https://doi.org/10.26442/20795696.2022.6.201990

ОРИГИНАЛЬНАЯ СТАТЬЯ

Применение методов интерпретации нейронных сетей для анализа гистологических изображений патологических процессов молочной железы

А.В. Фомина[1], А.М. Борбат[2], Е.А. Карпулевич✉[3], А.Ю. Наумов[3]

[1]ФГАОУ ВО «Московский физико-технический институт (национальный исследовательский университет)», Москва, Россия;
[2]ФГБУ «Государственный научный центр Российской Федерации – Федеральный медицинский биофизический центр им. А.И. Бурназяна» ФМБА России, Москва, Россия;
[3]ФГБУН «Институт системного программирования им. В.П. Иванникова» РАН, Москва, Россия

Аннотация

Обоснование. Нейронные сети активно используются в цифровой патологии для анализа гистологических снимков и поддержки принятия врачебных решений. Популярным подходом является решение задачи классификации, где в качестве ответов модели выдают только метки классов. Однако полезно понимать, какие области изображения сильнее всего влияют на ответ модели. Эту проблему помогают решить методы интерпретации машинного обучения.

Цель. Выяснить, насколько согласованы друг с другом разные методы интерпретации нейронных сетей в задаче классификации гистологических изображений молочной железы, и получить экспертную оценку результатов исследуемых методов.

Материалы и методы. Проведены предварительный анализ и предобработка имеющегося набора данных, на которых обучены заранее выбранные нейросетевые модели. Применены существующие методы визуализации областей внимания обученных моделей на простых для понимания данных, после чего стало возможным убедиться в правильности их использования. Те же нейросетевые модели обучены на гистологических данных, и выбранные методы интерпретации применены к задаче классификации гистологических изображений, после чего получена оценка согласованности результатов использованных методов между собой и экспертная оценка результатов.

Результаты. В данной работе исследовано несколько методов интерпретации машинного обучения на примере двух различных архитектур нейронных сетей и наборе гистологических изображений патологических процессов молочной железы. Результаты обучения моделей ResNet18 и ViT-B-16 на наборе гистологических изображений на тестовой выборке: метрика Accuracy 0,89 и 0,89, метрика ROC_AUC 0,99 и 0,96 соответственно. Оценка полученных результатов проводилась экспертом при помощи инструмента Label Studio. Для каждой пары картинок эксперту предлагалось выбрать один наиболее подходящий, по его мнению, ответ – «да» или «нет» – на вопрос «Соответствует ли большинство выделенных областей классу Malignant?». Доля ответов «да» для категории ResNet_Malignant – 0,56; для ViT_Malignant – 1,0.

Заключение. Проведены эксперименты по интерпретируемости с двумя различными архитектурами: сверточной сетью ResNet18 и сетью с механизмом внимания ViT-B-16. Результаты обученных моделей визуализированы с помощью методов GradCAM и Attention Rollout соответственно. Сначала эксперименты проведены на простом для интерпретации наборе данных с целью убедиться в правильности их использования. Затем методы применены к набору гистологических изображений. На простых для понимания снимках (изображениях котов) сверточная сеть больше согласована с восприятием человека, а на гистологических изображениях рака молочной железы – наоборот, ViT-B-16 дал сильно более близкие к восприятию эксперта результаты.

Ключевые слова: цифровая патология, интерпретируемость, нейронная сеть

Для цитирования: Фомина А.В., Борбат А.М., Карпулевич Е.А., Наумов А.Ю. Применение методов интерпретации нейронных сетей для анализа гистологических изображений патологических процессов молочной железы. Гинекология. 2022;24(6):529–537. DOI: 10.26442/20795696.2022.6.201990

Введение

Современные методы машинного обучения становятся неотъемлемым инструментом в работе с медицинскими данными. Они помогают врачам быстрее и точнее анализировать снимки, записи и другие объекты. Одним из наиболее распространенных типов данных, к которым применяются методы глубокого обучения, являются изображения – медицинские снимки. Один из основных способов применения инструментов искусственного интеллекта для медицины – это помощь врачам в принятии решений путем обучения моделей и получения предсказаний. Без какого-либо объяснения этого вывода полезность модели ограничена, поскольку она не раскрывает процесс рассуждений. Интерпретируемость систем глубокого обучения помогает отследить причины ошибок моделей, а также позволяет обнаружить другую важную информацию в данных, которая могла бы остаться незамеченной.

В нашем исследовании рассматривается проблема выявления опухолевой патологии молочной железы с помощью гистологического исследования. Процедура является рутинной в медицинской диагностической практике и представляет собой получение образцов ткани методом биопсии или при оперативном вмешательстве, их технологическую обработку с целью получения срезов толщиной 5 мкм,

Информация об авторах / Information about the authors

✉**Карпулевич Евгений Андреевич** – науч. сотр. ФГБУН «ИСП РАН им. В.П. Иванникова». E-mail: karpulevich@ispras.ru; ORCID: 0000-0002-6771-2163

Фомина Анна Владимировна – студентка ФГАОУ ВО МФТИ. ORCID: 0000-0002-2269-0271

Борбат Артем Михайлович – канд. мед. наук, доц. каф. патологической анатомии ФГБУ «ГНЦ ФМБЦ им. А.И. Бурназяна». ORCID: 0000-0002-9699-8375

Наумов Антон Юрьевич – мл. науч. сотр. ФГБУН «ИСП РАН им. В.П. Иванникова». ORCID: 0000-0003-4851-7677

✉**Evgeny A. Karpulevich** – Res. Officer, Ivannikov Institute for System Programming. E-mail: karpulevich@ispras.ru; ORCID: 0000-0002-6771-2163

Anna V. Fomina – Student, Moscow Institute of Physics and Technology (National Research University). ORCID: 0000-0002-2269-0271

Artyom M. Borbat – Cand. Sci. (Med.), Russian State Research Center – Burnasyan Federal Medical Biophysical Center. ORCID: 0000-0002-9699-8375

Anton Yu. Naumov – Res. Assist., Ivannikov Institute for System Programming. ORCID: 0000-0003-4851-7677

https://doi.org/10.26442/20795696.2022.6.201990

ORIGINAL ARTICLE

Neural network interpretation techniques for analysis of histological images of breast abnormalities

Anna V. Fomina[1], Artyom M. Borbat[2], Evgeny A. Karpulevich[✉3], Anton Yu. Naumov[3]

[1]Moscow Institute of Physics and Technology (National Research University), Moscow, Russia;
[2]Russian State Research Center – Burnasyan Federal Medical Biophysical Center, Moscow, Russia;
[3]Ivannikov Institute for System Programming, Moscow, Russia

Abstract

Background. Neural networks are actively used in digital pathology to analyze histological images and support medical decision-making. A common approach is to solve the classification problem, where only class labels are the only model responses. However, one should understand which areas of the image have the most significant impact on the model's response. Machine learning interpretation techniques help solve this problem.

Aim. To study the consistency of different methods of neural network interpretation when classifying histological images of the breast and to obtain an expert assessment of the results of the evaluated methods.

Materials and methods. We performed a preliminary analysis and pre-processing of the existing data set used to train pre-selected neural network models. The existing methods of visualizing the areas of attention of trained models on easy-to-understand data were applied, followed by verification of their correct use. The same neural network models were trained on histological data, and the selected interpretation methods were used to systematize histological images, followed by the evaluation of the results consistency and an expert assessment of the results.

Results. In this paper, several methods of interpreting machine learning are studied using two different neural network architectures and a set of histological images of breast abnormalities. Results of ResNet18 and ViT-B-16 models training on a set of histological images on the test sample: accuracy metric 0.89 and 0.89, ROC_AUC metric 0.99 and 0.96, respectively. The results were also evaluated by an expert using the Label Studio tool. For each pair of images, the expert was asked to select the most appropriate answer ("Yes" or "No") to the question: "The highlighted areas generally correspond to the Malignant class." The "Yes" response rate for the ResNet_Malignant category was 0.56; for ViT_Malignant, it was 1.0.

Conclusion. Interpretability experiments were conducted with two different architectures: the ResNet18 convolutional network and the ViT-B-16 attention-enhanced network. The results of the trained models were visualized using the GradCAM and Attention Rollout methods, respectively. First, experiments were conducted on a simple-to-interpret dataset to ensure they were used correctly. The methods are then applied to the set of histological images. In easy-to-understand images (cat images), the convolutional network is more consistent with human perception; on the contrary, in histological images of breast cancer, ViT-B-16 provided results much more similar to the expert's perception.

Keywords: digital pathology, interpretability, neural network

For citation: Fomina AV, Borbat AM, Karpulevich EA, Naumov AYu. Neural network interpretation techniques for analysis of histological images of breast abnormalities. Gynecology. 2022;24(6):529–537. DOI: 10.26442/20795696.2022.6.201990

с последующим окрашиванием и сканированием (рис. 1). В результате получаются цифровые фотографии микропрепаратов (whole slide image), пригодные как для исследования врачом, так и для работы с компьютерным зрением. Задача врачей – проанализировать эти снимки и найти на них признаки патологических процессов, если таковые имеются. Для человека эта задача достаточно трудоемкая и требует много времени, а также сопряжена с известной степенью субъективности и естественным утомлением, поэтому целесообразно привлекать инструменты искусственного интеллекта для помощи в анализе таких изображений.

Для моделей машинного обучения задача ставится похожим образом: нужно найти злокачественные области на входном снимке. Однако полученные с микроскопов сканы имеют большой размер, что замедляет работу моделей, а разметка полноразмерных снимков является трудозатратной для экспертов. В силу этих причин исходные снимки часто разделяют на небольшие поля зрения (рис. 2), на которых решается задача классификации. Таким образом получается разметка целого слайда, состоящая из классифицированных полей зрения [1].

Далее под входным изображением будем понимать именно небольшое поле зрения от исходного слайда.

Модель-классификатор принимает на вход изображение и выдает в качестве ответа только метку класса, к которому принадлежит входное изображение. Однако для последующего анализа и отладки моделей полезно понимать, какие объекты на снимке больше всего повлияли на ответ. Выделить эти объекты помогают различные методы интерпретации алгоритмов машинного обучения. Чаще всего для классификации изображений используют нейронные сети, поэтому в данной работе рассмотрены методы интерпретации именно для нейронных сетей.

В контексте гистологии проблема интерпретации нейронных сетей освещена далеко не для всех архитектур и методов, поэтому важно проводить дополнительные исследования для поиска наилучших решений.

Задача классификации

Задача классификации в машинном обучении ставится следующим образом. На вход модели подается изображение (или другой объект), на выходе получается метка класса, к которому оно принадлежит. Множество классов определяется заранее, и каждому изображению из набора данных присваивается истинная метка класса. В данной работе рассматривается задача бинарной классификации, т.е. на два класса.

Задачи бинарной классификации могут отличаться постановкой. Назовем задачами 1-го типа такие, в которых требуется определить, к какому из двух классов относится объект, изображенный на снимке. Пример: определить, кто на фото – кошка или собака. При этом данные должны состоять только из снимков кошек и собак. Задачами 2-го типа назовем такие, в которых требуется определить, присутствует ли на снимке объект одного определенного класса. Пример: определить, есть ли на фото кошка. Данные для такой задачи состоят из снимков кошек и снимков с любыми другими объектами, на которых нет кошек. В нашей работе рассмотрена задача 2-го типа.

Сверточные нейронные сети. ResNet

Нейронная сеть – это математическая модель, построенная по принципу работы нервных систем живых организ-

Рис. 1. Схема проведения гистологического анализа.
Fig. 1. Diagram of histological analysis.

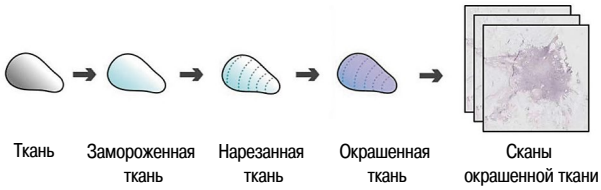
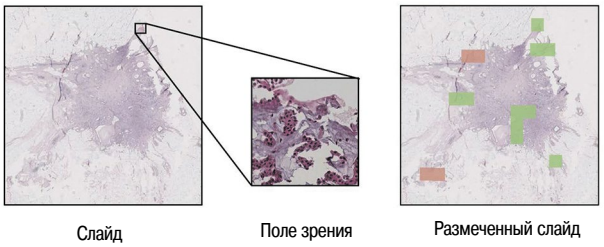
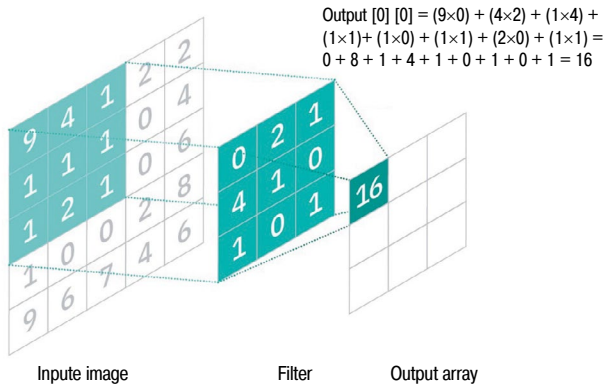


Рис. 2. Выделение полей зрения на слайде. На последнем рисунке зеленым выделены участки, классифицированные как доброкачественные, красным – злокачественные.
Fig. 2. Highlight the fields of view on the slide. In the last figure, the areas categorized as benign are highlighted in green and malignant in red.



Примечание. Данная разметка является искусственным примером, иллюстрирующим метод.

Рис. 3. Операция свертки. Изображение взято из источника [3].
Fig. 3. Convolution operation. Image taken from [3].

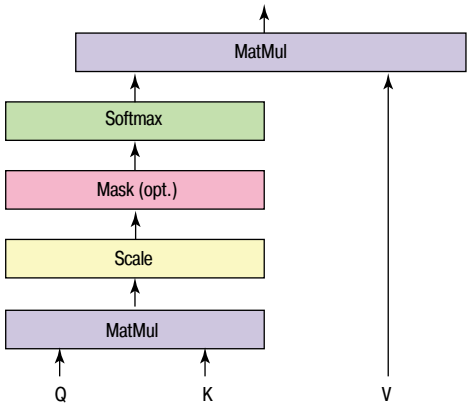


мов, и ее программная реализация. В самом простом случае нейронная сеть включает несколько слоев, каждый из которых состоит из вычислительных единиц – нейронов. Каждый нейрон с предыдущего слоя передает информацию всем нейронам следующего слоя с некоторыми весами. Эти веса настраиваются во время обучения. Таким образом сеть учится предсказывать ответ на поставленную задачу по входным данным.

В задачах классификации изображений наиболее часто используются сверточные нейронные сети. Введем необходимые определения:

Операция свертки [2] – это операция вычисления нового значения заданного пикселя, в которой учитываются значения его соседних пикселей. Принцип работы показан на рис. 3 [3].

Рис. 4. Схема блока внимания. Изображение взято из статьи [5].
Fig. 4. Attention block diagram. Image taken from [5].



Сверточный слой [2] – это слой нейронной сети, в котором к входным данным применяется операция свертки с некоторым шагом. Такие слои обычно состоят из нескольких сверток, каждая из которых отвечает за свой канал.

Сверточная нейронная сеть [2] – это нейронная сеть для работы с изображениями, в архитектуре которой содержатся сверточные слои.

Соединение быстрого доступа [4] – это соединение между несоседними слоями сети, которое складывает результаты этих слоев.

Полносвязный слой – это слой, каждый выходной нейрон которого связан со всеми входными нейронами.

Одна из самых популярных и эффективных моделей в задаче классификации изображений – это ResNet [4].

ResNet [4] – это архитектура сверточной нейронной сети, состоящей из сверточных слоев, соединений быстрого доступа и полносвязного слоя. У этой сети несколько версий: ResNet18, ResNet34, ResNet50, которые отличаются глубиной, т.е. количеством сверточных слоев.

Нейронные сети с механизмом внимания. Vision Transformer

Механизм внимания [5] – это техника, используемая в нейронных сетях для поиска взаимосвязей между различными частями входных данных. Изначально она применялась только для анализа текстов, однако выяснилось, что этот метод подходит и для работы с изображениями.

Проиллюстрируем принцип работы на примере: допустим, есть набор входных векторов v_i. Значение внимания вычисляется следующим образом. Сначала из матрицы v, составленной из векторов v_i, получаются три матрицы: значение V, ключ K и запрос Q, умножая v на матрицы весов W_V, W_K и W_Q соответственно. Эти веса выучиваются в процессе обучения модели. Далее перемножаются матрицы Q и K_T, таким образом, учитывается связь каждого ключа со всеми запросами. Полученное произведение делится на $\sqrt{d_k}$, где dk – размерность матриц Q и K, однако возможны и другие варианты. Затем применяется функция softmax, и результат перемножается с матрицей V (формула 1).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Процедура вычисления внимания реализована в виде блока внимания – это структурная единица модели, которая принимает на вход набор значений, ключей и запросов и считает взвешенную сумму значений с весами, кото-

https://doi.org/10.26442/20795696.2022.6.201990

рые получаются из ключей и запросов (рис. 4, формула 1). Таким образом вычисляется степень важности друг для друга ключей и значений, в роли которых могут быть закодированные слова или части изображения.

Vision Transformer [6] – это архитектура нейронной сети, состоящей из нескольких слоев параллельных блоков внимания, соединений быстрого доступа и полносвязного слоя. Параллельно используемые блоки внимания часто называют головами трансформера.

Материалы и методы

Методы интерпретации нейронных сетей

В данной работе рассмотрены два основных метода интерпретации для соответствующих архитектур: GradCAM для ResNet и Attention Rollout для ViT. Оба эти метода принимают на вход обученную модель и изображение и выдают тепловую карту, на которой более высокие значения соответствуют большему вкладу пикселей в предсказание модели. Участки с наиболее высокими значениями на тепловых картах будем называть «областями внимания» модели для данного изображения.

GradCAM

Gradient Class Activation Mapping (GradCAM) [7] – это метод интерпретации нейронных сетей, который использует информацию о градиенте, поступающую в последний слой модели для присвоения значения важности каждого нейрона для интересующего нас класса. Метод GradCAM принимает на вход метку целевого класса и слои, для которых будут рассчитаны градиенты. На выходе вычисляется маска, на которой большие значения соответствуют большему влиянию пикселей на ответ сети. Метод работает по следующим формулам: сначала считаются веса α_k^c (2) для класса c и канала k, затем L_GradCAM^c (3).

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2)$$

$$L_{GradCAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (3)$$

Здесь $1/Z\ \Sigma_i\Sigma_j$ – это операция глобального среднего пулинга, y^c – выход сети до softmax для класса c, A^k – матрица активаций слоя k.

Attention Rollout

Attention Rollout [8] – это метод интерпретации для моделей архитектуры Vision Transformer (см. секцию 1.4), который рекурсивно перемножает карты внимания из каждого слоя и агрегирует их значения по головам одним из следующих способов: берет максимальные значения, средние или минимальные. Выбран способ, который используется в оригинальной статье – брать среднее значение по головам трансформера. Метод работает по следующей формуле:

$$V_{l+1}=(W+I)\ V_l \quad (4)$$

Здесь V_l – маска внимания для слоя l, W – матрица внимания, I – единичная матрица, которую необходимо прибавить к W, чтобы учесть соединения быстрого доступа между слоями.

Ниже Rollout будет означать то же, что Attention Rollout.

Обзор существующих решений

Для интерпретации нейронных сетей, классифицирующих гистологические снимки, существует большое количество различных методов. В данной работе рассматриваются только те из них, которые с помощью тепловых карт выделяют на входных изображениях области, в том или ином смысле повлиявшие на ответы обученных моделей. Напомним, что под входными изображениями подразумеваются небольшие поля зрения исходных снимков ткани.

Рис. 5. Примеры изображений.

Fig. 5. Examples of images.

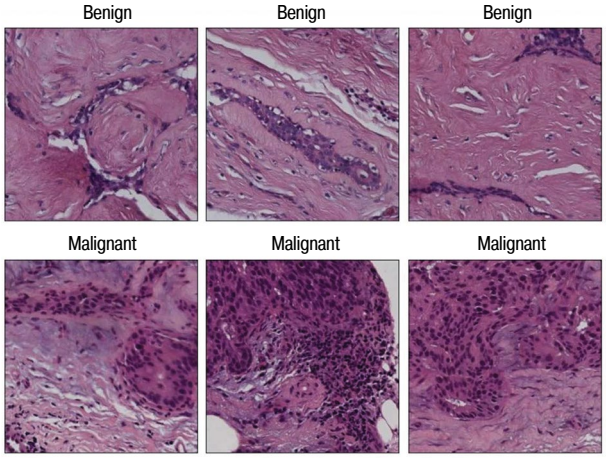
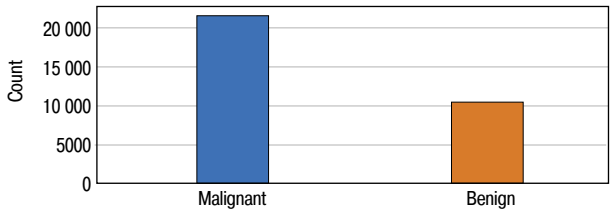


Рис. 6. Распределение данных по классам.

Fig. 6. Distribution of data by class.



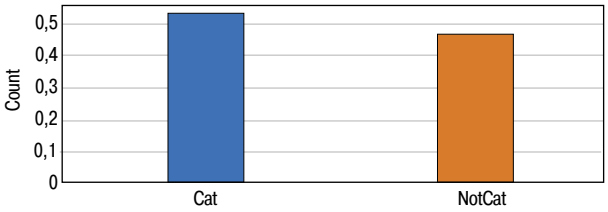
Интерпретация сверточных сетей в задаче классификации гистологических снимков рака кожи

В статье [9] решается задача классификации образований на злокачественные – меланомы, и доброкачественные – невусы. В качестве набора данных использовались патчи размером 256×256 пикселей, полученные из целых слайдов. Для обучения выбраны модели Vgg19, состоящая из 19 сверточных слоев, и ResNet50, состоящая из 50 сверточных слоев и быстрых соединений. В ходе обучения достигнуты достаточно высокие значения метрики: AUC=0,98 для ResNet50, и AUC=0,95 для Vgg19.

Для интерпретации полученных результатов использовали метод Classification Activation Map – CAM (карта активаций классификации) [10]. CAM – это метод, который генерирует тепловую карту активации класса для обученной сверточной сети. Созданная тепловая карта представляет собой двумерную дробную сетку, связанную с определенной категорией выходных данных. Метод вычисляет, насколько важна каждая позиция входного изображения для конкретного класса. Например, для обученной нейронной сети, которая различает меланому и невус, для каждого входного изображения с помощью CAM-визуализации может быть сгенерирована тепловая карта, показывающая, насколько каждая часть изображения похожа на особенности меланомы.

Рис. 7. Примеры изображений.
Fig. 7. Examples of images.

Рис. 8. Распределение данных по классам.
Fig. 8. Distribution of data by class.



В выводе статьи [9] сказано, что выделенные области внимания сетей почти пересекаются с областями внимания экспертов-морфологов, однако не приведено никаких численных результатов.

Описанное решение очень близко к нашему, здесь используются сверточная нейронная сеть ResNet и метод интерпретации CAM, который схож с методом GradCAM, но в отличие от него не вычисляет градиенты с нижних слоев. Другое отличие – в типе тканей: в приведенной статье используются ткани кожных покровов. В статье также нет численных результатов сравнения областей внимания сетей и экспертов.

Построение интерпретируемой модели для задачи классификации гистологических слайдов рака груди

В статье [11] описан метод интерпретации двух построенных на основе архитектуры AlexNet моделей. В качестве данных используются целые гистологические слайды из набора TCGA-BRCA [12]. Решение состоит из двух этапов: на I этапе слайд разделяется на патчи меньшего размера (аналогично нашим полям зрения), из которых первая сверточная сеть выделяет наиболее важные, т.е. те, в которых содержатся признаки доброкачественных или злокачественных образований. Затем на отобранных патчах размером 224×224 обучается вторая сверточная сеть, которая определяет стадию заболевания. Первая сеть обучилась до точности примерно 0,82, вторая – до точности 0,45.

Таблица 1. Значения метрик для ResNet18
Table 1. Metric values for ResNet18

Выборка	Accuracy	ROC AUC
Валидационная	0,9600	0,9936
Тестовая	0,9235	0,9931

Таблица 2. Значения метрик для ViT-B-16
Table 2. Metric values for ViT-B-16

Выборка	Accuracy	ROC AUC
Валидационная	0,9437	0,9858
Тестовая	0,9235	0,9790

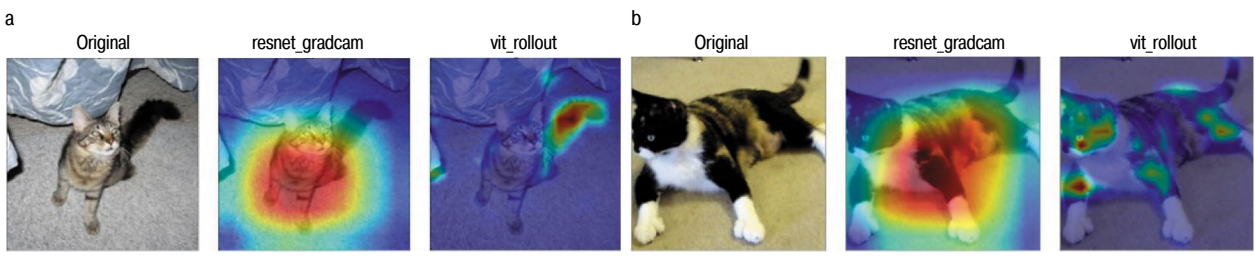
С помощью метода CAM [10] авторы генерируют тепловые маски для патчей, на которых обучалась сеть, и таким образом выделяют области наибольшего «интереса» модели. Из выделенных областей извлекаются 3 типа признаков: цвет, размер клеток и структура ткани. Результаты данного исследования показывают, что модели действительно «обращают внимание» на выделенные признаки, а значит, поддаются интерпретации.

Описанный метод, в отличие от нашего, автоматизирует отбор значимых полей зрения – из нашего набора данных уже исключены экспертом неинформативные изображения. Метод использует архитектуры исключительно на основе сверточных нейронных сетей. В данном исследовании также проводилась оценка согласованности с мнением экспертов, по результатам которой сделан вывод о частичном пересечении некоторых морфологических признаков.

Выводы

В найденных статьях на тему интерпретации нейронных сетей в задаче классификации гистологических снимков используются преимущественно сверточные сети и такие методы интерпретации, как CAM и GradCAM. В этих статьях не исследована интерпретируемость архитектур на основе

Рис. 9. Визуализация ResNet18+GradCAM и ViT-B-16+Rollout: *a* – пример 1; *b* – пример 2. Теплые цвета означают большее внимание сети к выделенной области, холодные – меньшее.
Fig. 9. Visualization ofResNet18+GradCAM and ViT-B-16+Rollout: *a* – example 1; *b* – example2. Warm colors indicate more network attention to the highlighted area, and cold colors indicate less attention.



https://doi.org/10.26442/20795696.2022.6.201990

Рис. 10. Визуализация методов на гистологических снимках: *a* – класс Benign; sIoU=0,1575; *b* – класс Malignant; sIoU=0,4116.
Fig. 10. Visualization of methods on histological images: *a* – class: Benign; sIoU=0.1575; *b* – class: Malignant; sIoU=0.4116.

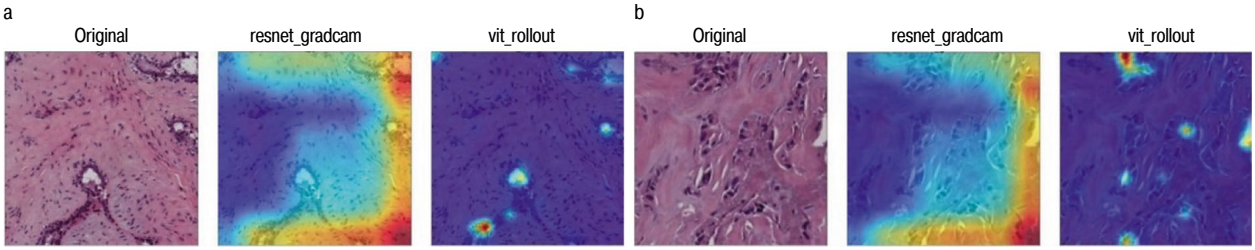
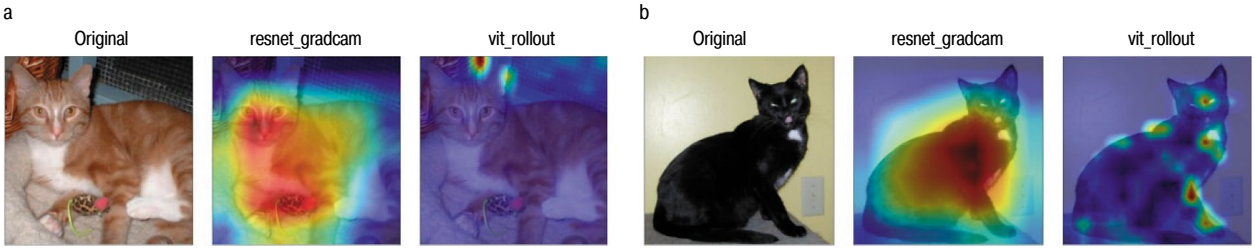


Рис. 11. Примеры значений функции sIoU для изображений класса Cat.
Fig. 11. Examples of sIoU function values for Cat class images.



трансформеров, а также отсутствуют численные результаты сравнения с мнением экспертов.

Результаты

Обзор набора данных

Данные для обучения взяты из первого российского набора данных гистологических изображений патологических процессов молочной железы [13]. Набор состоит из изображений небольших полей зрения – частей больших сканов образцов тканей (рис. 5). Для каждого изображения экспертами зафиксированы метки классов, к которым принадлежат соответствующие поля зрения.

В работе используются только те изображения, которые получены с помощью сканера с увеличением ×10.

Набор содержит 32 260 изображений:

- 22 370 – размером 500×500 пикселей;
- 9890 – размером 300×300 пикселей.

Для обучения изображения приводились к одному размеру 256×256.

В качестве классов для задачи классификации использовались метки категории «Dia2» из статьи [13]: «Benign» – доброкачественный участок, «InSitu» – неинвазивная опухоль, «Invasive» – инвазивная опухоль. Однако в силу вырожденности второго класса «InSitu» и «Invasive» объединены в один класс «Malignant» – злокачественный участок. Таким образом, поставлена задача бинарной классификации изображений на те, которые содержат признаки злокачественной опухоли, и те, которые не содержат такие признаки.

- Изображений класса «Benign» – 10 516.
- Изображений класса «Malignant» – 21 744

Заметим, что выборка не сбалансирована по классам (рис. 6).

Обучение моделей и визуализация их областей внимания на легко интерпретируемом наборе данных

Напомним, что эксперимент с набором простых для интерпретации данных решено провести, чтобы убедиться в правильности использования исследуемых методов.

Таблица 3. Значения метрик для ResNet18
Table 3. Metric values for ResNet18

Выборка	Accuracy	ROC AUC
Валидационная	0,9482	0,9874
Тестовая	0,8936	0,9919

Таблица 4. Значения метрик для ViT-B-16
Table 4. Metric values for ViT-B-16

Выборка	Accuracy	ROC AUC
Валидационная	0,8915	0,9628
Тестовая	0,8936	0,9628

Обзор вспомогательного набора данных

В качестве такого набора взяты изображения кошек из набора Cats vs Dogs [14] и изображения мест из набора Places [15], в которых нет кошек. Набор состоит из 23 272 картинок, принадлежащих к одному из двух классов:

- «NotCat» – на картинке отсутствует кот;
- «Cat» – на картинке присутствует кот.

Примеры картинок – на рис. 7.

Изображений класса «NotCat» – 10 867, изображений класса «Cat» – 12 405. Распределение по классам показано на рис. 8.

Таким образом, поставлена задача бинарной классификации, аналогичная задаче на гистологических снимках.

Выбор и обучение моделей. Визуализация

Для данной задачи выбраны непредобученные модели ResNet18 и ViT-B-16 (см. секции 1.3, 1.4). Веса инициализированы случайным образом. Подробнее о выборе моделей и обучении см. раздел «Описание практической части».

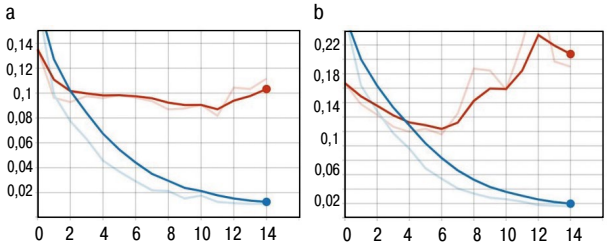
Данные разбиты на 3 части: тренировочная, валидационная и тестовая в соотношении 0,8:0,1:0,1 соответственно.

Метрики, достигнутые в ходе обучения моделей, приведены в табл. 1, 2.

Из табл. 1, 2 видно, что модели достигли достаточно высокого качества, ResNet18 достиг чуть более высоких значений

Рис. 12. Кривые обучения моделей на простом наборе данных: *a* – ResNet18; *b* – ViT-B-16.

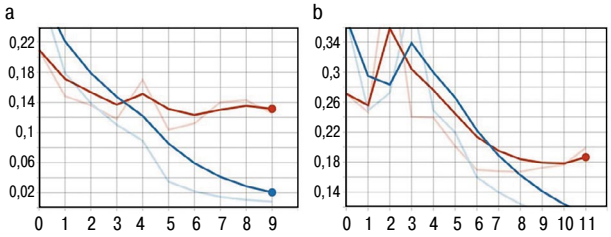
Fig. 12. Model training curves on a simple dataset: *a* – ResNet18; *b* – ViT-B-16.



Примечание. Здесь и на рис. 13: голубым обозначена тренировочная часть, красным – валидационная.

Рис. 13. Кривые обучения моделей на гистологическом наборе данных: *a* – ResNet18; *b* – ViT-B-16.

Fig. 13. Model training curves on histological dataset: *a* – ResNet18; *b* – ViT-B-16.



метрик, чем ViT-B-16, но разница незначительная. Подробнее о подборе гиперпараметров и версий моделей, а также их обучении см. раздел «Описание практической части».

Для визуализации выбраны методы GradCAM и Rollout (см. секции 1.5.1, 1.5.2).

Из примеров (рис. 9) видно, что ResNet18+GradCAM склонен выделять объект целиком, а ViT-B-16+Rollout делает акцент на деталях искомого объекта.

Обучение моделей на гистологических изображениях. Визуализация результатов

Результаты обучения моделей ResNet18 и ViT-B-16 на наборе гистологических изображений приведены в табл. 3, 4.

Из табл. 3, 4 видно, что модели также достигли неплохого качества, однако оно ниже, чем у тех же моделей на простом наборе данных. Это можно связать с большей зашумленностью гистологических данных.

Пример на рис. 10 аналогично показывает, что ViT-B-16 больше сосредоточен на небольших деталях изображения, тогда как ReNet18 «ищет» объект целиком.

Оценка результатов визуализации

Оценка полученных результатов в нашей работе делится на две части: сравнение карт внимания от ResNet18 и ViT-B-16 между собой, а также оценка согласованности с интерпретацией независимых экспертов.

Сравнение карт внимания

Для реализации сравнения карт внимания используется функция sIoU [16] (5):

$$sIoU(A^1, A^2) = \frac{\sum_{ij} min(A^1_{ij}, A^2_{ij})}{\sum_{ij}(A^1_{ij} + A^2_{ij})} \quad (5)$$

Здесь A^1 и A^2 – маски, которые планируется сравнить. Значения функции лежат в диапазоне [0, 1], чем меньше пересечение высоких значений масок, тем ниже значение функции.

Использование этой функции обосновывается тем, что поставлена задача получить численную оценку согласованности методов между собой. Высокая согласованность могла бы помочь в поиске общих паттернов в изображениях или других целях дальнейших исследований. Привычным способом сравнения выделенных областей в машинном обучении является функция IoU – Intersection over Union, которая обычно используется в задачах детекции в качестве метрики. Однако эта функция предназначена только для бинарных масок. Поэтому выбрана ее версия, адаптированная для небинарных карт, – sIoU. Примеры распределения значений функции sIoU приведены на рис. 11.

Значение sIoU посчитано для 200 картинок из тестовой выборки простого набора данных и 200 для гистологического датасета. На первом наборе среднее значение функции составило 0,33. На втором наборе – 0,21. Эти значения подтверждают наше предположение о том, что методы плохо согласованы между собой. Причем на гистологических изображениях разница усиливается.

Экспертная оценка

Оценка полученных результатов проводилась экспертом при помощи инструмента Label Studio. Создано 2 проекта: ViT_Malignant и ResNet_Malignant – названия составлены из модели и класса, который она предсказала для изображения. В каждом проекте содержится 100 пар изображений: оригинал и изображение с наложенной маской. Для каждой пары картинок эксперту предлагалось выбрать один наиболее подходящий, по его мнению, ответ – «да» или «нет» – на вопрос «Соответствует ли большинство выделенных областей классу Malignant?».

Доля ответов «да» для категории ResNet_Malignant – 0,56; для ViT_Malignant – 1,0.

Аналогично проведена оценка визуализации на простом датасете, для категории ResNet_Cat получено 0,98 положительного ответа, для ViT_Cat – 0,72.

Заметим, что на наборе изображений класса Cat доля положительных ответов для ResNet18 больше, чем для ViT-B-16. Это означает, что области наибольшего внимания человека больше пересекаются с областями внимания ResNet18. Однако для гистологических снимков наблюдается противоположная картина: ViT-B-16 дает наиболее близкие к восприятию эксперта результаты.

Описание практической части

Практическая часть реализована на языке Python с помощью jupyter notebook.

Выбор версий моделей, обучение и подбор гиперпараметров

В ходе эксперимента обучены сверточные модели ResNet18, ResNet34, ResNet50. Использование более глубоких архитектур, чем ResNet18, не дало улучшения в качестве, поэтому решено оставить наименее глубокую версию.

Аналогично из семейства архитектур ViT-B-16, ViT-B-32, ViT-L-16, ViT-L-32, которые отличаются глубиной и количеством голов внимания, выбрана архитектура ViT-B-16, так как более глубокие не давали явного улучшения качества.

В качестве финального слоя модели выбран полносвязный слой с 1 выходом, из которого получается величина вероятности присутствия объекта на входном изображении.

Модели обучались на тренировочной части набора данных. Во время обучения после завершения каждой эпохи

сохранялись значения лосса и некоторых метрик с помощью инструмента tensorboard [17]. На рис. 12, 13 приведены кривые обучения моделей.

Для ResNet18 на простом датасете выбрана 11-я эпоха, для ViT-B-16 – 6-я эпоха. На гистологическом наборе данных для ResNet18 на простом датасете выбрана 5-я эпоха, для ViT-B-16 – 6-я эпоха. Выбор эпох осуществлялся по кривой обучения на валидационной части набора данных: бралась низшая точка, так как далее модель переобучается.

Графики обучения моделей на простом наборе данных.

Реализация методов

Модели и инструменты их обучения реализованы с помощью библиотеки Pytorch. Для обучения моделей использовались следующие функции и объекты:

- ImageDataset – класс датасета, унаследованный от Dataset из библиотеки torch.utils.data, предназначенный для работы с гистологическими изображениями;
- CatsDataset – класс датасета для работы с изображениями из простого набора данных;
- init_weights – функция для инициализации весов моделей перед обучением. В данной функции использовался метод torch.nn.init.xavier_uniform;
- tensorboard [17] – инструмент для логирования промежуточных результатов (метрик, функции потерь) во время обучения;
- BCEWithLogitsLoss – функция из модуля torch.nn для расчета функции потерь;
- optim.Adam – использовался в качестве оптимизатора при обучении.

Обучение проводилось на GPU – NVIDIA Corporation GP104.

Для реализации методов интерпретации использовались следующие модули:

- pytorch_gradcam_master, взятый из репозитория [18];
- vit_rollout, взятый из репозитория [19].

Заключение

В целях исследования проблемы интерпретации методов глубокого обучения на медицинских данных в статье проведены эксперименты с двумя различными архитектурами: сверточной сетью ResNet18 и сетью с механизмом внимания ViT-B-16. Результаты обученных моделей визуализированы с помощью методов GradCAM и Attention Rollout соответственно. Сначала эксперименты проведены на простом для интерпретации наборе данных с целью убедиться в правильности их использования. Затем методы применены к набору гистологических изображений. В результате получены наборы изображений с тепловыми картами, которые сравнены между собой при помощи функции sIoU, чтобы узнать, насколько согласованы методы между собой. По результатам сравнения сделан вывод, что на обоих наборах данных области внимания ResNet18 и ViT-B-16 имеют пересечения, но сильно различаются по объему выделенных участков. Далее проведена экспертная оценка исследуемых методов интерпретации, которая показала, что на простых для понимания снимках (изображениях котов) сверточная сеть больше согласована с восприятием человека. А на гистологических изображениях – наоборот, ViT-B-16 дал сильно более близкие к восприятию эксперта результаты.

Разницу между картами двух методов можно объяснить склонностью GradCAM для ResNet выделять объекты целиком, тогда как Attention Rollout для ViT, как правило, выделяет небольшие детали объекта. Первый подход лучше работает на привычных человеку изображениях, однако для интерпретации в задаче классификации гистологических снимков больше подходит второй.

Раскрытие интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Disclosure of interest. The authors declare that they have no competing interests.

Вклад авторов. Авторы декларируют соответствие своего авторства международным критериям ICMJE. Все авторы в равной степени участвовали в подготовке публикации: разработка концепции статьи, получение и анализ фактических данных, написание и редактирование текста статьи, проверка и утверждение текста статьи.

Authors' contribution. The authors declare the compliance of their authorship according to the international ICMJE criteria. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published and agree to be accountable for all aspects of the work.

Источник финансирования. Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (соглашение №075-15-2022-294 от 15 апреля 2022 г.).

Funding source. This work was supported by the Ministry of Science and Higher Education of the Russian Federation, agreement №075-15-2022-294 dated 15 April 2022.

Литература/References

1. Hou L, Samaras D, Kurc TM, et al. Patch-based convolutional neural network for whole slide tissue image classification. *arXiv.* 2016;1504.07947.
2. O'Shea K, Nash R. An introduction to convolutional neural networks. *arXiv.* 2015;1511.08458.
3. ROBINVC. Popular ML/NN/CNN/RNN Model code snippets. Available at: https://www.kaggle.com/code/nsff591/popular-ml-nn-cnn-rnn-model-code-snippets/notebook. Accessed: 9.11.2022.
4. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *arXiv.* 2015;1512.03385.
5. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *arXiv.* 2017;1706.03762.
6. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv.* 2021;2010.11929.
7. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv.* 2019;1610.02391.
8. Abnar S, Zuidema W. Quantifying attention flow in transformers. *arXiv.* 2020;2005.00928.
9. Xie P, Zuo K, Zhang Y, et al. Interpretable classification from skin cancer histology slides using deep learning: A retrospective multicenter study. *arXiv.* 2019;1904.06156.
10. Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization. *arXiv.* 2015;1512.04150.
11. Srivastava A, Kulkarni C, Huang K, et al. Imitating pathologist based assessment with interpretable and context based neural network modeling of histology images. *Biomed Inform Insights.* 2018;10:1178222618807481.
12. Thennavan A, Beca F, Xia Y, et al. Molecular analysis of TCGA breast cancer histologic types. *Cell Genom.* 2021;1(3):100067.
13. Борбат А.М., Лищук С.В. Первый российский набор данных гистологических изображений патологических процессов молочной железы. *Врач и информационные технологии.* 2020;3:25-30 [Borbat AM, Lishchuk SV. The first russian breast pathology histologic images data set. *Vrach i informatsionnie tekhnologii.* 2020;3:25-30 (in Russian)].
14. Golle P. Machine learning attacks against the Asirra CAPTCHA. Proceedings of the 15th ACM conference on Computer and communications security. 2008:535-42.

15. Bitton A, Esling P. ATIAM 2018-ML Project Regularized auto-encoders (VAE/WAEs) applied to latent audio synthesis. Available at: https://esling.github.io/documents/mlProj_bitton.pdf. Accessed: 9.11.2022.
16. Wang L, Wu Z, Karanam S, et al. Reducing visual confusion with discriminative attention. *arXiv.* 2019;1811.07484.
17. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. *arXiv.* 2016;1603.04467.
18. Gildenblat J. Class Activation Map methods implemented in Pytorch. Available at: https://github.com/jacobgil/pytorch-grad-cam. Accessed: 9.11.2022.
19. Gildenblat J. Explainability for Vision Transformers (in PyTorch). Available at: https://github.com/jacobgil/vit-explain. Accessed: 9.11.2022.

Статья поступила в редакцию / The article received: 13.11.2022
Статья принята к печати / The article approved for publication: 16.12.2022