Известия СПбГЭТУ «ЛЭТИ». 2024. Т. 17, № 6. С. 65–77
LETI Transactions on Electrical Engineering & Computer Science. 2024. Vol. 17, no. 6. P. 65–77

УДК 004.891.2 Научная статья

https://doi.org/10.32603/2071-8985-2024-17-6-65-77

Исследование модели машинного обучения на основе архитектуры сверточной нейронной сети методами объяснимости

И. А. Уткин✉, Д. С. Нагорный

Военно-космическая академия им. А. Ф. Можайского, Санкт-Петербург, Россия

✉ ewanytken@mail.ru

Аннотация. Методами объяснимости проведено исследование модели машинного обучения на основе архитектуры сверточной нейронной сети. В качестве методов использовались карты активации классов, которые рассчитывались посредством применения алгоритмов на основе прямого и обратного распространения тензоров изображения через составные части сети. Также осуществлен анализ избыточности карт активаций классов и статистический анализ весов сети при прохождении изображений.
Цель работы – повысить объяснимость внутренних процессов функционирования сверточной нейронной сети на базе модели ResNet50. Результатом исследования являются закономерности, следствия отражающие механизмы работы сверточной нейронной сети при решении задачи классификации изображений.

Ключевые слова: объяснимость, сверточная нейронная сеть, карты активации классов, метод главных компонент, методы статистического анализ, плотность распределения

Для цитирования: Уткин И. А., Нагорный Д. С. Исследование модели машинного обучения на основе архитектуры сверточной нейронной сети методами объяснимости // Изв. СПбГЭТУ «ЛЭТИ». 2024. Т. 17, № 6. С. 65–77. doi: 10.32603/2071-8985-2024-17-6-65-77.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

Study of a Machine Learning Model Based on a Convolution Neural Network by Explainability Methods

I. A. Utkin✉, D. S. Nagorny

Mozhaisky Military Space Academy, Saint Petersburg, Russia

✉ ewanytken@mail.ru

Abstract. Explainability methods are used to investigate a machine learning model based on the convolution neural network architecture. Class activation maps are calculated by applying algorithms based on forward and backward propagation of image tensors through the network components. A redundancy analysis of class activation maps and a statistical analysis of network weights during image propagation are also performed.
The purpose of the work is aimed at increasing the explainability of internal processes of convolutional neural network functioning on the basis of the ResNet50 model. As a result, regularities and consequences reflecting the mechanisms of convolutional neural network operation when solving the problem of image classification are presented.

Keywords: explainability, convolution neural network, class activation maps, method of principal components, statistical analysis methods, distribution density

For citation: Utkin I. A., Nagorny D. S. Study of a Machine Learning Model Based on a Convolution Neural Network by Explainability Methods // LETI Transactions on Electrical Engineering & Computer Science. 2024. Vol. 17, no. 6. P. 65–77. doi: 10.32603/2071-8985-2024-17-6-65-77.

Conflict of interest. The authors declare no conflicts of interest.

Введение. Модели машинного обучения (ММО) способны решать широкий класс задач, связанных с классификацией, генерацией, прогнозированием и т. д. При решении задач с применением ММО возникает ряд вопросов, связанных с обоснованностью принимаемого ею реше-

ния. В особенности это касается областей, связанных с здравоохранением, автоматизацией процессов управления в различных сферах деятельности, финансовых решений, где каждое ошибочное решение может привести к гибели людей или неэффективному расходованию ресурсов. Сложность в понимании происходящих процессов внутри ММО связана с тем, что большинство моделей представляют собой «черный ящик», полученный в ходе вычисления алгоритма и подбора параметров математическими методами или опытным путем.

При изучении процесса функционирования «черного ящика» ММО [1]–[3] выделены два основных направления исследования: объяснимость и интерпретируемость. Первое направление ищет причины, на основе которых ММО принимает именно такое решение. Второе направление исследования рассматривает построение достоверно интерпретируемых моделей на основе изучаемого «черного ящика», которые способны решать поставленные задачи с тем же качеством. Общую схему описанных исследований можно представить следующим образом (рис. 1).

В каждом из рассматриваемых направлений разработаны группы методов и алгоритмов, которые позволяют оценить степень влияния определенных параметров на качество решения задачи, выявить закономерности, а также попытаться раскрыть базовые принципы функционирования «черного ящика» ММО. В данной статье рассмотрены вопросы, связанные с объяснимостью.

Существующие алгоритмы и методы объяснимости частично позволяют раскрыть принципы функционирования ММО. В частности, если рассмотреть ММО на базе архитектуры сверточной нейронной сети (НС), применяемой для классификации объектов, на изображении можно выделить перечень методов, реализующих построение карт активации прямого и обратного распространения [4]. При прямом распространении анализируются веса слоев НС при прохождении через них изображения, а при обратном распространении оцениваются значения, полученные в ходе расчета градиентов, как при обучении НС.

Базовые версии этих методов имеют аббревиатуру CAM (Class activation map) и gradCAM (gradient-weighted CAM). Далее представлены данные методы в различной интерпретации, а также новые, основанные на построении карт активаций и статистическом анализе.

В качестве исходной ММО использовалась сверточная нейронная сеть архитектуры ResNet50, а в качестве датасета – ImageNet1000.

Анализ сверточной нейронной сети за счет построения карт активации классов (CAM) и послойного представления. Идея метода CAM заключается в получении каналов с последнего сверточного слоя нейронной сети в виде относительно небольших по размерности двумерных объектов (каналов после активации) с глобальным усреднением и последующим перемножением с активирующими весами последнего слоя [5], [6]. Данные операции можно представить в качестве следующего выражения:

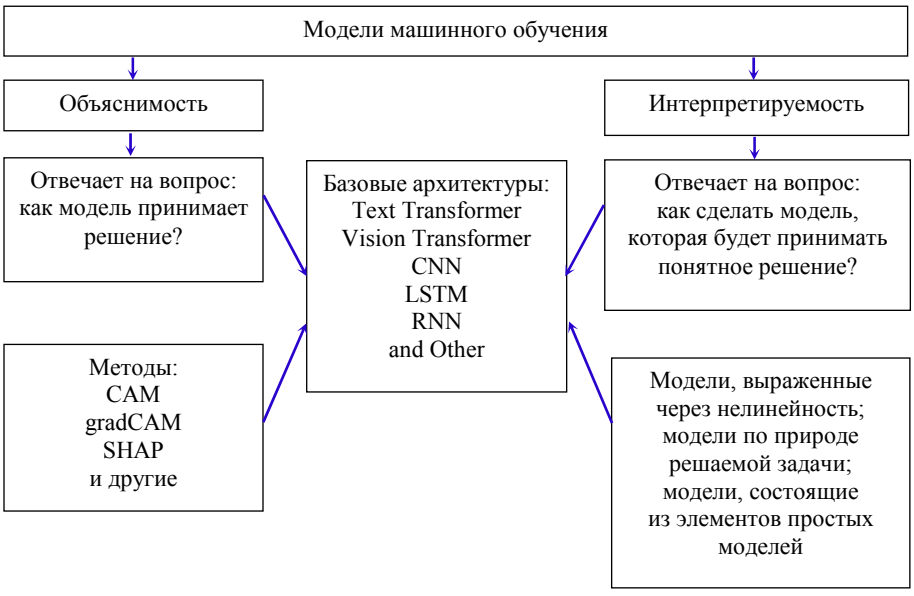

Рис. 1. Схематичное представление объяснимости и интерпретируемости
Fig. 1. Schematic representation of explainability and interpretability

$$y^{\text{class}} = \sum_k w^{\text{class}} \frac{1}{N} \sum_i H^k \sum_j W^k,$$

где y^{class} – карта активации класса; w^{class} – веса с последнего слоя прямого распространения активирующие наибольшее значение в слое класси-

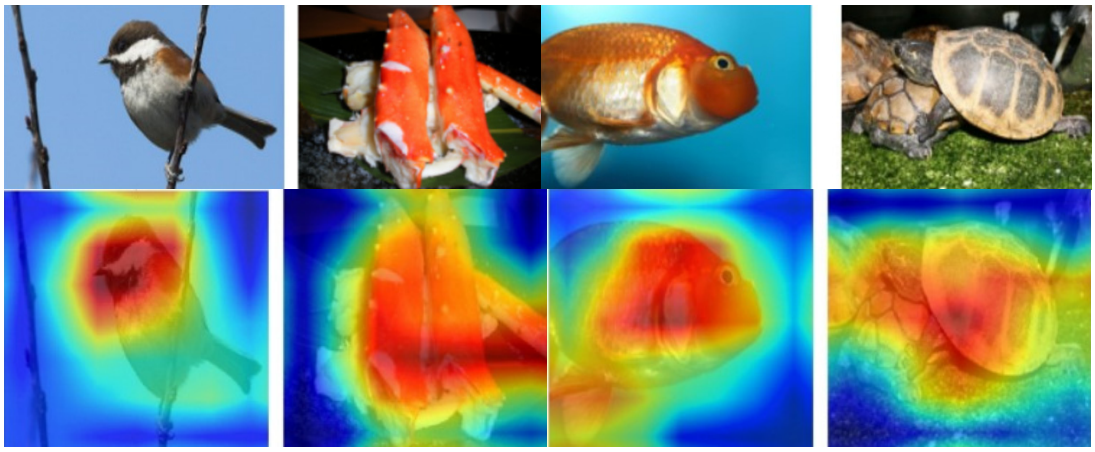

Рис. 2. Карты активации нескольких случайных изображений из различных классов.
Верхний ряд – исходные изображения, нижний ряд – изображения с картами активации классов
Fig. 2. Activation map for coincidental samples from different classes.
Top row – original images, bottom row – images with class activation maps

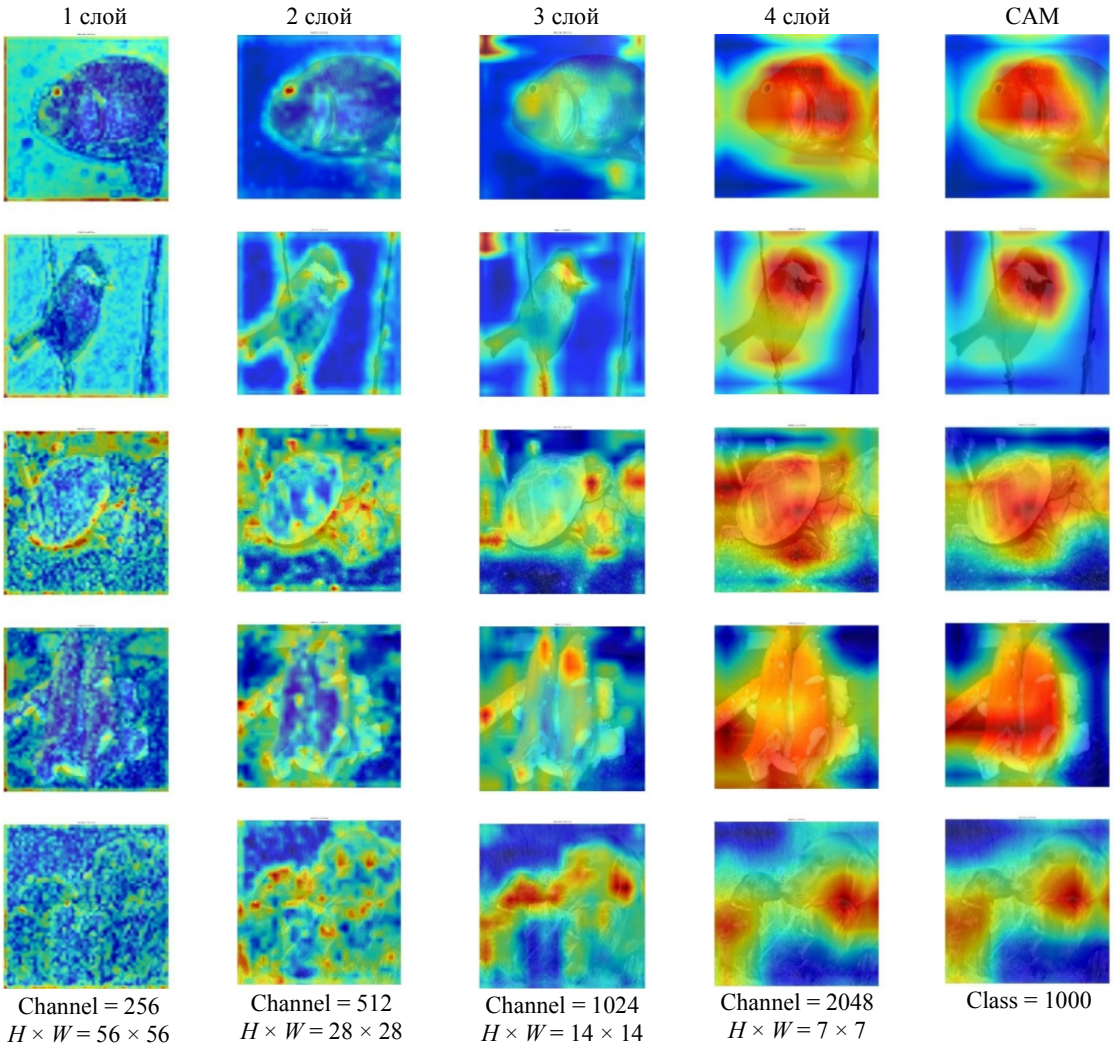

Рис. 3. Послойное отображение серии картинок при прохождении через слои НС
Fig. 3. Layer by layer mapping for a set of images

фикации; N – общее количество пикселей в канале; H^k, W^k – размеры канала после активации (прямом прохождении).

Правая часть выражения состоит из весов, принадлежащих классу, выбранному нейронной сетью и усреднением каналов. Количество каналов совпадает с количеством весов перед слоем классификации.

Реализация CAM и алгоритмов, которые будут использоваться далее, представлена в [7].

Карты активации для нескольких случайных изображений из различных классов имеют вид, представленный на рис. 2.

На рисунке показаны области признаков, которые сверточная нейронная сеть выделила в качестве значимых, – чем ярче тепловая карта, тем более значима область.

Полученные области признаков позволяют оценить конечный результат принятия решения нейронной сетью, однако не дают объяснения, как сеть пришла к этому результату. Для более глубокого понимания происходящих процессов внутри НС при прямом прохождении изображения был проведен послойный анализ. Данный анализ основывается на алгоритме построения карт активации классов, но ввиду отсутствия последнего слоя НС не учитывается индекс, относящий изображение к активированным нейронам, определяющим класс.

Выражение для послойного отображения примет вид

$$y = \sum_k w \frac{1}{N} \sum_i H^k \sum_j W^k,$$

где w – веса с промежуточных слоев прямого распространения; N – общее количество пикселей в одном канале; H^k, W^k – размеры канала после активации.

Послойное прохождение изображения и реализация алгоритма CAM отображены на рис. 3.

После анализа послойного представления сверточной НС был сделан ряд выводов:

1. При увеличении числа слоев НС растет ее обещающая способность, а это указывает на то, что от частных признаков каждого изображения класса НС стремится к выявлению базовых признаков целевого класса.

2. Изображения с ярко выраженными признаками, которые НС обобщает с начальных слоев, при приближении к последним слоям может сдвигать область признаков и захватывать части изображения, не принадлежащие к классу объекта. Это свидетельствует о способности НС к переобучению, что при значительных размерах сети может привести к неправильной классификации «простых» изображений, поэтому размер сети должен подбираться, исходя строго из решаемой задачи.

3. Размеры сверток в слоях НС зависят от разрешения входного изображения. При больших свертках на маленьких изображениях значительно усложняется выявление признаков, присущих целевому классу, и наоборот.

4. Увеличение числа каналов от слоя к слою также повышает обобщающую способность НС.

Анализ сверточной нейронной сети методом карт активации обратного распространения. Помимо построения карт активации классов и послойного отображения нейронной сети, часто используется метод построения карт активации обратного построения – gradCAM [8]. Данный метод включает в себя построение аналогичных карт активации, но относительно градиента выходного значения целевого класса. Вместо использования весов, как в предыдущем способе, используются весовые коэффициенты, которые представляют собой усредненный градиент выходного слоя относительно каналов последнего сверточного слоя после его активации.

В формализованном виде алгоритм можно представить следующими выражениями:

$$\alpha_k^{\text{class}} = \frac{1}{N} \sum_i \sum_j \frac{dy^{\text{class}}}{dA_{ij}^k},$$

$$L^{\text{class}} = \text{ReLU}\left(\sum_k \alpha_k^{\text{class}} A^k \right),$$

где A^k – размер канала ($H \times W$) последнего сверточного слоя сети после активации; α_k^{class} – весовой коэффициент; L^{class} – линейная комбинация весового коэффициента и каналов после активации; $\frac{dy^{\text{class}}}{dA_{ij}^k}$ – градиент выходного слоя сверточной НС относительно каналов после активации.

Реализация алгоритма на модели ResNet50 дает результаты, представленные на рис. 4.

Проанализировав результаты, можно увидеть, что обученная модель при расчете градиентов

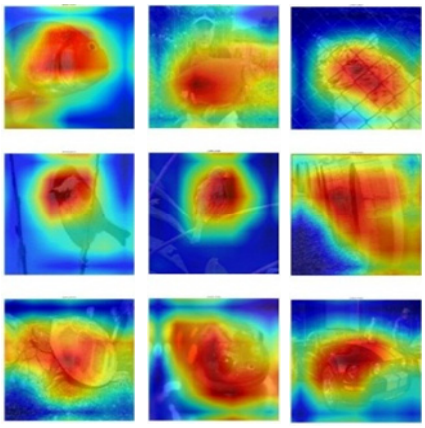

Рис. 4. Отображение gradCAM на изображения из датасета ImageNet1000
Fig. 4. GradCam mapping to images from ImageNet1000 datase

выделяет такие же признаки изображения, как и расчет карты активации классов. Схожесть выделенных областей для обученной модели указывает на то, что сверточная НС как при прямом прохождении, так и при обратном обращает внимание на одинаковые базовые признаки целевого класса. Связывая данный факт с принципом расчета градиентов при обучении, а также решения оптимизационной задачи при нахождении функции потерь можно утверждать, что алгоритм обучения ищет глобальный минимум в заданном признаковом пространстве. В некоторых случаях на изображении выделяется несколько областей, отображающих минимум, что связано с нахождением не одного глобального минимума, а нескольких.

Существуют также другие типы алгоритмов градиентных весов, например Guided GradCAM, GradScore, GradCAM++. Данные методы в том или ином виде основываются на расчете градиентов различного порядка.

В частности, если рассмотреть GradCAM++, при расчете которого учитывались дифференциалы второго порядка, полученная область признаков на изображении будет иметь меньшую площадь и более точно выделять базовые признаки [9]. Однако в большинстве случаев обучение сверточной НС основывается на алгоритмах с расчетом дифференциалов первого порядка и применении данного метода, а также аналогичных с их особенностями теряет смысл, так как выделение признаковых областей на изображении связано с более точным алгоритмом расчета градиента [10].

Анализ избыточности сверточной нейронной сети. Применение методов прямого и обратного прохождения с активацией сверток позволяет сделать заключение о том, какие области изображения модель выделяет при классификации, а также частично ответить на вопросы объяснимости ее функционирования.

Как было показано в подпункте «*Анализ сверточной нейронной сети за счет построения карт активации классов* (*CAM*) *и послойного представления*», при прохождении изображения через слои сверточной НС были получены карты активации признаков. Полученные алгоритмом CAM карты признаков представляют собой двумерные плоскости, размер которых равен размеру канала после его активации. В случае использования модели ResNet50 и изображения с разрешением 224 × 224 область составляет 7 × 7 (рис. 5).

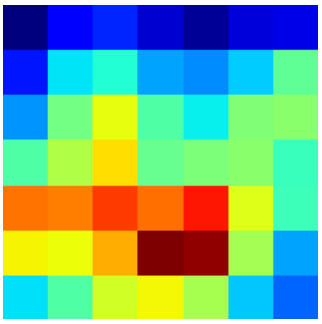

Рис. 5. Пример карты активации без обработки
Fig. 5. Example of a class activation map without processing

Рассчитанные карты активации с помощью 49 точек содержат информацию о наиболее важных признаках целевого класса при прямом прохождении изображения. Чтобы понять, достаточно ли этого для решения задачи классификации, следует применять методы их разложения.

Для применения алгоритмов исследования необходимо предварительно обработать имеющиеся карты активации классов. По своей сути они содержат информацию о признаках в сжатом двумерном виде, данные карты можно представить в виде векторов, т. е. изменить размер в приведенном выше случае с 7 × 7 на 1 × 49 (рис. 6).

Вектор в единичном экземпляре дает мало информации и, по сути, может быть интерпретирован, как одно измерение (случай) в 49-мерном признаковом пространстве. Для большего количества измерений необходимо рассчитать CAM для выборки изображений, принадлежащих одному целевому классу.

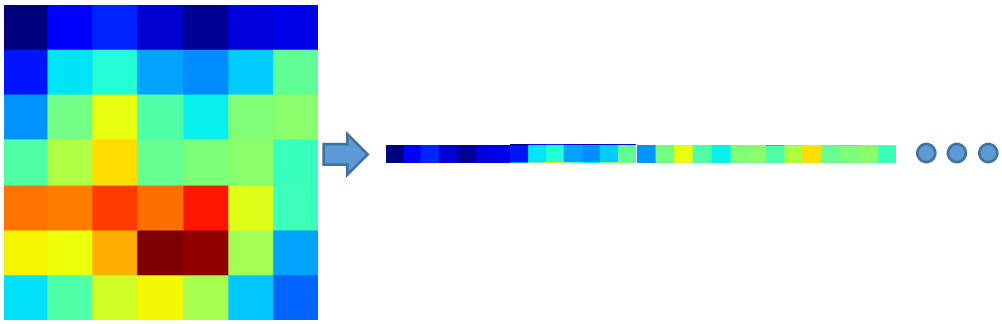

Рис. 6. Представление карты в виде вектора признаков
Fig. 6. Representation of a class activation map as a vector of features

В качестве примера был использован произвольный класс из датасета ImageNet1000.

После расчета карт активаций и представления их в виде вектора была получена матрица размером 49 × 41, где количество столбцов соответствует количеству изображений в однородном классе. Графически это представлено на рис. 7.

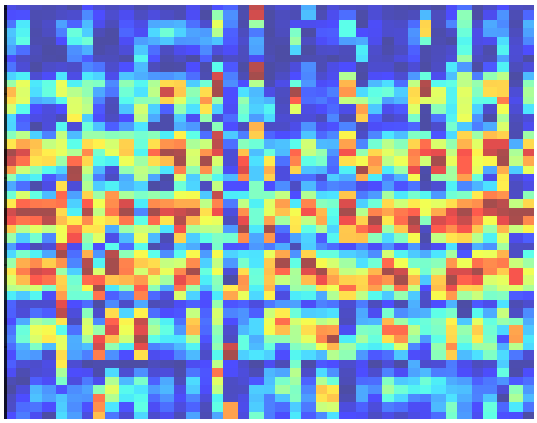

Рис. 7. Пространство признаков размером 49 (ось ординат) на 41 (ось абсцисс)
Fig. 7. Feature space with a dimension of 49 (ordinate axis) to 41 (abscissa axis)

Оценивание избыточности или дефицита размеров пространства признаков можно осуществить за счет его сжатия. Признаком дефицита размеров послужит значительная потеря информации при минимальном сжатии, избыточности, соответственно, обратное.

На рис. 8 представлена схема оценивания избыточности/дефицита размеров признакового пространства.

В качестве алгоритма сжатия размеров был использован метод главных компонент. Его суть заключается в построении ковариационной матрицы признаков и расчета на основе айген-векторов, айген-значений с последующим матричным произведением и сжатием размеров. Реализация метода была взята из библиотеки sklearn [11].

Изображения после расчета карт активации для целевого класса представлены на рис. 9 (5 изображений из 41).

Далее осуществлено сжатие размеров признакового пространства до 36 компонент (6 × 6) и отображение его на исходные изображения датасета (рис. 10).

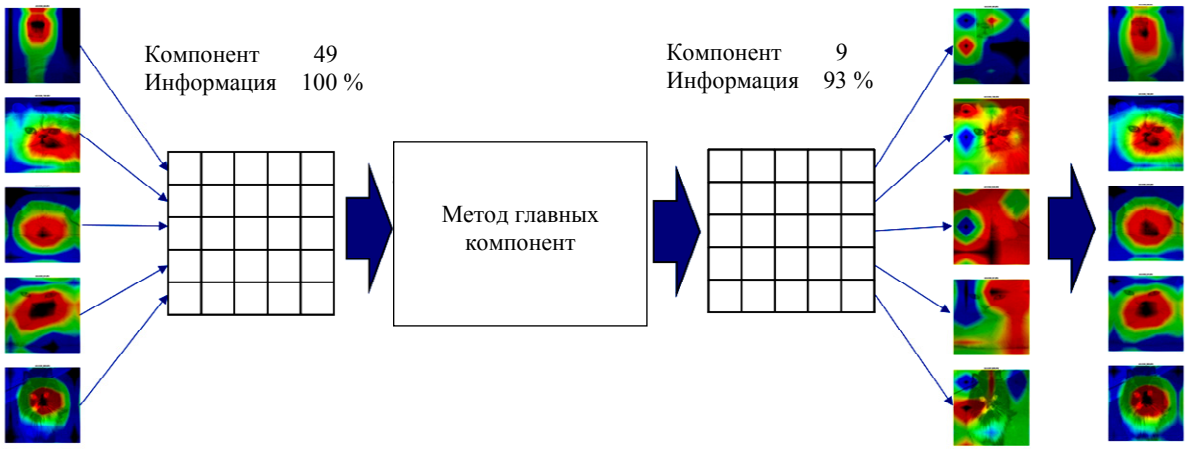

Рис. 8. Схема оценивания размеров признакового пространства
Fig. 8. Scheme for estimating the dimensionality of the feature space

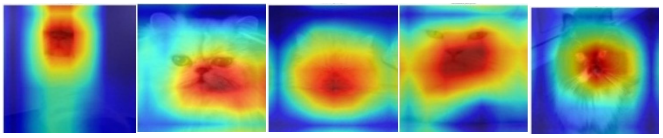

Рис. 9. Построение карт активации классов для выборки из целевого класса
Fig. 9. Construction of class activation maps for sampling from the target class

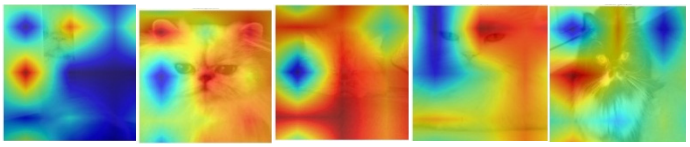

Рис. 10. Отображение на исходные изображения датасета сжатых карт активации
Fig. 10. Mapping compressed activation maps to the original dataset images

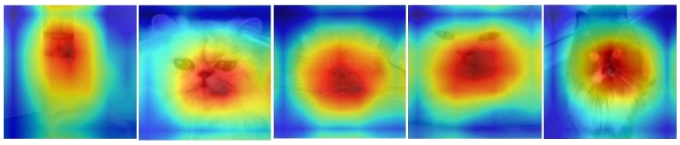

Рис. 11. Восстановленные карты активации целевого класса
Fig. 11. Recovered class activation maps

```
[5.6252072e-15 1.0156708e-05 2.7649643e-05 3.7482449e-05 4.2952932e-05
 6.3283624e-05 8.0670885e-05 1.0771373e-04 1.4873820e-04 2.1942118e-04
 2.6044162e-04 3.1393403e-04 3.4259117e-04 3.7550344e-04 5.1631185e-04
 6.2412251e-04 7.8092172e-04 8.0532720e-04 9.4386912e-04 9.9355390e-04
 1.2137018e-03 1.6810955e-03 1.9495980e-03 2.1692021e-03 2.5554232e-03
 3.0200800e-03 3.7222896e-03 4.6060761e-03 6.4338180e-03 8.6312983e-03
 1.0153794e-02 1.1521587e-02 1.4861445e-02 2.1814652e-02 4.7785360e-02
 5.5464689e-02 7.1629591e-02 1.3610165e-01 1.5274547e-01 1.8200956e-01
 2.5323495e-01]
```

Рис. 12. Матрица объясняемой дисперсии с выделенными значениями
Fig. 12. Matrix of explainable with underlying variation

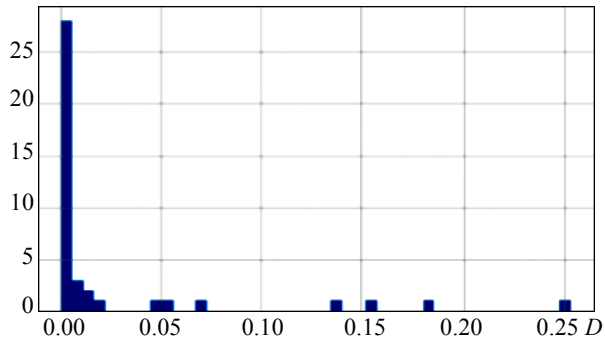

Рис. 13. Гистограмма распределения компонент
Fig. 13. Histogram of component distribution

Карты активации имеют смазанный вид и не отображают областей признаков, как в предыдущем случае, поэтому их графическая объяснимость достаточно затруднительна.

Для более наглядного представления сжатые карты признаков с потерей информации были восстановлены к исходным размерам 7×7 или 49 компонентам (рис. 11).

Карты активации с потерей информации приняли аналогичный вид карт с полным ее содержанием.

Сжатие 13 компонент оказало слабое влияние на потерю признаков при классификации изображений. Для анализа распределения информации между компонентами была получена матрица объясняемой дисперсии для 41 компоненты, представленная на рис. 12, а на ее основе была построена гистограмма распределения компонент (рис. 13).

Исходя из матрицы объяснимости дисперсии и гистограммы распределения компонент следует, что основная информация о признаках, необходимых для классификации сверточной НС, распределена между 9 компонентами, которые выделены на рис. 12, и составляет порядка 93.73 %.

Данные результаты указывают на то, что модель ResNet50 при классификации изображений создает значительную избыточность в признаковом пространстве. Это означает, что для обучения на датасете ImageNet1000 возможно использование ММО с меньшим количеством параметров или измененной архитектурой, за счет чего сократится время обучения модели и ее размеры.

Рис. 14. Выборки из датасета ImageNet1000
Fig. 14. Samples from the ImageNet1000 dataset

Применение метода статистического анализа значений сверточной нейронной сети. Расчет карт активации классов и их сжатие основывается на использовании весов с различных слоев сверточной НС. Веса, полученные в ходе обучения НС, также могут интерпретироваться в другом контексте. Например, таким контекстом может служить область математической статистики, в частности решение задачи статистической проверки гипотез.

Если рассматривать массивы весов НС в качестве совокупности случайных значений, становится возможным применить методы математической статистики, что позволяет углубленно изучить механизмы выбранной архитектуры.

Для более разностороннего анализа результатов, полученных в ходе применения методов математической статистики, были использованы две произвольные выборки из датасета ImageNet1000. Первая выборка: n02123394 – «Персидская кошка», вторая выборка: n03100240 – «Кабриолет») (рис. 14).

Статистическая проверка гипотезы осуществлялась согласно следующим этапам:

1. Расчет гистограммы распределения весов из слоев НС после прямого прохождения изображения из выборок.

2. Построение гистограмм значений весов и плотности их распределения.

3. Оценка достоверности по критерию согласия Колмогорова–Смирнова.

Внешний вид гистограмм значений весов из слоев НС позволил выдвинуть гипотезу, что их распределение соответствует нормальному и логнормальному законам распределения.

В качестве критерия согласия использовался критерий Колмогорова–Смирнова [12]. Его расчет выполнялся согласно следующим выражениям:

$$D = \max\left|F^{*}(x) - F(x)\right|,$$

$$D\sqrt{n} \geq \lambda.$$

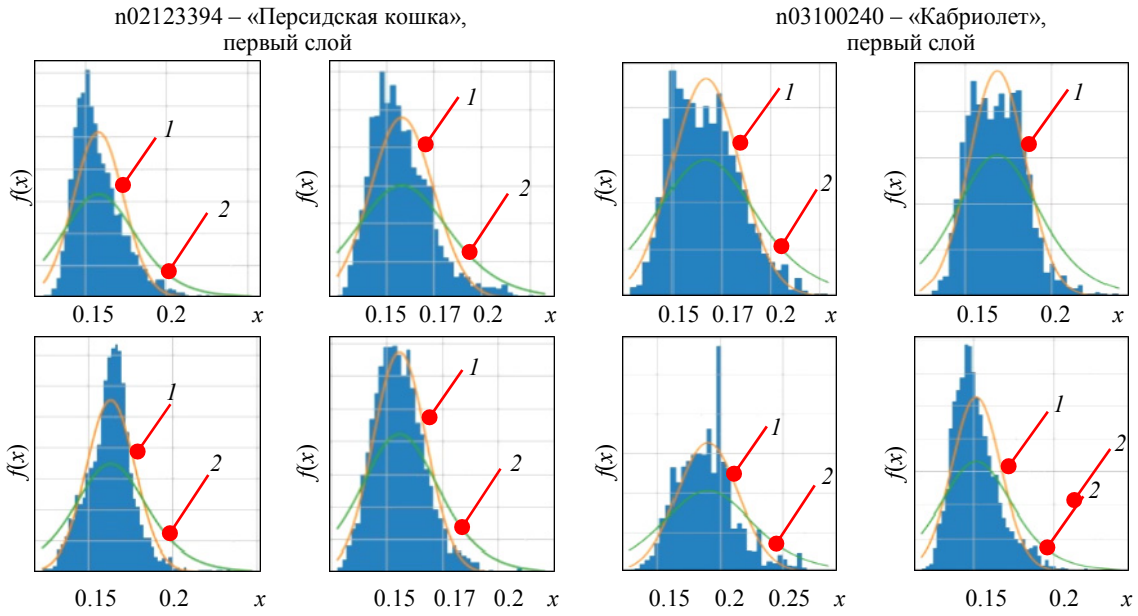

Рис. 15. Проверка статистических гипотез значений первого слоя
Fig. 15. Statistical hypothesis testing of the first layer values

Реализация алгоритмов статической обработки данных выполнена с использованием функций библиотеки scipy [13].

Размерность карты активации на выходе каждого слоя НС представляет собой четырехмерный тензор: батч, канал, высота, ширина, где под батчем понимается число изображений из выборок датасета. Первая выборка состоит из 41 изображения и классифицируется как «283: 'Persian cat'», вторая выборка состоит из 26 изображений и классифицируется как «511: 'convertible'». Высота и ширина распрямляются из двумерной матрицы в вектор по аналогии при сжатии размерности, но с усреднением значений по каналу. Указанные преобразования выполнялись для всех слоев, за исключением полносвязного. Количества значений в слоях составили: 3136 для 1-го слоя, 784 для 2-го слоя, 196 для 3-го слоя, 49 для 4-го слоя и 1000 для полносвязного. Для наглядности были представлены по 4 изображения с выборки.

Расчет проверки статистических гипотез значений согласно перечисленным этапам для первого слоя представлен на рис. 15.

Внешний вид гистограмм имеет форму нормальной или логнормальной плотности распределения. Согласно выбранному критерию согласия, достоверность изменяется в пределах 0.73…0.8 (табл. 1) для нормального закона распределения (кривая *1*) и 0.84…0.95 (табл. 1) для логнормального закона распределения (кривая *2*). Данные факты говорят о том, что НС стремится привести значения весов к некоторой средней вне зависимости от того, что изображено на приходящей на вход картинке. Преимущество логнормального закона возможно объяснить за счет применения *функции активации выпрямителя ReLU*, которая отсекает отрицательные значения перед попаданием в выборку.

Табл. 1. Значение статистических гипотез первого слоя
Tab. 1. Statistical hypothesis values of the first layer

Наименование изображения в датасете ImageNet1000	Значение критерия при гипотезе нормального закона распределения	Значение критерия при гипотезе логнормального закона распределения
n02123394_1036	0.7971	0.9563
n02123394_1086	0.7774	0.9419
n02123394_1205	0.8022	0.9014
n02123394_1305	0.8026	0.9426
n03100240_1132	0.7452	0.9247
n03100240_12194	0.7413	0.9107
n03100240_12316	0.7187	0.8475
n03100240_12995	0.8083	0.9547

Результаты проверки статистических гипотез значений второго слоя представлены на рис. 16.

Исходя из полученных данных на втором слое, НС также стремится сдвинуть значения к среднему, достоверность по критерию согласия повысилась относительно результатов первого слоя (табл. 2).

Гистограммы весов и плотности распределения по выбранным законам для третьего слоя представлены на рис. 17.

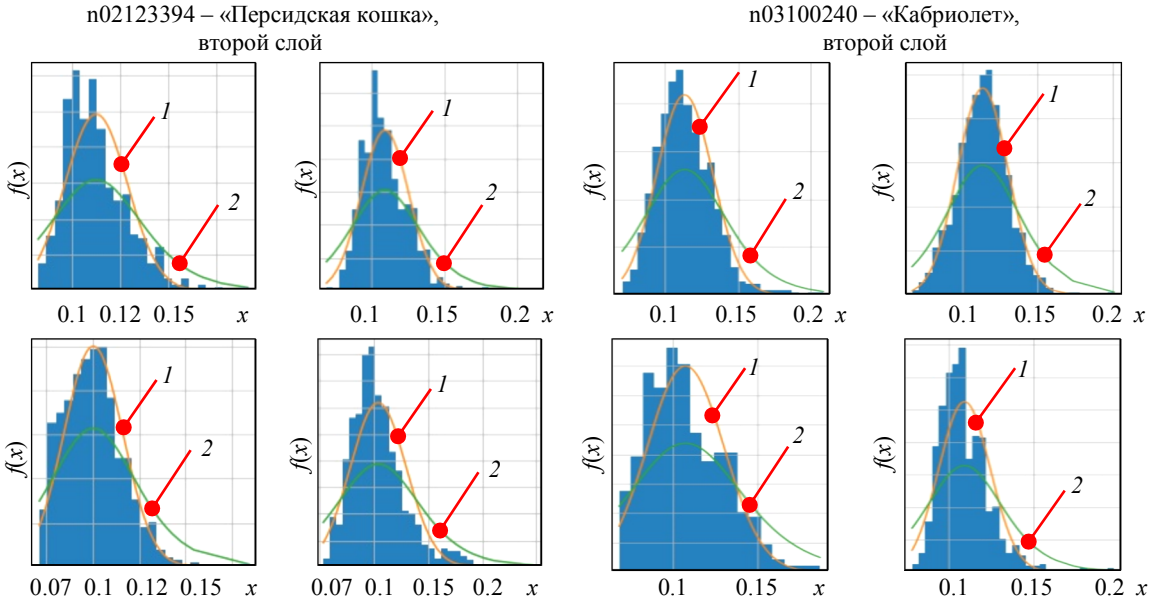

Рис. 16. Проверка статистических гипотез значений второго слоя
Fig. 16. Statistical hypothesis testing of the second layer values

Табл. 2. Значение статистических гипотез второго слоя
Tab. 2. Statistical hypothesis values of the second layer

Наименование изображения в датасете ImageNet1000	Значение критерия при гипотезе нормального закона распределения	Значение критерия при гипотезе лог-нормального закона распределения
n02123394_1036	0.8098	0.9795
n02123394_1086	0.8724	0.9744
n02123394_1205	0.8354	0.9923
n02123394_1305	0.8852	0.9910
n03100240_1132	0.8265	0.9591
n03100240_12194	0.8380	0.9451
n03100240_12316	0.8278	0.9579
n03100240_12995	0.8724	0.9846

Табл. 3. Значение статистических гипотез третьего слоя
Tab. 3. Statistical hypothesis values of the third layer

Наименование изображения в датасете ImageNet1000	Значение критерия при гипотезе нормального закона распределения	Значение критерия при гипотезе лог-нормального закона распределения
n02123394_1036	0.9795	1.0
n02123394_1086	0.9183	1.0
n02123394_1205	0.9693	1.0
n02123394_1305	0.9489	1.0
n03100240_1132	0.9591	1.0
n03100240_12194	0.9489	1.0
n03100240_12316	0.9336	1.0
n03100240_12995	0.9948	1.0

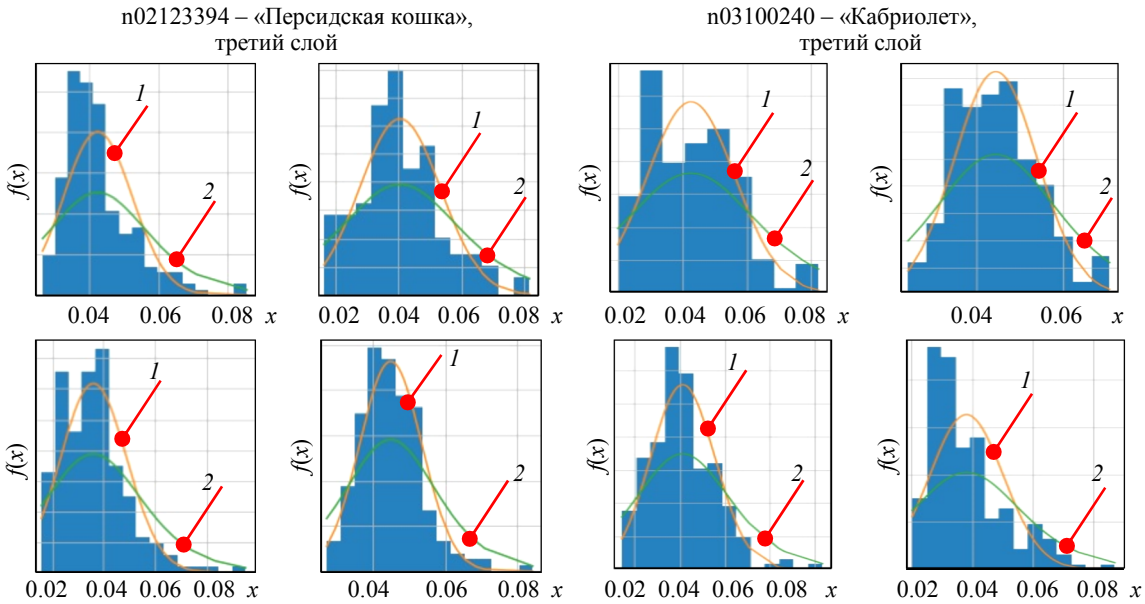

Рис. 17. Проверка статистических гипотез значений третьего слоя
Fig. 17. Statistical hypothesis testing of the third layer values

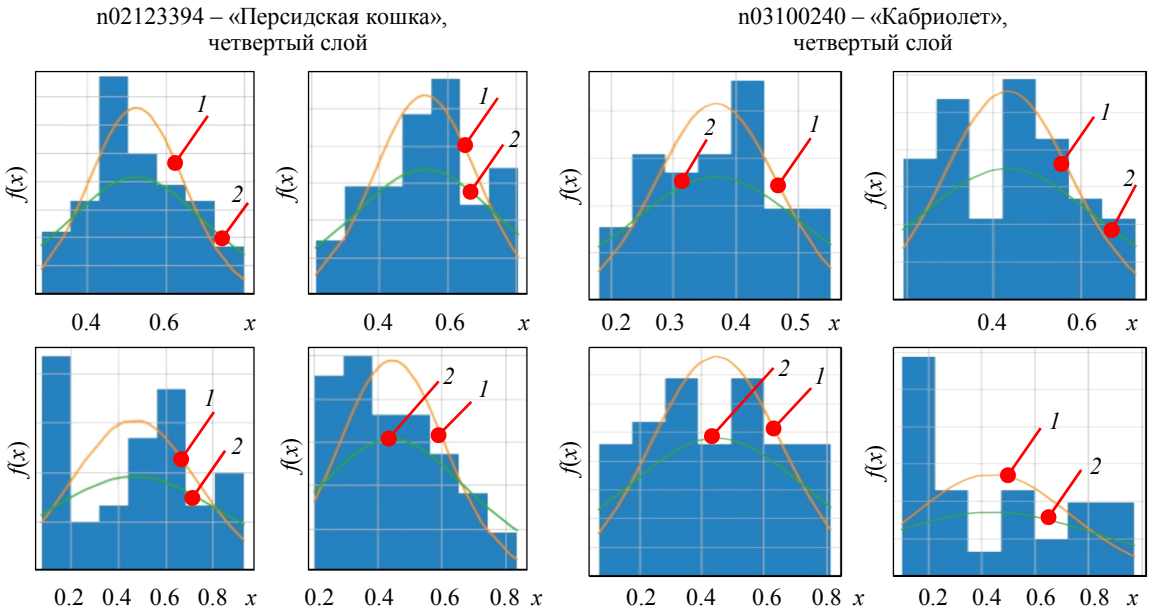

Рис. 18. Проверка статистических гипотез значений четвертого слоя
Fig. 18. Statistical hypothesis testing of the fourth layer values

Табл. 4. Значение статистических гипотез четвертого слоя
Tab. 4. Statistical hypothesis values of the fourth layer

Наименование изображения в датасете ImageNet1000	Значение критерия при гипотезе нормального закона распределения	Значение критерия при гипотезе лог-нормального закона распределения
n02123394_1036	0.3469	0.2653
n02123394_1086	0.2653	0.2040
n02123394_1205	0.2040	0.2653
n02123394_1305	0.2857	0.1836
n03100240_1132	0.4489	0.4489
n03100240_12194	0.3265	0.2653
n03100240_12316	0.2448	0.1632
n03100240_12995	0.1632	0.3877

При обработке третьего слоя значение достоверности для логнормального закона распределения значений всех изображений обоих выборок стало единицей. Достоверность для нормального распределения также в среднем выросла до 0.95 (табл. 3).

Несмотря на растущее значение критерия согласия от слоя к слою на четвертом слое достоверность снижается до менее 0.5 (табл. 4), что свидетельствует о неправильной гипотезе выбранных законов распределения для данных выборок (рис. 18). Это говорит о том, что прохождение определенного порога удвоения числа каналов относительно каждого слоя влияет на достоверность распределения данных по выбранным законам распределения. Также на низкую достоверность влияет небольшое число значений при проверке критерия согласия.

Табл. 5. Значение статистических гипотез полносвязного слоя
Tab. 5. Statistical hypothesis values of the dense layer

Наименование изображения в датасете ImageNet1000	Значение критерия при гипотезе нормального закона распределения
n02123394_1036	0.563
n02123394_1086	0.554
n02123394_1205	0.535
n02123394_1305	0.561
n03100240_1132	0.563
n03100240_12194	0.554
n03100240_12316	0.535
n03100240_12995	0.561

Гистограммы и плотности распределения значений, полученные в результате проверки статистических гипотез значений полносвязного слоя (рис. 19).

Значения полносвязного слоя также не прошли критерий согласия Колмогорова–Смирнова, так как достоверность находится в интервале 0.5…0.6 (табл. 5), хотя внешний вид гистограммы достаточно сильно напоминает нормальное распределение аналогично 1–3-му слоям. Ввиду того, что веса последнего слоя – основополагающие при классификации изображения, НС стремится свести их к нулю, чтобы отделить ключевое значение, соответствующее классу, т. е. как можно правее сдвинуть по оси абсцисс.

После проведения проверки гипотез был сделан ряд выводов:

1. Изменение датасетов не влияет на распределение значений в слоях НС. Достоверность за-

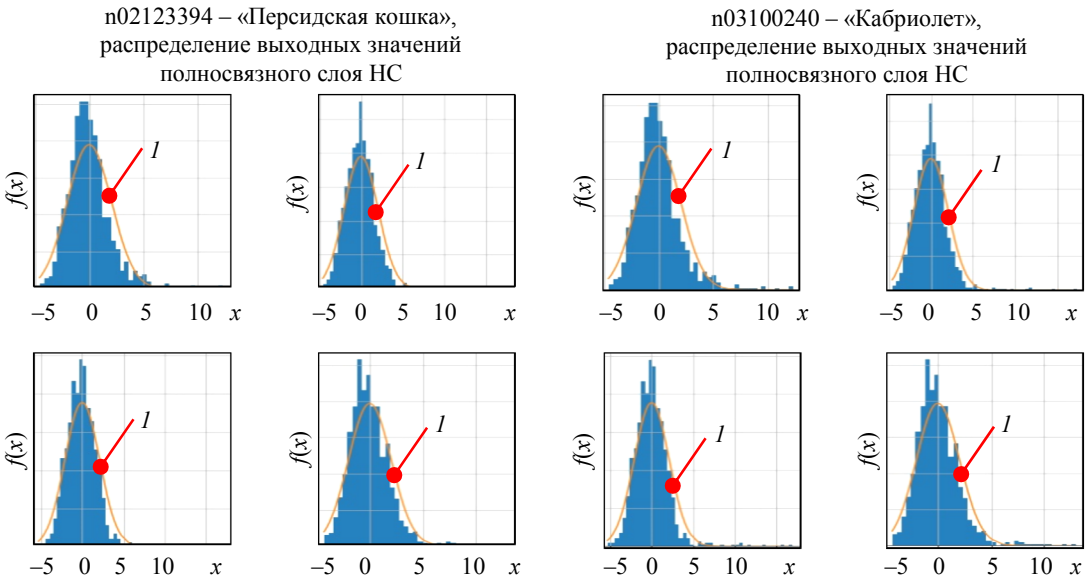

Рис. 19. Проверка статистических гипотез значений выходного слоя
Fig. 19. Statistical hypothesis testing of output layer values

конов распределения незначительно различима между выборками.

2. При обучении НС стремится усреднить значения к некоторому числу. Это особенно важно в последнем слое прямого распространения, на основе значений которого классифицируют изображения.

3. Веса первых трех слоев изучаемой НС распределяются по логнормальному закону распределения с достоверностью в интервале от 0.8 до 0.99;

4. Веса четвертого слоя не проходят критерий Колмогорова–Смирнова по анализируемым распределениям (нормальному, логнормальному).

Заключение. В ходе исследования объяснимости модели ResNet50 на базе архитектуры сверточной нейронной сети были рассмотрены методы и алгоритмы, часть из которых уже активно используется для анализа ее функционирования, а часть была предложена в данной статье как дополнение к имеющимся. Это позволило более глубоко понять механизмы функционирования выбранной модели.

В частности, за счет применения вышеописанных методов было продемонстрировано, что с ростом числа слоев НС растет ее обобщающая способность, а изображения с ярко выраженными признаками класса могут приводить к переобученности больших моделей.

Методы с расчетом градиентов стремится выделить аналогичные признаки изображения при условии предобученной модели. Также можно отметить, что существующие алгоритмы, основанные на подсчете градиентных весов карт активации классов, позволяют определить более точные области признаков, но не за счет проводимой классификации НС, а за счет более продвинутых алгоритмов.

При анализе избыточности сверточной НС было выявлено, что основная информация о признаках, необходимых для классификации распределена между 11 компонентами и составляет 93.73 %.

На заключительном этапе проведено исследование с применением методов статического анализа. В результате исследования были выявлены следующие основные закономерности: изменение датасетов не влияет на распределение значений в слоях НС, а при обучении нейронная сеть стремится усреднить значения к некоторому числу в зависимости от рассматриваемого слоя.

Список литературы

1. Four principles of explainable artificial intelligence / P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki. doi: 10.6028/NIST.IR.8312.

2. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. URL: arxiv.org/pdf/1811.10154v3.pdf (дата обращения: 25.09.2023).

3. Miller T. Explanation in artificial intelligence: Insights from the social sciences. URL: https://arxiv.org/pdf/1706.07269v3.pdf (дата обращения: 28.09.2023).

4. Интерпретация моделей или как заглянуть в черный ящик. URL: https://habr.com/ru/articles/677940/ (дата обращения: 18.09.2023).

5. Class activation map generation by multiple level class grouping and orthogonal constraint / K. Huang, F. Meng, H. Li, Sh. Chen, Q. Wu, K. N. Ngan. URL: https://arxiv.org/pdf/1909.09839v1.pdf (дата обращения: 12.10.2023).

6. Learning deep features for discriminative localization / B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. URL: https://arxiv.org/pdf/1512.04150v1.pdf (дата обращения: 16.10.2023).

7. Repository with code snippet. URL: https://github.com/ewanytken/moduleInterpret (дата обращения: 26.11.2023).

8. Grad-CAM: Visual explanations from deep networks via gradient-based localization / R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. URL: https://arxiv.org/pdf/1610.02391v4.pdf (дата обращения: 04.10.2023).

9. Grad-CAM++: Improved visual explanations for deep convolutional networks / A. Chattopadhyay, A. Sarkar, P. Howlader, V. Balasubramanian. URL: https://arxiv.org/pdf/1710.11063v3.pdf (дата обращения: 11.11.2023).

10. Sanity checks for saliency maps / Ju. Adebayo, Ju. Gilmer, M. Muelly, Ian Goodfellow, M. Hardty, B. Kim. URL: https://arxiv.org/pdf/1810.03292v3.pdf (дата обращения: 06.11.2023).

11. Principal component analysis (PCA). URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (дата обращения: 01.10.2023).

12. Kolmogorov–Smirnov test. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html (дата обращения: 21.09.23).

13. Statistical functions. URL: https://docs.scipy.org/doc/scipy/reference/stats.html (дата обращения: 07.10.2023).

Информация об авторах

Уткин Иван Алексеевич – канд. техн. наук, преподаватель кафедры информационно-аналитической работы Военно-космической академии им. А. Ф. Можайского, ул. Ждановская, 13, г. Санкт-Петербург, 197198, Россия.
E-mail: ewanytken@mail.ru

Нагорный Дмитрий Сергеевич – преподаватель кафедры информационно-аналитической работы Военно-космической академии им. А. Ф. Можайского, ул. Ждановская, 13, г. Санкт-Петербург, 197198, Россия.
E-mail: hillpskov@rambler.ru

References

1. Four principles of explainable artificial intelligence / P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, M. A. Przybocki. doi: 10.6028/NIST.IR.8312.

2. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. URL: arxiv.org/pdf/1811.10154v3.pdf (data obrashhenija: 25.09.2023).

3. Miller T. Explanation in artificial intelligence: Insights from the social sciences. URL: https://arxiv.org/pdf/1706.07269v3.pdf (data obrashhenija: 28.09.2023).

4. Interpretacija modelej ili kak zagljanut' v chernyj jashhik. URL: https://habr.com/ru/articles/677940/ (data obrashhenija: 18.09.23).

5. Class activation map generation by multiple level class grouping and orthogonal constraint / K. Huang, F. Meng, H. Li, Sh. Chen, Q. Wu, K. N. Ngan. URL: https://arxiv.org/pdf/1909.09839v1.pdf (data obrashhenija: 12.10.2023).

6. Learning deep features for discriminative localization / B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. URL: https://arxiv.org/pdf/1512.04150v1.pdf (data obrashhenija: 16.10.2023).

7. Repository with code snippet. URL: https://github.com/ewanytken/moduleInterpret (data obrashhenija: 26.11.2023).

8. Grad-CAM: Visual explanations from deep networks via gradient-based localization / R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. URL: https://arxiv.org/pdf/1610.02391v4.pdf (data obrashhenija: 04.10.2023).

9. Grad-CAM++: Improved visual explanations for deep convolutional networks / A. Chattopadhyay, A. Sarkar, P. Howlader, V. Balasubramanian. URL: https://arxiv.org/pdf/1710.11063v3.pdf (data obrashhenija: 11.11.2023).

10. Sanity checks for saliency maps / Ju. Adebayo, Ju. Gilmer, M. Muelly, Ian Goodfellow, M. Hardty, B. Kim. URL: https://arxiv.org/pdf/1810.03292v3.pdf (data obrashhenija: 06.11.2023).

11. Principal component analysis (PCA). URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (data obrashhenija: 01.10.2023).

12. Kolmogorov–Smirnov test. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html (data obrashhenija: 21.09.23).

13. Statistical functions. URL: https://docs.scipy.org/doc/scipy/reference/stats.html (data obrashhenija: 07.10.2023).

Information about the authors

Ivan A. Utkin – Cand. Sci. (Eng.), teacher of the Information-Analitical Department of Mozhaisky Military Space Academy, Zhdanovskaya St., 13, Saint Petersburg, 197198, Russia.
E-mail: ewanytken@mail.ru

Dmitry S. Nagorny – teacher of the Information-Analitical Department of Mozhaisky Military Space Academy, Zhdanovskaya St., 13, Saint Petersburg, 197198, Russia.
E-mail: hillpskov@rambler.ru

Статья поступила в редакцию 29.02.2024; принята к публикации после рецензирования 25.04.2024; опубликована онлайн 21.06.2024.

Submitted 29.02.2024; accepted 25.04.2024; published online 21.06.2024.