Evolution of Convolutional Neural Network Architecture in Image Classification Problems

Andrey Arsenov[1], Igor Ruban[1], Kyrylo Smelyakov[1], Anastasiya Chupryna[1]

[1] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
andrii.arsienov@nure.ua, ruban_i@ukr.net,
kirillsmelyakov@gmail.com, anastasiya.chupryna@nure.ua

Abstract. At present, the models and computer vision algorithms are increasingly used in various fields of activity. For example, in systems of sample analysis in medicine and pharmacology, in identification of individuals by a fingerprint, iris or face, in video surveillance security systems and in many other systems and applications. In connection with the growth of computing power and the emergence of big databases of images, it became possible to learn and use deep neural networks for solving the problems of classification and recognition. As to the image classification problem, the Convolutional Neural Networks showed themselves best of all; every year since 2012, they won the prestigious international contest – the ImageNet Large Scale Visual Classification Challenge (ILSVRC), in which such giants as Google and Microsoft participated. Thanks to the revealing of their capabilities, the convolutional neural networks are increasingly used for pattern recognition, image classification, object detection, semantic segmentation, and solving many other problems. The paper examines the evolution of the most efficient models and trends in development of architecture of convolutional neural networks, which are currently used for classification of images that have been included in the list of winners of this international competition, ILSVRC. More precisely, the key features of architecture and its annual variations are revealed on the background of increasing efficiency of practical application of these networks. The data of numerous experiments conducted over the past few years are summarized, classes of applied problems are analyzed, and estimates are given for an effectiveness of use of the considered convolutional neural networks. In fact, these performance estimates are based on evaluation of probability of adequate classification of images. On this basis, a generalized algorithm is formulated, and practical recommendations are proposed taking into account the problem features.

Keywords: Convolutional Neural Network, Image Classification, Neural Network Architecture, Efficiency.

1 Introduction

Computer vision technologies are becoming increasingly popular. They are used in data analysis systems in medicine and pharmacology, in personal identification tasks, by face, by fingerprint, by iris, in security video surveillance systems, for example,

for identifying vehicles by their license plates and in many other systems and applications [1-3]. In connection with the growth of computing power and the emergence of huge image bases, it became possible to train deep neural networks for solving problems in the field of computer vision, such as classification and recognition. Convolutional Neural Networks showed themselves best in the image classification task [4-5], which since 2012 each year won the competition of the international competition ImageNet Large Scale Visual Classification Challenge (ILSVRC), in which such giants took part and Microsoft.

A convolutional neural network is a neural network with a convolutional layer. Usually in the convolutional neural networks there are also a sub-sampling layer (pooling layer) and a fully connected layer. Convolutional neural networks are used for pattern recognition, object detection, image classification, semantic segmentation, and other tasks. In convolutional neural networks, layers of convolution and subsampling consist of several "levels" of neurons, called feature maps, or channels. Each neuron of this layer is connected to a small section of the previous layer, called a receptive field. In the case of an image, a feature map is a two-dimensional array of neurons, or simply a matrix. Other measurements can be used if another kind of data is taken as input, for example, audio data (one-dimensional array) or volume data (three-dimensional array) [6-7].

At the same time, although such networks are used quite successfully, the question of choosing the optimal architecture and setting the parameters of the neural network are remains unresolved. In this regard, the task of the work is to analyze the available experimental data using the most efficient convolutional neural networks used to classify images, in order to develop a general algorithm and practical recommendations on choosing the best architecture and setting the parameters of the neural network, according to the specifics of the problem.

2 The Effectiveness of the Use of the Convolutional Neural Network for Image Classification

This section presents the most efficient and widely used architectures of convolutional neural networks for classifying images that are arranged in chronological order.

2.1 Convolutional Neural Network AlexNet

The first neural network that won the ILSVRC image classification competition was AlexNet, in 2012, reaching a top-5 classification error of 15.31%. For comparison, the method that does not use convolutional neural networks received a classification error of 26.1%. AlexNet collected the latest technology at the time to improve the network. The architecture of this network is shown in Fig. 1.

Training network AlexNet due to the large number of network parameters occurred on two graphics processors (abbreviated GPU – Graphics Processing Unit), which reduced training time in comparison with learning based on the central processor (abbreviated CPU – Central Processing Unit). It also turned out that using the Recti-

fied Linear Unit (ReLU) activation function instead of more traditional functions (sigmoids and hyperbolic tangent) made it possible to reduce the number of learning epochs by six times. This is due to the fact that the function of network activation Rectified Linear Unit allows you to overcome the problem of gradient attenuation inherent in other activation functions. Graphically, the activation function of the Rectified Linear Unit is shown in Fig. 2.

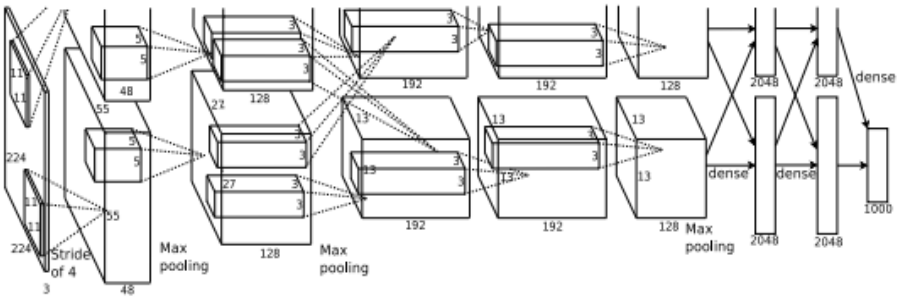

Fig. 1. Architecture of convolutional neural network AlexNet [8].

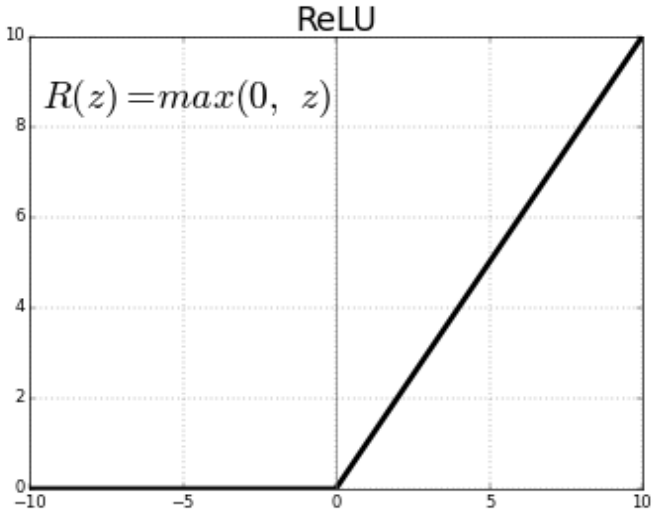

Fig. 2. Network activation function Rectified Linear Unit [9].

Also, a dropout technique (Dropout) was used in AlexNet, which randomly turns off each neuron on a given layer with a probability p at each epoch. Then, after learning the network, at the recognition stage, the weights of the layers to which the dropout was applied should be multiplied by 1/p. Technology Dropout acts as a regularizer, not allowing the network to retrain. To understand the effectiveness of this technique, there are several interpretations. First, this dropout causes neurons not to rely on neighboring neurons, but to learn to recognize more persistent signs. And the

second, later, is that learning a network with a dropout is an approximation of learning a network of ensembles, each of which represents a network without some neurons. As a result, the probability of error is reduced, since the final decision is made not by one network, but by an ensemble, each network of which is trained differently.

2.2 Convolutional Neural Network ZF Net

The convolutional neural network ZF Net is the winner of ILSVRC 2013 with a top-5 classification error of 14.8%. The main achievement of this architecture is the creation of a filter visualization technique - a sweep network (deconvolutional network), consisting of operations, in a sense, reverse operations of the network. As a result, the network sweep displays a hidden layer of the network on the original image.

To study the behavior of the filter on a particular image using a trained neural network, you must first make a network output, then in the layer of the studied filter zero all weights, except the weights of the filter itself, and then apply the resulting activation to the network of the sweep network. The network sweeps consistently used operations Unpooling ReLU and filtering. The Unpooling operation partially restores the input of the corresponding sub-sampling layer by remembering the coordinates that the sub-sampling layer has selected. The ReLU operation is a regular layer that uses the ReLU function. The filtering layer performs the convolution operation with the weights of the corresponding convolution layer, but the weights of each filter are "inverted" vertically and horizontally. Thus, the initial activation of the filter moves in the opposite direction until it is displayed in the original image space. The architecture of the considered network is shown in Fig. 3.

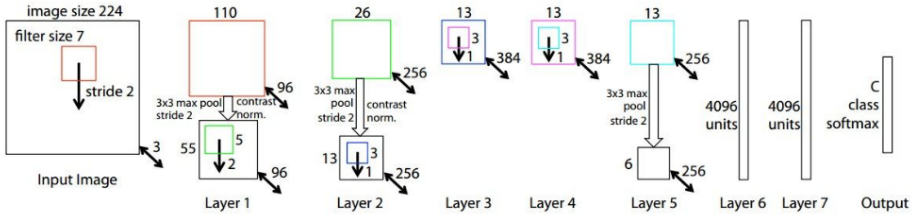

Fig. 3. Network activation function ZF Net [10].

2.3 Convolutional Neural Network VGG Net

VGG Net is a convolutional neural network model that won the 2014 image classification competition. In this network, they refused to use filters larger than 3x3. Since the authors proved that the 7x7 filter layer is equivalent to three layers with 3x3 filters, and in this case 55% less parameters are used. Similarly, a 5x5 filter layer is equivalent to two layers with a 3x3 filter, which saves 22% of network parameters.

Features of the architecture and internal organization of this neural network are shown in Fig. 4.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Number of parameters (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

Fig. 4. Different variations of the convolutional neural network architecture VGG Net [11].

2.4 Convolutional Neural Network Inception

The Inveption-v1 convolution neural network is the winner of the ILSVRC 2014 competition with a top-5 error of 6.7%, also known as GoogleNet. The creators of this network, led by Christian Szegedy, proceeded from the fact that after each layer of the network it is necessary to make a choice whether the next layer will be a convolution with a 3x3, 5x5, 1x1 filter or a subsampling layer. Each of these layers is useful – a 1x1 filter reveals a correlation between channels, while larger filters respond to more global features, and a subsampling layer reduces dimensionality without large loss of information. Instead of choosing which layer should be next, it is proposed to use all layers at once, parallel to each other, and then merge the results into one. To avoid an increase in the number of parameters, a 1x1 convolution is used in front of each convolution layer, which reduces the number of feature maps. Such a block of layers was called an Inception module. The architecture features of this neural network are shown in Fig. 5.

Fig. 5. The architecture of the convolutional neural network GoogleNet [12].

Also, GoogLeNet abandoned the use of a fully connected layer at the end of the network, using the Average Pooling layer instead, which drastically reduced the number of parameters in the network. Thus, GoogLeNet, consisting of more than one hundred basic layers, has almost 12 times fewer parameters than AlexNet (about 7 million parameters against 138 million).

In the next iteration of the Inception module, called Inception-v2, the authors, as was done on the VGG network, decomposed the 5x5 layer into two 3x3 layers. Next, the Batch Normalization technique was used, which allows to multiply the learning speed by means of normalizing the distribution of layer outputs within the network.

In the same article [12], the authors proposed Inception-v3. In this model, they developed the idea of filter decomposition, proposing to decompose the NxN filter with two successive 1xN and Nx1 filters. Also in Inception-v3, RMSProp is used instead of the standard gradient descent and truncated gradients are used to increase the learning stability. An ensemble of four Inception-v3 received a top-5 error of 3.58% at ILSVRC 2015, losing to ResNet.

2.5 Convolutional Neural Network ResNet

The winner of the ILSVRC 2015 competition with a top-5 error of 3.57% was an ensemble of six networks of the ResNet (Residual Network) type, developed at Microsoft Research. The authors of ResNet have noticed that with the addition of new layers, the quality of the model grows to a certain limit (see VGG-19), and then begins to fall. This problem is called the degradation problem, a decrease in accuracy on the validation set.

The authors were able to find such a topology in which the quality of the model grows with the addition of new layers. A neural network can approximate almost any function, for example, some complex function H (x). Then it is true that such a network will easily learn the residual function: F (x) = H (x) – x. Obviously, that our initial objective function will be H (x) = F (x) + x. If we take a certain network, for example, VGG-19, and add twenty layers to it, we would like the deep network to behave at least as good as its shallow analogue.

The problem of degradation implies that a complex nonlinear function F (x), obtained by adding several layers, must learn the same transformation, if the previous layers had reached the quality limit. But this does not happen; it is possible that the optimizer simply cannot cope with adjusting the weights so that a complex non-linear hierarchical model does the same transformation. In order to "help" the network, it was proposed to introduce a missing connection (Shortcut Connections). The architecture features of this neural network is shown in Fig. 6.

2.6 Convolutional Neural Networks Inception-v4 and Inception-ResNet

After the success of applying the ResNet convolutional neural network, the following versions of the Inception network were introduced: Inception-v4 and Inception-ResNet. In both cases, the Inception module was divided into modules A, B, and C for inputs with dimensions of 35x35, 17x17, and 8x8, respectively. Reduction blocks

were also identified, in which the dimensionality decreases and the depth of the data inside the network increases. In Inception-v4, the main innovations are the replacement of Max Pooling with Average Pooling in the Inception modules themselves.

For Inception-ResNet, skipping connections have been added to the Inception modules. Two versions of the network were designed – Inception-ResNet-v1, which requires less computation, and Inception-ResNet-v2.

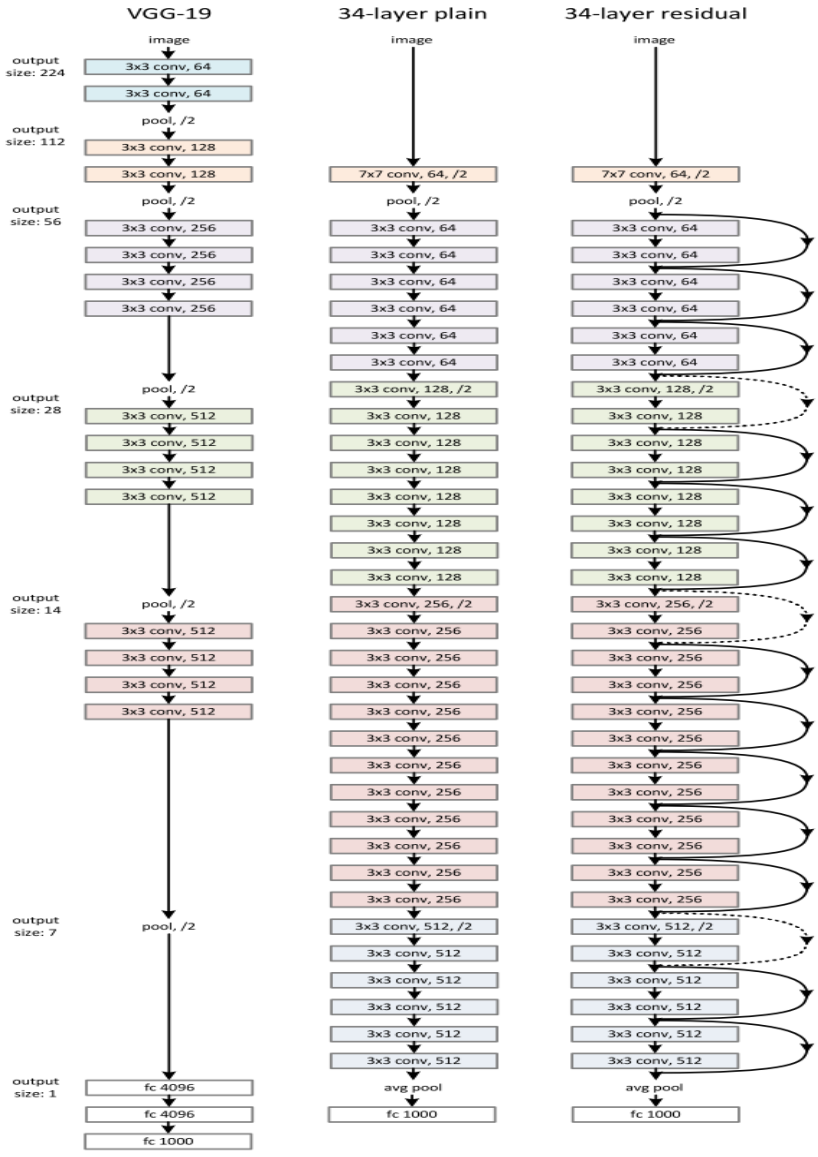

Fig. 6. The architecture of the convolutional neural network Residual Network [13].

2.7 Obtaining of Estimates and Analysis of Effectiveness Using of the Considered Models of Convolutional Neural Networks

To estimate the convolutional neural network models in addition to the type of errors usually indicate the number of models in the ensemble and the number of notches images that were fed to the input of each model. For example, 10 notches means that four notches are made at the corners of the image, one notch in the center, and each notch is additionally horizontally inverted.

According to numerous experiments [10-14], the generalization and analysis of the obtained results were made.

In the Tab. 1 shown the results of the considered neural networks with one model and one cutout based on ImageNet images (except ResNet-152, for which the result for 10 notches is indicated).

Table 1. Network efficiency for a single cut-out model.

Neural network	Top-1	Top-5	Number of layers	Number of operations (G-Ops)
AlexNet	39,7 %	18,9 %	8	70 M
ZF Net	37,50 %	14,8 %	8	70 M
VGG Net	25,60 %	8,10 %	19	155 M
GoogLeNet	29,00 %	9,20 %	22	10 M
Inception-v3	21,20 %	5,60 %	101	35 M
Inception-v4	20,00 %	5 %	152	35 M
Inception-ResNet-v2	19,90 %	4,90 %	467	65 M
ResNet-152	19,38 %	4,49 %	152	65 M

In Tab. 2 shown the results of using ensembles of models with many cutouts based on ImageNet images.

As can be seen from these tables, for five years, from 2012 to 2016, the Top-5 error on ImageNet for single models decreased almost four times (from 17% to 4.49%), and for the ensemble – almost five times (from 15.40% to 3.10%).

Analyzing the experimental data (Tab. 1, Tab. 2), we can conclude that the choice of network architecture is made according to the following criteria: classification errors, performance, and the complexity of learning a neural network. For this, the following algorithm is usually used.

Initially, guided by certain requirements, they set a permissible classification error. For example, it is currently believed that a classification error when using human vision is in the range from 5% to 10%. If you look at the classification error of the latest convolutional neural networks, you can see that they are coping with this task as

well as a human. This means that in the classification problems that a person solved classically, you can choose any network with an error not higher than the specified one. Based on the analysis of data in Tab. 2, we can conclude that the last five networks will suit us. But we need one. What should be done?

Table 2. The effectiveness of the network for ensembles with many notches.

Neural network	Models	Notches	Top-1	Top-5
AlexNet	7	1	36,70 %	15,31 %
ZF Net	6	10	36 %	14,70 %
VGG Net	2	150	23,70 %	6,80 %
GoogLeNet	7	144	—	6,67 %
Inception-v3	4	144	17,20 %	3,58 %
ResNet-152	6	144	—	3,57 %
Inception-v4 + 3x Inception-ResNet	4	144	16,50 %	3,10 %

Next, the selection of an admissible network is made in order to satisfy the specified restrictions on labor intensity (estimates of labor intensity are given in Tab. 1), taking into account the available hardware capacities. This choice is also made taking into account the time constraints on the network learning process, since with the increase in the number of layers and network parameters, the training time will also increase.

The choice of complexity can also be ambiguous, since at the first stage five networks were chosen. In such a situation, the most important criterion is usually identified and the best network is selected by this criterion.

To improve the quality of the classification results, it is planned to use specialized frame preprocessing models and algorithms [15-19] in addition to developing of network ensembles.

3 Conclusion

In the course of considering the most effective models of convolutional neural networks used in our time for the purposes of image classification, an analysis of their architectural features was performed. According to numerous experiments, a generalization and analysis of the results of the efficiency of using neural networks for image classification (Tab. 1, Tab. 2) was made. On this basis, a generalized algorithm is formulated and practical recommendations are given regarding the choice of the best architecture of a neural network, respectively, the specifics of the problem.

References

1. Rafael C. Gonzalez, Richard E. Woods Digital Image Processing, 4th edition Pearson/Prentice Hall, 2018. – 1168p.
2. David A. Forsyth, Jean Ponce Computer Vision: A Modern Approach (2nd ed.). – Pearson Education Limited, 2015. – 792p.
3. Milan Sonka, Vaclav Hlavac, Roger Boyle, Image Processing, Analysis, and Machine Vision (4th ed.). – Cengage Learning, 2014. – 896p.
4. Ian Goodfellow, Yoshua Bengio, Aaron Courville Deep Learning. – MIT Press, 2016. – 787p.
5. Peter Norvig , Stuart Russell Artificial Intelligence: A Modern Approach, Global Edition. – Pearson Education Limited, 2016. – 1152p.
6. Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. International journal of computer vision 88(2), pp. 303-338.
7. Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik Rich feature hierarchies for accurate object detection and semantic segmentation. The IEEE Conference on Computer Vision and Pattern Recognition. – 2014. – pp. 580-587.
8. Papers, https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.
9. Medium, https://medium.com/@kanchansarkar/relu-not-a-differentiable-function-why-used-in-gradient-based-optimization-7fef3a4cecec.
10. Arxiv, https://arxiv.org/pdf/1311.2901v3.pdf.
11. Arxiv, https://arxiv.org/pdf/1409.1556v6.pdf.
12. Cv-foundation, https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf.
13. Arxiv, https://arxiv.org/pdf/1512.03385v1.pdf.
14. Ross Girshick Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision. – 2015. – pp. 1440-1448.
15. I. Ruban, K. Smelyakov,V Martovytskyi, D. Pribylnov and N. Bolohova Method of neural network recognition of ground-based air objects // IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT), 24-27 May 2018. – P. 589-592. DOI: 10.1109/DESSERT.2018.8409200
16. K. Smelyakov, D. Pribylnov, V. Martovytskyi, A. Chupryna Investigation of network infrastructure control parameters for effective intellectual analysis // IEEE 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), 20-24 Feb. 2018. – P. 983-986. DOI: 10.1109/TCSET.2018.8336359
17. K. Smelyakov, A. Chupryna, D. Yeremenko, A. Sakhon, V. Polezhai Braille Character Recognition Based on Neural Networks // IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 509-513.
18. S. Mashtalir, O. Mikhnova, M. Stolbovyi Sequence Matching for Content-Based Video Retrieval// IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 21-25 August 2018. – P. 549-553.
19. G. Churyumov, V. Tokarev, V. Tkachov and S. Partyka, "Scenario of Interaction of the Mobile Technical Objects in the Process of Transmission of Data Streams in Conditions of Impacting the Powerful Electromagnetic Field", 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 2018. – DOI: 10.1109/DSMP.2018.8478539.