https://doi.org/10.31891/csit-2026-1-14

Oleksandr DORENSKYI
Docent, PhD, Associate Professor of Cybersecurity and Software Department,
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine, https://orcid.org/0000-0002-7625-9022
e-mail: dorenskyiop@kntu.kr.ua

Oleksandr DRIEIEV
Docent, PhD, Associate Professor of Cybersecurity and Software Department,
Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine, https://orcid.org/0000-0001-6951-2002
e-mail: drey.sanya@gmail.com

Hanna DRIEIEVA
Docent, PhD, Senior Lecturer of Cybersecurity and Software Department, Central Ukrainian National Technical University, Kropyvnytskyi, Ukraine,
https://orcid.org/0000-0002-8557-3443
e-mail: gannadreeva@gmail.com

Received: 05/01/2026
Accepted:08/03/2026
Published:26/03/2026

UDC 004.932:004.032

THE METHOD OF IDENTIFYING KEY ELEMENTS OF A DIGITAL IMAGE IN THE DECISION-MAKING PROCESS OF CLASSIFICATION BY A NEURAL NETWORK

This paper addresses the problem of evaluating the quality of data used for training neural networks by identifying significant elements of digital images, which the neural network algorithm relies on for classification. As is well known, image classification systems are widely used in computer vision, where entire images are assigned to a single class without distinguishing individual objects within them. This is a typical problem in computer vision. For classification tasks, pre-trained convolutional neural networks (CNNs) are often employed, trained on labelled datasets. However, the unresolved issue remains as to which specific elements the neural network relies on when making a particular decision. The paper presents a method based on a competitive gradient descent process to extract details (elements) that were significant during the classification process, i.e., key elements. This method involves the competition between the process of image detail degradation and the preservation of classification results. Using a self-trained neural network, the authors analyse the presence of details in a classified digital image by visually assessing the elements directly related to the classified object after applying the proposed method. Thanks to this approach, the degradation of digital image details while preserving classification quality is achieved, and the network architecture may be arbitrary. This allows for a comparison of attention areas in neural networks with different architectures: convolutional architecture and mixing architecture. Based on the research findings, a method for localizing neural network attention with arbitrary architecture is proposed. The preserved elements on the degraded image can provide additional information about the validity of the classification performed by a specific neural network. This can be assessed by localizing the preserved elements. For example, the presence of these elements in the classified object indicates a high probability of correct neural network performance. In other words, if the preserved elements do not belong to the classified object, it can be concluded that the training data is not representative (for example, one of the objects may more frequently appear against a characteristic background). Experimental studies demonstrated the advantages of the proposed method over existing alternatives: accuracy in localizing significant details (elements), the presence of information about global significant elements of the digital image and their shape, as well as its applicability for both convolutional and other types of neural networks.

Keywords: computer vision, neural networks, attention localization, images, data quality

Introduction

The modern development of computer vision systems has led to the widespread use of neural networks for object classification in images. Many studies focus on the correct operation of neural networks during image classification. For this, techniques such as augmentation [1], validation [2, 3], testing [3], methods for identifying neural network attention areas [4], and methods for controlling the spread of intermediate interlayer data in neural networks are employed. Some of these methods are designed to prevent incorrect training of the network, some visualise the attention areas of the neural network, while others can identify features of unstable neural network operation in specific classification cases. However, these methods have their own limitations in terms of usage and architectural constraints. For instance, augmentation methods can create the illusion of image variety for training the neural network, but this leads to the recurrent inclusion of background elements during training, which is detrimental to the learning process. Validation during training checks the network's performance on labelled data that does not

belong to the training dataset. This allows for a comparison of classification accuracy between the training and validation datasets. If the accuracy on the validation data is significantly lower than on the training data, it is concluded that the neural network is overfitted. Testing is designed to verify the network's performance on data that was not used for either training or validation. In cases where there is insufficient data for training the network, validation results may show favourable outcomes by chance. Additional testing on test data helps reduce the likelihood of incorrectly drawing conclusions about the classification quality of the neural network when using only validation.

Image classification by neural networks is carried out on the entire image. Therefore, the analysis concerns not only the regions with the object but also background elements. As a result, it is common to encounter cases where the neural network classifies based on elements that do not belong to the object being classified. This case may indicate that the set of images selected for training the neural network was improperly chosen. Moreover, such an error affects both the test and validation parts of the training data. Therefore, it is not possible to automatically determine the training error that takes background elements into account. This makes it crucial to detect the elements of a digital image on which the classification by the neural network is based.

Related works

The scientific task of identifying (locating) the regions of an image that have the greatest impact on the classification result by a neural network has been the focus of numerous studies. The beginning of the solution to this problem was laid out in the scientific work [5], where the authors improved the system for identifying the attention centres of the neural network by highlighting the components of the image that contributed to the classification features. This was achieved through an optimisation process on the coefficients of the input image, where the loss function increases in accordance with the sharpness evaluation of the input image and rises as the output values of the convolutional layer deviate from the reference obtained during the initial classification [5]. Furthermore, the research in [6] proposed a method for detecting parts of the image that stand out from the overall background based on texture and can be interpreted as deviations from the typical signal of local brightness and colour distribution on the image. Additionally, the study proposed a classification of the requirements for image segmentation systems and generalised the process of feature detection on a digital image step by step. Moreover, the paper [4] is dedicated to the location of key features during classification, while the study [7] focuses on the architecture and training methods in the process of creating optimised convolutional neural networks (CNNs) for object and scene detection in images.

In [8], based on the results of analysing CNN architectures for handwritten symbol and digit recognition, the authors demonstrate how the structure of the network affects recognition accuracy. The direct link to the research on localization of key elements in digital images is found in [9], which proposes techniques for visualising which specific parts (regions) of the image are used by the neural network to make classification decisions. At the same time, the works [10, 11] are valuable in the context of our research. The first study explores data augmentation techniques to improve image classification accuracy using neural networks, which is important for understanding image quality in classification regarding the identification of key image elements [10]. The second work is dedicated to the well-known Grad-CAM method for generating pixel importance maps for convolutional neural networks, which allows visualising the parts of the image the model focuses on; the method for evaluating significant elements combines the process of detail degradation [11].

A review of the research and literature shows that existing methods are only effective for convolutional neural networks, where the convolution processes involve only parts of the image [12]. Additionally, access to the last tensor, which is the result of the convolution, is required. Such access to the structure of the neural network is not always available, so this method is not always applicable. The discussed methods also do not provide information on the shape of the image elements that served as the attention centre for the neural network during classification (only their approximate location is known).

Therefore, the task of identifying (outlining) the regions important for making the classification decision (for example, for detecting people in an image in [13] and the shape of significant elements of that image to confirm the adequacy of the response of the trained neural network) remains unresolved. The goal of the research is, therefore, to determine the correct attention of the neural network during the classification of digital images. Achieving this will ensure the development of methods for searching significant elements in digital images.

Identification of key elements of a digital image in the process of decision-making for classification by a neural network

The scientific and technical task of identifying the details that had the most significant impact on the classification result

The identification of neural network attention focal points during image classification enables the determination of whether these focal points are localized on classification objects or the background, allowing for an evaluation of the neural network's performance. To determine the attention focal points of the neural network, the authors chose a modification of the competitive learning method [14]. Unlike the well-known competitive learning methods, the proposed method also consists of two processes, but the first one is responsible for destroying the details in the image, while the competitive process preserves the probability vector indicating the image's membership in a set of classes.

For the destruction process, linear smoothing filters can be used, specifically averaging around each pixel, or a Gaussian filter. An important aspect is determining the loss function for the image, which provides a scalar that should be larger for an image with a greater number of details. Several measures can be proposed for this function: 1) the mean value of the absolute differences between adjacent pixels; 2) the root mean square deviation of the differences between adjacent pixels; 3) the mean deviation between the input image pixels and its blurred copy, and many others. What is common among these methods is that the value is higher for images with more significant brightness changes. The measure of detail presence in an image is zero when the image is monochromatic, i.e., when the brightness or colour of all pixels is the same. This value is denoted as $Loss_d$ in the text.

The process of preserving the classification vector is more complex. It requires determining the loss in the probability distribution for the image's membership to the simplified version of the image. To find differences in classification, the probability vector of the image's membership to the classes is preserved for the original image. The simplified image is also classified by the neural network under study, providing a probability vector indicating the membership of the current image to the classes. The similarity between the obtained vectors can then be assessed using several methods: 1) as the sine of the angle between these vectors; 2) as the Euclidean distance between points in the multidimensional space representing the obtained vectors; 3) as the sum of squared differences between corresponding values in the vectors (the squared Euclidean distance), and others. What is common among these measures is that the obtained scalar is higher for vectors that are more separated in space. This value is denoted as $Loss_c$ in the text.

The method of degrading insignificant details in the image using the described quantities is illustrated by the diagram shown in Fig. 1.

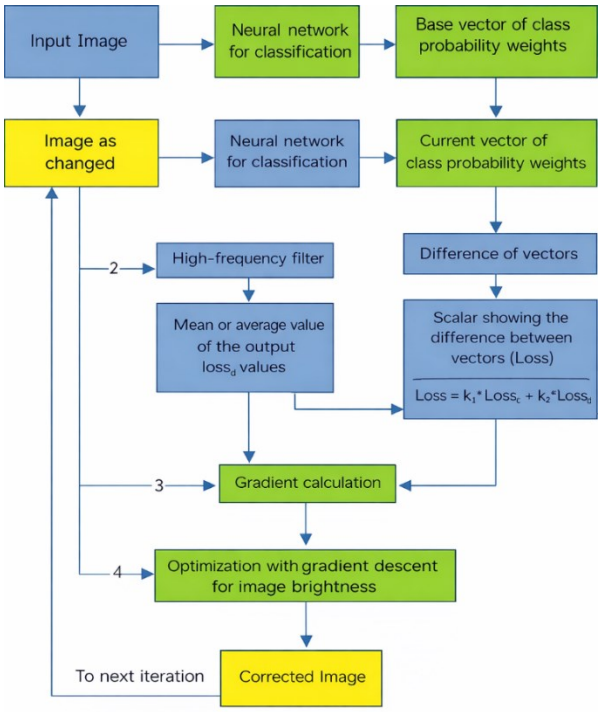


Fig. 1. Sequence diagram of actions for enhancing image details that are important for the classification result

For the purpose of mathematical formalisation of the calculations, the following notations are introduced:

$Loss_d$ – the loss or penalty for the existing details in the image;

$Loss_c$ – the loss or penalty for the difference between the current classification and the initial one;

I – the three-dimensional tensor of the input image with dimensions;

w – the width of the image in pixels;

h – the height of the image in pixels;

$I_{x,y}$ – a triplet of numbers representing the colour of the pixel in the x-th column and y-th row;

$Loss$ – the value of the loss function for the current image, for which the minimisation process is carried out to reduce the number of details while preserving class membership;

v_b – the initial or baseline classification vector of the input image, where the number of elements matches the number of classes defined by the neural network;

v – the current classification vector of the input image, which needs to be kept close to the vector vb.

As shown in the diagram in Fig. 1, the process of degrading insignificant image elements begins with the input image. First, the input image undergoes classification by the neural network (the movement in the diagram to the right of the input image). As a result of the neural network's operation, a vector of coefficients vb is obtained. Each

element of the class membership vector contains the degree of membership of the input image to the corresponding class. Often, the neural network provides such a vector in a normalized form by using a 'softmax' output layer, where the sum of all coefficients is equal to one, and these coefficients are interpreted as the probability of the image belonging to each of the classes. For a high-quality process, it is preferable to obtain this vector before normalisation, where the class membership features are absolute values. The class membership vector for the input image is recorded at the beginning of the process and remains unchanged during the operation, so this vector is considered as the baseline.

The next step of the method is to transfer the image into a tensor, where the image will now be altered. This path is marked with the number 1 in the diagram. Next, moving to the right along the diagram, the image undergoes classification by the neural network, where the current class membership vector v is obtained.

The next step involves assessing the number of details on the image $Loss_d$, as well as evaluating the deviation of the classification result of the altered image $Loss_c$. The processing path of the image for evaluating the number of details is marked as 2. The image is filtered to highlight the high-frequency components of the image. In the simplest case, the following process (1) can be used:

$$Loss_d = \frac{1}{(w-2)(h-2)} \sum_{x=0}^{w-2} \sum_{y=0}^{y-2} \left(|I_{x,y}| - |I_{x+1,y}| \right)^2. \quad (1)$$

The deviation from the initial class membership can be evaluated using the following expression (2):

$$Loss_c = \frac{|v - v_b|}{|v||v_b|}. \quad (2)$$

An alternative method for evaluating the deviation of classification vectors can also be used, based on the cosine angle between the vectors:

$$Loss_c = 1 - \frac{\langle v, v_b \rangle}{|v||v_b|}. \quad (3)$$

In the formulas, the straight brackets denote the operation of determining the length of the vectors, and the angular brackets denote the scalar product of the vectors.

However, for calculating the gradient, by which the images will be adjusted, a single loss coefficient is required. Therefore, it is proposed to use a weighted loss coefficient (4):

$$Loss = k_1 Loss_c + k_2 Loss_d. \quad (4)$$

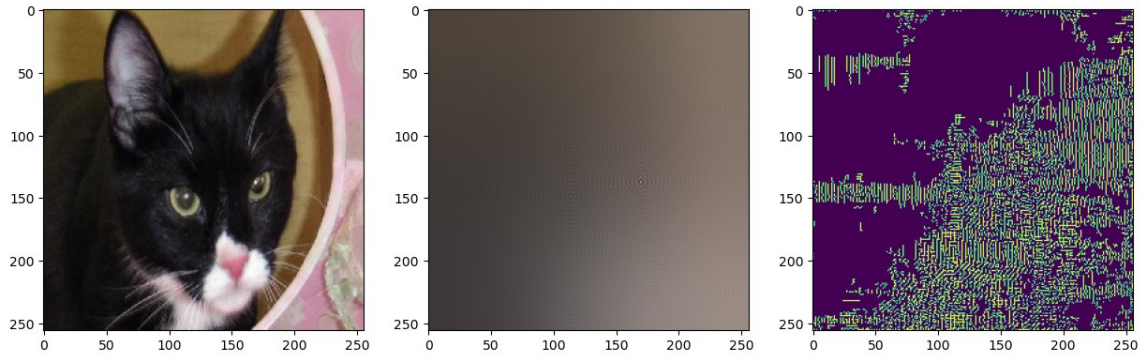
The subsequent stages involve, through the internal tools of the neural network development and training library, in the given example *TensorFlow* [9], the calculation of the gradient and the adjustment of the image pixels. These actions are repeated until the changes in the image remain noticeable. Further optimisation becomes unnecessary. The specific number of iterations in the process strongly depends on the nature of the image and the neural network used for analysis. Therefore, the degradation of details in the image should be monitored visually. After several experiments, the number of iterations can be fixed.

Improvement of the method for determining the attention areas of neural networks

The algorithm used in this work to assess the performance of neural networks for classification is based on convolutional layers and a neural network architecture called the 'Mixer' [16], which has not gained widespread popularity but has its own advantages and disadvantages. The main distinction of the Mixer architecture is that the information from each zone of the image is mixed with information from other zones. In contrast, the convolutional architecture processes information only from a limited area around each pixel, although the complete feature set is used when determining the class.

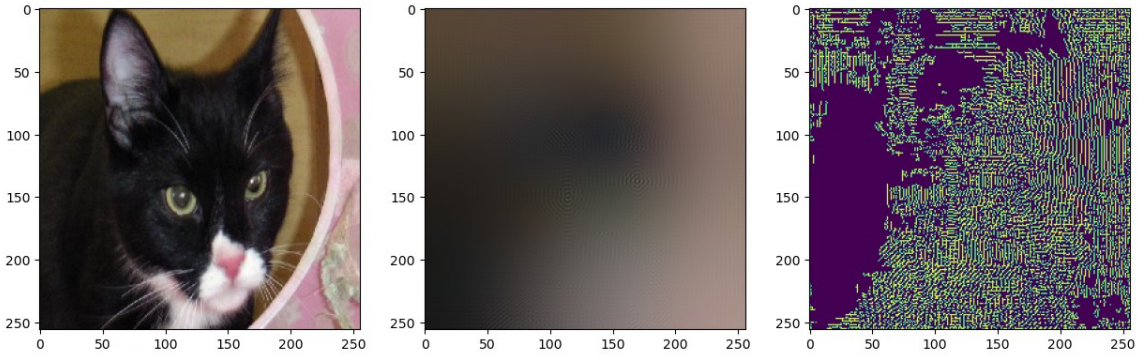
For training binary classification models, the data of cats and dogs [17] were used. The model complexity was tuned to achieve a validation accuracy of over 90%. For the convolutional neural network, more than 200,000 training coefficients were used, while for the MLP model, around 2 million coefficients were required, with a lower accuracy result.

The results of image detail regression for the cat are shown in Fig. 2 for the convolutional network and Fig. 3 for the MLP-Mixer network. In the figures, the following are indicated: a) – the input image; b) – the image with degraded details; c) – the result of applying the detail extraction filter.



a) **b)** **c)**

Fig. 2. Result of image detail degradation for the cat using the convolutional network and loss evaluation based on methods (1) and (2)



a) **b)** **c)**

Fig. 3. Result of image detail degradation for the cat using the MLP-Mixer network and loss evaluation based on methods (1) and (2)

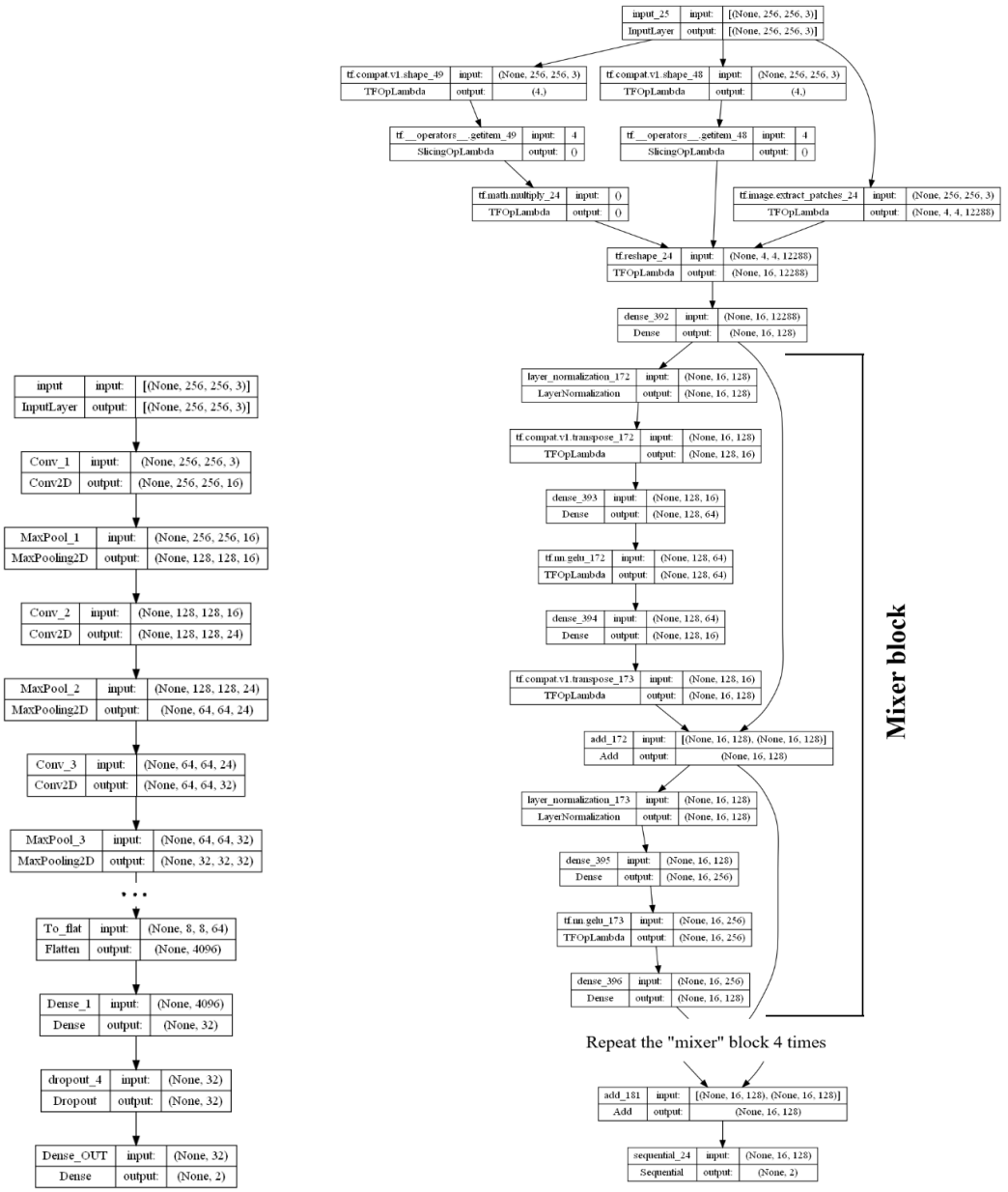
In comparing the results of image detail degradation, it can be observed that both neural networks focus on the presence of small image elements, not only in the cat's muzzle area but also around the eyes and nose. A significant number of details were retained in the background, which did not belong to the cat. At the same time, the degraded image better preserves the overall silhouette of the animal, particularly for the MLP-Mixer neural network, which leads to the conclusion that this neural network focuses more on the overall shape of the animal than the convolutional neural network.

The presence of a significant number of details in the background, which does not belong to the animal, indicates a difference in dominant backgrounds for cats and dogs. Therefore, the authors focused on the specifics of the locations where the cat and dog photos were taken in the training dataset. Visual inspection of the dataset revealed that cats appear outdoors in only about 20 out of 12,500 images. In contrast, dogs appear outdoors in about 50% of the 12,500 images. This directly confirms the conclusions regarding the difference in background for cats and dogs.

The architecture of the convolutional neural network and the Mixer architecture is shown in Fig. 4.

The study also further improves the method for determining the attention areas of neural networks [18]. Unlike the use of a fully connected layer in the form of a one-dimensional feature vector before the final classification, it is proposed to use convolutional features, as shown in Fig. 5.

According to the used architecture, the features across the image are convolved from the initial image size down to 8x8 coefficient maps. Due to the use of ReLU neuron activation, all features are non-negative. Therefore, each of the non-zero coefficients contributes to the confirmation of class membership, with the position of this coefficient on the 8x8 map corresponding to the position of significant elements in the classification process by this neural network. Consequently, during the validation of the neural network, the locality of the features on which the class membership decision was made can be determined. The degradation method for the input image, proposed in this work, is also applicable to this neural network to extract significant image details.



a) Convolutional **b) 'mixer'**

Fig. 4. Used network architectures for highlighting significant image details

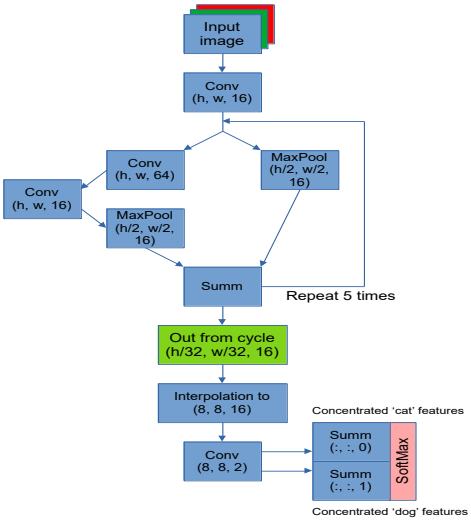


Fig. 5. RESNET neural network with a 'greedy' output classifier architecture

The implementation of such an architecture using Keras is shown in Fig. 6.

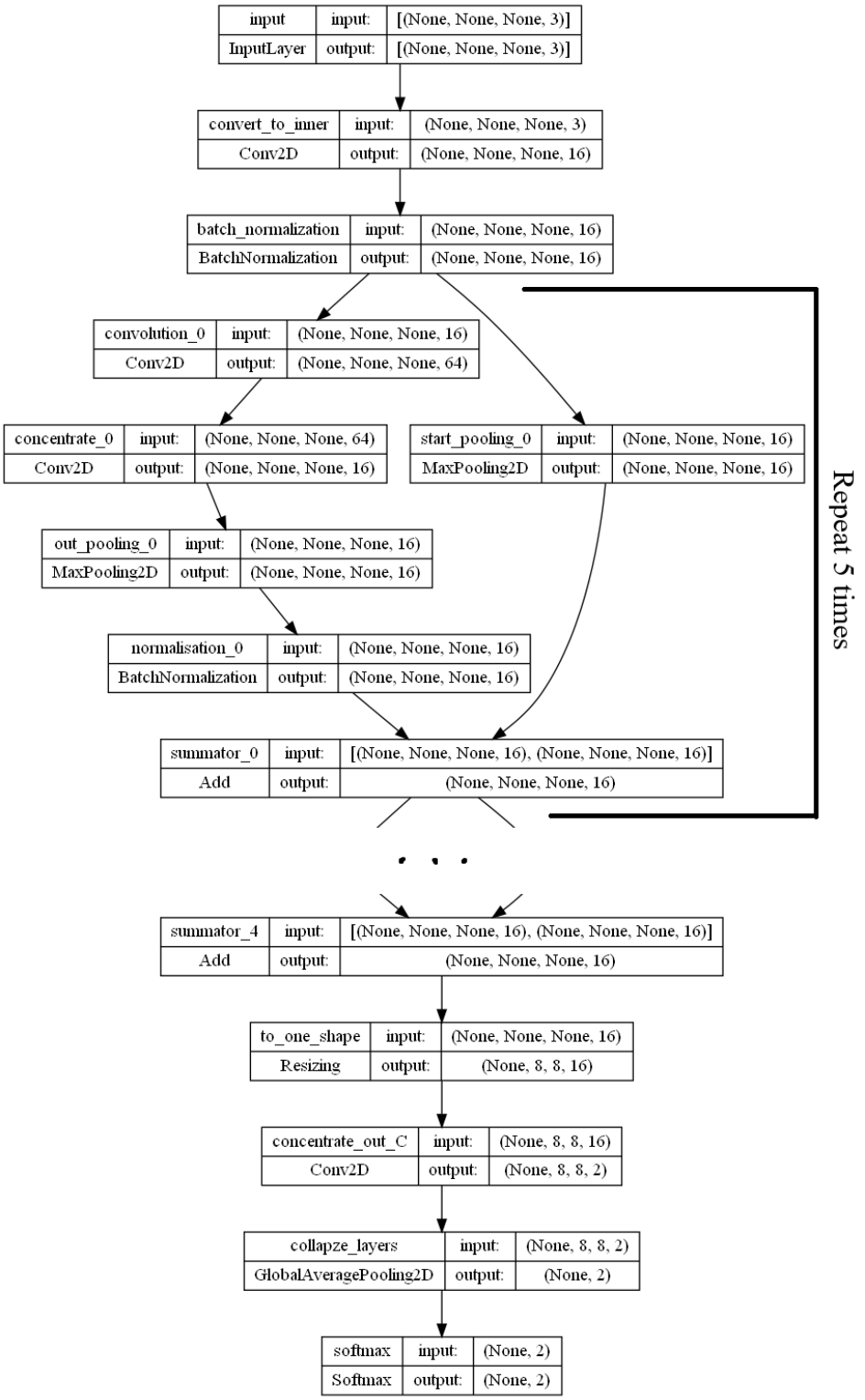


Fig. 6. Structure of the implemented RESNET with a 'greedy' classifier

Additionally, based on the proposed method of image degradation (for the classification process, Fig. 1), the authors enhanced the features for classifying cats from the same initial noise. The result obtained is shown in Fig. 7.

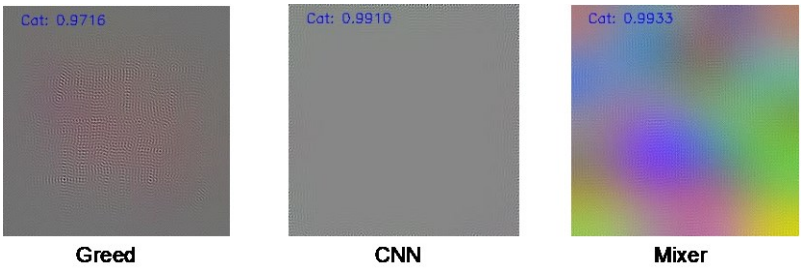


Fig. 7. Result of enhancing the 'cat' features for the initial random noise

According to the result shown in Figure 6, the RESNET network with a "greedy" classifier showed weak dependency on the image's colour and a concentration of features in the central part of the image. The convolutional neural network (CNN) uses details located near the edges of the image. This can easily be explained by the fact that dog images in the training dataset are mostly photographed against a background of grass and other outdoor objects, while cat photos are predominantly taken indoors. Therefore, this type of neural network is better avoided for practical tasks. For the MLP-Mixer neural network, as expected from its name, the placement of features important for classification is evenly distributed across the entire image, and a significant influence of slow colour changes is observed.

Experiments

Unlike the proposed method, the approach in [18] uses a reverse classification method, where neurons from the fully connected layer to the last convolutional layer with the largest coefficients are identified for a given class.

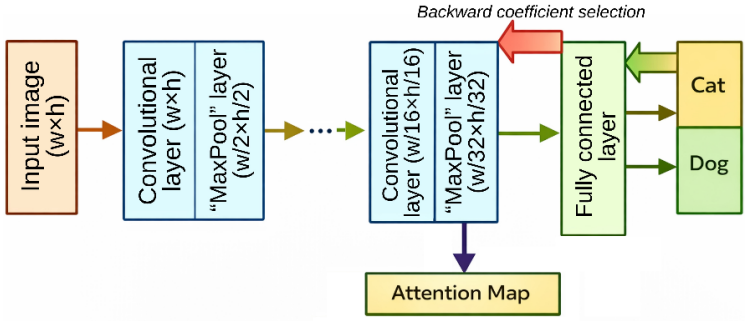


Fig. 8. Scheme for obtaining the attention distribution map from the last convolutional layer of the neural network based on the weight coefficients of the output fully connected layer (based on [18])

For the convolutional neural network used in this study, the method provided in [18] was also applied (Fig. 8). The result of constructing the attention map is shown in Figure 8. Since the convolutional neural network used in this study has 5 convolutional layers with max pooling, the resolution of the map in the last convolutional layer is 2^5 times smaller in both horizontal and vertical dimensions. That is, one pixel of focused attention is spread over a radius of 32 pixels without identifying specific elements to which attention is directed. Therefore, such an attention map is enlarged to the size of the input image using bicubic interpolation (Fig. 9, a) and overlaid on the input image as a heat map (Fig. 9, b).

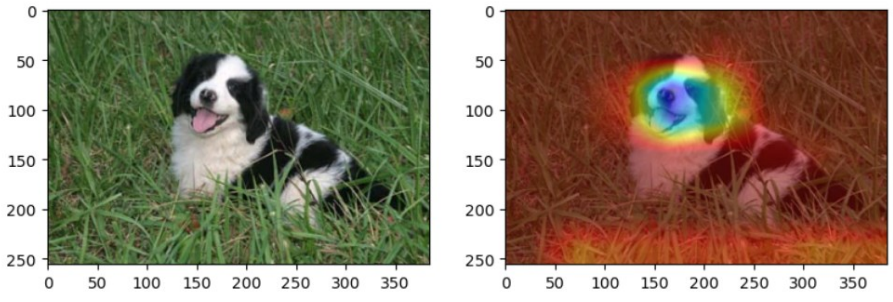


Fig. 9. Photo and identification of attention areas during classification using the reverse coefficient tracing method [18]

As a result of these actions, it is clear that the neural network placed significant focus on the animal's head area, but there is also considerable attention to the presence of grass in the background. Therefore, it is important to note the disadvantages of this method: low accuracy of localization of significant details; lack of information about

global significant image elements; absence of information on the shape of significant image elements; applicability of the method only for convolutional neural networks. The method developed in this study does not have these shortcomings but requires significantly more computational resources for its application.

Based on the experiment, the following directions for further research and scientific inquiry can be identified: the development of the input image element degradation algorithm in order to prevent the emergence of new details or changes in their interpretation by the neural network; the theoretical substantiation of the applicability of the method to arbitrary classifier architectures; the creation of a universal software tool for detecting problematic images using more powerful neural networks; the implementation of automated training procedures for simplified neural networks with the selection of architectural hyper parameters; as well as the construction of datasets containing only the most significant image details in order to accelerate learning towards the identification of meaningful patterns within digital images.

Improving the degradation algorithm for elements of the input image represents a logical continuation of the proposed competitive process, in which the destruction of image details is performed in parallel with the preservation of the classification outcome. This approach makes it possible to avoid the introduction of artificial patterns or distortions in the network's interpretation of image features, thereby retaining only genuinely significant elements. Since the approach considered in the paper is not restricted solely to convolutional neural networks (CNNs), but may also be applied to other types of architectures, a theoretical justification of its applicability to arbitrary classifier designs is both necessary and well grounded. This addresses the limitations of existing methods, which typically require direct access to convolutional tensors.

The software implementation for identifying problematic images follows from the assumption that stronger neural networks are capable of localizing attention and determining whether classification decisions are based on the object itself or on background artefacts. If the preserved key elements do not correspond to the classified object, this indicates that the training data may be non-representative.

The study demonstrates distinct behavioural differences between architectures (such as CNNs and MLP-Mixer models) in terms of their sensitivity to fine-grained details and global shape information. Therefore, automated training of simplified neural networks with hyper parameter optimisation is a justified direction. Such automation would support the construction of efficient models without unnecessary architectural complexity. For faster and more reliable learning, it is also important to construct datasets that contain only significant image details, as the degradation-based method enables the extraction of elements that truly influence the neural network's decision-making process. This reduces the impact of background bias and encourages the learning of meaningful patterns rather than incidental correlations.

Conclusions

The result of this study is the determination of the correct attention of the neural network during the classification of digital images. By using competition between the processes of detail destruction and preservation in the image, a method has been developed for localizing significant image elements, which allows the identification of details (elements) of a digital image that are key (most significant) for classification by the neural network. The task of identifying preserved elements as indicators of classification correctness is considered: the attention localization method enables the assessment of which parts of the image the neural network uses for classification. If the preserved elements belong to the object, it indicates the correct operation of the network.

Experimental results for convolutional and MLP neural networks showed that the MLP-Mixer focuses more on the overall shape of the object, while the convolutional network focuses on smaller elements, particularly in the background. It was also determined that background elements can significantly affect the classification process, and it is important to consider the location where the photos were taken for the training data. For example, cat photos often have a characteristic background, which impacts classification.

Thus, enhancing classification features based on image degradation improves classification accuracy for neural networks with different architectures. The example demonstrates the applicability of the image degradation method for neural networks with a more general architecture, distinguishing this method from known ones, as most of them are tied to the architecture of convolutional neural networks.

In this research, the set of methods for assessing the quality of a trained neural network has been expanded, and tools for additional evaluation of training data quality for neural networks have been obtained. The results of the study can be used to assess the performance of neural networks for image classification and to identify deficiencies in the training data for neural networks.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, O.Dr., H.D. and O.D.; methodology, O.Dr.; software, O.Dr.; validation, H.D., O.Dr. and O.D.; formal analysis, H.D. and O.D.; investigation, O.D., O.Dr. and H.D; resources, O.Dr.; data curation, O.Dr.; writ-ing – original draft preparation, O.Dr. and O.D; writing – review and editing, O.D.; visualization, H.D. and O.Dr.; su-

pervision, O.Dr.; project administration, O.D. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the authors used Gemini 3 Pro and ChatGPT 5.4 for: retrospective analysis, text translation, grammar and spelling checks. After using this tool/service, the authors reviewed and edited the content and takes full responsibility for the content of this publication.

REFERENCES

1. Rama J., Nalini C., Kumaravel A. Image pre-processing: enhance the performance of medical image classification using various data augmentation techniques. *ACCENTS Transactions on Image Processing and Computer Vision*. 2019. Vol. 5(14). P. 7-14. https://doi.org/10.19101/TIPCV.2018.413001

2. Kattenborn T., Schiefer F., Frey J., Feilhauer H., Mahecha M.D., Dormann C.F. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*. 2022. Vol. 5. P. 100018. https://doi.org/10.1016/j.ophoto.2022.100018

3. Kahloot K.M., Ekler P. Algorithmic Splitting: A Method for Dataset Preparation. *IEEE Access*. 2021. Vol. 9. P. 125229-125237. https://doi.org/10.1109/ACCESS.2021.3110745

4. Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. Learning Deep Features for Discriminative Localization. *CVPR*. 2016. P. 2921-2929. https://doi.org/10.1109/CVPR.2016.319

5. Dorenskyi O.P., Drieiev O.M., Mynailenko R.M. Method for Determining Features on Which a Neural Network Makes Classification Decisions. *Information Security and Computer Technologies*. 2023. P. 55. URL: dspace.kntu.kr.ua/server/api/core/bitstreams/a471dabd-8eb1-4ee8-a9c1-5d0b2d1b4c2f/content#page=55.

6. Drieiev O.M., Dorenskyi O.P., Drieieva G.M. Neural Network Method for Detecting Textural Anomalies in Digital Images. *Central Ukrainian Scientific Bulletin. Technical Sciences*. 2022. Issue 5(36). Part 2. P. 335-346. https://doi.org/10.32515/2664-262X.2022.5(36).2.335-346

7. Orlov R., Taboransky S. Training Models of Convolutional Neural Networks for Object, Scene, and Context Detection in Images. Challenges and Issues of Modern Science. 2024. Vol. 3. P. 150–156. URL: https://cims.fti.dp.ua/j/article/view/236.

8. Chychkaryov E.A., Zintshenko O.V., Lysenko M.M. Information Technology for Recognizing Handwritten Ukrainian Letters and Digits Using Synthetic Datasets. *Connection*. 2023. No. 1. https://doi.org/10.31673/2412-9070.2023.013237

9. Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. Learning Deep Features for Discriminative Localization. *CVPR*. 2016. P. 2921-2929. https://doi.org/10.1109/CVPR.2016.319

10. Rama J., Nalini C., Kumaravel A. Image pre-processing: enhance the performance of medical image classification using various data augmentation techniques. *ACCENTS Transactions on Image Processing and Computer Vision*. 2019. Vol. 5(14). P. 7-14. https://doi.org/10.19101/TIPCV.2018.413001

11. Selvaraju R.R., Cogswell M., Das A. et al. Grad CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *arXiv*. 2016. URL: https://arxiv.org/abs/1610.02391.

12. Khan S., Rahmani H., Shah S.A.A., Bennamoun M. Convolutional Neural Network. *A Guide to Convolutional Neural Networks for Computer Vision*. Synthesis Lectures on Computer Vision. Springer, Cham. 2018. https://doi.org/10.1007/978-3-031-01821-3_4

13. Al-Oraiqat A.M., Drieiev O., Drieieva H. et al. Spatiotemporal crowds features extraction of infrared images using neural network *J Ambient Intell Human Comput*. 2024. Vol. 15. P. 2543–2556. https://doi.org/10.1007/s12652-024-04771-5

14. Shinohara T., Xiu H., Matsuoka M. Point2color: 3d point cloud colorization using a conditional generative network and differentiable rendering for airborne lidar. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021. P. 1062-1071. URL: Point2color paper.

15. Catch up on the latest ML and AI developer updates from Google I/O “tf.gradients” / Google. URL: https://www.tensorflow.org/api_docs/python/tf/gradients.

16. Tolstikhin I., Houlsby N., Kolesnikov A. et al. MLP-Mixer: An all-MLP Architecture for Vision. 2021. https://doi.org/10.48550/arXiv.2105.01601 URL: https://github.com/rish-16/mlp-mixer-tf.

17. Microsoft and PetFinder.com have created this dataset in collaboration. *Kaggle: Cat VS Dog Dataset*. URL: https://www.kaggle.com/datasets/karakaggle/kaggle-cat-vs-dog-dataset.

18. An J., Joe I. Attention Map-Guided Visual Explanations for Deep Neural Networks. *Applied Sciences*. 2022. Vol. 12(8). P. 3846. https://doi.org/10.3390/app12083846

19. Liang J. Image classification based on RESNET. *Journal of Physics: Conference Series*. 2020. Vol. 1634. P. 012110. https://doi.org/10.1088/1742-6596/1634/1/012110

20. Su J., Vargas D.V., Sakurai K. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation*. 2019. Vol. 23(5). P. 828-841. https://doi.org/10.1109/TEVC.2019.2890858

Олександр ДОРЕНСЬКИЙ, Олександр ДРЄЄВ, Ганна ДРЄЄВА
Центральноукраїнський національний технічний університет, м. Кропивницький, Україна

МЕТОД ІДЕНТИФІКАЦІЇ КЛЮЧОВИХ ЕЛЕМЕНТІВ ЦИФРОВОГО ЗОБРАЖЕННЯ В ПРИЙНЯТТІ РІШЕННЯ КЛАСИФІКАЦІЇ НЕЙРОННОЮ МЕРЕЖЕЮ

Дослідження присвячене вирішенню проблеми оцінювання якості даних для тренування нейронних мереж шляхом виділення значущих елементів цифрового зображення, на які спирається нейромережевий алгоритм для класифікації. Адже, як відомо, в реалізації комп'ютерного зору широко застосовуються системи класифікації зображення, які цілковите зображення відносять до окремого класу без виокремлення на ньому окремих об'єктів. Це – задача комп'ютерного зору. Для задач класифікації часто використовують заздалегідь навчені на тренувальних даних згорткові нейронні мережі. При цьому невирішеним залишається завдання, на які саме елементи спирається нейронна мережа для прийняття конкретного рішення. В статті представлено метод конкурентного процесу градієнтного спуску для виділення деталей (елементів), які в процесі класифікації були значимими, тобто ключовими. Цей метод передбачає конкуренцію процесу знищення деталей зображення з процесом збереження результатів класифікації. На прикладі самостійно навченої нейронної мережі автори проаналізували наявність деталей цифрового зображення, яке отримало класифікацію, на візуальну оцінку експертом наявності елементів безпосередньо об'єкта класифікації після застосування розробленого метода. Завдяки цьому методу результати деградації деталей цифрового зображення із збереженням якості класифікації, архітектура використаної нейронної мережі є довільною. Це дало змогу порівняти зони акценту уваги нейронних мереж різної архітектури: згорткової архітектури та архітектури мікшування. За результатами дослідження запропоновано метод локалізації уваги нейронної мережі з довільною архітектурою. Збережені елементи на деградованому зображенні можуть дати додаткову інформацію про валідність класифікації конкретною нейронною мережею. Це можна оцінити за локалізацією збережених елементів. Наприклад, приналежність елементів об'єкту класифікації свідчить про високу ймовірність правильної роботи нейронної мережі. Тобто якщо збережені елементи не належать об'єктові класифікації, то робиться висновок про відсутність репрезентативності даних для навчання (наприклад, один з об'єктів частіше зустрічається на характерному фоновому тлі). Проведене експериментальне дослідження показало переваги запропонованого методу порівняно з аналогами: досягається підвищення точності локалізації значущих деталей (елементів), забезпечується наявність інформації про глобальні значущі елементи цифрового зображення та їх форму, а також придатність методу для застосування для інших, окрім згорткових, типів нейронних мереж.

Ключові слова: комп'ютерний зір, нейронні мережі, локалізація уваги, зображення, якість даних