METHOD OF CONTROLLED IMAGE DEGRADATION FOR DETECTING ATTENTION CENTERS IN CLASSIFICATION NEURAL NETWORKS

Oleksandr Dorenskyi
Central Ukrainian National Technical University, PhD in Information Technology, Associate Professor, Cybersecurity and Software Department, dorenskyiop@kntu.kr.ua
Oleksandr Drieiev
Central Ukrainian National Technical University, PhD in Telecommunications Systems and Networks, Associate Professor, Cybersecurity and Software Department, drey.sanya@gmail.com
Hanna Drieieva
Central Ukrainian National Technical University, PhD in Computer Engineering, Associate Professor, Cybersecurity and Software Department, gannadreeva@gmail.com

The paper discusses a method of controlled image degradation for detecting attention centers in neural networks for classification tasks. A method is proposed that preserves classification logits during image degradation, using a uniform background and optimizing the 'transparency' map. This approach allows working with modern neural network architectures such as MIXER and ViT and improves training by identifying significant elements for classification.

Keywords: computer vision, attention localization, data quality.

Implementation of image classification systems in relatively low-powered computational systems often requires the use of simplified or custom-trained neural networks to meet the computational system's requirements. Additionally, the limited availability of training data with known classifications can have a significant impact. In such cases, a neural network trained on the available data may provide satisfactory results, but the features on the images, on which the network bases its classification decisions, remain unknown. The authors have conducted a study on existing methods for identifying attention centers in classification networks for specific examples, such as CAM, Grad-CAM, and similar methods. However,

most of these methods work only for convolutional neural network-based models, including CNN, MobileNET, ResNET, and VGG. The works reviewed by the authors leave open questions for more modern architectures, such as MIXER and ViT. Therefore, the authors propose a process of controlled image degradation with the preservation of the original classification logits before the softmax activation function.

The proposed controlled degradation process for the input image is based on constants: the input image, the neural network, and a uniform background derived from the input image with an average color, where the "transparency" map between the input image and the uniform background is optimized. The mixing of the input image is represented by a set of coefficients pi,j ranging from 0 to 1, which indicates how much of the original image should be taken, while (1-pi,j) represents the background. The loss function takes into account: the total sum of pi,j values; the presence of sharp transitions between neighboring pi,j, which may otherwise lead to the generation of details not present in the input image; the difference between the original classification logits of the image and the current logits resulting from the classification of the degraded image. This process is illustrated in the following figure (Fig. 1).

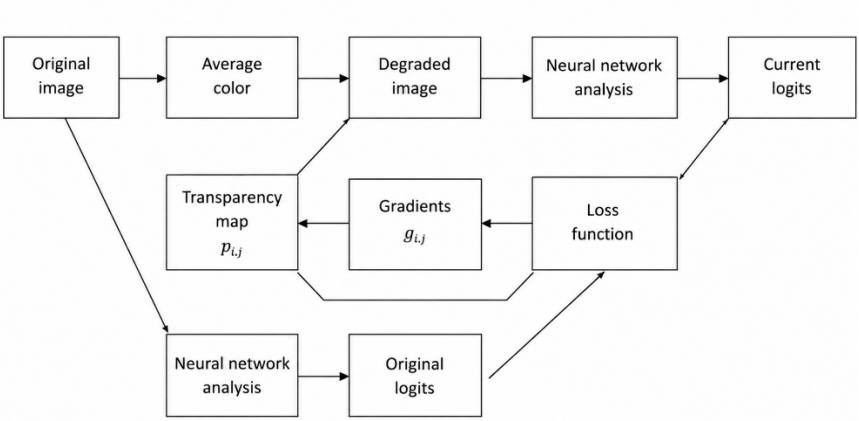
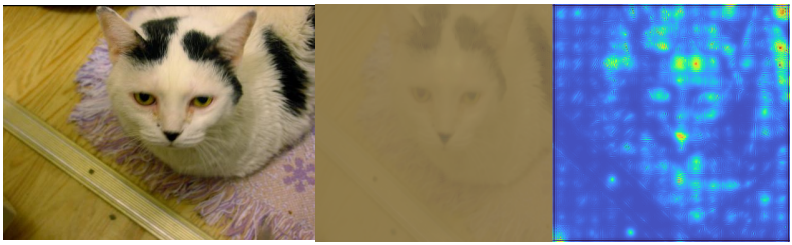


Figure 1. Diagram of Information Flow in the Image Degradation Process

Figure 2 shows the result of applying the proposed image degradation method to the ViT network.



a) *b)* *c)*

Figure 2. Result of Image Degradation for the ViT Network: a) original image, b) вegraded image with preserved classification result and network confidence level, c) Normalized 'transparency' map

As shown in the example in Figure 2, the neural network focuses on the animal's eyes and fur color. A drawback is the non-zero weight given to the floor texture. Therefore, it can be concluded that for better network training, it is necessary to diversify and balance the background in the training data.

The use of the controlled image degradation method for detecting attention centers in neural networks holds significant potential in improving the image classification process. The proposed combination of transparency map optimization and classification logits preservation allows for identifying key elements on which the network's decisions are based and works effectively with neural networks of various architectures, including modern models. Improving the quality of training data, particularly by diversifying backgrounds, is crucial as it will enhance the accuracy and stability of classification while minimizing the impact of background elements. Therefore, the proposed method can be an important tool for further research and applications in the field of computer vision and image classification.

References

1. Rama, J., Nalini, C., & Kumaravel, A. (2019). Image pre-processing: Enhance the performance of medical image classification using various data augmentation techniques. ACCENTS Transactions on Image Processing and Computer Vision, 5(14), 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. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. ISPRS Open Journal of Photogrammetry and Remote Sensing, 5, 100018. https://doi.org/10.1016/j.ophoto.2022.100018.

3. Drieiev, O. M., Dorenskyi, O. P., & Drieieva, G. M. (2022). Neural network method for detecting textural anomalies in digital images. Central Ukrainian Scientific Bulletin. Technical Sciences, 5(36), Part 2, 335-346. 10.32515/2664-262X.2022.5(36).2.335-346.

4. Dorenskyi, O. P., Drieiev, O. M., & Mynailenko, R. M. (2023). Method for determining features on which a neural network makes classification decisions. Information Security and Computer Technologies, 55. Retrieved from https://dspace.kntu.kr.ua/items/519383c1-1857-4761-8868-ef3b9bacf292.

5. Dorenskyi, O., Drieiev, O., & Drieieva, H. (2026). The method of identifying key elements of a digital image in the decision-making process of classification by a neural network. Computer Systems and Information Technologies, (1), 145–155. https://doi.org/10.31891/csit-2026-1-14.