Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Detecting Images That Have a Destructive Impact on Users on the Internet

Anastasia Iskhakova*
V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
iao@ipu.ru

Roman Meshcherykov
V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
Moscow, Russia
mrv@ipu.ru

***Abstract*—This article deals with the current problem of automated detection of facts of destructive influence of images on the Internet on the person. It should be noted that detecting aggressive content in images is more difficult than detecting an object and defining its category. The reason for this is that aggressive content has no specific color or object parameters, there are no common features for classification. In this case the use of convolutional neural networks with pre-learning is a successful solution. Due to increase in aggressive content on the Internet, the problem of identifying such objects in order to minimize its impact on users is acute. The paper presents the developed architecture of the neural network for solving the problem of image with aggressive content recognition. Experiments during 180 epochs showed that at the number of epochs ~ 95-100 during training and ~ during testing it is possible to achieve the classification accuracy 91.8% taking into account the top-5 results.**

Keywords—neural network, aggressive content, convolutional neural network, residual neural network, object, destructive impact, image processing, control and security for critical infrastructure systems

I. INTRODUCTION

Countering aggression, extremism, suppression, humiliation of human dignity in a virtual environment is a pressing task today. The popularity of various virtual portals, social networks, the expansion of their functionality and accessibility to all segments of the population, together with anonymity, the possibility to act on another name, the emergence of many fraudulent and criminal schemes, creates a problem of ensuring security of stay in a virtual environment for a person. The most appropriate environment for attacks on protected psychophysiological resources is currently the Internet 's global computing network. This is due to a number of features:

- scale;
- heterogeneity;
- decentralization;
- lack of censorship;
- possibility to transmit multimedia data of any kind.

Thus, not only the problem of protection of the information is extremely actual, but also the problem of protection of the person from the information. And the last task has recently acquired an international scale and strategic character.

The tasks of safe interaction between man and machine are extremely relevant and important for society in recent years. A more private task in this aspect is the problem of safe interaction of the person with the virtual world - the content of the world-wide Internet. After all, today virtual tools allow to significantly simplify many processes within the society, and the undeniable benefit of the Internet forces users to trust it, to spend a long time in the virtual environment, to receive information without checking it in additional sources, which can be used for illegal purposes.

To solve this complex, multifaceted and ambiguous problem, it is necessary to consider a number of smaller private tasks, among which it is important to monitor the content of the virtual environment, create automatic means of recognizing and detecting possible destructive effect on the user and respond to such content in a timely manner. An example of a text material analysis is proposed in the work [1-3], the processing of audio materials (or audio parts of it) to detect perception anomalies is presented according to [4, 5]. In this paper the authors give the results of research of graphic images, as an integral part of Internet content, which very often occurs on web resources of any orientation and can also negatively affect the condition of a person, change his mood, strengthen the impression of accompanying material of another type, and so on.

The main stage of solving this problem issue is planned to introduce intelligent analysis technologies with dynamic training elements based on the use of artificial neural networks to detect signs of aggression, psychological pressure, and destructive impact on individual and group consciousness of users. The scientific novelty of the study consists in the subsequent development of a system of social process management, characterized by a new approach to data analysis, decision-making and management of social phenomena in a virtual environment by modeling a socio-cyber physical system.

II. APPLYING NEURAL NETWORKS IN TASKS OF AUTOMATIC PROCESSING OF DIGITAL IMAGES

The task of finding as well as recognizing objects in images is becoming increasingly urgent due to the development of digital technologies and the spread of fixing video devices everywhere. Previously, methods using Histogram of Oriented Gradients (HOG), described in [6], Haar features, described in [7], Viola-Jones methods, as in works [8-11] have been proposed.

Artificial neural networks trained by the inverse error propagation mechanism have virtually displaced other approaches in many parameter recognition and estimation tasks. Convolutional neural networks are well established in

solving problems of image recognition and classification according to [12].

Nowadays, there are ready-made models of neural networks trained on a rather massive set of objects' images belonging to different categories. Availability of ready-made models allows taking into account a large number of classes, not to waste time for database formation and model training, not to look for features for classification. You can learn such models from your dataset. A disadvantage of such models is the need to bring the source data to a predetermined pattern (limited to the first layer of the model network), and in the additional training case, the need for a sufficient set of training data so that the necessary classes are equally recognizable by the network. For example, there are VGG Net models with 16 or 19 layers, Alex Net, ZF Net, ResNet. All these models were ILSVRC winners in different years, were trained based on the ImageNet dataset with over 15 million images belonging to 1000 categories. The author of [13] proposed comparative studies of the above-described networks, as well as their combined application for object classification.

2.1 Convolutional Neural Network

A convolution network is a multi-layered perceptron specifically designed to recognize two-dimensional surfaces with a high degree of invariance to transformations, scaling, distortions, and other types of deformation. Each neuron of the convolutional layer receives an input signal from a local receptor field in the previous layer, thereby extracting its local features. Once a feature is retrieved, its exact location is irrelevant because it is approximately positioned relative to other features.

Each computing layer of the network consists of a plurality of feature maps, each of which has the shape of a plane on which all neurons must share the same plurality of synaptic scales.

Such neural networks structure provides invariance to displacement realized by feature maps using convolution with a small size nucleus and reduction in the number of free parameters realized by sharing synaptic weights.

However, the task requires the ability to recognize thousands of different objects, often depicted in conjunction with each other. This fact significantly increases the requirements for the computational ability of the model. In flat neural networks, this is solved by increasing depth (i.e., increasing the number of layers), but this approach leads to new problems. Virtually all modern models of neural networks are trained based on gradient methods. Because of their specificity, increasing the number of layers leads to the problem of gradient degradation, and therefore the impossibility of quality learning. Residual neural networks may be the solution in this situation.

2.2 Residual Neural Network

It is known that a neural network can approximate almost any function, such as some complex function $H(x)$. Then it is fair that such network will easily learn residual one: $F(x) = H(x) – x$. Clearly, our original objective function will be equal to H(x) = F(x) + x. If we take some network, and add a few more layers to it, we would like the deep network to behave at least as well as its shallow counterpart.

The degradation problem implies that a complex nonlinear function F(x), obtained by joining multiple layers, must learn the identical transformation, if a quality limit has been reached on previous layers. However, this does not happen for reasons that the optimizer simply fails to adjust weights so that a complex nonlinear hierarchical model makes an identical transformation. Adding a direct connection will allow the optimizer to make all weights close to zero rather than creating an identical transformation.

The main building element of the classical residual neural network is the residual block (Fig. 1). It is a few interconnected neural layers with a parallel short-circuited connection. In addition to all the positives described above, such structure is still good in that it imposes no restrictions on the type of internal neural layers, except that they should not change the dimension of the data. The residual neural network itself consists of input and output layers, between which residual blocks are located.

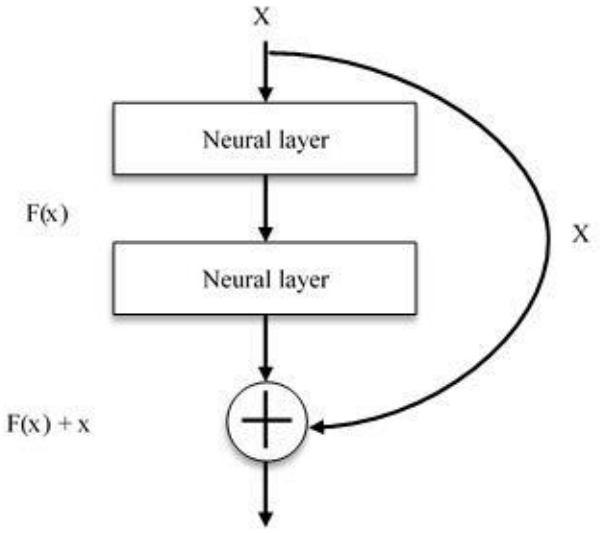

Fig. 1 Schematic diagram of the residual block structure

Built on this type neural networks can contain tens and hundreds of layers and retain the ability to learn, as opposed to deep convolutional neural networks.

III. NEURAL NETWORK'S ARCHITECTURE FOR SOLVING THE TARGET TASK OF DETECTING IMAGES WITH FEATURES OF AGRESSION

In order to solve the problem of recognition of aggressive images of social media, it was decided to use 50- layer residual neural network, consisting of convolutional and averaging layers. The main part of such network consists of 16 residual blocks.

The input layer consists of three matrices of neurons with size 224x224 (for each color channel of the picture), to which image data is supplied. The layer has no non-linear activation function and enters data into the model.

The input layer is followed by a first convolution layer having 64 feature maps, an unsaturated ReLU activation function, described in [14, 15] and capable of normalizing data by batch normalization, as in works [16, 17]. The layer

extracts the primary features from the image, which are 112x112 matrices, by zero-indentation matching. To reduce the size of the obtained data, an averaging layer with a 3x3 core and a step 2 is used.

The input neural cascade is followed by 4 groups of discriminators, each with several residual blocks. Each one consists of three convolutional layers and a short-circuited joint (Fig. 2).

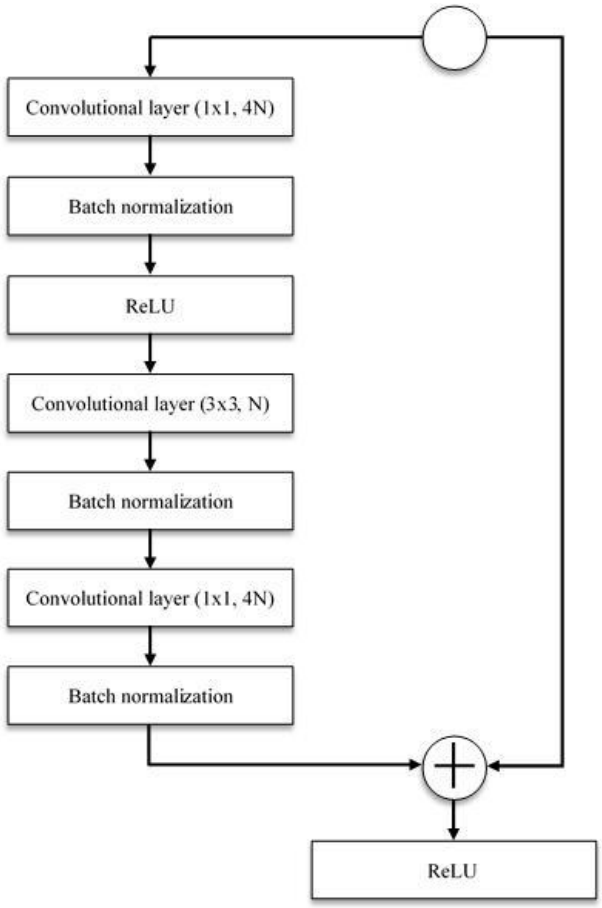

Fig. 2 Overall structure of the residual block

Each convolutional layer extracts features from the image and simultaneously acts as a discriminator. It is worth noting that the convolutional layers in the residual blocks have no activation functions. This is necessary for the successful implementation of package normalization. A layer of packet normalization is necessary to reduce internal covariant shift, which increases convergence rate and allows higher learning ratios to be used. It is only after the normalization step that a nonlinear transformation occurs in the ReLU layers where rectified linear neurons are used.

Weighted links are only present on the convolution layers because the remaining layers only convert the resulting data. All connections of neurons in the residual block are parallel. The residual block itself is fully bonded to the preceding layer.

The residual blocks of the 2nd, 3rd and 4th subgroups have a similar structure except that the first subgroup block reduces the size of the data. This is done by convolution with step 2. In order to be able to use short-circuited connections later, it is necessary to bring the data to the same dimension. This can be done by adding the same convolutional layer in front of the adder. It is also worth noting that when you move to a new subgroup, the number of convolution cores doubles.

The output stage of neural network consists of averaging layer and output one. The averaging layer has a recipe field size 7x7, operates with step 7, and has no activation function, in which case the output of the residual block with a dimension 7x7xN is converted to 1x1xN. The output layer consists of 1000 neurons, each corresponding to a class. The layer has softmax activation function.

This structure allows the network to dynamically increase its depth in the learning process, as the gradient will tend to spread along the shortest path from exit to entrance. As a result, the training will affect the layers that contribute to the network as much as possible, and if the necessary discriminatory ability is achieved, the weights of the excess layers will be close to zero.

The overall architecture of the neural network is shown in Table 1.

TABLE 1. RESIDUAL NEURAL NETWORK'S ARCHITECTURE

Name of block	***Output size of data***	***Description of layer***
Input block	(224x224, 3)	Input layer (224x224, 3)
	(112x112, 64)	Convolutional layer (7x7, 64) with step 2
	(56x56, 64)	Averaging layer (3x3, 64) with step 2
Residual blocks 1-3	(28x28, 256)	[1x1, 64; 3x3, 64; 1x1, 256] x 3
Residual blocks 4-7	(14x14, 512)	[1x1, 128; 3x3, 128; 1x1, 512] x 4
Residual blocks 8-13	(14x14, 1024)	[1x1, 256; 3x3, 256; 1x1, 1024] x 6
Residual blocks 14-16	(7x7, 2048)	[1x1, 512; 3x3, 512; 1x1, 2048] x 3
Output block	(1x1, 2048)	Averaging layer (7x7, 2048) with step 7
	(1x1000)	Full-coherent layer (softmax)

The activity of each neuron of the output layer reflects the probability of the presence of an object of a certain class on the image. Previously, all 1,000 classes were divided into neutral and aggressive classes. The 5 most dominant classes are used for overall image evaluation, if at least one of them is in the group of aggressive, then the image is considered aggressive.

IV. RESULTS OF EXPERIMENTAL STUDIES AND EVALUATION OF THE DEVELOPED MODEL'S EFFICENCY

The implementation of the model for solving the aggressive images recognition problem is as follows. The original images are transformed to a pixel dimension 224x224, for each pixel normalization is performed by subtracting the average value, after which data is supplied to the neural net's input layer. Batch normalization layers are used to normalize data within the model, which are between each convolution and activation layers. For all convolution cores L2 regularization is applied with degradation factor equal to 0.0001. The weights initialization is randomly. Learning is done using the reverse error propagation method

in the 128 size mini-batch mode. Optimization was carried out by Adam method, as in work [18] with the following parameters: $\alpha = 0.001$, $\beta 1 = 0.9$, $\beta 2 = 0.999$ and $\varepsilon = 10e^{-8}$. The duration of the full stage of training is 180 epochs.

For the model estimation was used the ImageNet 2012 dataset, which contains images divided into 1000 classes. Training was done on 1,280,000 training images. Accuracy was evaluated on a test sample of 50,000 images. The accuracy of classification of 5 dominant classes on the image was evaluated. The results are shown below (Fig. 3), where the bold line indicates the error on the test set and the thin one on the training set. Classification error was 8.2%. On average, the learning process converged to the 110 epoch, after which the increase in accuracy almost stopped.

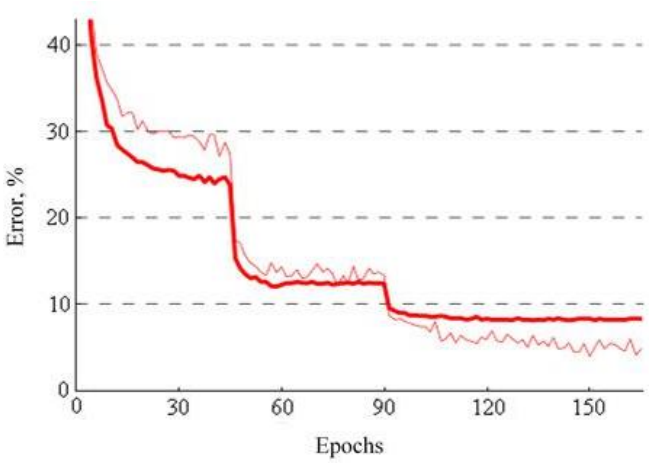

Fig. 3 Classifier error estimate (5 classes)

The results of detection of aggressive images by use of the model depend entirely on the accuracy of the classifier and the categorization of object classes. Examples of processed images are shown below (see Fig. 4, Fig.5).

Fig. 4 Experimental results (part 1)

Fig. 5 Experimental results (part 2)

V. Conclusion

The work proposed an architecture of a residual neural network consisting of convolutional and averaging layers. Experiments during 180 epochs showed that at the number of epochs ~ 95-100 during training and ~ 110 during testing it is possible to achieve the classification error 8.2% taking into account the top 5 results. It is worth noting that detecting of aggressive content in images is more difficult than searching an object and defining its category, because aggressive content has no natural color or object parameters, there are

no generally accepted features for classification. In this case, the use of convolutional neural networks with pre-learning is a successful solution. For existing networks that are used to classify images into categories of depicted objects, not to classify the emotional component, the result of the 91.8% classification accuracy with a relatively small test sample size is high.

The results of the work are significant both in terms of basic science and as practical research work. They are of great State and national importance because they can be used in automated systems of critical facilities, as well as in the telecommunications sector, in power ministries and departments, in administrative-State structures, where computer systems and networks are widely used and there are increased requirements to protect against unwanted and harmful information. The results can be implemented in products of companies engaged in development and implementation of promising tools of parental control, spam filtering and analytical processing in social networks, in particular, and in information and communication systems, in general.

ACKNOWLEDGMENT

The reported study was partially funded by RFBR, project number 18-29-22104.

REFERENCES

[1] I. Kulagina, A. Iskhakova, and R. Galin, “Modeling the practice of aggression in the socio-cyber-physical systems,” Tomsk State University Journal of Philosophy, Sociology and Political Science, vol. 52, pp.147-161, 2019.

[2] E. Garin, and R. Meshcheryakov, “Method for determination of the social graph orientation by the analysis of the vertices valence in the connectivity component,” Bulletin of the South Ural State University», series «Mathematics. Mechanics. Physics, vol. 9(4), pp. 5-12, 2017.

[3] A. Iskhakova, A. Iskhakov, and R. Meshcheryakov, “Research of the estimated emotional components for the content analysis,” Journal of Physics Conference Series, vol. 1203(1), article no. 012065, April 2019.

[4] A. Iskhakova, M. Alekhin, and A. Bogomolov, “Time-Frequency Transforms in Analysis of Patterns of Non-Stationary Quasi-Periodic Biomedical Signals for Acoustic Anomalies Identification,” Information and Control Systems, vol. 1, pp. 15-23, February 2020.

[5] D. Levonevskii, O. Shumskaya, A. Velichko, M. Uzdiaev, and D. Malov, “Methods for Deter-mination of Psychophysiological Condition of User Within Smart Environment Based on Complex Analysis of Heterogeneous Data,” Proceedings of 14th International Conference on Electromechanics and Robotics “Zavalishin's Readings”. Smart Innovation, Systems and Technologies, vol. 154, pp. 511-523, August 2019.

[6] Y. Bardhan, T. Fulzele, P. Ranjan, S. Upadhyway, and V. Bharate, “Emotion Recognition using Image Processing,” International Journal of Trend in Scientific Research and Development (IJTSRD), vol. 2, issue 3, pp. 1523-1526, March-April 2018.

[7] L.A.B. Prieto, and Z. Kominkova-Oplatkova, “A performance comparison of two emotion-recognition implementations using OpenCV and Cognitive Services API,” MATEC Web Conf., vol. 125, acticle no. 02067, October 2017.

[8] F.Z. Salmam, A. Madani, and M. Kissi, “Emotion Recognition from Facial Expression Based on Fiducial Points Detection and using Neural Network,” International Journal of Electrical and Computer Engineering, vol. 8, issue 1, pp. 52-29, February 2018.

[9] Kh. Lekdioui, Y. Ruichek, R. Messoussi, Y. Chaabi, and R. Touahni, “Facial Expression Recognition Using Face-Regions,” Proceedings of the 3rd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP’2017 (2017), pp. 1-6, May 2017.

[10] K. A. Verma, “Facial expression recognition using Gabor filter and multi-layer artificial neural network,” Proceedings of the International Conference on Information, Communication, Instrumentation and Control (ICICIC), pp. 1-5, August 2017.

[11] S. Ren, “Object detection networks on convolutional feature maps,” IEEE transactions on pattern analysis and machine intelligence, vol. 39, issue 7, pp. 1476–1481, 2017.

[12] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images,” IEEE Transactions on Geoscience and Remote Sensing, vol. 54, issue 12, pp. 7405–7415, 2016.

[13] O.S. Sikorskij, “Review of convolutional neural networks in the image classification task,” New Information Technologies at Automated Systems, vol. 20, pp. 1-5, 2017.

[14] Y. Liu, J. Zhang, Ch. Gao, J. Qu, and L. Ji, ”Natural-Logarithm-Rectifed Actvation Function in Convolutional Neural Nets,” ArXiv, vol. abs/1908.03682, 2019.

[15] Md. I. Quraishi, J.P. Choudhury, and P. Chakraborty, “A Framework for the Recognition of Human Emotion using Soft Computing Models,” International Journal of Computer Applications, vol. 40, issue 17, pp. 50-55, 2012.

[16] S. Ioffe, and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,”. Proceedings of Machine Learning Research, vol. 37, pp. 448-456, July 2015.

[17] N. Dalal, and B. Triggs, “Histograms of oriented gradients for human detection,” International Conference on Computer Vision & Pattern Recognition (CVPR ’05), pp. 886-893, June 2005.

[18] D. Kingma, and J. Ba. Adam, “A Method for Stochastic Optimization,” 3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings, pp. 1-15, May 2015.