

DOI: 10.15514/ISPRAS-2026-38(6)-20



Интерпретируемость визуального трансформера при классификации бактерий на изображениях сканирующей электронной микроскопии

¹В.Н. Гридин, ORCID: 0000-0002-6361-9113 <info@ditc.ras.ru>¹Б.Р. Салем, ORCID: 0000-0003-3348-9407 <isub97@gmail.com>¹В.И. Солодовников, ORCID: 0000-0001-5533-214X info@ditc.ras.ru^{1,2}И.А. Новиков, ORCID: 0000-0003-4898-4662 <i.novikov@niigb.ru>²Е.С. Родина, ORCID: 0000-0002-5525-1793 <liza.akatova.la@gmail.com>¹ Центр информационных технологий в проектировании РАН, Россия, 143003 Московская область, г. Одинцово, ул. Маршала Бирюзова д. 7А.² Научно-исследовательский институт глазных болезней имени М.М. Краснова, Россия, 119021, г. Москва, ул. Россолимо, дом 11, корпус А, корпус Б.

Аннотация. Рассматривается задача многоклассовой однометочной классификации бактериальных микроорганизмов на одноканальных изображениях, полученных методом сканирующей электронной микроскопии с лантаноидным контрастированием. Цель работы состоит в сопоставлении трёх способов интерпретации решения Vision Transformer: базовой карты внимания последнего блока (AttentionCAM, CLS→фрагменты), класс-специфической карты CDAM и метода R-Cut, формирующего связную область локализации. Анализ выполнен для используемой в работе модели ViT-Base/16 адаптированной к одноканальным изображениям и работающей с входом 384×384. Для оценки классификационных показателей использована независимая тестовая подвыборка из 371 изображения. На ней получены следующие интегральные показатели: доля верных классификаций 0.784, макроусреднённая F1-мера 0.790 и взвешенная F1-мера 0.785. Для количественного сопоставления объяснений без ручной масочной разметки использована отдельная стратифицированная подвыборка из 160 корректно классифицированных изображений. На этой подвыборке AttentionCAM показал лучшие средние значения по площади под кривыми поэтапного удаления и добавления значимых фрагментов, по среднему снижению целевого отклика и по доле интенсивности карты внутри автоматически выделенной области объекта. В то же время CDAM и R-Cut дают иной тип интерпретации: первый выделяет локальные дискриминативные участки, второй стремится к формированию более связной области, однако чувствителен к выбору функции нормировки выходов модели и параметра ϕ .

Ключевые слова: Vision Transformer; интерпретируемость; карты внимания; класс-специфическая атрибуция; связная локализация; классификация бактерий; сканирующая электронная микроскопия.

Для цитирования: Гридин В.Н., Салем Б.Р., Солодовников В.И., Новиков И.А., Родина Е.С. Интерпретируемость визуального трансформера при классификации бактерий на изображениях сканирующей электронной микроскопии. Труды ИСП РАН, 2026, том 38, вып. 6, часть 1, стр. 311–322 (на английском языке). DOI: 10.15514/ISPRAS-2026-38(6)-20.

Благодарности: Исследование выполнено в рамках темы государственного задания № FFNR-2024-0003.

Interpretability of Vision Transformer in Bacterial Classification on SEM Images

¹V.N. Gridin, ORCID: 0000-0002-6361-9113 <info@ditc.ras.ru>¹B.R. Salem, ORCID: 0000-0003-3348-9407 <isub97@gmail.com>¹V.I. Solodovnikov, ORCID: 0000-0001-5533-214X info@ditc.ras.ru^{1,2}I.A. Novikov, ORCID: 0000-0003-4898-4662 <i.novikov@niigb.ru>²E.S. Rodina, ORCID: 0000-0002-5525-1793 <liza.akatova.la@gmail.com>¹ Design Information Technologies Center, Russian Academy of Sciences, 7A, Marshal Biryuzov st., Odintsovo, Moscow region, 143003, Russia.² Krasnov Research Institute of Eye Disease, 11A, B, Rossolimo st., Moscow, 119021, Russia.

Abstract. The paper addresses multiclass single-label classification of bacterial microorganisms on single-channel images obtained by scanning electron microscopy with lanthanide contrast enhancement. The study compares three approaches to interpreting Vision Transformer decisions: the baseline attention map of the last block (AttentionCAM, CLS-to-patches), the class-specific CDAM map, and the R-Cut method, which yields a connected localization region. The analysis was carried out for the current version of the ViT-Base/16 model operating on 384×384 images. An independent test subsample of 371 images was used to evaluate classification performance. The following integrated performance indicators were obtained: a correct classification rate of 0.784, a macro-averaged F1 score of 0.790, and a weighted F1 score of 0.785. For quantitative comparison of explanations without manual mask annotation, a stratified subset of 160 correctly classified images was used. On this subset, AttentionCAM showed the best mean values in terms of the area under the deletion and insertion curves, the average decrease in the target response, and the proportion of map intensity inside the automatically extracted object region. At the same time, CDAM and R-Cut provide a different type of interpretation: the former highlights local discriminative regions, whereas the latter tends to form a more connected localization area, although it is sensitive to the choice of output normalization and the parameter ϕ .

Keywords: Vision Transformer; interpretability; attention maps; class-specific attribution; connected localization; bacterial classification; scanning electron microscopy.

For citation: Gridin V.N., Salem B.R., Solodovnikov V.I., Novikov I.A., Rodina E.S. Interpretability of Vision Transformer in bacterial classification on SEM images. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 6, part 1, pp. 311-322. DOI: 10.15514/ISPRAS-2026-38(6)-20.

Acknowledgements. This study was supported as a part of state assignment topic no. FFNR-2024-0003.

1. Введение

В задачах биомедицинского анализа изображений высокие интегральные метрики качества не гарантируют, что модель принимает решение по содержательно релевантным признакам. При наличии фоновых артефактов, неоднородной пробоподготовки и шумовых структур классификатор может опираться на побочные корреляции, которые не связаны с морфологией объекта. Поэтому анализ интерпретируемости в данной постановке выступает не дополнительной иллюстрацией, а необходимым этапом проверки корректности поведения модели.

Архитектура Vision Transformer [1] предоставляет доступ к внутренним матрицам самовнимания и тем самым позволяет сопоставлять несколько уровней объяснения: локализацию внимания, класс-специфическую атрибуцию и локализацию связной области. Вместе с тем сами карты внимания не следует отождествлять с причинным объяснением решения, поэтому в работе намеренно сравниваются методы, различающиеся по своей интерпретационной роли.

Цель настоящей работы состоит в сравнительном анализе трёх способов интерпретации решений модели Vision Transformer при классификации бактерий на одноканальных

изображениях, сделанных сканирующим электронным микроскопом (СЭМ): базовой карты внимания последнего блока, класс-специфической карты CDAM и метода R-Cut, формирующего связную область локализации. Для этого рассматривается актуальная постановка ViT-384, приводятся результаты полного прогона модели по текущему набору данных и выполняется количественное сопоставление методов без привлечения ручной масочной разметки.

2. Обзор литературы

Задача интерпретируемости нейронных сетей в медицинской визуализации активно исследуется в последние годы. В обзоре [2] систематизированы методы интерпретации и визуализации для моделей глубокого обучения в медицинской визуализации, рассмотрены метрики их оценки, а также выделены текущие вызовы и перспективные направления. Авторы указывают, что выбор метода интерпретации должен определяться типом решаемой задачи и характером данных, поскольку универсального подхода не существует. В частности, для задач с ограниченным числом обучающих примеров и специфичными доменными признаками методы, опирающиеся на градиентную информацию, могут давать нестабильные объяснения.

Вопрос адекватности методов интерпретации был поставлен в работе [3], где предложены так называемые “проверки вменяемости” (sanity checks) карт значимости: если карта объяснения существенно не меняется при замене весов модели случайными значениями или при перемешивании целевых меток, то она отражает свойства входных данных, а не поведение конкретной модели. Этот подход ставит под сомнение ряд широко применяемых методов объяснения и показывает, что визуальное правдоподобие карты не является достаточным критерием её качества.

Наиболее широко применяемым методом интерпретации для свёрточных сетей остаётся GradCAM [4], который вычисляет карту значимости как взвешенную сумму карт признаков, где веса определяются через градиент целевого класса по соответствующему слою. Его преимуществом является класс-специфичность и вычислительная эффективность. В контексте медицинских изображений GradCAM применялся для задач анализа рентгенограмм, гистологических препаратов и МРТ-снимков. Методы LIME [5] и SHAP [6], основанные на локальной аппроксимации и теории игр Шепли соответственно, предоставляют объяснения на уровне признаков (суперпикселей или фрагментов изображения), но требуют значительных вычислительных ресурсов при работе с изображениями высокого разрешения.

Для Vision Transformer методы интерпретации отличаются от CNN-ориентированных подходов в силу архитектурных особенностей: отсутствия пространственной иерархии свёрточных слоёв и наличия механизма самовнимания. Работа [7] предложила метод Transformer Interpretability Beyond Attention Visualization (Transformer Attribution), распространяющий понятие релевантности через слои с учётом как матриц внимания, так и остаточных связей. Помимо рассмотренных в настоящей работе CDAM [8] и R-Cut [9], существуют методы, использующие усреднение голов внимания по нескольким слоям (Attention Rollout), а также подходы, основанные на интегральных градиентах применительно к токенам. Все перечисленные методы не являются причинными объяснениями в строгом смысле: они описывают согласованность между входными областями и выходом модели, но не устанавливают направленную причинно-следственную связь.

В отличие от большинства цитируемых работ, настоящее исследование направлено не на разработку нового метода интерпретации, а на их сравнительный анализ применительно к специфическому типу данных – одноканальным СЭМ-изображениям бактерий. Особенность постановки состоит в отсутствии ручной масочной разметки, что не позволяет применить стандартные метрики попиксельной точности или отношения пересечения к объединению

областей для оценки пространственного качества карт. Вместо этого используется протокол, основанный на кривых вставки и удаления (deletion/insertion) [10] и автоматически выделяемой маске объекта – подход, близкий по духу к оценке верности объяснений (faithfulness) в том смысле, в каком это понятие используется в [10].

3. Данные и постановка эксперимента

Рассматривается многоклассовая одноклассовая постановка: каждому изображению соответствует ровно один класс. Текущий набор содержит 1852 изображения 13 классов, включая класс normal (является контролем). Источником изображений служат СЭМ-снимки, полученные на электронном микроскопе Zeiss EVO 10 с детектором обратно-рассеянных электронов; предварительная пробоподготовка выполнялась с использованием лантаноидного контрастирования. В анализируемой версии модели используются одноканальные изображения.

В табл. 1 приведено фактическое распределение текущего набора по классам. Дальнейшим развитием результатов исследования, представленных в [11], является существенное увеличение общего объёма выборки, а также включение в текущий набор данных дополнительного класса normal.

Табл. 1. Распределение изображений по классам в текущем наборе данных.

Table 1. Distribution of images across classes in the current dataset.

Класс	Количество изображений
Achromobacter	105
Acinetobacter	71
Burkholderia	119
Escherichia	192
Klebsiella	131
Pseudomonas	230
Rothia	90
Salmonella	108
Serratia	156
Shigella	146
Staphylococcus	173
Stenotrophomonas	174
normal	157
Итого	1852

В текущей схеме анализа используется сохранённая обученная модель ViT-Base/16 (vit_base_patch16_clip_384), адаптированная к одноканальным изображениям размера 384×384. На этапе интерпретации вход приводится к тем же параметрам предобработки, что и при обучении: среднее значение 0.20967 и стандартное отклонение 0.10449. Все тепловые карты строятся по одному прямому проходу модели, что исключает дополнительное слаживание объяснений и обеспечивает сопоставимость методов.

4. Методы интерпретируемости

Базовый метод AttentionCAM в настоящей работе строится не по схеме последовательного перемножения матриц внимания, а по матрице внимания последнего блока ViT. Из тензора самовнимания размера $N \times T \times T$, где N — число параллельных ветвей механизма самовнимания, T — число токенов, выбирается строка CLS→фрагменты, далее выполняется усреднение по всем таким ветвям, преобразование в сетку фрагментов и интерполяция до размера входного изображения. Получаемая карта отражает распределение внимания CLS-токена на завершающем этапе агрегирования признаков и рассматривается как базовый ориентир для последующего сопоставления.

Метод CDAM формирует класс-специфическую карту, объединяя внимание и градиентную релевантность токенов. В текущей реализации используются токены после слоя `postml` последнего блока. Для каждого токена фрагмента вычисляется вклад в ненормированный выход модели для целевого класса (логит), после чего пространственное распределение релевантности взвешивается вниманием CLS→фрагменты. В статье рассматриваются как знаковая карта, так и вариант с выделением только положительной части, более удобный для наглядного сопоставления с другими методами.

Метод R-Cut включает два этапа. На первом этапе по токенам фрагментов оцениваются веса компонент векторного представления на серии прогонов с искусственно модифицированным входом; это позволяет получить класс-ориентированное представление токенов. На втором этапе строится граф фрагментов по косинусной близости и выполняется нормализованный разрез, выделяющий связную область. В отличие от AttentionCAM и CDAM, данный подход ориентирован не столько на локальную атрибуцию, сколько на получение пространственно целостной области. При этом результат заметно зависит от выбора функции нормировки выходов модели и параметра ϕ .

5. Результаты

5.1. Интегральные показатели классификации

При оценке обученной модели на независимой тестовой подвыборке из 371 размеченного изображения получены интегральные показатели, представленные в табл. 2.

Табл. 2. Интегральные метрики классификации на независимой тестовой подвыборке из 371 изображения.

Table 2. Aggregate classification metrics on an independent test subset of 371 images.

Показатель	Значение
Доля верных классификаций	0.784
Точность, макроусреднение	0.808
Полнота, макроусреднение	0.785
F1-мера, макроусреднение	0.790
F1-мера, взвешенное усреднение	0.785

Близость макроусреднённых и взвешенных показателей указывает на то, что итоговое качество не определяется одним-двумя преобладающими классами. Среднее значение максимального отклика модели составило 0.872 для верных классификаций и 0.591 для

ошибочных, что свидетельствует о заметном, но не полном разделении корректных и некорректных решений по уровню уверенности.

5.2. Показатели по отдельным классам и основные смещения классов

Табл. 3 детализирует результаты по отдельным классам. Она показывает, что интегральные показатели скрывают заметную неоднородность качества: наряду с классами, распознаваемыми устойчиво, присутствуют классы с более низкой полнотой и выраженными смещениями.

Табл. 3. Показатели классификации по отдельным классам на независимой тестовой подвыборке. Table 3. Class-wise classification performance on the independent test subset.

Класс	Число изображений	Точность	Полнота	F1-мера
Achromobacter	21	0.682	0.714	0.698
Acinetobacter	14	0.824	1.000	0.903
Burkholderia	24	0.739	0.708	0.723
Escherichia	39	0.897	0.897	0.897
Klebsiella	26	0.750	0.692	0.720
Pseudomonas	46	0.635	0.870	0.734
Rothia	18	1.000	0.778	0.875
Salmonella	22	0.857	0.818	0.837
Serratia	31	0.688	0.710	0.698
Shigella	29	0.812	0.897	0.852
Staphylococcus	35	0.970	0.914	0.941
Stenotrophomonas	35	0.656	0.600	0.627
normal	31	1.000	0.613	0.760

Наиболее высокие значения F1-меры получены для классов Staphylococcus (0.941), Acinetobacter (0.903), Escherichia (0.897) и Rothia (0.875). Наиболее трудными в текущей конфигурации оказались Stenotrophomonas (0.627), Achromobacter (0.698), Serratia (0.698), Klebsiella (0.720) и Burkholderia (0.723).

При интерпретации показателей отдельных классов необходимо принимать во внимание ограниченный объём тестовой выборки. Ряд классов представлен в тесте 14–21 изображениями, что делает соответствующие показатели по отдельным классам статистически нестабильными: отдельные ошибки классификации могут существенно смещать значение F1-меры. В частности, значение $F1=1.0$ для Rothia при точности 1.000 и полноте 0.778 получено по 18 тестовым изображениям и должно рассматриваться как предварительный результат, требующий подтверждения на расширенной выборке.

Повышение статистической надёжности метрик по отдельным классам является одним из приоритетных направлений дальнейшего развития работы.

Наибольшее число ошибок связано со смещениями *Stenotrophomonas*→*Pseudomonas* (7 случаев), *normal*→*Pseudomonas* (5 случаев), *Achromobacter*→*Pseudomonas* (4 случая), *Stenotrophomonas*→*Serratia* (4 случая), *Burkholderia*→*Achromobacter* (3 случая), *Shigella*→*Pseudomonas* (3 случая), *Salmonella*→*Shigella* (3 случая) и *Klebsiella*→*Serratia* (3 случая). Эти соотношения (рис. 1) указывают на наличие группы классов, для которых морфологические признаки и фоновые текстуры пока разделяются моделью менее уверенно.

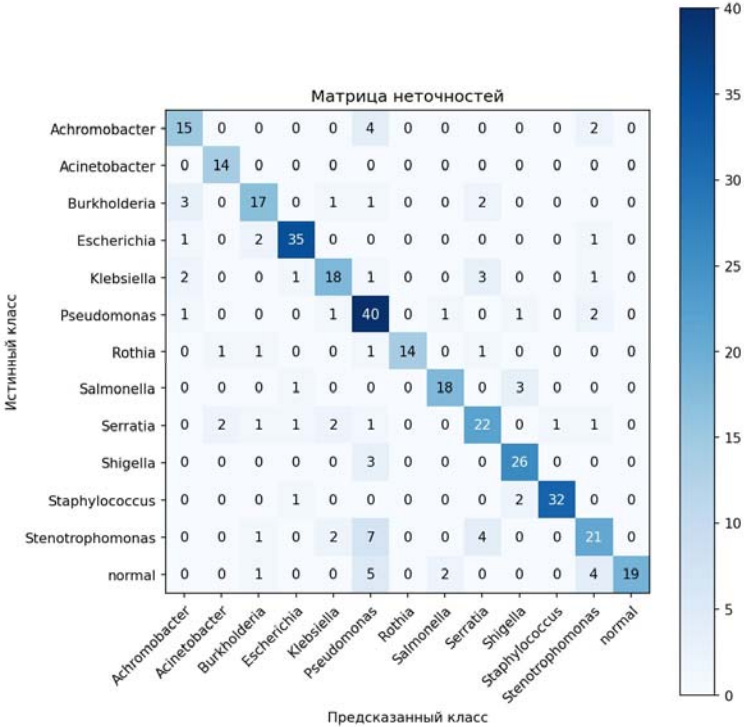


Рис. 1. Матрица неточностей результатов классификации по отдельным классам на независимой тестовой подвыборке.
Fig. 1. Confusion matrix of classification results for individual classes on an independent test subsample.

5.3. Качественное сопоставление методов

На рис. 2 приведён пример сопоставления трёх методов на одном и том же изображении. В данной конфигурации AttentionCAM охватывает значительную часть бактериального кластера и одновременно сохраняет чувствительность к ряду фоновых структур. CDAM выделяет компактные участки, дающие наибольший вклад в ненормированный выход для целевого класса, что делает карту более разреженной. R-Cut формирует связную область, однако при выбранных параметрах имеет тенденцию к расширению области локализации за пределы самого кластера.

Для оценки воспроизводимости наблюдаемого различия на серии изображений на рис. 3 приведены несколько примеров одного класса. На этих примерах AttentionCAM повторяемо охватывает значительную часть кластера, CDAM сохраняет точечный характер локализации,

а R-Cut формирует более связную область, но нередко включает участки фона по границе кластера.

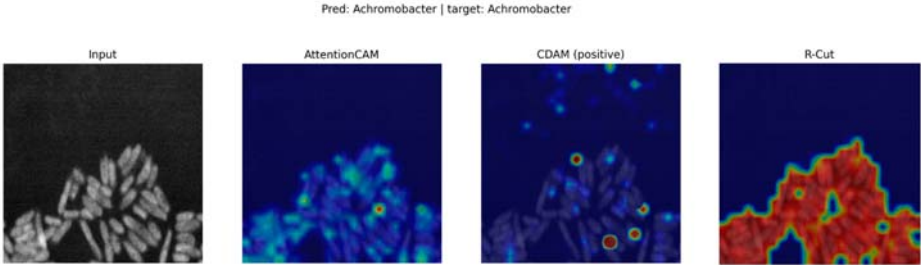


Рис. 2. Сопоставление исходного изображения, базовой карты AttentionCAM, положительной части карты CDAM и карты R-Cut для одного примера.
Fig. 2. Comparison of the input image, the baseline AttentionCAM map, the positive part of the CDAM map, and the R-Cut map for a single example.

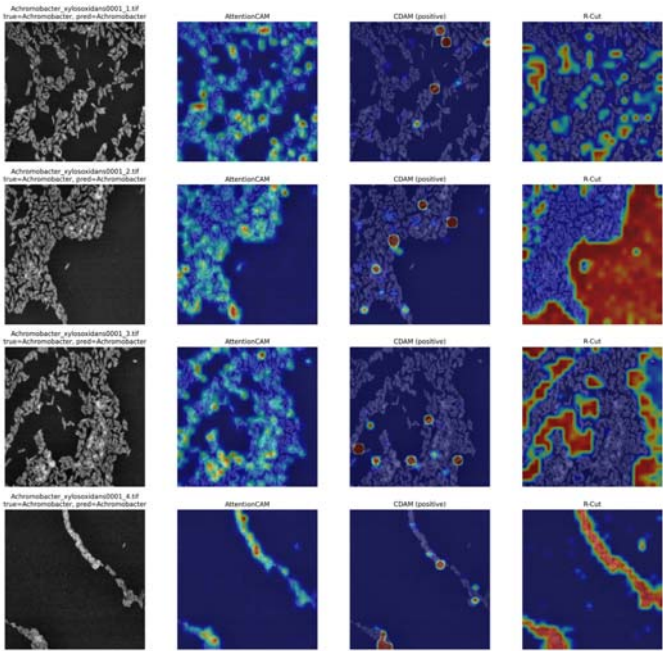


Рис. 3. Серия качественных примеров для одного класса: сопоставление исходных изображений и карт, построенных методами AttentionCAM, CDAM и R-Cut.
Fig. 3. A series of qualitative examples for one class: comparison of input images and maps produced by AttentionCAM, CDAM, and R-Cut.

5.4. Количественная оценка согласованности объяснений с решением модели без ручной масочной разметки

Для количественного сопоставления методов использована стратифицированная подвыборка из 160 корректно классифицированных изображений, включающая по 12–13 примеров на

каждый класс. Для каждого метода рассчитывались: площадь под кривой поэтапного удаления значимых фрагментов, площадь под кривой их поэтапного добавления, среднее снижение целевого отклика при сохранении только наиболее значимых областей, доля случаев повышения целевого отклика при поэтапном добавлении наиболее значимых областей, а также доля суммарной интенсивности карты внутри грубо выделенной области объекта, полученной автоматически без ручной масочной разметки (табл. 4, рис. 4). Меньшие значения первого и третьего показателей и большие значения второго, четвертого и пятого интерпретируются как более тесная согласованность объяснения с решением модели.

Табл. 4. Средние значения показателей согласованности объяснений с решением модели на 160 стратифицированно отобранных изображениях (среднее $\pm$ стандартное отклонение, где применимо).

Table 4. Mean values of explanation-to-decision agreement metrics on 160 stratified images (mean $\pm$ standard deviation, where applicable).

Метод	Площадь под кривой удаления $\downarrow$	Площадь под кривой добавления $\uparrow$	Среднее снижение целевого отклика $\downarrow$	Доля случаев повышения целевого отклика $\uparrow$	Доля интенсивности карты внутри области объекта $\uparrow$
AttentionCAM	0.257 ± 0.284	0.499 ± 0.252	0.697 ± 0.371	0.081	0.548 ± 0.122
CDAM	0.278 ± 0.282	0.383 ± 0.285	0.715 ± 0.372	0.075	0.415 ± 0.190
R-Cut	0.357 ± 0.282	0.407 ± 0.282	0.726 ± 0.367	0.062	0.320 ± 0.299

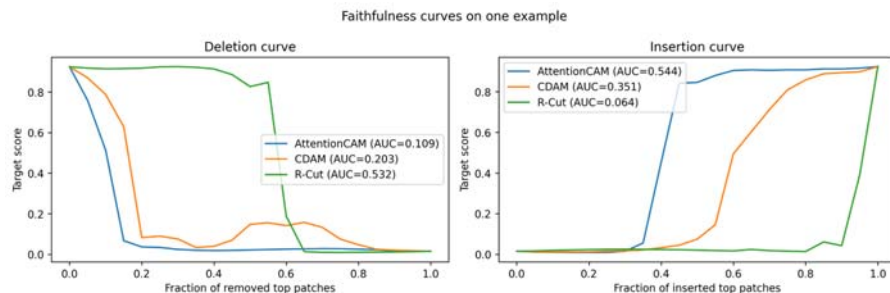


Рис. 4. Кривые поэтапного удаления и добавления значимых фрагментов для одного изображения. На данном примере базовая карта AttentionCAM быстрее снижает целевой отклик при удалении и быстрее восстанавливает его при добавлении, чем CDAM и R-Cut.

Fig. 4. Deletion and insertion curves for a single image. In this example, the baseline AttentionCAM map decreases the target response faster during deletion and restores it faster during insertion than CDAM and R-Cut.

В использованном протоколе базовая карта AttentionCAM показала лучшие средние значения по всем рассматриваемым показателям: площадь под кривой удаления 0.257, площадь под кривой добавления 0.499, среднее снижение целевого отклика 0.697, доля случаев повышения отклика 0.081 и доля интенсивности карты внутри грубо выделенной области объекта 0.548. CDAM уступает AttentionCAM по интегральным количественным критериям, но чаще выделяет локальные дискриминативные зоны. R-Cut демонстрирует наиболее связанные карты, однако в данной конфигурации оказывается наименее выгодным по количественным показателям, что согласуется с его чувствительностью к настройкам.

5.5. Чувствительность R-Cut к выбору функции нормировки выходов и параметра ϕ

Дополнительный анализ чувствительности показал, что результат R-Cut существенно зависит от способа преобразования выходов модели и от значения параметра ϕ . На рис. 5 видно, что при использовании ненормированных выходов модели форма карты в диапазоне $\phi=0.02-0.10$ остаётся относительно устойчивой, тогда как использование нормировки выходов функцией softmax (преобразование выходов в вероятности) при малом ϕ способно качественно изменить распределение отклика и фактически поменять местами предполагаемую область объекта и фон. Это наблюдение подтверждает необходимость фиксировать конфигурацию R-Cut в протоколе эксперимента и отдельно проверять её на нескольких контрольных изображениях до проведения массового прогона.

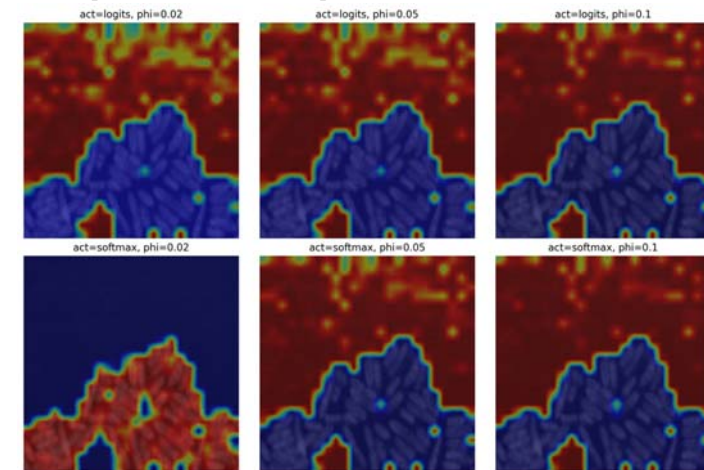


Рис. 5. Анализ чувствительности метода R-Cut: зависимость карты от выбора функции нормировки выходов модели (logits/softmax) и параметра ϕ .

Fig. 5. Sensitivity analysis of R-Cut: dependence of the map on the choice of output normalization (logits/softmax) and on the parameter ϕ .

Полученные результаты показывают, что внешне более «объектная» карта не обязательно обеспечивает лучшее соответствие решению модели по количественным показателям. В рассматриваемой конфигурации базовая карта AttentionCAM систематически превосходит методы CDAM и R-Cut по показателям удаления и добавления значимых областей, по среднему снижению целевого отклика и по доле интенсивности карты внутри автоматически выделенной области объекта. Это означает, что для текущей версии модели именно карта внимания последнего блока в среднем лучше согласуется с поведением классификатора при контролируемом вмешательстве во вход.

Вместе с тем такие показатели не исчерпывают вопрос интерпретируемости. Доля интенсивности карты внутри области объекта опирается на грубую автоматическую полученную маску, а не на ручную экспертную разметку. Также методы CDAM и R-Cut предназначены для выделения иного типа информации: CDAM – для локальных участков, наиболее поддерживающих целевой класс, а R-Cut – для связанной области локализации. Поэтому количественное преимущество AttentionCAM не отменяет прикладной ценности двух других методов, особенно в ситуациях, когда исследователю важна не интегральная метрика, а характер пространственной структуры объяснения.

С практической точки зрения сравнение трёх методов целесообразно использовать раздельно. AttentionCAM подходит для быстрой проверки того, на каких участках изображения концентрируется внимание модели; CDAM – для выделения компактных фрагментов изображения, повышающих ненормированный выход для целевого класса; R-Cut – для получения связной области локализации при условии предварительной настройки параметров. Такое разделение ролей позволяет избежать некорректного ожидания, что один метод одновременно обеспечит и наилучшие количественные показатели, и наиболее наглядную пространственную карту.

При интерпретации количественных результатов важно разграничивать два аспекта оценки. Используемые метрики вставки и удаления deletion/insertion и снижение целевого отклика характеризуют согласованность объяснения с поведением модели (faithfulness), а не его соответствие биологической морфологии бактерий. Соответственно, более высокая согласованность AttentionCAM указывает на то, что карта внимания ближе к признакам, на которые опирается классификатор; оценка же биологической значимости выделяемых областей выходит за рамки метрик достоверности и требует привлечения экспертной разметки. Проверка устойчивости наблюдаемых различий на расширенном наборе данных и с участием эксперта составляет естественное продолжение работы.

6. Заключение

В работе рассмотрена актуальная задача интерпретируемости и построения карт внимания модели ViT для одноканальных СЭМ-изображений бактерий в режиме 384×384, а также приведены результаты оценки модели ViT-Base/16 на независимой тестовой подвыборке из 371 изображения 13 классов. Интегральные показатели классификации составили: доля верных классификаций 0.784, макроусреднённая F1-мера 0.790 и взвешенная F1-мера 0.785.

Для количественного сопоставления объяснений без ручной масочной разметки была сформирована стратифицированная подвыборка из 160 корректно классифицированных изображений. По большинству рассматриваемых показателей на этой подвыборке лучшие средние значения показала базовая карта AttentionCAM. CDAM оказался полезен для выделения наиболее дискриминативных локальных зон, тогда как R-Cut продемонстрировал связный характер локализации, но одновременно и высокую чувствительность к выбору параметров.

Дальнейшее развитие работы связано с использованием независимой тестовой выборки, подготовленной заранее и неизменной для всех последующих сравнений; расширением анализа ошибочных случаев и результатов по отдельным классам; а также привлечением ручной экспертной разметки для прямой проверки пространственной корректности карт.

Список литературы / References

- [1]. Dosovitskiy A. et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. ICLR, 2021.
- [2]. Tschandl P., Rinner C., Apalla Z. et al. Human-computer collaboration for skin cancer recognition. *Nature Medicine*, 2020. DOI: 10.1038/s41591-020-0942-0. (Review of interpretability and visualization methods for deep learning models in medical imaging, including evaluation metrics, current challenges, and future directions. PubMed PMID: 39252818.)
- [3]. Adebayo J., Gilmer J., Muelly M., Goodfellow I., Hardt M., Kim B. Sanity Checks for Saliency Maps. *NeurIPS*, 2018. arXiv preprint v3, 2020. DOI: 10.48550/arXiv.1810.03292.
- [4]. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *ICCV*, 2017.
- [5]. Ribeiro M.T., Singh S., Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD*, 2016.
- [6]. Lundberg S.M., Lee S.-I. A Unified Approach to Interpreting Model Predictions. *NeurIPS*, 2017.
- [7]. Chefer H., Gur S., Wolf L. Transformer Interpretability Beyond Attention Visualization. *CVPR*, 2021.

- [8]. Brocki M. et al. Class-Discriminative Attention Maps (CDAM) for Vision Transformers. *ICLR*, 2024.
- [9]. Huang Y. et al. R-Cut: Relationship Weighted Cut for Denoising ViT Attention Maps. 2024.
- [10]. Wu Z. et al. SaCo: Salience-guided Faithfulness Coefficient for Evaluating Explanations. *CVPR*, 2024.
- [11]. Гридин В.Н., Новиков И.А., Салем Б.Р., Солодовников В.И. Сравнение интерпретируемости моделей ResNet50 и ViT-224 в задаче классификации бактерий на снимках сканирующего электронного микроскопа. *Труды ИСП РАН*, том 37, вып. 6, часть 1, 2025 г., стр. 233-242. DOI: 10.15514/ISPRAS-2025-37(6)-15. / Gridin V.N., Novikov I.A., Salem B.R., Solodovnikov V.I. Comparison of the interpretability of ResNet50 and ViT-224 models in the classification task is erroneous on images of a scanned microscope object. *Trudy ISP RAN/Proc. ISP RAS*, vol. 37, issue 6, part 1, 2025, pp. 233-242 (in Russian). DOI:10.15514/ISPRAS-2025-37(6)-15.

Информация об авторах / Information about authors

Владимир Николаевич ГРИДИН – доктор технических наук, профессор. Научный руководитель Центра информационных технологий в проектировании Российской академии наук. Область научных интересов: информационные технологии, искусственный интеллект, CAD – системы, численно-аналитические методы.

Vladimir Nikolaevich GRIDIN – Dr. Sci. (Tech.), Prof. Scientific Director of the Design Information Technologies Center Russian Academy of Sciences. Research interests: information technology, artificial intelligence, and CAD systems, including analytical methods.

Басим Раед САЛЕМ – научный сотрудник Центра информационных технологий в проектировании Российской академии наук. Область научных интересов: искусственный интеллект, системы поддержки принятия решений.

Basim Raed SALEM – researcher of the Design Information Technologies Center Russian Academy of Sciences. Research interests: rtificial intelligence, decision support systems.

Владимир Игоревич СОЛОДОВНИКОВ – кандидат технических наук. Директор Центра информационных технологий в проектировании Российской академии наук. Область научных интересов: информационные технологии, методы машинного обучения и искусственного интеллекта применительно к самоорганизующимся системам поддержки принятия решений и автоматизированным средствам анализа данных.

Vladimir Igorevich SOLODOVNIKOV – Cand. Sci. (Tech.). Director of the Design Information Technologies Center Russian Academy of Sciences. Research interests: information technology, methods of machine learning and artificial intelligence as applied to self-organizing decision support systems and automated data analysis tools.

Иван Александрович НОВИКОВ – старший научный сотрудник Центра информационных технологий в проектировании Российской академии наук, старший научный сотрудник ФГБНУ научно-исследовательский институт глазных болезней им. М.М. Краснова. Область научных интересов: информационные технологии, анализ данных, биомеханика тканей глаза, микробиология.

Ivan Aleksandrovich NOVIKOV – senior researcher of the Design Information Technologies Center Russian Academy of Sciences, senior researcher of the Krasnov Research Institute of Eye Disease. Research interests: information technology, data analysis, biomechanics of eye tissues, microbiology.

Елизавета Сергеевна Родина – младший научный сотрудник ФГБНУ научно-исследовательский институт глазных болезней им. М. М. Краснова. Область научных интересов: диагностика и лечение микробных кератитов.

Elizaveta Sergeevna Rodina – research assistant of the Krasnov Research Institute of Eye Disease. Research interests: Microbial keratitis diagnosis and treatment.