Локализованная SHAP-объяснимость для детекции объектов на основе параллельного анализа атрибуций

Ю. В. Трофимов
Университет «Дубна»
ura_trofim@bk.ru

А. Н. Аверкин
Университет «Дубна»
averkin2003@inbox.ru

А. В. Шевченко
Университет «Дубна»
leviathan0909@gmail.com

И. А. Трусов
Университет «Дубна»
trusov.iva@yandex.ru

А. К. Алексеев
Университет «Дубна»
aak.24@uni-dubna.ru

Е. С. Кондрашова
Университет «Дубна»
kes.th@uni-dubna.ru

***Аннотация*. Современные методы объяснимого искусственного интеллекта (XAI) в задачах детекции объектов, как правило, формируют карты атрибуции для всего изображения или для целого класса, что ограничивает их способность объяснять решения модели на уровне отдельных детектированных объектов. В работе предложен метод построения локализованных SHAP-объяснений для детекционных нейронных сетей, основанный на разделении области прямого вывода модели и области атрибутивного сэмплирования. Прямой проход нейронной сети выполняется по полному изображению, тогда как возмущения, используемые для оценки вклада признаков, формируются только в окрестностях детектированных bounding box. Такой подход позволяет сохранить контекстно-зависимый характер детекционного решения, одновременно ограничивая вычисления релевантными пространственными областями. Для повышения вычислительной эффективности предложена локализованная схема атрибутивного анализа, ориентированная на маммографические изображения, в которой процедуры объяснения выполняются только для детектированных областей интереса. Такой подход сокращает пространство атрибутивного анализа и позволяет строить более интерпретируемые побоксовые объяснения для детекционных моделей.**

Ключевые слова: искусственный интеллект; объяснимый искусственный интеллект; XAI; SHAP; локализованные объяснения

I. Введение

Глубокие нейронные сети демонстрируют высокую точность в задачах детекции объектов на изображениях [1, 2], однако их практическое внедрение в чувствительных областях – медицинской диагностике, промышленном контроле качества, автономном транспорте – сдерживается отсутствием прозрачности принимаемых решений. Задача объяснимого ИИ (XAI) состоит в том, чтобы сделать логику таких решений понятной специалисту-эксперту.

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2 Применение объяснительного искусственного интеллекта для интерпретации алгоритмов машинного обучения).

Существующие методы атрибуции признаков для задач компьютерного зрения – Grad-CAM [3], LIME [4], RISE [5] и другие – были изначально разработаны для задач классификации и формируют тепловую карту атрибуции для всего входного изображения целиком. Применение таких методов к детекционным моделям приводит к принципиальному противоречию: детекционная сеть предсказывает набор локализованных объектов с индивидуальными оценками уверенности, тогда как глобальная карта атрибуции смешивает вклады признаков, относящихся к различным детектируемым объектам. В результате интерпретация объяснений применительно к конкретному боксу становится ненадёжной.

Проблема усугубляется при работе с высокоразрешёнными медицинскими изображениями, такими как маммограммы. Размер типичного маммографического снимка составляет 2000–5000 пикселей по каждой оси, тогда как область патологии (микрокальцинаты, образования) может занимать несколько десятков пикселей. Глобальная карта атрибуции, вычисленная для такого изображения, содержит огромное количество нерелевантной информации и не позволяет надёжно локализовать признаки, обусловившие детекцию конкретного объекта.

В работе предлагается метод локализованной SHAP-атрибуции для детекционных моделей. Основной вклад состоит в следующем: (1) введён принцип разделения области прямого вывода и области атрибутивного сэмплирования; (2) предложен алгоритм побоксовой SHAP-атрибуции с локальными масками возмущений; (3) описана вычислительно эффективная схема анализа, ориентированная на маммографические изображения.

II. Связанные работы

Методы интерпретации нейронных сетей для задач компьютерного зрения можно разделить на два класса: методы, основанные на градиентах, и методы, основанные на возмущениях.

Методы градиентной атрибуции используют обратное распространение для оценки чувствительности выходного сигнала к входным пикселям. Grad-CAM [3] формирует тепловую карту на основе взвешенных

активаций последнего сверточного слоя и не требует модификации архитектуры модели. Guided Backpropagation [6] фильтрует отрицательные градиенты на ReLU-активациях, получая более чёткие карты признаков. Integrated Gradients [7] усредняет градиенты вдоль прямолинейной траектории в пространстве входных данных, гарантируя выполнение аксиом полноты и нечувствительности. Общий недостаток градиентных методов применительно к детекционным задачам — они оптимизированы под скалярный выход классификатора и не содержат механизма привязки атрибуции к конкретному детектированному боксу.

Методы на основе возмущений (perturbation-based methods) оценивают вклад пикселей или суперпикселей путём их систематической маскировки или замены и анализа изменения предсказания. LIME [4] аппроксимирует локальное поведение модели интерпретируемой линейной моделью, обученной на возмущённых вариантах входного изображения. RISE [5] использует случайные двоичные маски для оценки значимости каждого пикселя. Применение методов этого класса к детекционным задачам требует определения целевой функции, характеризующей детекцию конкретного бокса.

SHAP (SHapley Additive exPlanations) [8] представляет собой унифицированный теоретико-игровой подход к атрибуции признаков, гарантирующий выполнение ряда желательных аксиом: локальной точности, отсутствия влияния признаков-«пустышек» и симметрии. Попытки применить SHAP к детекционным задачам предпринимались в [9], однако в них не решается задача локализованного анализа по отдельным боксам, что и является предметом настоящего исследования.

Grad-CAM++ [14] предложил взвешенное усреднение градиентов по пространственным позициям карт активации, что частично решает проблему множественных объектов одного класса на изображении. Однако метод остаётся привязан к последнему сверточному слою сети и не обеспечивает количественно интерпретируемые значения вклада в смысле значений Шепли.

EigenCAM [15] применяет разложение по сингулярным числам (SVD) к картам активации детекционной головки, получая объяснение без использования меток классов и без обратного прохода. Подход вычислительно эффективен, однако не позволяет разграничить вклады признаков для перекрывающихся объектов и не обеспечивает аддитивного разложения вклада в стиле SHAP.

В контексте медицинской визуализации обзор [10] подчёркивает ограничения интерпретации saliency-карт без строгой локализации и клинической валидации. В работе [11] проводится бенчмаркинг методов интерпретации для рентгеновских снимков грудной клетки, показывающий высокую вариативность объяснений и чувствительность выводов к выбранному методу. Эти результаты дополнительно мотивируют разработку локализованных объяснений для детекционных моделей.

III. Предлагаемый метод

Рассмотрим детекционную нейронную сеть, для которой каждому детекционному объекту k сопоставляется ограничивающий прямоугольник b_k, предсказанный класс c_k и оценка уверенности s_k. Целью SHAP-атрибуции является вычисление вектора вкладов признаков $\varphi^k \in \mathbb{R}^M$, характеризующего влияние суперпикселей на детекцию объекта k.

A. Значение Шепли для детекционной задачи

Значение Шепли для i-го признака определяется как справедливый вклад этого признака в предсказание модели в теоретико-игровом смысле. Для детекционной модели и k-го бокса значение Шепли вычисляется по формуле:

$$\phi_i^k = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \cdot [v_k(S \cup \{i\}) - v_k(S)]$$

где F – множество всех признаков (суперпикселей), S – подмножество признаков, $v_k(S)$ – целевая функция ценности, характеризующая детекцию k-го бокса при активных признаках из S. В качестве функции детекции предлагается использовать оценку уверенности детекции при условии применения маски возмущений:

$$v_k(S) = f_{score}(x \odot m_S^k)[k]$$

где m_S^k – бинарная маска, равная единице для пикселей, принадлежащих признакам из множества S в окрестности k-го бокса, и нулю – в остальных регионах. При этом прямой проход сети выполняется по полному изображению x: возмущение применяется только внутри расширенного bounding box B'^k, тогда как остальная часть изображения остаётся нетронутой.

B. Маска атрибуций и расширенная область интереса

Для k-го бокса $b_k = (x_k, y_k, w_k, h_k)$ определяется расширенная область интереса с контекстным отступом α:

$$B'^k = (x_k - \alpha \cdot w_k, y_k - \alpha \cdot h_k, (1 + 2\alpha) \cdot w_k, (1 + 2\alpha) \cdot h_k),$$

где α ∈ [0, 1] – параметр контекстного расширения. Значение α = 0 соответствует атрибуции строго внутри бокса, тогда как $\alpha \in [0.25, 0.5]$ позволяет учитывать ближайщий контекст влияющий на детекционное решение. Маска возмущений формируется исключительно в пределах B'^k:

$$m_S^{k[p]} = \begin{cases} 1 \text{ если } p \in B'^k \text{ and } seg(p) \in S \\ 0 \text{ иначе} \end{cases}$$

Такое определение обеспечивает контекстную зависимость детекционного решения: прямой проход по полному изображению сохраняет все межобъектные зависимости, улавливаемые рецептивным полем сети. Одновременно атрибуция локализована в пределах B'^k, что резко сокращает пространство признаков и вычислительные затраты.

C. Алгоритм локализованного SHAP

Для каждого детектированного бокса предлагаемый алгоритм включает следующие шаги: (1) в пределах расширенной области R_k^α строится SLIC-сегментация на M суперпикселей; (2) генерируется N случайных бинарных коалиций признаков; (3) для каждой коалиции

формируется возмущенное изображение и выполняется прямой проход сети; (4) по полученным значениям целевой функции решается задач взвешенной линейной регрессии с SHAP-весами; (5) вектор вкладов проецируется на соответствующие суперпиксеои, формируя карту атрибуции для бокса k.

Ключевое отличие от стандартного SHAP состоит в том, что целевая функция $v_k(S)$ явно привязана к конкретному детектированному боксу посредством IoU-сопоставления предсказаний до и после возмущения. Если после возмущения сопоставленный бокс отсутствует, значение целевой функции полагается равным нулю. Такая постановка обеспечивает корректную побоксовую атрибуцию в многообъектных сценах.

IV. Иллюстрация подходов

На рис. 1 показан результат применения стандартного подхода – атрибуционная маска строится для всего изображения, что затрудняет интерпретацию вклада в конкретный бокс. На рис. 2 приведён результат предлагаемого побоксового метода: карта атрибуции сформирована отдельно для каждого детектированного бокса, что обеспечивает точную пространственную локализацию активаций.

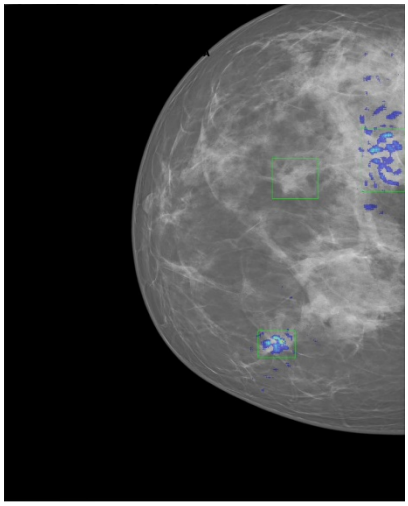

Рис. 1. Стандартный подход: глобальная SHAP-атрибуция по всему изображению

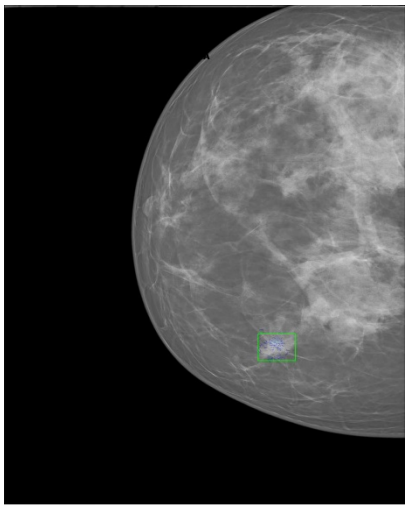

Рис. 2. Предлагаемый метод: побоксовая SHAP-атрибуция. Карты активаций сформированы отдельно для каждого bounding box

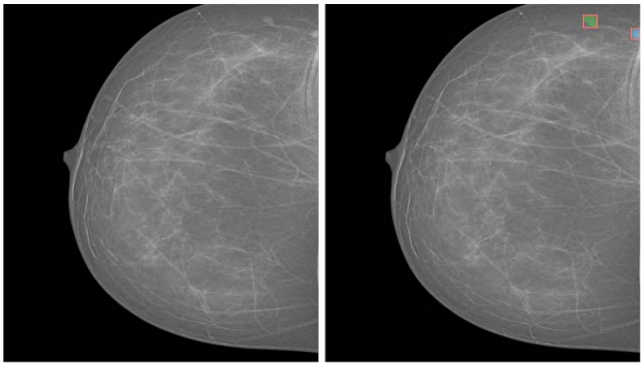

Рис. 3. Пример работы побоксовой SHAP-атрибуции

V. Заключение

В работе предложен метод локализованных SHAP-объяснений для детекционных нейронных сетей, основанный на принципе разделения области прямого вывода модели и области атрибутивного сэмплирования. Ключевая идея состоит в выполнении прямого прохода нейронной сети по полному изображению при одновременном ограничении пространства возмущений расширенными bounding box детектированных объектов. Предложенный подход обеспечивает: (1) корректную побоксовую атрибуцию в многообъектных сценах; (2) учёт контекстных зависимостей в пределах расширенного ROI; (3) вычислительную эффективность за счёт сокращения пространства признаков.

Таким образом, предложенный подход может рассматриваться как практический способ построения локализованных и вычислительно эффективных объяснений для детекционных моделей в задачах медицинской визуализации.

Список литературы

[1] Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016. P. 779–788.

[2] Ren S., He K., Girshick R., Sun J. Faster R-CNN: towards real-time object detection with region proposal networks // Advances in Neural Information Processing Systems (NeurIPS). 2015. Vol. 28. P. 91–99.

[3] Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, 2017. P. 618–626.

[4] Ribeiro M.T., Singh S., Guestrin C. «Why should I trust you?»: explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016. P. 1135–1144.

[5] Petsiuk V., Das A., Saenko K. RISE: randomized input sampling for explanation of black-box models [Electronic resource] // arXiv. 2018. arXiv:1806.07421. URL: https://arxiv.org/abs/1806.07421 (20.03.2026).

[6] Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M. Striving for simplicity: the all convolutional net [Electronic resource] // arXiv. 2014. arXiv:1412.6806. URL: https://arxiv.org/abs/1412.6806 (20.03.2026).

[7] Sundararajan M., Taly A., Yan Q. Axiomatic attribution for deep networks // Proceedings of the 34th International Conference on Machine Learning (ICML). Sydney, 2017. Vol. 70. P. 3319–3328.

[8] Lundberg S.M., Lee S.-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30. P. 4765–4774.

[9] Kakogeorgiou I., Karantzalos K. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing // International Journal of Applied Earth Observation and Geoinformation. 2021. Vol. 103. P. 102520.

[10] Reyes M., Meier R., Pereira S., Dahlweid M., von Tengg-Kobligk H., Summers R. M., Lengeling A. On the interpretability of artificial intelligence in radiology: challenges and opportunities // Radiology: Artificial Intelligence. 2020. Vol. 2, No. 3. e190043.

[11] Saporta A., Gui X., Agrawal A., Lortie M., Bhatt D. L., Rajpurkar P. Benchmarking saliency methods for chest X-ray interpretation // Nature Machine Intelligence. 2022. Vol. 4. P. 867–878.

[12] Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017. P. 2117–2125.

[13] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. Learning deep features for discriminative localization // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016. P. 2921–2929

[14] Chattopadhay A., Sarkar A., Howlader P., Balasubramanian V. N. Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks // Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). 2018. P. 839–847.

[15] Muhammad M. B., Yeasin M. Eigen-cam: Class activation map using principal components //2020 international joint conference on neural networks (IJCNN). IEEE, 2020. C. 1-7.