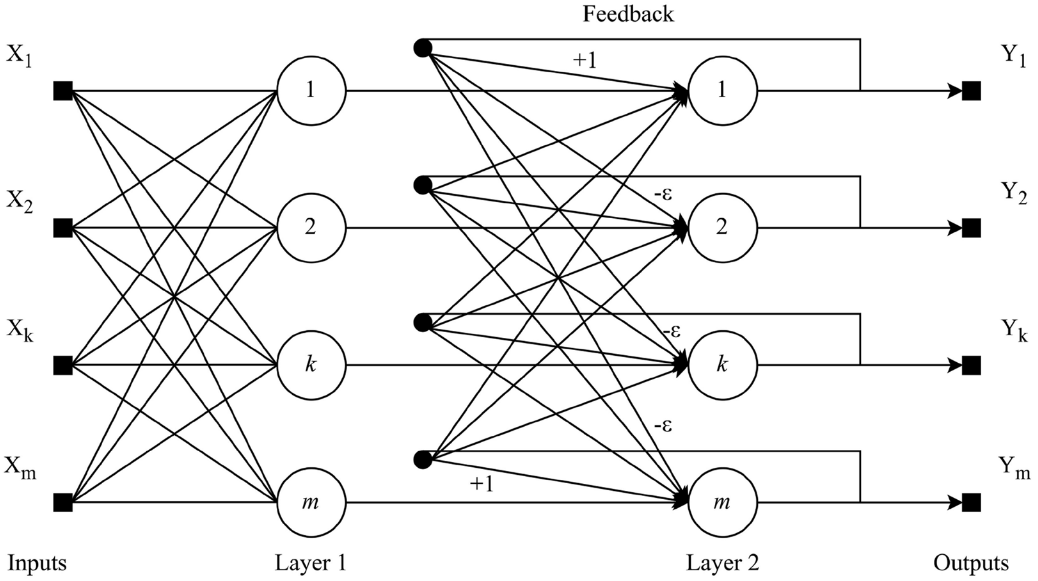

Fig. 5. Structural diagram of the Hamming network [5]

The proposed stereo matching method uses the characteristic points highlighted by the Moravec interest operator as the basis for matching stereo images. They used horizontal displacement, vertical displacement and diagonal displacement to display both left and right images. The advantage of using an artificial neural network is that global matching is achieved automatically, since all neurons (processors) are interconnected in a feedback loop, and the output of one of them affects the input of all others. The analysis of the work [5] revealed several unresolved issues: limited performance of the Hopfield network due to the rigid threshold activation function, significant computational resources for training the Hamming network due to the complexity of calculating the Hamming distance, low efficiency of the stereo matching method in the presence of noise due to the sensitivity of the Moravec operator, and the risk of "getting stuck" in local minima during the automation of global matching. The reasons are the shortcomings of the network architecture, high computational complexity and insufficient adaptability of algorithms to complex images, which requires further improvements.

In [6], the results of assessing the quality of images in the case of deterioration of their contrast (CDI – Contrast Distorted Images) are presented. A systematic and up-to-date review of the perceptual visual quality metrics (PVQMs – Perceptual Visual Quality Metrics) for assessing image quality according to human perception was conducted. The study used convolutional neural networks for automatic assessment of image quality in the absence of a basic reference image for comparison (NR-IQA – No-Reference Image Quality Assessment). The image databases TID2013, CID2013, CSIQ were used. The analysis of [6] revealed the following unresolved issues: limited effectiveness of convolutional neural networks for assessing image quality in the case of complex contrast distortions due to the absence of a basic reference image (NR-IQA); the need for large data sets for training networks while maintaining high accuracy of assessment; dependence on the specifics of the databases (TID2013, CID2013, CSIQ), which may limit the generalizability of the results to other types of images. The main reasons are the complexity of modeling human perception of image quality, insufficient resistance of neural networks to new types of distortions and high computational costs for training on large amounts of data.

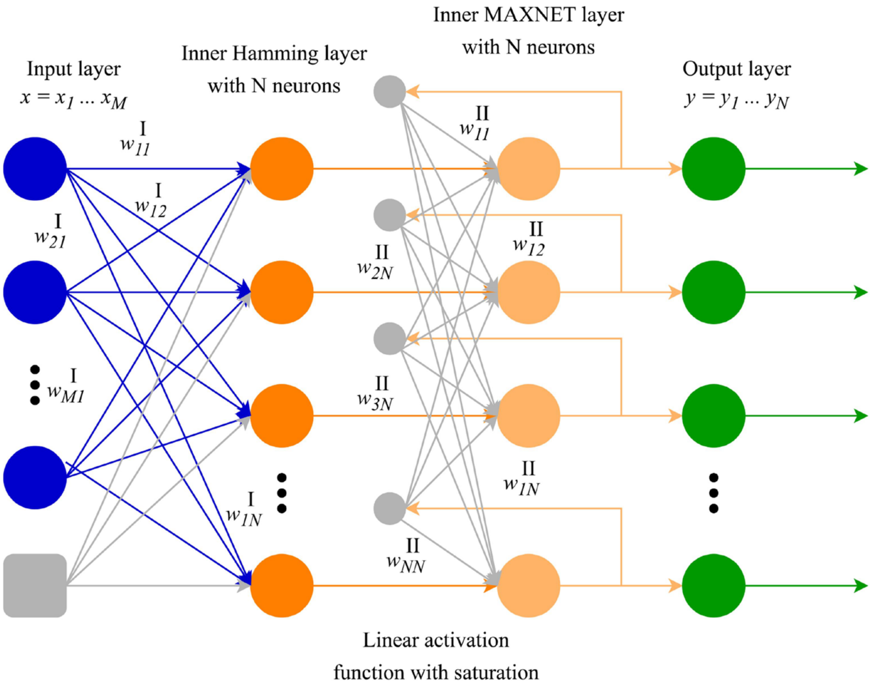

Fig. 6. Three-layer structure of the Hamming network [5]

In [7], the results of research on image quality assessment, which are currently used in solving the problem of automatic image processing, are presented. An artificial neural network was built to assess image quality using the Image Quality Index.

In [8], the results of using specific structures of artificial neural networks and recognizing graphic images using them are presented. The application of transformers (encoders and decoders) for assessing image quality based on reference images using convolutional neural networks to highlight image features is described. The study considers the features of the application of image quality assessment within the framework of the task of machine learning with reinforcement for ultrasound and X-ray medical images. The analysis of the work [8] revealed the following unresolved issues: limited accuracy of image quality assessment when processing ultrasound and X-ray medical data due to the heterogeneity of structures and high noise levels; the complexity of adapting transformers and convolutional neural networks to medical images with different levels of detail; significant computational resources required for training models with reinforcement on large medical datasets. The reasons are the difficulty of extracting relevant features in low-quality medical images, the lack of universal metrics for assessing the quality of graphic images within different tasks, and the need to optimize neural network architectures to reduce computational complexity.

In [9], the results of using trained convolutional structures to build an artificial neural network for image quality assessment and the study of multilayer artificial neural networks for automatic feature extraction when solving the problem of image recognition are presented. According to the results of the study, a convolutional neural network was built for image quality assessment using a pre-trained convolutional basis of the network trained on the basis of ImageNet images and using the LIVE database, which contains 29 basic RGB images distorted by five types of distortions with several of their levels. Various types of convolutional neural networks trained on large image databases were considered for further use as pre-trained convolutional structures. The analysis of [9] revealed the following unresolved issues: limited generalizability of the constructed convolutional neural networks for image quality assessment due to the use of a limited LIVE database of 29 basic images and distortions of only five types; insufficient adaptability of models to new or specific types of distortions not represented in the training data; dependence on the pre-trained basis of the ImageNet network, which is focused on natural images, which reduces efficiency in narrow-profile tasks. The reasons are the limited and unrepresentative nature of the training data, the lack of flexibility of neural network architectures to handle various distortions, and the need to refine models for specialized applications, such as medical or technical images.

There are frequent attempts to use recurrent neural networks (RNNs) for working with images. A recurrent neural network works on the principle of storing the output of a layer and feeding it back to the input to help predict the result of the layer. For example, in [10] the use of the Deep Voice network (developed in the Baidu laboratory, California) was considered. However, when training deep RNNs, problems arose with vanishing gradients, which made it difficult to train the network to capture long-term dependencies between pixels in the image. There is also a limited ability to capture long-term dependencies, especially in the basic LSTM and GRU, which is critical for the task of compressing large images.

Although neural networks have made significant progress in image compression, a critical analysis of the reviewed sources confirms the presence of certain problems and limitations that require further research.

First, the work [5] indicates the limited effectiveness of Hopfield neural networks due to the rigid threshold activation function, which limits the performance of compressing images with high detail. Also, significant resources are required to train the Hamming network, which makes it difficult to use in real time.

Second, it was shown in the work [6] that convolutional neural networks for image quality assessment have limited robustness to complex contrast distortions, especially in the absence of reference images (NR-IQA). The main problem is the need for large amounts of data for training the models and the difficulty of accurately modeling human quality perception.

Third, research [8] revealed the difficulties of processing medical images, in particular ultrasound and X-ray images, due to the high level of noise and heterogeneity of structures. The insufficient adaptability of transformers and convolutional networks limits their effectiveness, especially for specific tasks.

Fourth, work [9] highlighted the problem of limited generalizability of models trained on the LIVE basis, which does not cover all types of distortions. The dependence on the pretrained basis on ImageNet leads to a decrease in efficiency on highly specialized tasks.

Thus, the main challenges include: significant computational resources for training models, limited ability to generalize to different datasets, sensitivity to noise and the appearance of artifacts at high compression ratios. Further research is needed to develop more robust and universal neural network models that can effectively work with different types of images and minimize quality losses.

## 3. The aim and objectives of the study

The aim of the study is to achieve better performance in image compression using artificial neural networks. This will make it possible to optimize information on the issue of image compression using artificial neural networks and will allow finding practical use of the developed approach.

To achieve this aim, the following objectives need to be solved:

– justify the choice of the network that will be used for image compression;

– describe the compression algorithm that will be used when solving the problem;

– simulate an artificial neural network in the MATLAB environment to perform image compression.

## 4. Materials and methods

The object of the study is artificial neural networks in image processing.

The subject of the study is a set of necessary conditions that provide the best approach to optimizing image compression using artificial neural networks.

The hypothesis of the study assumed that a continuous network for recognizing dynamic modes, consisting of two parallel modules, each of which is a modified ART-2 network, will allow image compression without loss of quality and eliminates the drawback associated with the need to pre-select the number of code words that determine the size of the Kohonen map.

The following research methods were used in the work: theoretical analysis of scientific literature by research direction; statistical methods of analyzing literary data. The study was based on methods of comparative analysis and classification.

The current use of discrete ART networks for recognizing the operating modes of the unit for only five measured variables required the use of more than fifteen thousand binary neurons. Replacing the discrete ART artificial neural network with a continuous ART network can significantly reduce the number of neurons in recognition systems and expand the scope of application of continuous ART artificial neural networks. Continuous ART artificial neural networks cannot be directly used for recognizing the modes of functioning of dynamic objects. This is due to the presence of the continuous ART-2 neural network in the basic architecture due to the following features: in the basic architecture of the ART-2 network there is no possibility of simultaneous comparison of the input image with two or more images stored in the network memory; the definition of the image similarity parameter has been changed and these changes have led to the transformation of the architecture and algorithms of the ART artificial neural network.

A new network of adaptive resonance theory ART-2 is proposed – a continuous network for recognizing dynamic modes. The network consists of two modules operating in parallel, each of which is a modified ART-2 network, unlike the Kohonen and Grossberg star artificial neural network algorithms. The Kohonen network, as one of the many variations of neural networks today, is often used for lossy image compression. It allows to allocate similar data fragments into classes. The class number usually takes up much less memory space than the class kernel. If to transfer to the recipient all the class kernels and class numbers encoding each data fragment, the data can be restored. The proposed algorithms solve the problem of atypical behavior when increasing the layers in the autoencoder structure and increasing image quality (the use of additional differential image coding in the algorithms, i.e., initial coding errors). To create a correspondence between the input and output spaces consisting of code elements – code words, or neurons, an artificial neural network ART was used. The algorithms used should not have the drawback associated with the need to pre-select the number of code words that determine the size of the Kohonen map.

Comparison of the proposed algorithms with existing ones in terms of the ratio of the maximum possible signal level to the level of distorting noise (PSNR) and the mean square error (MSE) was performed by modeling in the Matlab environment.

The following hardware and software were used in the study.

Windows 10/11 operating system. PowerShell/Git-Bash. 100 MBit Internet connection. Availability of Google Chrome browser not lower than version 87.0 or Mozilla Firefox not lower than version 83.0.

Quad-core processor with a clock frequency of 1.8 GHz. GPU – 1 Nvidia K80 with 12 GB of memory. Hard Drive – 375 GB. RAM must have a capacity of at least 8 GB. Windows 10 x64 operating system.

General parameters of the artificial neural network used in the study are given in Table 1.

Table 1

General parameters of the artificial neural network

| Name | Value |
|---|---|
| Worker | 16 |
| Quantizer levels | [3, 5, 7] |
| Sample size | 32 |
| Encoder learning rate | 0.0001 |
| Decoder learning rate | 0.0001 |
| Entropy learning rate | 0.0001 |
| Quantizer learning rate | 0.00005 |
| Milestone | [200, 300, 350] |
| Number of epochs | 400 |
| Gamma | 0.2 |

## 5. Results of the analysis of image compression methods using artificial neural networks

### 5. 1. Choosing the type of neural network for image compression

In addition to these three types of artificial neural networks, an ART network (adaptive resonance theory) can be used to solve the image compression problem.

ART artificial neural networks classify input images by assigning them to one of the known classes if they have sufficient similarity to the prototype of this class. If the input image matches the prototype with a given accuracy, the prototype is updated to better match the new data. If none of the prototypes stored in the neural connections match the input image, the network creates a new class based on this image. This is possible due to the presence of reserve neurons that are activated only when necessary. If there are no free neurons and the input image does not fall into any of the existing classes, the network does not respond. Thus, ART artificial neural networks are able to store new information without disturbing the already existing data and without causing retraining. The algorithm of the ART artificial neural network operation is presented in Fig. 7, 8 [11].

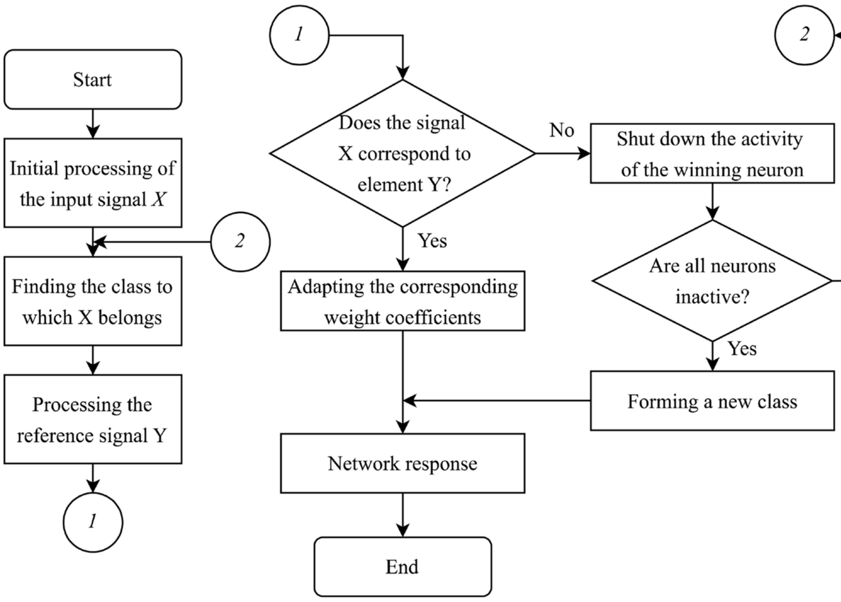
Start
Initial processing of the input signal $X$
2
Finding the class to which X belongs
Processing the reference signal Y
1
1
Does the signal X correspond to element Y?
No
Yes
Adapting the corresponding weight coefficients
Shut down the activity of the winning neuron
Are all neurons inactive?
No
Yes
2
Forming a new class
Network response
End

Fig. 7. The algorithm of the ART network operation [11]

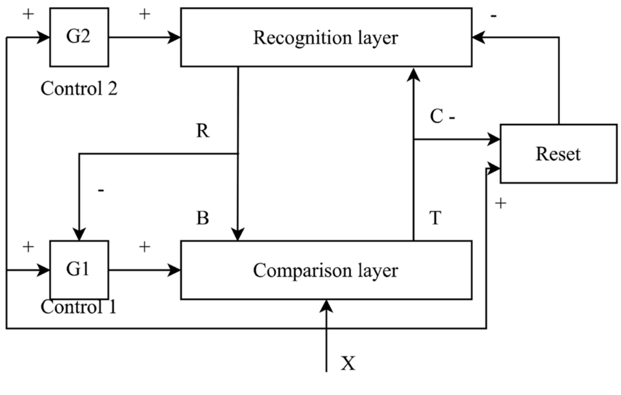

Fig. 8. Schematic representation of the ART network [11]

Both discrete and continuous adaptive resonance theory networks (ART-1 and ART-2) can work effectively in recognition systems, especially under conditions of significant uncertainty, when dozens or hundreds of different images need to be recognized.

However, the application of such networks in real control systems, where it is necessary to recognize dynamic modes of objects based on numerous variables, is complicated by the large variety of measurement data for the same mode of the object (thousands or even tens of thousands of variations of one mode). This creates a serious problem with the selection and preservation of relevant information, since the direct use of ART networks in such situations becomes problematic due to the need for an excessively large number of neurons.

Thus, the ART artificial neural network became the first artificial neural structure that implemented a binary information model of adaptive resonance. In contrast to the direct focus on image compression observed in the Kohonen network and autoencoder, ART networks specialize in unsupervised learning for image recognition and classification tasks.

It is also worth noting that the ART network cannot operate under interference conditions, while the Kohonen network can, since the number of sessions is fixed, the weights change slowly, and weight adjustments are completed after training. A schematic representation of the architecture of the ART 2 network is shown in Fig. 9.

The initial processing in any ART network is performed in a module that represents a trained competitive network. The inputs t of neurons of layer F1 store the input image $I=(t_1, t_2,..., t_n)$. Each neuron of the output layer F2 receives the upstream network activity $t_j$, which is formed from all outputs S=I of layer F1.

The elements of the vector T=$(t_1, t_2,...., t_n)$, which are calcu lated as $t_j=\sum_{i=1}^{m} wi_{j.}\, i_i$, can be considered as the result of comparisons of the input image I with the prototypes $W_1=(w_{11},..., w_{1m}),..., W_n=(w_{n1},..., w_{nm})$ – the weights of synaptic connections between layers F1 and F2. The output of only one neuron J of layer F2, which received the largest ascending network activity $t_j$, is set to one, while the outputs of the other neurons remain zero.

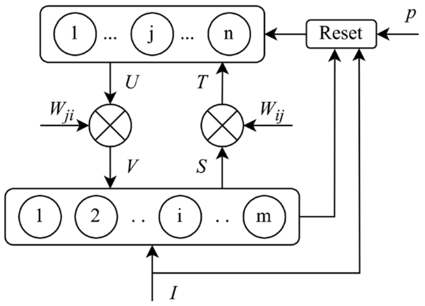

Fig. 9. Schematic representation of the ART-2 network architecture [11]

After the winning neuron J of layer F2 is determined, the corresponding prototype $Wj=(w_{1j},..., w_{mj})$ is adapted to the input image I according to the formula:

$$w_j^{(new)} = \eta \cdot I + (1-\eta) \cdot w_j^{(old)}, \tag{1}$$

where η ϵ [0, 1] – the learning index.

The synaptic downward weights $W_{ji}$ of the feedback connections, except for the possible scaling factor, are identical to the upward weights $W_{ji}$, and the downward network activity V is determined by analogy with the upward one:

$$v_i = \sum\nolimits_{j=1}^{n} u_j \cdot w_{ji}. \tag{2}$$

Since all outputs of layer F2, except one – $u_j$, are zero, the input layer Fl receives the prototype $W_j$, which represents the current class *J*, which won the competition. Next, the most complex part of data processing in ART networks is performed – comparison of the prototype $W_j$ with the input image *I*. The sensitivity *p* set by the researcher determines the required minimum similarity between the input image and the prototype of the corresponding class. If the degree of coincidence is less than ρ, the current winning neuron of layer F2 is eliminated from the competition by a reset signal. The reset signal sets the active neuron *J* of layer F2 to zero and thus enables another neuron to win the competition, thus at the output of layer F2 let's obtain the ascending network activity $t_j$ of the not yet reset output neurons. As soon as a prototype is found, the degree of similarity with the input image *I* is at least the same as the similarity parameter ρ, the reset signal will not occur and the network will reach resonance. The position of the last winning neuron of the F2 layer indicates the class of the input image I, after which the corresponding prototype is adapted [12].

### 5. 2. Image compression algorithm

The image compression scheme includes the following stages: discrete cosine transform (DCT), vector quantization, differential coding and entropy coding.

DCT decomposes image areas into amplitudes of some frequencies, taking into account that many coefficients in the frequency matrix are either close to each other or equal to zero. Vector quantization determines the degree of compression and information loss by increasing the number of zero elements in the vector matrix. A limited number of codewords are selected for the most accurate representation of the distribution of the output image vectors, and each output vector is replaced by the closest codeword, which allows transmitting fewer bits for encoding information [13].

After vector quantization, differential coding "compresses" the codes by taking into account that most parts of the image have smooth transitions. Entropy coding uses variable-length codes, where the length of the symbol code depends on the probability of this symbol appearing in the message. Run-length coding is a simple form of data compression, where sequences containing the same values are replaced by a single value and the number of times it occurs. This is effective for data with many such series, for example, simple graphic images where certain elements are repeated [14].

The work of any artificial neural network, including this network, begins with the training phase. During training, the artificial neural network optimizes its internal parameters to reflect the input data as accurately as possible. After this process is completed, the network becomes capable of efficient operation and processing of new data [15].

Depending on the nature of the input data and the methods of processing them, there are different models of ART networks, one of which is the ART 2 network, which is used to solve certain problems (Fig. 10).

Assuming that the transitions in the image are smooth, the direction in which the difference between the codes of two already encoded blocks is minimal will be the same as the direction in which the difference is minimal for the new encoded block.

If the image is characterized by smooth transitions (which is typical for most images, except for areas with sharp changes, where the differential scheme does not provide advantages), the choice of four possible directions provides a smaller difference compared to the standard scheme. In this case, the direction does not require additional coding, since the codes of blocks *a*–*h* have already been transmitted by the time block *i* is encoded, and the direction with the smallest difference between the codes can be determined based on the already encoded blocks, which eliminates the need to transmit additional data. In other words, it is assumed that the minimum difference between the codes of blocks *i* and *b*, *i* and *d*, *i* and *f*, and i and h will be in the same direction (D1, D2, D3 and D4), respectively, as the difference between the codes of the encoded blocks *b* and *a*, *d* and *c*, *f* and *e*, *h* and *g*.

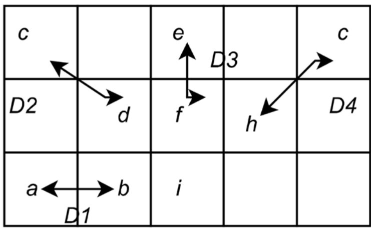

Fig. 10. Schematic representation of the approach to choosing the best direction for coding between block codes

Algorithm 1 is complex and includes the following steps:

1) partitioning the source image into square blocks of a given size (e.g., 4×4 or 8×8 pixels);

2) performing a discrete cosine transform (DCT) for each block;

3) representing each block as a vector in 16- or 64-dimensional space;

4) low-pass filtering to eliminate high-frequency components;

5) training the ART artificial neural network;

6) obtaining the neuron indices corresponding to each input vector;

7) creating a correspondence table between the index and the average cluster vector;

8) compressing the sequence of indices using series length coding and the Huffman algorithm.

The architecture of the proposed artificial neural network is shown in Fig. 11.

This algorithm takes into account the features of images with smooth transitions, which allows to reduce the code size and uses a pre-trained artificial neural network to optimize the compression process.

As follows from the description of the algorithm, in it, unlike JPEG, the ART network (steps 5–7) is used, the results of which are used for vector quantization. The corresponding operations are used for decoding, but in the reverse order.

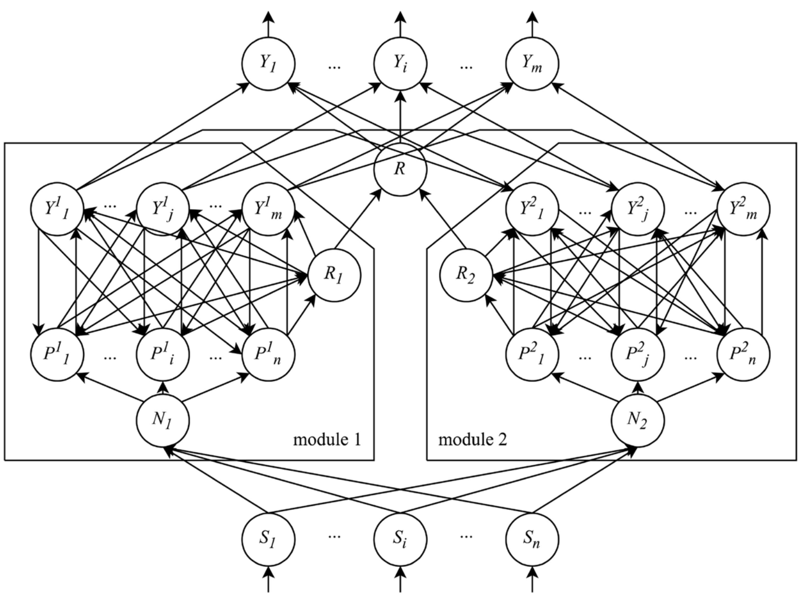

Fig. 11. Artificial neural network architecture

Algorithm 2. Includes the same steps as Algorithm 1, except for steps 2 and 4. Thus, in this algorithm, the network performs its main function - performs vector quantization without calculating DCT and low-pass filtering.

Algorithm 3. This algorithm is based on Algorithm 2, but to improve image quality, another compression cycle is added, in which the image error is encoded, i.e. from the difference between the original image and the image obtained after the first encoding cycle. During decoding, the second image is summed with the first, thus correcting the coding error [16].

### 5. 3. Artificial neural network modeling

When modeling the compression process in the Matlab environment, the PSNR (Peak Signal-to-Noise Ratio) indicators were used to assess the compression quality (since most signals have a very wide dynamic range, PSNR is usually represented on a logarithmic scale) and MSE (Mean Squared Error) – the mean square error.

MSE (mean square error) is one of the most common indicators for assessing image quality. This is a full-scale metric, and the closer its value is to zero, the higher the image quality. MSE is the second moment of error, taking into account both the variance of the estimate and its bias. In the case of an unbiased estimate, MSE reflects the variance of the estimate. The units of measurement of MSE correspond to the square of the quantity being estimated, similarly to the variance. MSE is often confused with root mean square error (RMSE) or root mean square deviation (RMSD) and is sometimes referred to as the standard deviation of the variance. This metric is also known as the mean square deviation (MSD) of the estimator. Estimation is the process of measuring the invisible volume of an image. MSE or MSD measures the mean squared error, where the error is the difference between the estimate and the expected result. It is a risk function that estimates the average squared loss or loss from an error.

PSNR (peak signal-to-noise ratio) is used to determine the ratio between the maximum possible signal power and the noise power that distorts the signal quality. This ratio is calculated in decibels and compares two images. Because signals can have a wide dynamic range, PSNR is usually expressed on a logarithmic scale. Dynamic range can vary from minimum to maximum values, affecting image quality. PSNR is the most popular metric for evaluating the quality of restoration in lossy compression, where the signal represents the original data and the noise is the error that occurs due to compression or distortion. PSNR allows to evaluate how well the restoration matches human perception, especially when working with image compression codecs. In image and video compression, the PSNR value typically ranges between 30 and 50 dB for 8-bit data and between 60 and 80 dB for 16-bit data. In the case of wireless transmission, the acceptable level of quality loss is around 20–25 dB.

For two monochrome images (where one is a representation of the other), the MSE is calculated as follows [17]. The formula used to estimate the difference between two images or data sets by calculating the mean square of their pixel values is:

$$MSE=\frac{1}{mn}\sum_{i=0}^{m-1}\cdot\sum_{j=0}^{n-1}\left\|I(i,j)-K(i,j)\right\|^{2}, \qquad (3)$$

where $I$ and $K$ – monochrome images, i – the input vector, $j$ – the output signal, ($m$×$n$) – the image dimension, $m$ – the number of rows, and $n$ – the number of columns.

For color images with three RGB components, the MSE is defined as the sum of all squared differences divided by the image size in formula (3).

The PSNR indicator is defined as:

$$PSNR=10\cdot\log_{10}\left(\frac{MAX^{2}}{MSE}\right)=20\cdot\log_{10}\left(\frac{MAX_{i}}{\sqrt{MSE}}\right), \qquad (4)$$

where PSNR (peak signal-to-noise ratio) – the ratio of the maximum possible signal level to the level of noise distorting it, MSE (mean squared error) – the mean square error, $MAX_i$ – the maximum value of a pixel in the image.

Compression level ($C_R$) is the simplest criterion for evaluating the efficiency of the algorithm, which is suitable for any type of data:

$$C_{R}=\frac{B}{A}, \qquad (5)$$

where $B$ – the size of the data before compression; $A$ – the size of the data after applying the compression algorithm.

If the pixels are represented by 8-bit values, $MAX_i$=255. In the general case, when used for representation in Bit, the maximum possible value for $MAX_i$ is $2^{B}$–1, where B – number of bits used to encode the pixel values. Table 2 presents the results of modeling algorithms 1–3 and the JPEG algorithm.

Typically, for compression algorithms, PSNR is in the range of 30–40 dB. Blocks of 8×8 points were used to simulate the compression process; quantization was performed using the ART 2A sub E network. The proposed compression algorithms showed similar results for different images, a common image is used to illustrate the results below.

Table 2

Results of modeling algorithms 1–3 and the JPEG algorithm

| Algorithm | Q/ρ | CR | PSNR | MSE | Algorithm | Q/ρ | CR | PSNR | MSE |
|---|---|---|---|---|---|---|---|---|---|
| JPEG | 0 | 57.115 | 24.120 | 252.389 | Algorithm 2 | 0.700 | 105.960 | 22.466 | 368.512 |
| | 10 | 31.512 | 29.577 | 71.851 | | 0.740 | 73.760 | 22.908 | 332.855 |
| | 20 | 21.396 | 31.796 | 42.490 | | 0.780 | 53.477 | 23.292 | 304.720 |
| | 30 | 15.750 | 33.042 | 32.274 | | 0.820 | 34.231 | 23.964 | 261.018 |
| | 40 | 12.998 | 33.814 | 27.023 | | 0.840 | 27.031 | 24.459 | 232.901 |
| | 50 | 11.114 | 34.444 | 23.371 | | 0.860 | 20.008 | 25.040 | 203.745 |
| | 60 | 9.599 | 35.059 | 20.284 | | 0.880 | 14.181 | 25.895 | 167.326 |
| | 70 | 7.853 | 35.889 | 16.755 | | 0.900 | 9.053 | 27.122 | 126.135 |
| | 80 | 6.120 | 37.046 | 12.837 | | 0.920 | 5.996 | 28.754 | 86.642 |
| | 90 | 3.811 | 39.440 | 7.398 | | 0.960 | 2.658 | 34.304 | 24.134 |
| | 100 | 1.481 | 58.440 | 0.093 | | 0.700 | 105.960 | 22.466 | 368.512 |
| Algorithm 1 | 0.961 | 101.902 | 26.395 | 149.127 | Algorithm 3 | 0.600 | 37.364 | 23.690 | 278.008 |
| | 0.959 | 82.296 | 27.356 | 119.519 | | 0.640 | 27.817 | 25.943 | 165.477 |
| | 0.968 | 49.770 | 28.821 | 85.298 | | 0.680 | 26.533 | 26.590 | 142.595 |
| | 0.973 | 42.242 | 29.467 | 73.513 | | 0.720 | 21.292 | 26.804 | 135.722 |
| | 0.979 | 33.566 | 30.184 | 62.331 | | 0.740 | 19.534 | 27.065 | 127.811 |
| | 0.981 | 21.199 | 31.590 | 45.090 | | 0.760 | 17.287 | 28.257 | 97.145 |
| | 0.983 | 17.809 | 32.113 | 39.971 | | 0.780 | 14.911 | 28.784 | 86.028 |
| | 0.985 | 14.288 | 32.691 | 34.990 | | 0.800 | 12.368 | 29.408 | 74.525 |
| | 0.987 | 12.596 | 32.931 | 33.110 | | 0.820 | 9.858 | 30.464 | 58.430 |
| | 0.989 | 7.849 | 34.330 | 23.994 | | 0.860 | 4.399 | 33.893 | 26.530 |

*Note: Q/ρ – parameter of the compression algorithm, CR (Compression Ratio) – compression ratio, PSNR (peak signal-to-noise ratio) – ratio of the maximum possible signal level to the level of the noise distorting it, MSE (mean squared error) – the mean square error.*

Fig. 12, 13 show generalized graphs of PSNR and MSE dependence on the compression ratio (CR) for all algorithms.

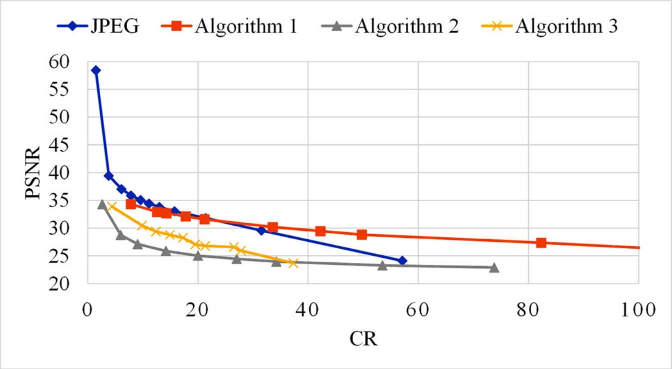

Fig. 12. PSNR curves depending on the degree of compression CR

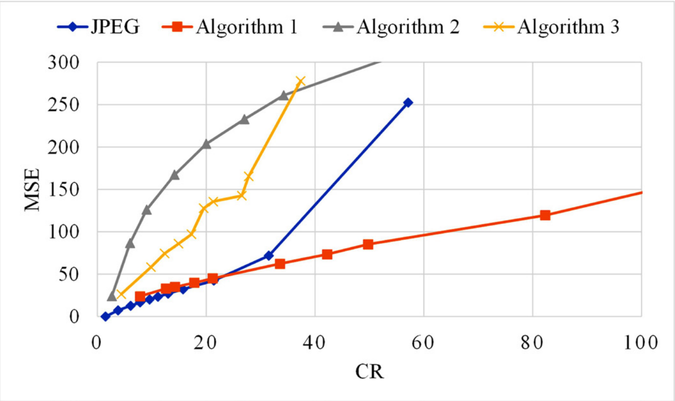

Fig. 13. MSE curves depending on the degree of compression CR

Image compression was performed using the ART 2A-E network with a similarity parameter of ρ=0.980 and a learning index of η=0.25. The degree of compression in this case was 33.6, and PNSR=30.2 dB.

## 6. Discussion of the results of using ART networks for image compression

Artificial neural networks based on the adaptive resonance theory (ART) demonstrate significant potential in image compression tasks due to the ability to classify input images and adapt prototypes in real time. The main advantage of ART networks is the ability to create new classes for atypical images without destroying existing information, which allows to avoid overtraining and maintain stability to new data. This is especially useful in the tasks of automatic compression and recognition of graphic images in systems with a large number of input variations. However, the analysis of the results of using ART networks reveals certain limitations. First, in real systems where there are dynamic changes in the modes of the object or a large number of variations of one image, the efficiency of ART networks decreases due to the need for an excessively large number of neurons. This creates a significant load on computing resources, which is critical for real-time applications. Second, ART networks demonstrate sensitivity to interference and noise in the input data. Unlike Kohonen networks, where the number of neurons is fixed and the weights stabilize after training, ART networks require flexible control of the sensitivity of the parameter ρ to achieve a balance between classification accuracy and performance. Of particular importance is the ability of ART networks to compare prototypes with input images to achieve resonance. The approach based on the parameter ρ allows to control the degree of similarity, but at low values of this parameter the network may discover too many new classes, which reduces its generalization ability. At the same time, high values of ρ may lead to the omission of important features of new images. This indicates the need for adaptive tuning of the network sensitivity depending on the specifics of the input data. Thus, ART networks are effective for unsupervised learning and classification tasks, which is important for compressing images with high variability. However, their practical application in real conditions requires optimization of the architecture to reduce computational complexity and increase resistance to noise and interference. Further research may be aimed at integrating ART networks with other learning methods or hybrid neural structures to improve their performance in automatic image compression tasks.

The proposed image compression algorithm combines classical methods, such as discrete cosine transform (DCT), vector quantization, differential and entropy coding, using an ART artificial neural network. DCT effectively decomposes the image into frequencies, allowing to isolate low-energy components, and vector quantization reduces the data size by replacing vectors with their prototypes. The advantage of the algorithm is the use of an ART network at the quantization stage, which provides adaptive grouping of similar image blocks and optimization of the compression process. As a result, this approach achieves a reduction in data volume without significant loss of quality, especially for images with smooth transitions. Unlike the traditional JPEG format, where DCT and entropy coding are key components, the algorithm with ART networks can work more flexibly by adapting prototypes to new input data, which allows to avoid overtraining and increases the robustness to heterogeneous images. However, the described algorithms have certain limitations. First, the complexity of training the ART network and the computational cost with a large number of blocks can become a bottleneck in real compression systems. Second, Algorithm 2 simplifies the calculations by eliminating DCT and low-pass filtering, but this may reduce the efficiency for images with sharp boundaries or textures. Algorithm 3 adds an error correction loop, which improves the quality of image restoration, but requires additional computational resources at the decoding stage. Thus, the proposed algorithms are promising for adaptive compression, but their further optimization remains an important task to reduce the computational complexity and ensure universality for different types of images.

The results of the simulation of the image compression process (Table 2) demonstrate the effectiveness of the ART 2A-E network in combination with classical data processing methods. The use of PSNR and MSE metrics to assess the compression quality allowed to objectively compare the proposed algorithms. In particular, the proposed algorithm 1, which combines the discrete cosine transform (DCT), low-pass filtering and vector quantization with the ART network, showed stable results: at a compression ratio of CR=33.6, the PSNR value=30.2 dB, which is acceptable for image compression with minimal quality loss. Algorithm 2, which simplifies the calculations by eliminating DCT and low-pass filtering, demonstrates lower PSNR values at similar compression ratios due to the absence of initial filtering of high-frequency components. In contrast, Algorithm 3, due to an additional error correction cycle, achieves higher restoration accuracy, which is confirmed by better PSNR indicators and lower MSE values compared to Algorithm 2. The general trends presented in the graphs of the dependence of PSNR and MSE on the compression ratio (CR) (Fig. 12, 13) indicate a decrease in image quality with increasing CR, which is typical for all compression algorithms. However, unlike JPEG, where the PSNR value decreases sharply at high compression ratios, the use of the ART network provides a smoother transition and adaptation of the model to new data due to its ability to classify and maintain stability. At the same time, there is a dependence of the results on the settings of the similarity parameter ρ and the learning index η, which determine the balance between compression and image restoration quality. Thus, the results of the study confirm that the use of the ART network in the compression process is a promising direction, however, optimization of computational costs and parameter tuning remain key tasks for improving the efficiency of the algorithm. Thus, for two monochrome and color images with three RGB MSE components, a continuous network for dynamic mode recognition was applied, consisting of two parallel modules, each of which is a modified ART-2 network.

Summarizing the results of the study, it can be stated that the proposed image compression algorithms using the ART network in combination with classical methods of discrete cosine transform, vector quantization and error correction provided an effective solution to the tasks. The developed approaches made it possible to achieve an optimal balance between the compression ratio (CR) and the quality of the restored image, which is confirmed by the PSNR and MSE indicators. In particular, the use of the adaptive capabilities of the ART network made it possible to eliminate the problem of overtraining, ensure high generalization ability of models and effectively classify input data with minimal information loss. As a result, the algorithms demonstrated stable results on test images, proving their effectiveness in the tasks of automatic compression and processing of graphic images, which makes them promising for further implementation in real data compression systems.

From Fig. 12 it is seen that PSNR is directly proportional to ρ, that is, this parameter directly affects the quality of the compressed image. As the simulation showed, at η=0.25 the efficiency of training of the artificial neural network reaches its maximum, which corresponds to the optimal ratio of the components of the training vector and the class prototype vector at the adaptation stage.

This approach allows for image compression without loss of quality and eliminates the disadvantage associated with the need to pre-select the number of code words that determine the size of the Kohonen map.

During testing of the information technology of image compression, the values of the following general training parameters were set: show=25, epochs=300, time=inf, goal=1e-5, where show – the step of outputting intermediate information; epochs – the maximum number of training cycles; time – the limit training time; goal – the limit value of the training criterion.

For the practical implementation of the formulated hypothesis, PSNR and MSE indicators were used in the Matlab environment, which indicate that the studied compression algorithms show similar results for different images.

The obtained experimental results indicate an effective degree of image compression, because unlike the algorithms used for image compression (Kohonen artificial neural network and Grossberg star algorithms), the compression process in the work was lossy and images with rather large deformations were obtained, our algorithms did not have such shortcomings.

A limitation of this study is the dependence of the quality of the results on the training parameters of the ART network, in particular the similarity index ρ and the learning rate η, which require careful tuning for each specific set of images. The proposed algorithms demonstrate efficiency for images with smooth transitions, but their performance may decrease when processing images with complex texture or sharp boundaries, where standard methods such as JPEG may be more stable. In addition, the computational complexity of the ART network increases with a large amount of input data, which may limit the application of the algorithms in real-time. Another important condition is the use of a fixed image block size (e.g., 8×8 pixels), which affects the overall compression efficiency at different resolutions.

Further research directions are to study the process of compressing images and videos using deep learning and improve the quality by reducing the data rate and increasing the efficiency of memory and computation in realistic photos. Currently, the biggest obstacle to the widespread implementation of deep learning technology is the significant computational load and often limited memory capacity. To improve performance, larger artificial neural networks with more levels and nodes should be used.

## 7. Conclusions

1. The choice of the network recommended for image compression is justified. To solve the image compression problem, the ART network (adaptive resonance theory) is chosen due to its ability to unsupervised learning, adaptive clustering of input data, and the unique property of preserving existing information when adding new classes. The ART network effectively solves the problem of overtraining, since it can create new classes only when necessary, which is especially important for processing images with significant variability. Unlike other networks, such as autoencoders or Kohonen networks, ART allows prototypes to be automatically adapted to new data, which makes it flexible for use in image compression tasks with different characteristics. In addition, its architecture provides high resistance to changes in input data and reduces the risk of losing relevant information, which is critical for achieving the optimal ratio between the degree of compression and the quality of the restored image. The used continuous network, consisting of two ART-2 modules, is chosen due to its ability to effectively process continuous data and adapt to changes in input images due to the stable resonance mechanism. Its advantages are high noise immunity, flexible clustering of complex images, and the ability to work with large amounts of data without destroying already trained prototypes. However, for effective use, it is necessary to ensure careful tuning of parameters such as ρ (similarity threshold) and learning rate η, as well as stable computing resources to support parallel operation of modules in real time.

2. Three algorithms for compressing digital images using artificial neural networks, in particular the ART-2 network, are proposed, which provide adaptive clustering of image blocks for effective reduction of data volume with minimal loss of quality. and their testing is carried out. It is found that the compression algorithms used show similar results for different images, because in their work they take into account the features of images with smooth transitions, which allows to reduce the code size and uses a pre-trained artificial neural network to optimize the compression process. The use of additional coding of the difference image in the algorithms, i.e. the initial coding errors, can improve the quality of the resulting image.

3. An artificial neural network is simulated in the MATLAB environment (with a similarity parameter of $\rho$=0.980 and a learning index of $\eta$=0.25) to perform image compression using the PSNR, PSNR and MSE indices. The compression ratio in this case is 33.6, and PNSR=30.2 dB. The application of the developed algorithms is most effective when compressing images with repeating areas.

## Conflict of interest

The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, that could affect the research and its results presented in this article.

## Funding

The study was conducted without financial support.

## Data availability

The manuscript has no linked data.

## Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the presented work.

## References

1. Ali, A. N. M., Ahmad, N., Noor, N. M., Aris, S. A. M. (2022). Image Compression Using AMBTC with Artificial Neural Networks. 2022 IEEE Symposium on Future Telecommunication Technologies (SOFTT), 78–82. https://doi.org/10.1109/softt56880.2022.10009930
2. Dashkevich, A. (2016). Study of multilayer neural networks for automatic feature extraction in solving the problem of pattern recognition. Naukovyi visnyk TDATU, 2 (6), 134–139. Available at: https://repository.kpi.kharkov.ua/server/api/core/bitstreams/883b0aec-89a9-48c9-bb07-3c515300dd80/content
3. Lesyk, V. O., Doroshenko, A. Yu. (2023). Image compression module based neural network autoencoders. Problems in Programming, 1, 48–57. https://doi.org/10.15407/pp2023.01.048

4. Bosse, S., Maniry, D., Wiegand, T., Samek, W. (2016). A deep neural network for image quality assessment. 2016 IEEE International Conference on Image Processing (ICIP), 3773–3777. https://doi.org/10.1109/icip.2016.7533065
5. Syzonenko, Yu. I. (2016). Systema stysnennia ta zakhystu zobrazhen za dopomohoiu neitronnoi merezhi. Aktualni zadachi ta dosiahnennia u haluzi kiberbezpeky: Materialy Vseukrainskoi naukovo-praktychnoi konferentsiyi. Kropyvnytskyi, 157–158. Available at: https://core.ac.uk/download/pdf/84825428.pdf
6. Atta, R. E., Kasem, H., Attia, M. (2020). A comparison study for image compression based on compressive sensing. Eleventh International Conference on Graphics and Image Processing (ICGIP 2019), 60. https://doi.org/10.1117/12.2557296
7. Netalkar, R. K., Barman, H., Subba, R., Preetam, K. V., Raju, U. S. N. (2021). Distributed compression and decompression for big image data: LZW and Huffman coding. Journal of Electronic Imaging, 30 (05). https://doi.org/10.1117/1.jei.30.5.053015
8. Hrytsyk, V. (2017). Basic image quality estimates methods are used today to solve the problem of automatic image processing. Shtuchnyi intelekt, 1, 38–44. Available at: http://dspace.nbuv.gov.ua/handle/123456789/132099
9. Myasischev, O. A., Lenkov, Ye. S., Bilik, O. M. (2016). Recognition of graphic images using neural networks. Collection of Scientific Works of the Military Institute of Kyiv National Taras Shevchenko University, 54, 143 149. Available at: https://miljournals.knu.ua/index.php/zbirnuk/article/view/174
10. Jalilian, E., Hofbauer, H., Uhl, A. (2022). Iris Image Compression Using Deep Convolutional Neural Networks. Sensors, 22 (7), 2698. https://doi.org/10.3390/s22072698
11. Yelahina, K., Zhukovska, D., Voropaeva, V. (2021). Use of neural network architecture based on adaptive resonance for speech signal recognition. Naukovyi Visnyk Donetskoho Natsionalnoho Tekhnichnoho Universytetu, 1 (6)-2 (7), 55–67. https://doi.org/10.31474/2415-7902-2021-1(6)-2(7)-55-67
12. Hussain, A. J., Al-Fayadh, A., Radi, N. (2018). Image compression techniques: A survey in lossless and lossy algorithms. Neurocomputing, 300, 44–69. https://doi.org/10.1016/j.neucom.2018.02.094
13. Sadeeq, H. T., Hameed, T. H., Abdi, A. S., Abdulfatah, A. N. (2021). Image Compression Using Neural Networks: A Review. International Journal of Online and Biomedical Engineering (IJOE), 17 (14), 135–153. https://doi.org/10.3991/ijoe.v17i14.26059
14. Ding, K., Ma, K., Wang, S., Simoncelli, E. P. (2021). Comparison of Full-Reference Image Quality Models for Optimization of Image Processing Systems. International Journal of Computer Vision, 129 (4), 1258–1281. https://doi.org/10.1007/s11263-020-01419-7
15. Feng, Y., Zhang, Y., Zhou, Z., Huang, P., Liu, L., Liu, X., Kang, J. (2024). Memristor-based storage system with convolutional autoencoder-based image compression network. Nature Communications, 15 (1). https://doi.org/10.1038/s41467-024-45312-0
16. Zhang, S., Zhao, C., Basu, A. (2024). Principal Component Approximation Network for Image Compression. ACM Transactions on Multimedia Computing, Communications, and Applications, 20 (5), 1–20. https://doi.org/10.1145/3637490
17. Yang, F., Herranz, L., Cheng, Y., Mozerov, M. G. (2021). Slimmable Compressive Autoencoders for Practical Neural Image Compression. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4996–5005. https://doi.org/10.1109/cvpr46437.2021.00496