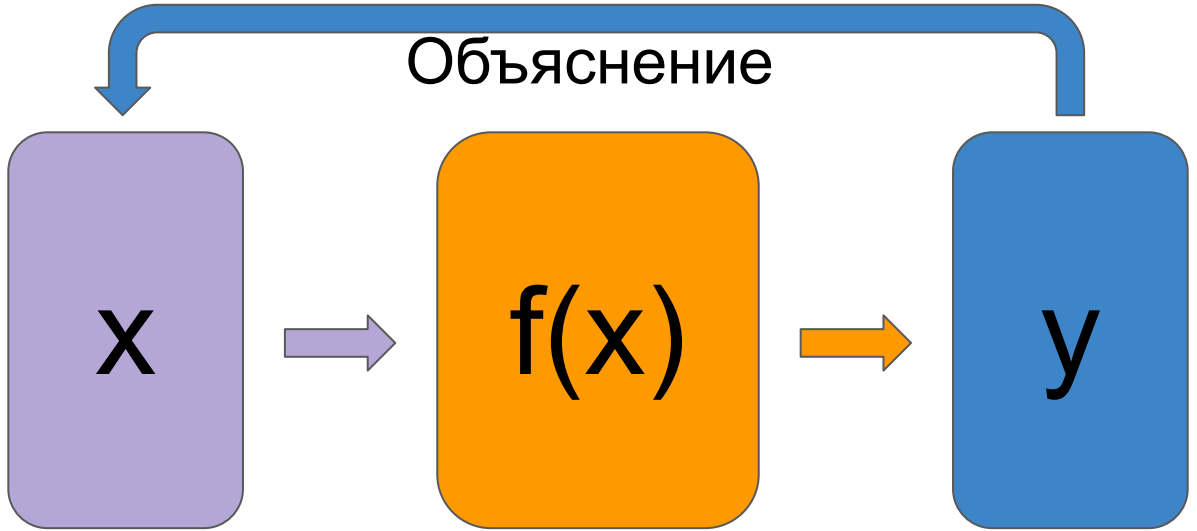
Модификация визуального трансформера методом b-cos для повышения интерпретируемости в задаче классификации изображений дерматоскопии.

Лукьянов Андрей Николаевич,
Волков Егор Николаевич,
Ярушев Сергей Александрович,
Аверкин Алексей Николаевич

Центр перспективных исследований в искусственном интеллекте,
РЭУ имени Г. В. Плеханова.

Объяснительный искусственный интеллект, b-cos сети.

Главная задача XAI в нейронный сетях - позволить человеку интерпретировать работу черного ящика.



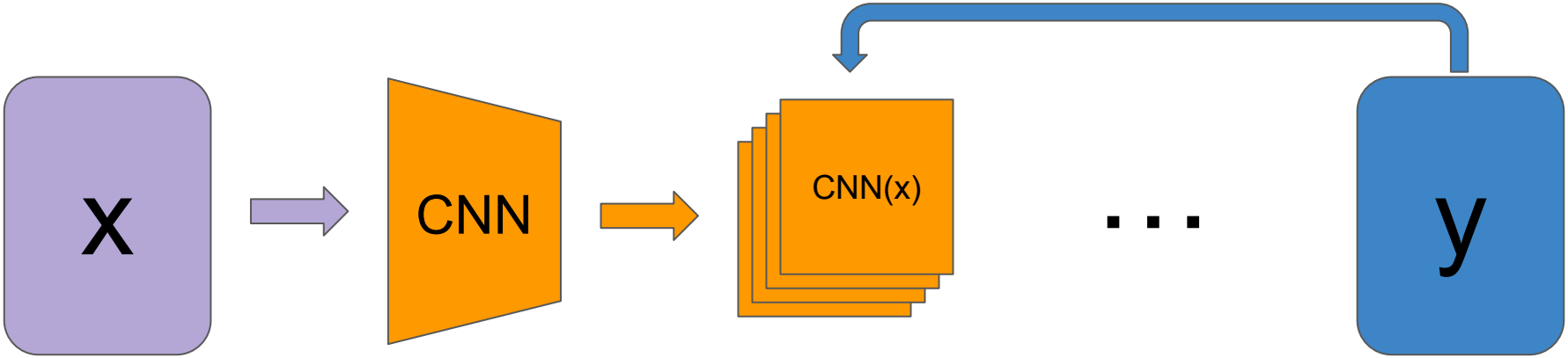
Объяснительный искусственный интеллект, b-cos сети.

В случае изображений объяснениями служат взвешенные последние карты признаков сети. Для сверточных нейронных сетей наиболее известным методом их получения является CAM (GradCam, HiResCam).

$$grad_{ck} = \frac{\partial y_c}{\partial A_k}$$

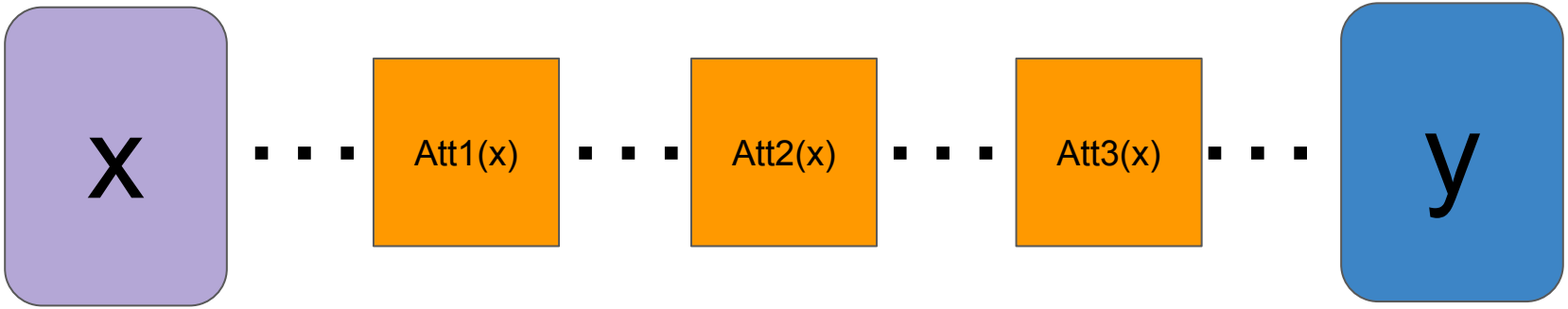
$$\alpha_{ck} = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} grad_{ck\,ij}$$

$$Expl_c = ReLU\left(\sum_k \alpha_{ck} \cdot grad_{ck} \right)$$



Объяснительный искусственный интеллект, b-cos сети.

Трансформерные модели вычисляют
попарные карты внимания внутри себя.



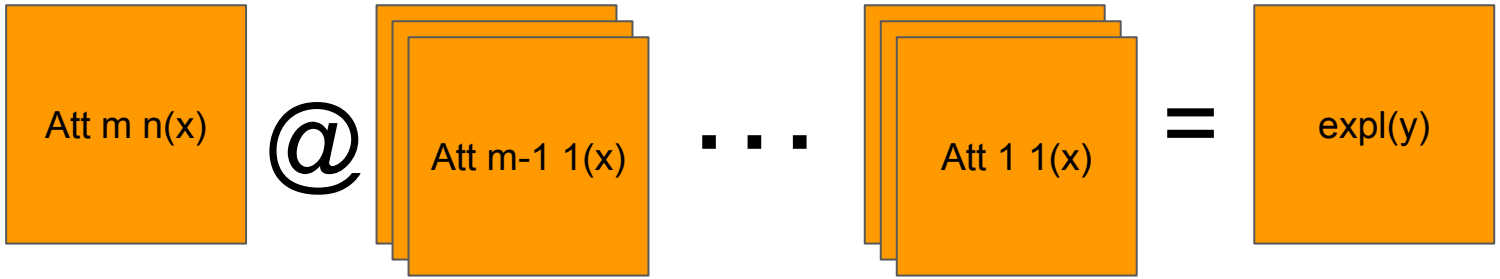
Объяснительный искусственный интеллект, b-cos сети.

Тогда, если решение сети определяется последним элементом последовательности, то можно использовать Attention Rollout.

n - длина последовательности;

m - число блоков внимания.

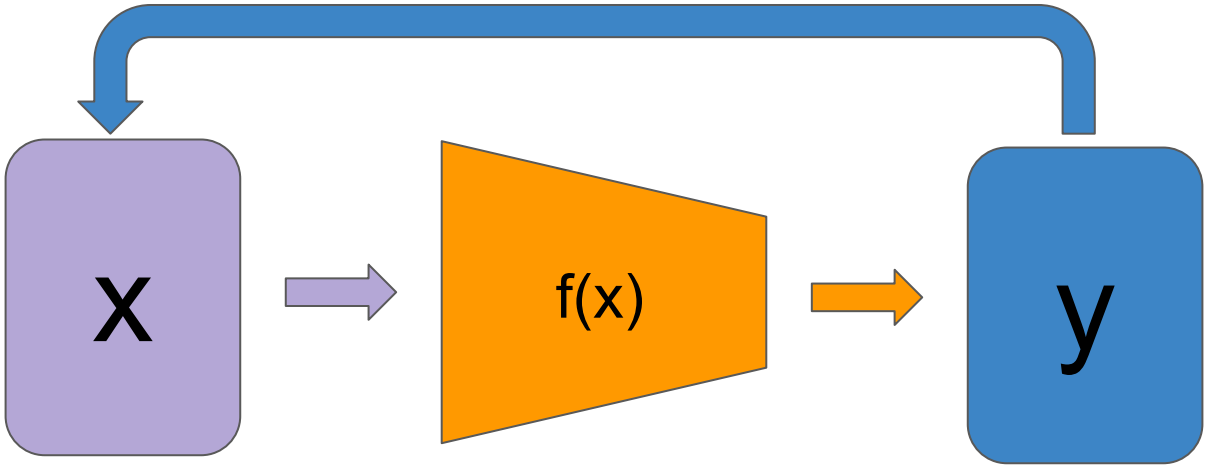
$$Att_1(x) \in R^{n \cdot n}$$
$$Att_2(x) \in R^{n \cdot n}$$
$$\dots$$
$$Att_m(x) \in R^{n \cdot n}$$
$$y = classifier(x_n) \implies$$
$$expl(y) = Att_{mn} Att_{m-1} \dots Att_1$$



Объяснительный искусственный интеллект, b-cos сети.

B-cos метод позволяет нам представить всю сеть как череду динамически линейных трансформаций, и таким образом суммаризировать вычисления единой матрицей.

$$Explain\left(\frac{\partial y}{\partial x}\right)$$



Объяснительный искусственный интеллект, b-cos сети.

Для представления нейронной сети чередой матричных умножений кусочно-линейные функции активации с двумя значениями могут быть представлены бинарными матрицами:

$$ReLU(x) = \begin{pmatrix} w_{11} & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & w_{nm} \end{pmatrix} x$$

$$w_{ij} \in \{0, 1\}$$

Объяснительный искусственный интеллект, b-cos сети.

$$f(x) = \tilde{W}_n \dots \tilde{W}_2 \tilde{W}_1 x$$

$$\tilde{y} = f(x)$$

$$grad_c = \frac{\partial \tilde{y}_c}{\partial x}$$

$$\dots$$

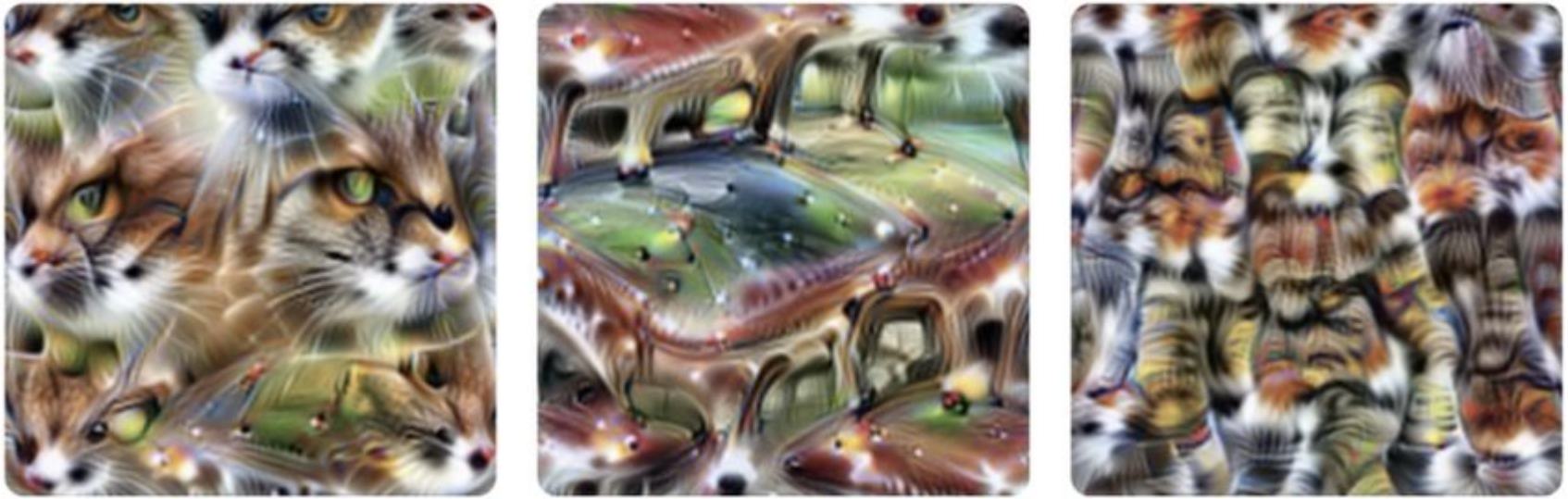
$$expl = (rgb_1, rgb_2, rgb_3, \alpha)$$

grad с 1-3 - объяснение для красного/синего/зеленого каналов

alpha - коэффициент прозрачности, скрывает те пиксели, которые внесли малый или отрицательный вклад в решение.

Полисемантика нейронов и слоев.

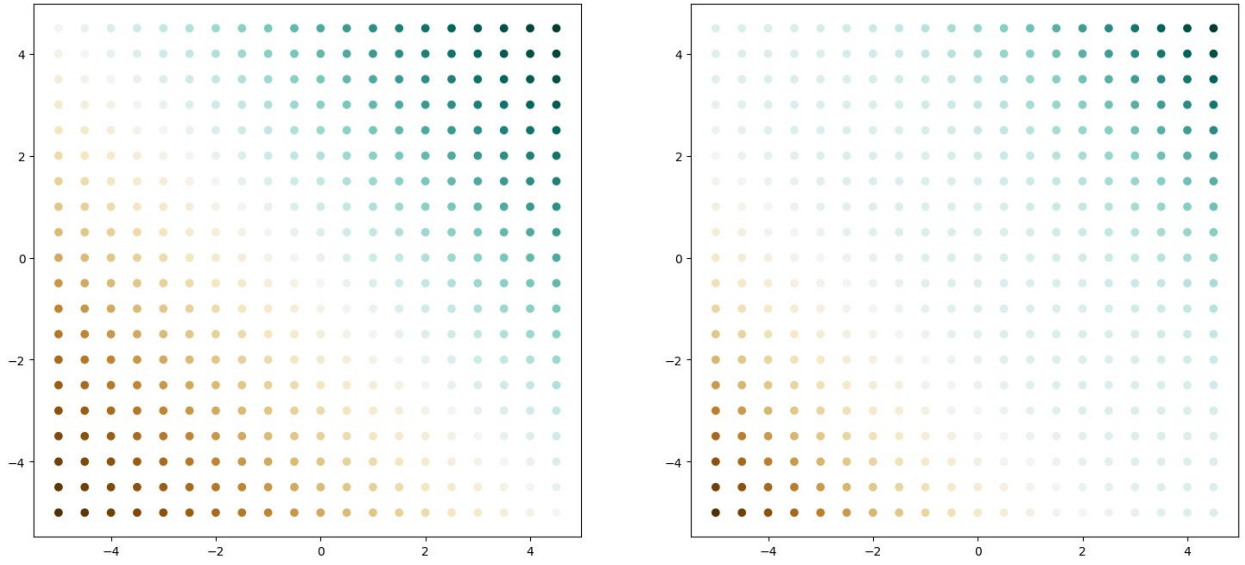
Последние исследования в сфере интерпретируемости нейронных сетей показывают, что для выполнения сложных задач нейроны обучаются **реагировать на несколько непересекающихся событий**.



Изображение взято из “Zoom In: An Introduction to Circuits” Olah et al.

Полисемантика нейронов и слоев.

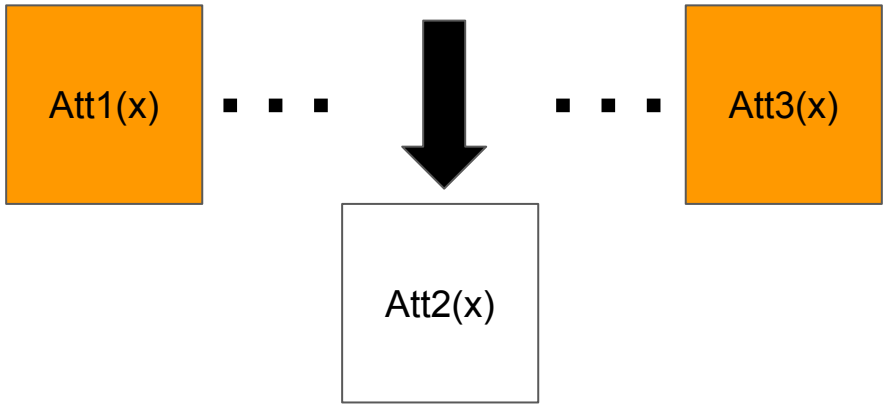
Мы предполагаем, что интерпретируемость b-cos сетей исходит благодаря наложению ограничений на диапазон возбуждения нейронов, что мешает развитию полисемантики.



Полисемантика нейронов и слоев.

Также большой проблемой интерпретируемости является коадаптация голов внимания, ибо они могут научиться подавлять работу друг друга.

Мы предлагаем прореживать половину голов во время обучения для избежания этого феномена.



Заключение и предварительные результаты.

Имея такого рода “распутанную” архитектуру мы стремимся изучить признаки которым обучается визуальный траснформер анализом нейронов перед классификацией.

Одно из направление исследования это использование описательных моделей для автоматической интерпретации полученных карт, а также обучение изучение эволюции полученных признаков и их взаимодействия.

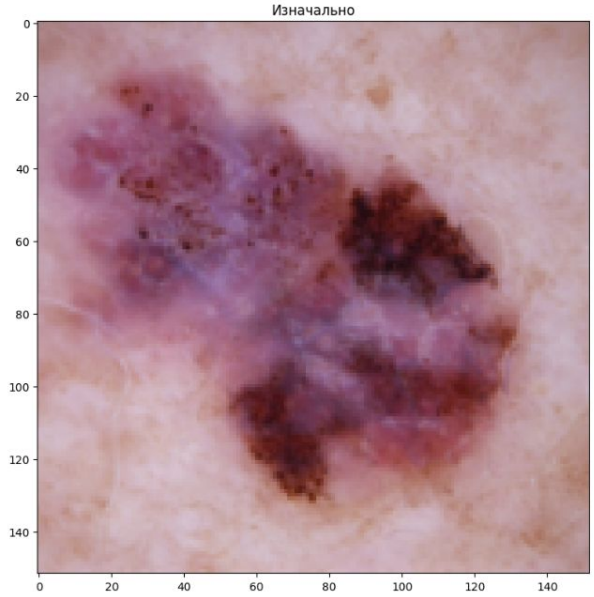
Исследование проведено за счет Российского научного фонда (грант № 22-71-10112)

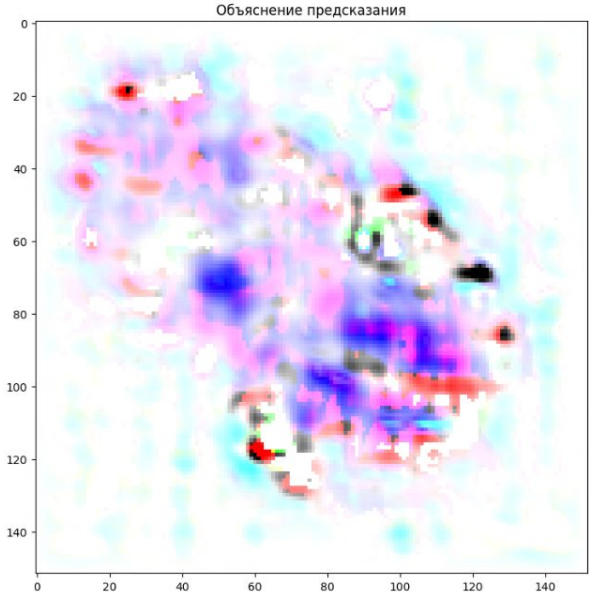
https://rscf.ru/project/22-71-10112

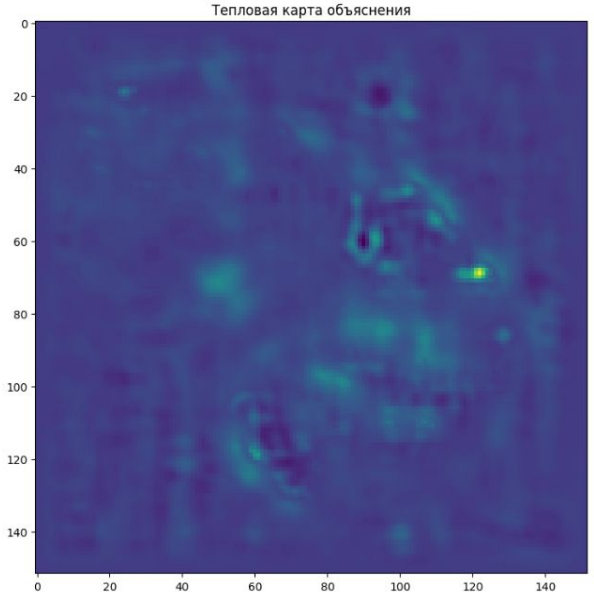
Истина: меланома

Предсказания:

меланома	[объяснение]	99.40%
меланоцитарные невусы		0.50%
базально-клеточная карцинома		0.00%
Актинические кератозы и внутриэпителиальная карцинома		0.00%
доброкачественные образования, похожие на кератоз		0.00%
дерматофиброма		0.00%
сосудистые поражения		0.00%



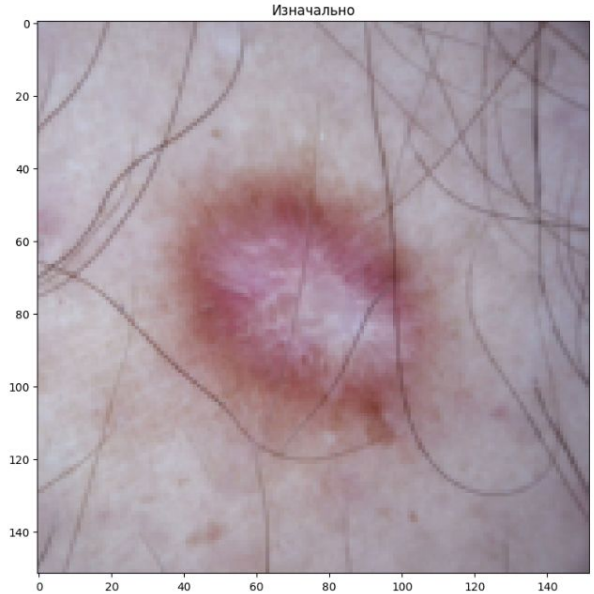


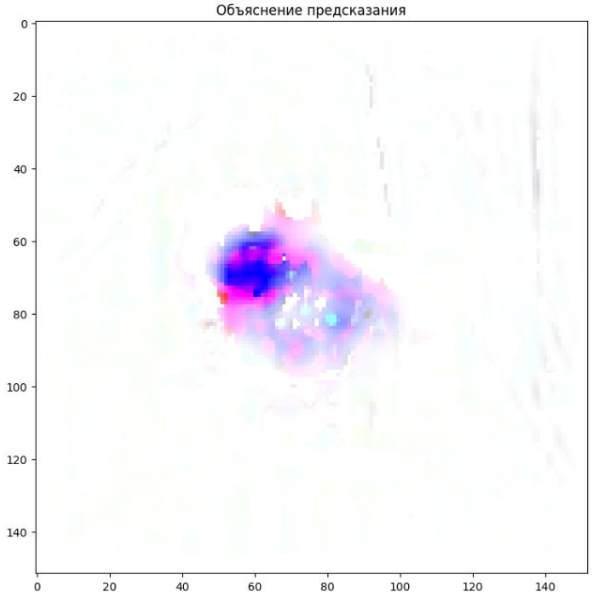


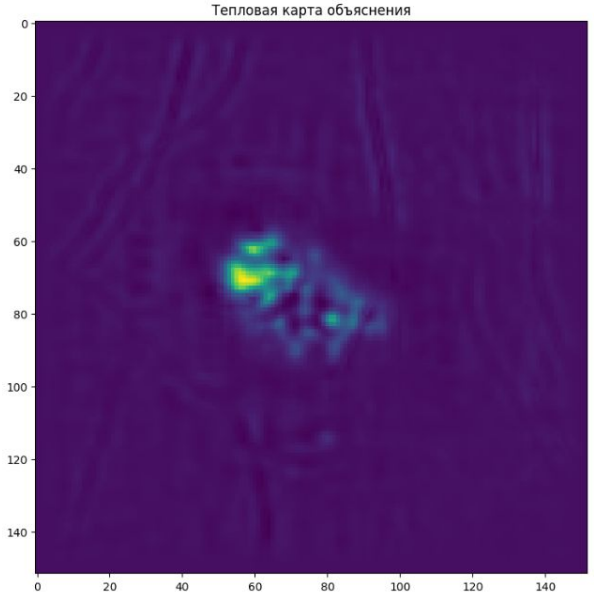
Истина: дерматофиброма

Предсказания:

меланома		0.10%
меланоцитарные невусы		0.80%
базально-клеточная карцинома		0.10%
Актинические кератозы и внутриэпителиальная карцинома		0.10%
доброкачественные образования, похожие на кератоз		0.10%
дерматофиброма	[объяснение]	98.80%
сосудистые поражения		0.00%



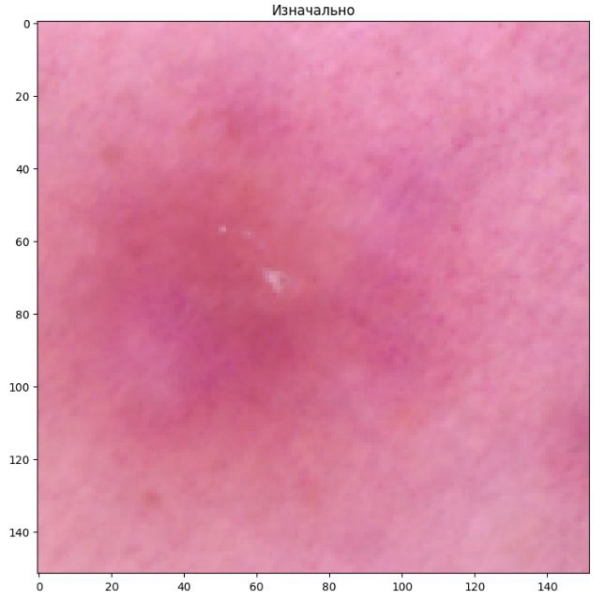


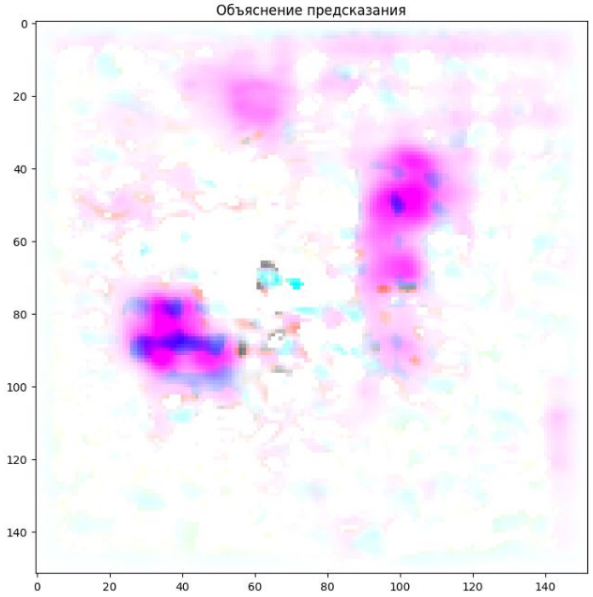


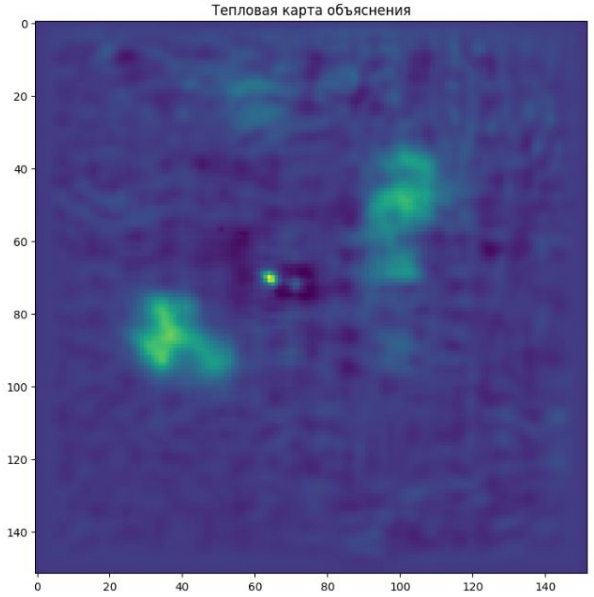
Истина: дерматофиброма

Предсказания:

меланома		0.10%
меланоцитарные невусы		3.80%
базально-клеточная карцинома		0.30%
Актинические кератозы и внутриэпителиальная карцинома		0.10%
доброкачественные образования, похожие на кератоз		0.20%
дерматофиброма	[объяснение]	95.50%
сосудистые поражения		0.10%



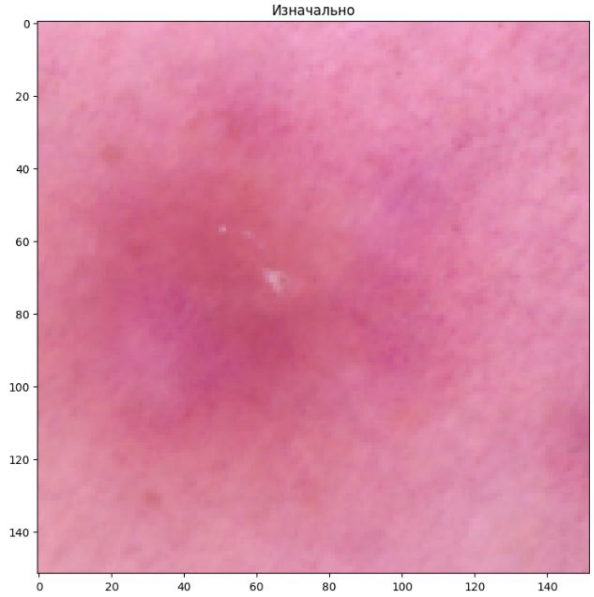


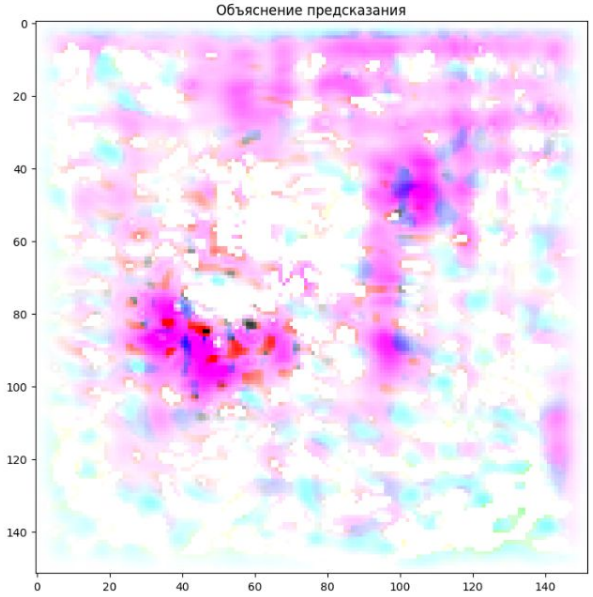


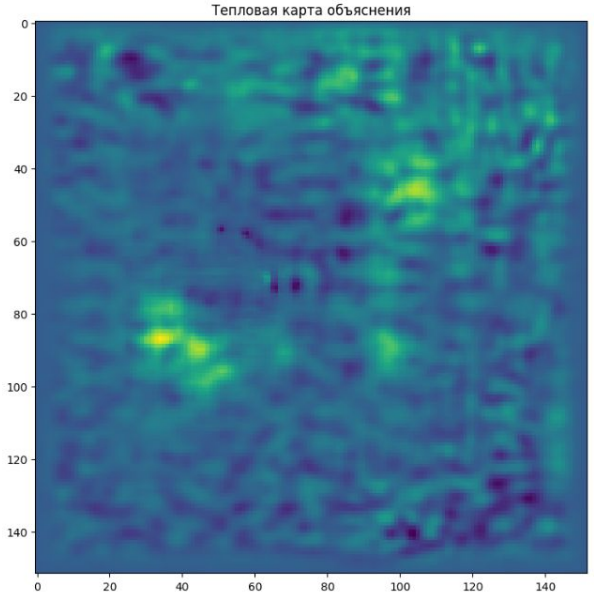
Истина: дерматофиброма

Предсказания:

меланома		0.10%
меланоцитарные невусы	[объяснение]	3.80%
базально-клеточная карцинома		0.30%
Актинические кератозы и внутриэпителиальная карцинома		0.10%
доброкачественные образования, похожие на кератоз		0.20%
дерматофиброма		95.50%
сосудистые поражения		0.10%







1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
2. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2020.
3. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization https://doi.org/10.1007/s11263-019-01228-7.
4. Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim. Sanity Checks for Saliency Maps. https://doi.org/10.48550/arXiv.1810.03292
5. Rachel Lea Draelos, Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. https://doi.org/10.48550/arXiv.2011.08891
6. Moritz Böhle, Mario Fritz, Bernt Schiele. B-cos Networks: Alignment is All We Need for Interpretability. https://doi.org/10.48550/arXiv.2205.10268
7. Moritz Böhle, Navdeeppal Singh, Mario Fritz, Bernt Schiele. B-cos Alignment for Inherently Interpretable CNNs and Vision Transformers. https://doi.org/10.48550/arXiv.2306.10898
8. Xianbiao Qi, Jianan Wang, Yihao Chen, Yukai Shi, Lei Zhang. LipsFormer: Introducing Lipschitz Continuity to Vision Transformers. https://doi.org/10.48550/arXiv.2304.09856
9. Moritz Böhle, Mario Fritz, Bernt Schiele. Holistically Explainable Vision Transformers. https://doi.org/10.48550/arXiv.2301.08669
10. Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, Guanghui Wang. Accumulated Trivial Attention Matters in Vision Transformers on Small Datasets. https://doi.org/10.48550/arXiv.2210.12333
11. Qihang Fan, Huaibo Huang, Mingrui Chen, Hongmin Liu, Ran He. RMT: Retentive Networks Meet Vision Transformers. https://doi.org/10.48550/arXiv.2309.11523
12. Peihao Wang, Wenqing Zheng, Tianlong Chen, Zhangyang Wang. Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice. https://doi.org/10.48550/arXiv.2203.05962
13. Yue Liu, Christos Matsoukas, Fredrik Strand, Hossein Azizpour, Kevin Smith. PatchDropout: Economizing Vision Transformers Using Patch Dropout. https://doi.org/10.48550/arXiv.2208.07220
14. Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, Ming Zhou. Scheduled DropHead: A Regularization Method for Transformer Models. https://doi.org/10.48550/arXiv.2004.13342
15. Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. https://doi.org/10.48550/arXiv.1207.0580
16. Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christopher Bregler. Efficient Object Localization Using Convolutional Networks. https://doi.org/10.48550/arXiv.1411.4280
17. Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. https://doi.org/10.48550/arXiv.1710.09412
18. Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, Quoc V. Le. Symbolic Discovery of Optimization Algorithms. https://doi.org/10.48550/arXiv.2302.06675
19. Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. https://doi.org/10.48550/arXiv.1909.13719
20. Philipp Tschandl, Cliff Rosendahl, Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. https://doi.org/10.48550/arXiv.1803.10417