

# Information Technologies

## Информационные технологии


Research article

DOI: <https://doi.org/10.18721/JCSTCS.17105>


UDC 681.3.05



### THE STUDY OF THE VISION TRANSFORMER ARCHITECTURE BY EXPLAINABILITY METHODS

*I.A. Utkin, V.V. Shkuropatsky*  ,  
*A.N. Pronikov, E.S. Rakov*

Mozhaisky Military Space Academy,  
St. Petersburg, Russian Federation

 [vitalius-47@mail.ru](mailto:vitalius-47@mail.ru)

**Abstract.** The article discusses issues of explainability of the operating principles of a machine learning model. As the architecture of the model, one of the types of transformer is considered, the task of which is to classify images based on the popular “ImageNet-1000” dataset. This type of transformer is also called vision transformer and can serve either as a standalone model or as part of a more complex architecture. The explainability methods included activation maps of classes, which were calculated by applying algorithms based on forward and backward propagation of image tensors through the components of the transformer: multi-head attention layers and fully connected multilayer networks. The aim of the work is to increase the explainability of the internal processes of the functioning of the vision transformer by analyzing the obtained activation maps and calculating a metric to evaluate their explainability. The results of the study reveal patterns that reflect the mechanisms of operation of the vision transformer in solving the image classification problem, as well as evaluating the importance of the identified classification features through the use of the explainability metric.

**Keywords:** machine learning model, explainability, visual transformer, encoder, attention mechanism, class activation maps, back propagation activation maps

**Citation:** Utkin L.A., Shkuropatsky V.V., Pronikov A.N., Rakov E.S. The study of the vision transformer architecture by explainability methods. Computing, Telecommunications and Control, 2024, Vol. 17, No. 1, Pp. 54–64. DOI: 10.18721/JCSTCS.17105

Научная статья

DOI: <https://doi.org/10.18721/JCSTCS.17105>

УДК 681.3.05



## ИССЛЕДОВАНИЕ АРХИТЕКТУРЫ ВИЗУАЛЬНОГО ТРАНСФОРМЕРА МЕТОДАМИ ОБЪЯСНИМОСТИ

*И.А. Уткин, В.В. Шкуропатский <sup>✉</sup>,  
А.Н. Проников, Е.С. Раков*

Военно-космическая академия имени А.Ф. Можайского,  
Санкт-Петербург, Российская Федерация

<sup>✉</sup> [vitalius-47@mail.ru](mailto:vitalius-47@mail.ru)

**Аннотация.** В статье рассматриваются вопросы объяснимости принципов функционирования модели машинного обучения. В качестве архитектуры модели рассмотрен один из видов трансформера, задача которого состоит в классификации изображений на базе популярного датасета «ImageNet-1000». Данный тип трансформера также называется визуальным трансформером и может служить, как отдельной моделью, так и составляющей более сложной архитектуры. Методами объяснимости являлись карты активации классов, которые рассчитывались посредством применения алгоритмов на основе прямого и обратного распространения тензоров изображения через составные части трансформера: слой механизма внимания и полносвязанные многослойные сети. Цель работы состоит в повышении объяснимости внутренних процессов функционирования визуального трансформера за счет анализа полученных карт активации и расчета метрики оценивания их объяснимости. Результатом работы являются закономерности, отражающие механизмы работы визуального трансформера при решении задачи классификации изображения, а также оценивание степени важности выделяемых признаков классификации за счет применения метрики объяснимости.

**Ключевые слова:** модель машинного обучения, объяснимость, визуальный трансформер, энкодер, механизм внимания, карты активации классов, карты активации обратного распространения

**Для цитирования:** Utkin L.A., Shkuropatsky V.V., Pronikov A.N., Rakov E.S. The study of the vision transformer architecture by explainability methods // Computing, Telecommunications and Control. 2024. Т. 17, № 1. С. 54–64. DOI: 10.18721/JCSTCS.17105

### Introduction

Technical solutions based on the architecture of various transformers are well established in many areas of science and technology. Today, this technology allows solving a wide range of problems: from object recognition to generating images and texts. Transformers have proven to be particularly effective in the field of natural language processing [2], which allowed a significant scientific leap with the development of large language models. However, in addition to natural language processing, similar architectures are also used in image classification. One of these models that allows solving the classification problem is the Vision Transformer (ViT) [3]. The architecture of the vision transformer is an encoder with 12 layers of multi-head attention [1] and a fully connected multilayer perceptron at the output (Fig. 1).

This diagram has a simpler structure than the one of the vanilla transformer model [10, 11]; in particular, there is no decoder unit. This and the fact that the input data are images, which are easier to visualize than, for example, text tokens, makes the vision transformer a good ‘candidate’ for studying explainability of the results of the functioning of models based on such solutions.

To date, existing explainability methods and algorithms allow revealing some aspects of the internal functioning of machine learning models based on various architectures. Many explainability approaches

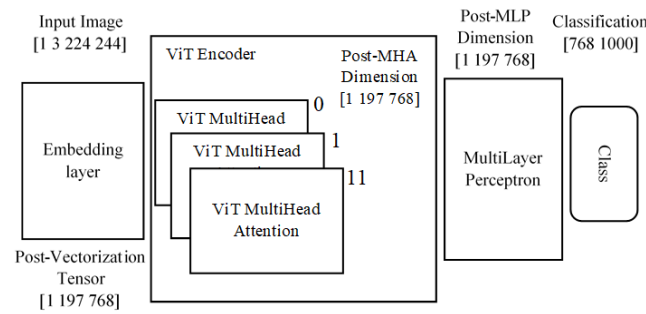


Fig. 1. ViT Architecture Diagram

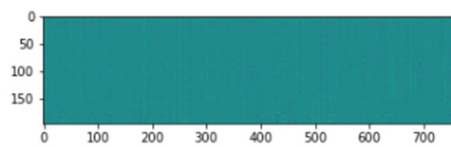


Fig. 2. Image view after passing through the embedding layer (1x197x768)

for transformer-based machine learning models are designed for large language models [16]. However, one explainability method for machine vision models is to construct class activation maps for images to identify areas with key features. In particular, this approach gives an explainable result for convolutional neural networks [5]. In addition to generating activation maps for direct passage of images, gradient-based methods [4] and their various modifications have been developed [17, 18], which take into account changes in weights when training machine learning models.

The study of a vision transformer with these explainability methods allow to partially reveal the mechanisms of its functioning and understand what elements of the image the model pays attention to during classification. Quantitative evaluation of how well certain activation maps display internal processes during classification was carried out by calculating explainability metrics.

The structure of the vision transformer allows to trace the image from the initial to the final layers and at each stage to track the changes occurring to it. When a classified image passes through the model, its dimension changes from 1x3x224x224 to 1x197x768 and then invariably spreads through all layers to the classification layer, where it expands or contracts depending on the number of classes.

The first way to analyze the principles of functioning of a vision transformer is to directly pass the image through its main layers and then restore to the original dimension, similar to models based on the convolutional neural network [15].

### Direct image passage through the layers of the model

The first layer of the model is the embedding layer, which vectorizes the input image and adds positional encoding to it. Since the input image is a three-dimensional tensor, it needs to be divided into smaller patches with a further vector representation, resulting in a dimension of 1x197x768. The image after passing through the embedding layer is shown in Fig. 2.

The vector appearance of the tensor makes it impossible to visually evaluate further processes taking place in the layers of the transformer. To visualize the results, the original picture dimension to 3x224x224 should be restored. This is achieved through a series of matrix transformations over the resulting tensor, where first the positional coding vector is removed, then the tensor is converted to 6-dimensional form (1x14x3x16x16) with further rearrangement and dimension change. The software implementation of the current transformations is presented in [6].

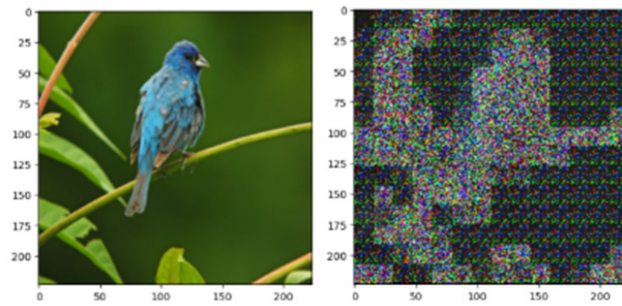


Fig. 3. Original image is restored after the embedding layer.  
On the left is the original image coming to the model input. On the right is the restored image

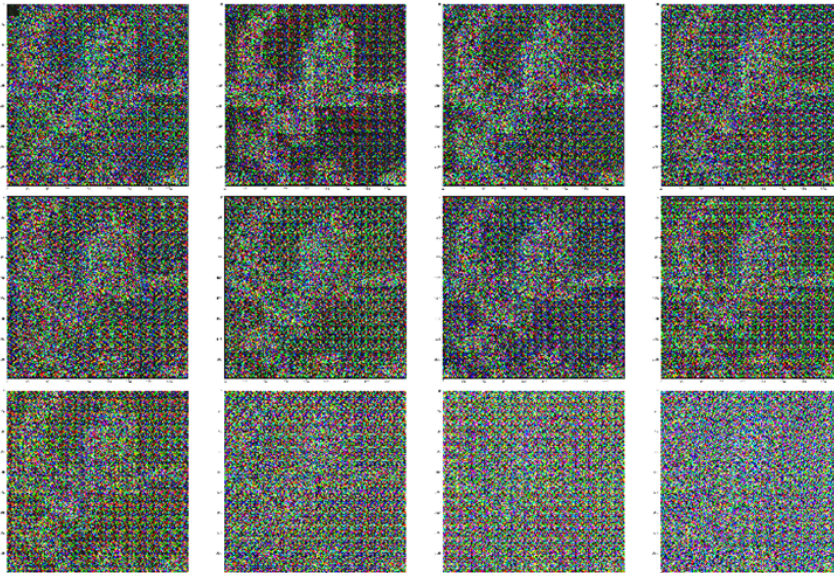


Fig. 4. Image after passing through the attention mechanism

The restored image is shown in Fig. 3.

Fig. 3 shows the structure of the original image with the loss of color and local features. After the embedding layer, the image passes through 12 multi-head attention layers [14], each of which is trained to identify image features (Fig. 4).

Fig. 4 shows the restored images after the attention mechanism. The weights of each multi-head attention layer are adjusted to separate their context from the vector view of the image, which qualitatively improves the ability of the model to classify. After multi-head attention layers, the generalized tensor enters the input of a fully connected neural network, alternating layer normalization and dropout-type regulation methods. The final dimension of the output layer of the model is 768x1000, where 1000 corresponds to the number of dataset classes (in this case ImageNet-1000 is considered).

#### Model activation maps calculation

In addition to the direct passage of the image with restoration, an analogue of activation maps for the transformer was obtained. The principle of the algorithm is based on the calculation of activation maps for convolutional neural networks according to the following expression:

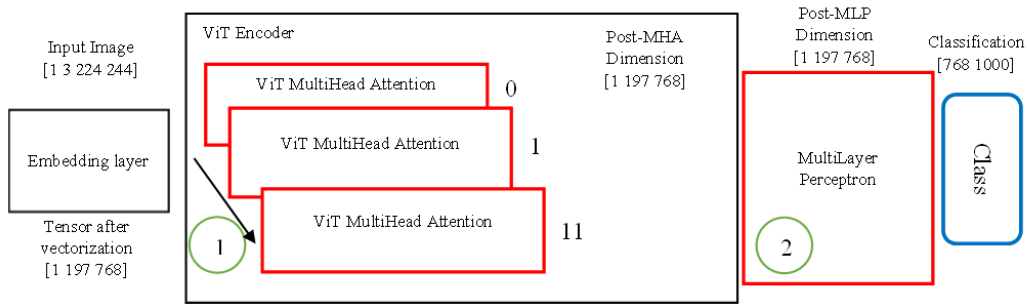


Fig. 5. Layers relative to which the activation maps were calculated are highlighted in red

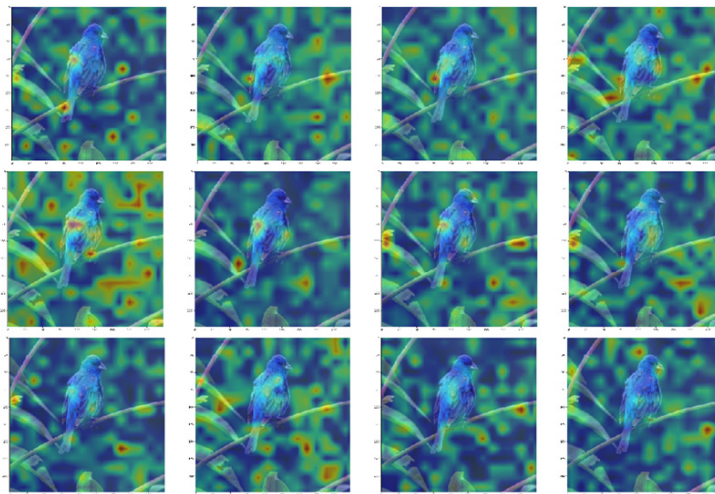


Fig. 6. ViT multihead attention activation maps

$$y_i^{class} = \sum_j w_{ij}^{class} * VitEncTensor_{ij}, \quad (1)$$

where  $y_i^{class}$  – class activation card, dimension  $3 \times 16 \times 16$ ;  $w_{ij}^{class}$  – weights from the last layer of direct distribution activating the maximum value in the classification layer;  $VitEncTensor_{ij}$  – encoder output tensor.

The consistency of the tensor dimension when passing through the layers of the model also allows activation maps to be calculated separately for each attention mechanism. Calculation expression is as follows:

$$y_i^{class^k} = \sum_j w_{ij}^{class} * MHA_{ij}^k, \quad (2)$$

where  $y_i^{class^k}$  – attention mechanism class activation map;  $w_{ij}^{class}$  – weights from the last layer of direct distribution activating the maximum value in the classification layer;  $MHA_{ij}^k$  – multi-head attention layers output tensor.

To perceive the obtained formulas on the general diagram of the vision transformer model more clearly, the layers used are highlighted in color, where the red color indicates the layers relative to which the output class layer is calculated (blue color) (Fig. 5).

Output tensors from layers of the attention mechanism (Fig. 6) were used as the first activation maps. The software implementation of the current transformations is presented in [6].

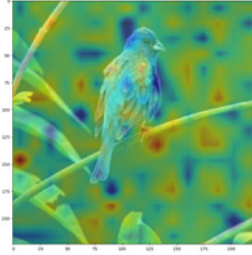


Fig. 7. Activation map of the penultimate layer and classification layer

The resulting activation maps in each layer have unique features, but do not have such a pronounced appearance as in the case of convolution neural networks [5], where strictly defined areas of features in the image were identified.

The activation map of the penultimate layer and classification layer is shown in Fig. 7.

Fig. 7 has a similar structure of a more vague nature. The resulting activation maps may indicate that the designer is actually looking for features across the entire image, without any specific areas of the feature space.

#### Calculation of back propagation activation maps

Another method of studying the explainability of transformer-based classification models is calculating of back propagation activation maps [5], that is, calculating the gradient relative to the selected layers in Fig. 5.

Back propagation activation maps were calculated using the following expressions:

$$L_i^{class} = GELU \left( \sum_j VitEncTensor_{ij} * \frac{dy^{class}}{d(VitEncTensor_{ij})} \right), \quad (3)$$

where  $VitEncTensor_{ij}$  – encoder output tensor;  $L_i^{class}$  – linear combination of weight coefficient and post-activation channels, dimensions 14x14;  $\frac{dy^{class}}{d(VitEncTensor_{ij})}$  – transformer output layer gradient as related to the encoder output tensor.

As in the case of activation maps for intermediate layers, a back propagation through the attention mechanism layers was calculated using the following expression:

$$L_i^{class^k} = GELU \left( \sum_j MHA_{ij}^k * \frac{dy^{class^k}}{d(MHA)_{ij}} \right), \quad (4)$$

where  $MHA_{ij}$  – multi-head attention tensors;  $L_i^{class^k}$  – linear combination of weight coefficient and post-activation channels;  $\frac{dy^{class^k}}{d(MHA)_{ij}}$  – transformer output layer gradient as related to the influence mechanism tensor.

The software implementation of the current algorithms is presented in [6].

According to (4) the following gradient images were obtained for the case of the multi-head attention layers (Fig. 8).

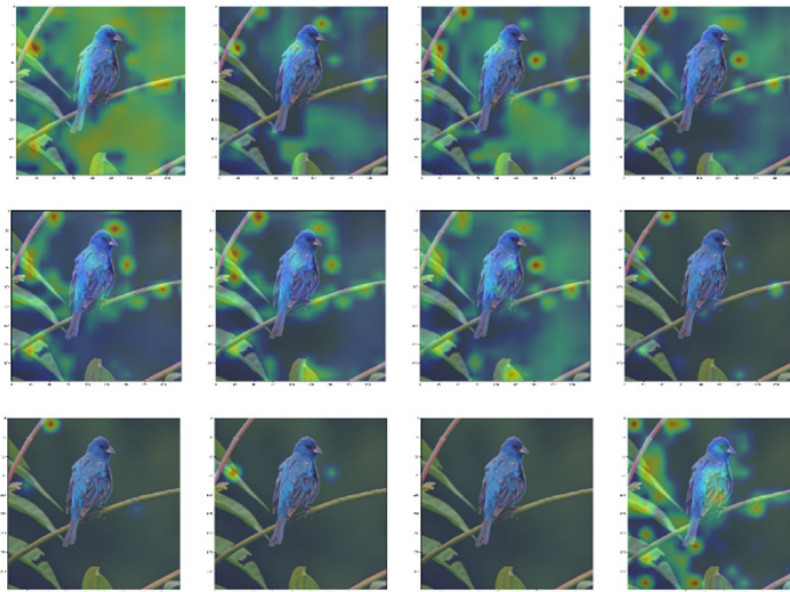


Fig. 8. Back propagation calculation for the case of the attention mechanism layers

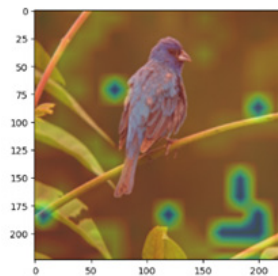


Fig. 9. Gradient calculation relative to the encoder output tensor

The obtained gradients on different layers of the attention mechanism indicate that each individual layer of attention allocates certain features on the image, for example, the background of the picture, some objects, etc.

The calculation of the gradient relative to the output tensor of the encoder according to the expression (3) allowed to obtain the following figure (Fig. 9).

The applied approach of calculating the gradient of the output class relative to the encoder tensor does not explicitly identify the features of the image.

#### **Metrics for evaluating the explainability of a transformer**

As was shown, the obtained transformer activation maps do not explicitly identify features of the image classification, and therefore their significance was evaluated using the explainability metrics.

The method for calculating metrics depends on the type of problem being solved, as well as the explainability technique used. Calculations of activation maps for the transformer were used as explainability methods. In turn, explainability evaluation metrics are numerical calculations based on derived expressions [7] or more visual implementations based on the removal of image patches by painting them in a certain color [8, 9].

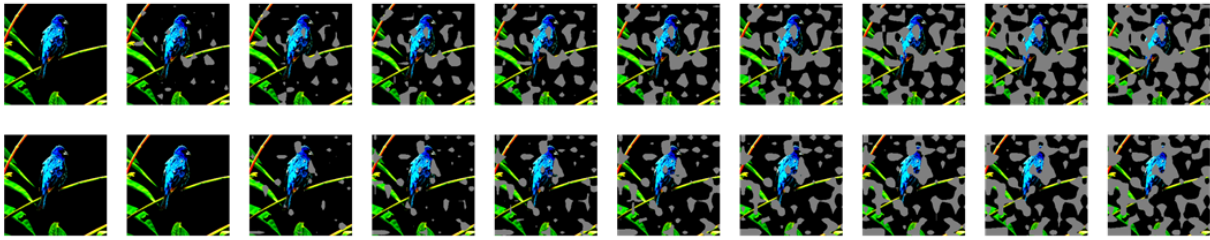


Fig. 10. Removing areas based on MoRF/LeRF evaluation metric.  
Top row refers to MoRF, bottom row refers to LeRF

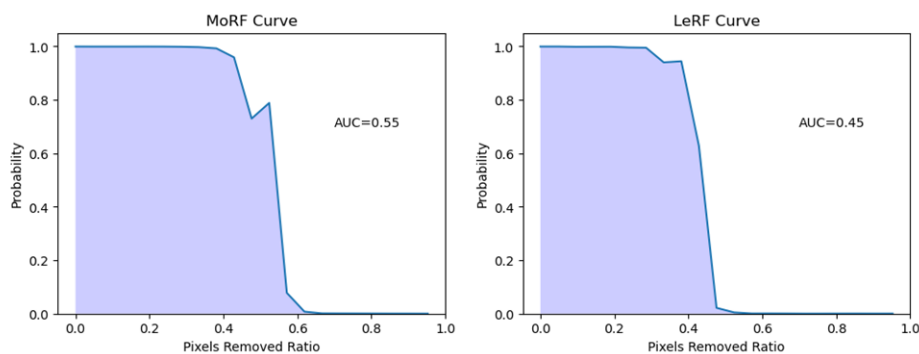


Fig. 11. Graphs showing the dependence of the probability of correct classification and removed elements relative to the whole image

One of the evaluation metrics associated with the removal of some information from the image is the algorithm MoRF/LeRF (most relevant first/least relevant first) [8]. It is based on the coloring in RGB colors (127, 127, 127) the most/least significant parts of the image according to the calculated activation maps and their further submission to the input of the transformer in order to obtain the probability of belonging to the target class.

The first 10 images for explainability based on (2) are shown in Fig. 10. A total amount of 20 images with removed areas were obtained.

The colored areas reflect the most/least important image patches with their accumulation. At the next stage of calculation, graphs of the dependence of the probability of correct classification and removed elements relative to the whole image were constructed (Fig. 11).

These graphs reflect the fact that the probability of correct classification decreases only when half of the image is removed. This is determined by the area below the curve (AUC – average under curve) which corresponds to 0.55 and 0.45 for MoRF and LeRF. Moreover, the MoRF graph decreases more slowly than the LeRF graph, which characterizes the independence of the activation map results from the features selected by the model, since the removal of more important patches affects the probability less than less important.

Similarly, the explainability metrics of MoRF/LeRF for the back propagation activation map based on the calculation of the output class gradient relative to the encoder tensor (Fig. 12, 13) have been calculated.

The AUC values for MoRF and LeRF graphs were 0.38 and 0.64, respectively. These values indicate that there is weak explainability basis for the use of back propagation activation maps. For the MoRF metric, the probability drops instantly after removing one third of the features, which may coincide with explainability evaluation metrics for images with many small features.

The software implementation of the current algorithms is presented in [6].



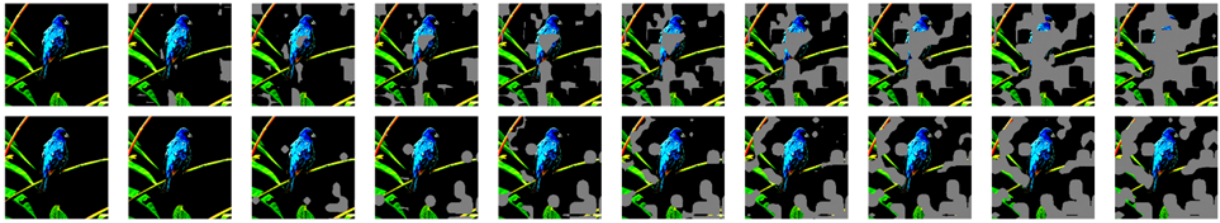


Fig. 12. Images with removed areas. The upper row refers to MoRF, the lower row refers to LeRF

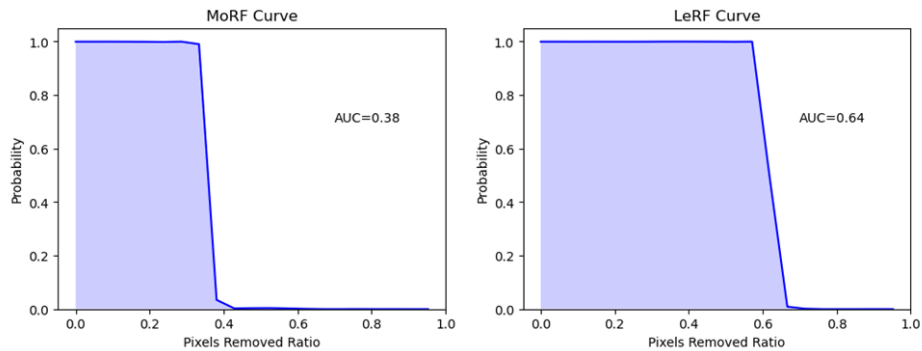


Fig. 13. Graphs showing the dependence of the probability of correct classification and removed elements relative to the whole image for the back propagation activation maps

### Conclusion

The principle of the transformer functioning, specifically its particular implementation, differs significantly from previous technologies used to solve the problem of image classification. Almost all layers of the image pass in a constant dimension, which, on the one hand, simplifies attempts to explain the transformer functioning, on the other hand, due to preliminary vectorization, complicates the process of analyzing its direct and reverse passage through the model.

Vectorization at the stage of passing the embedding layer significantly distorts the structure of the image and after its restoration only the main informative features are visible. Further passage through the layers of attention mechanism made it possible to see how the model selects certain features, then transferring them to the fully connected layers of the neural network.

The use of algorithms similar to the construction of activation maps, as in the case of convolutional neural networks, does not allow to unambiguously indicate the areas of features that the model turns to when classifying an image. The constant dimension when passing through the vision transformer made it possible to evaluate separately the output tensors from the encoder and layers of attention mechanism.

The algorithm based on the reverse passage or gradient calculation partially specified the different areas of features that the model indicates in the influence mechanisms. However, when considering gradients relative to the output tensor of the encoder, no obvious dependencies were established.

The calculated values of the MoRF/LeRF evaluation metric for two types of activation maps poorly characterized the significance of the features identified by these techniques. In the case of the activation map obtained from (2), the metric showed no distinguishing features detected by this explainability method, as well as the inverse AUC values of MoRF/LeRF. However, the values of the evaluation metric of the expression-based explainability technique [4] reflect more/less important features used by the model for correct classification (AUC for MoRF/LeRF is 0.38 and 0.64 respectively).

## REFERENCES

1. **Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I.** Attention is all you need. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010. DOI: 10.5555/3295222.3295349
2. OpenAI (2023), Available: <https://cdn.openai.com/papers/gpt-4.pdf> (accessed: 04.10.2023).
3. **Dosovitskiy A., Beyer L., Kolesnikov A., et al.** An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale. arXiv:2010.11929, 2021. DOI: 10.48550/arXiv.2010.11929
4. **Selvaraju R.R., et al.** Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision, 2019, Vol. 128, No. 2, pp. 336–359. DOI: 10.1007/s11263-019-01228-7
5. **Zhou B., et al.** Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016. DOI: 10.1109/cvpr.2016.319
6. Repository with code snippet, Available: <https://github.com/ewanytken/moduleInterpret> (accessed: 26.11.23).
7. **Yeh C.-K., Hsieh C.-Y., Suggala A.S., Inouye D.I., Ravikumar P.** On the (In)fidelity and Sensitivity of Explanations. arXiv:1901.09392, 2019. DOI: 10.48550/arXiv.1901.09392
8. **Samek W., et al.** Evaluating the visualization of what a Deep Neural Network has learned. IEEE Transactions on Neural Networks and Learning Systems, Vol. 28, No. 11, 2017, pp. 2660–2673. DOI: 10.1109/tnnls.2016.2599820
9. **Petsiuk V., Das A., Saenko K.** RISE: Randomized Input Sampling for Explanation of Black-box Models. arXiv:1806.07421, 2018. DOI: 10.48550/arXiv.1806.07421
10. **Radford A., Narasimhan K., Salimans T., Sutskever I.** Improving language understanding by generative pre-training, Available: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (accessed: 05.10.2023).
11. **Devlin J., et al.** Bert: Pretraining of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, 2019. DOI: 10.18653/v1/n19-1423
12. **Radford A., Kim J.W., Hallacy C., et al.** Learning Transferable Visual Models from Natural Language Supervision. arXiv:2103.00020, 2021. DOI: 10.48550/arXiv.2103.00020
13. **Wang J., Yang Z., Hu X., et al.** GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv:2205.14100, 2022. DOI: 10.48550/arXiv.2205.14100
14. **Cordonnier J.-B., Loukas A., Jaggi M.** Multi-Head Attention: Collaborate Instead of Concatenate. arXiv:2006.16362, 2021. DOI: 10.48550/arXiv.2006.16362
15. **Zeiler M.D., Fergus R.** Visualizing and Understanding Convolutional Networks. Lecture Notes in Computer Science, 2014, Vol. 8689, pp. 818–833. DOI: 10.1007/978-3-319-10590-1\_53
16. **Wu X., Zhao H., Zhu Y., et al.** Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era. arXiv:2403.08946, 2024. DOI: 10.48550/arXiv.2403.08946
17. **Smilkov D., Thorat N., Kim B., Viégas F., Wattenberg M.** SmoothGrad: removing noise by adding noise. arXiv:1706.03825, 2017. DOI: 10.48550/arXiv.1706.03825
18. **Chattopadhyay A., et al.** Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018. DOI: 10.1109/wacv.2018.00097

## INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

**Utkin Ivan A.**

**Уткин Иван Алексеевич**

E-mail: ewanytken@mail.ru

**Shkurovatsky Vitaly V.**

**Шкуропатский Виталий Владимирович**

E-mail: vitalius-47@mail.ru

**Pronikov Alexander N.**

**Проников Александр Николаевич**

**Rakov Evgeniy S.**

**Раков Евгений Сергеевич**

E-mail: djon.rus31@mail.ru

*Submitted: 20.03.2024; Approved: 04.05.2024; Accepted: 08.05.2024.*

*Поступила: 20.03.2024; Одобрена: 04.05.2024; Принята: 08.05.2024.*