DOI 10.37882/2223-2966.2024.1-2.05

МЕТОДЫ ОЦЕНКИ ИНТЕРПРЕТАЦИЙ МОДЕЛЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ

METHODS FOR EVALUATING INTERPRETATIONS OF COMPUTER VISION MODELS

I. Kapustin
O. Romashkova

Summary. The article examines the problem of interpretability of computer vision models. The author notes that although there are many methods designed to explain the decisions made by deep neural networks, little effort has been made to ensure that these explanations are objectively relevant. The article proves the correlation between the quality of a model and the quality of its interpretation, considers metrics for assessing interpretations taken from the field of algorithmic stability: average generalization (MeGe) and relative consistency (ReCo), and also proposes a new metric for assessing the quality of interpretations of computer vision models, taking into account quality of the original model.

Keywords: interpretation, machine learning, correlation, MeGe, ReCo, interpretation quality, model evaluation.

Капустин Илья Сергеевич
ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте РФ», г. Москва
ilya.kapustini@mail.ru
Ромашкова Оксана Николаевна
Доктор технических наук, профессор, профессор, ФГБОУ ВО «Российская академия народного хозяйства и государственной службы при Президенте РФ», г. Москва
ox-rom@yandex.ru

Аннотация. В статье исследуется проблема интерпретируемости моделей компьютерного зрения. Автор отмечает, что хотя существует множество методов, предназначенных для объяснения принимаемых глубокими нейронными сетями решений, мало усилий было приложено для обеспечения объективной релевантности этих объяснений. В статье доказана корреляция между качеством модели и качеством ее интерпретации, рассмотрены метрики для оценки интерпретаций, взятые из области алгоритмической стабильности: среднее обобщение (MeGe) и относительная согласованность (ReCo), а также предложена новая метрика для оценки качества интерпретаций моделей компьютерного зрения, учитывающая качество работы исходной модели.

Ключевые слова: интерпретация, машинное обучение, корреляция, MeGe, ReCo, качество интерпретации, оценка модели.

Введение

Несмотря на значительные усилия в разработке методов интерпретации моделей, проблема качественной оценки интерпретаторов остается актуальной в виду своей слабой изученности [1]. Проблема оценки интерпретации моделей машинного обучения является важным аспектом в контексте понимания и доверия к моделям, особенно в областях, чувствительных для качественной интерпретации, таких как медицина или финансы [2]. Интерпретация моделей обеспечивает объяснения и выводы о том, как модель принимает решения, что крайне полезно для понимания ее работы и принятия верных решений на основе этих моделей.

Существует множество методов интерпретации моделей [3, 4], таких как важность признаков, визуализации активаций нейронов, градиентные методы и др. Однако, оценка качества этих методов является сложной задачей. Понять, насколько интерпретация соответствует реальному поведению модели, требует систематического и объективного подхода к оценке.

Исследования в данной области стремятся:

— Разработать критерии и подходы, которые позволят объективно оценивать качество интерпретации и включать сопоставление интерпретаций с реальными данными или знаниями экспертов.
— Валидировать данные метрики посредством экспериментов, т.е. проведение экспериментов для оценки того, как различные методы интерпретации соотносятся с качеством модели и способностью человека понимать их объяснения.

Одним из сложных аспектов является неоднозначность и субъективность оценки интерпретации. То, что один человек считает хорошей интерпретацией, другой может оценить иначе. Поэтому создание объективных методов оценки становится ключевым аспектом в данной области исследований.

Статьи и исследования в этой области обычно проводят обзор существующих методов оценки интерпретации моделей и предлагают новые подходы или улучшения для более надежной и объективной оценки. Это помогает улучшить доверие к моделям машинного об-

учения и повысить их применимость в реальных условиях.

Постановка задачи

Пусть задана модель классификации $\varphi: X \to Y$, где X — пространство входных данных, а Y — множество классов.

Также у нас есть модель интерпретации $I: Y \to R^n$, которая отображает классы в векторное пространство R^n.

Пусть у нас есть метрика для модели классификации, обозначим её как $M_\phi\left(\hat{Y}, Y\right)$, где $\hat{Y}$ — предсказанные классы, Y — истинные классы.

Также у нас есть метрика для модели интерпретации, обозначим её как $M_I\left(V, \hat{V}\right)$, где V — истинные интерпретации (векторы, соответствующие истинным классам), $\hat{V}$ — предсказанные интерпретации.

Наша цель — выявить корреляцию между метриками. Корреляция означает, что с изменением одной метрики (точности модели предсказания) изменяется и другая метрика (точность интерпретации предсказаний модели).

Формально, корреляция между M_φ и M_I может быть выражена с помощью коэффициентов корреляции Пирсона p

$$p\left(M_\phi, M_I\right) = \frac{cov\left(M_\phi, M_I\right)}{\sigma M_\phi \cdot \sigma M_I},$$

где $\mathrm{cov}(M_\varphi, M_I)$ — ковариация между метриками M_φ и M_I;
σM_φ — стандартное отклонение метрики M_φ;
σM_I — стандартное отклонение метрики M_I.

Если $p > 0$, то это указывает на положительную корреляцию между метриками.

Обзор методов интерпретации и оценки их качества

В данном разделе мы рассмотрим два метода, которые предназначены для интерпретации предсказаний моделей машинного обучения — SHAP и LIME.

Концепция SHAP основана на значениях Шепли из теории кооперативных игр [3]. Для интерпретации вклада каждого признака в предсказание модели используется следующее выражение

$$\phi_i = \sum_{S \subseteq \mathbf{N} \setminus \{i\}} \frac{|S|!\left(|N| - |S| - 1\right)!}{|N|!}\left[f\left(S \cup \{i\}\right) - f\left(S\right)\right],$$

где φ_i — значение Шепли для признака i;
N — множество всех признаков;
S — подмножество признаков без признака i;
$f\left(S \cup \{i\}\right)$ — предсказание модели на выборе признаков $S \cup \{i\}$;
$f(S)$ — предсказание модели на выборе признаков S.

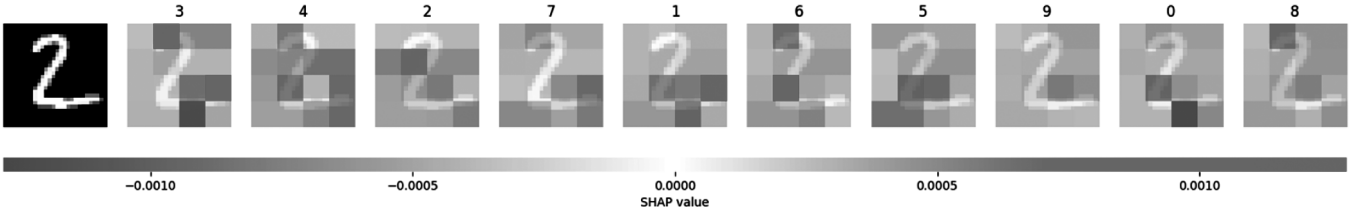


Рис. 1. Интерпретация SHAP для модели классификации, обученной на данных MNIST

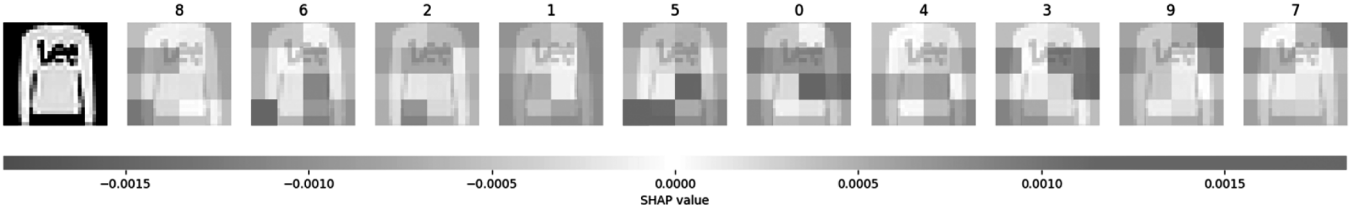


Рис. 2. Интерпретация SHAP для модели классификации, обученной на данных FASHION MNIST

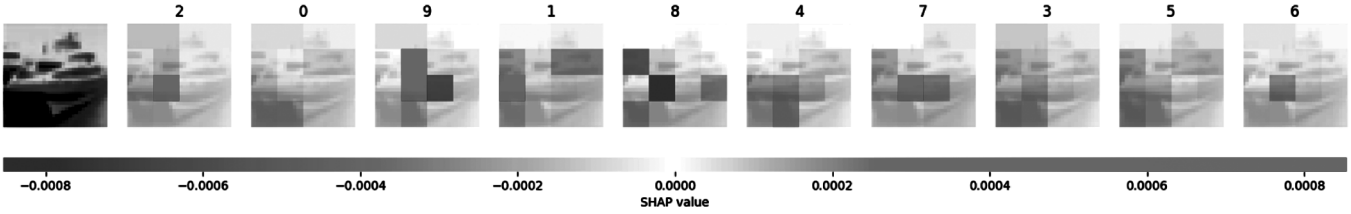


Рис. 3. Интерпретация SHAP для модели классификации, обученной на данных CIFAR10

Эта формула вычисляет вклад каждого признака *i* путем усреднения разниц в предсказаниях модели $f(S \cup \{i\}) - f(S)$ для всех возможных комбинаций признаков *S* без признака *i*. Значения Шепли показывают, насколько добавление конкретного признака меняет предсказание модели относительно других признаков.

Этот метод весьма прозрачен и точен, однако требует больших вычислительных ресурсов. В связи с этим обычно используется аппроксимация значения Шепли.

Примеры интерпретаций при помощи метода SHAP представлены на рисунках 1, 2 и 3.

Следующим мы рассмотрим метод LIME (Local Interpretable Model-agnostic Explanations). Данный метод объясняет предсказания модели путем аппроксимации ее поведения в небольших локальных областях данных. Основная идея LIME заключается в создании интерпретируемой модели (например, линейной модели) вокруг конкретного предсказания, которая приближает поведение исходной модели в окрестности этого предсказания [3].

Минусом данного подхода является то, что он весьма неустойчив к выбросам и сильно зависим от точности аппроксимации.

Примеры интерпретаций при помощи метода LIME представлены на рисунках 4, 5 и 6.

Теперь, рассмотрев популярные подходы к интерпретации алгоритмов машинного обучения, рассмотрим метрики, предложенные в работе [1] для оценки интерпретации.

Данные метрики берут свое начало из области алгоритмической стабильности: среднее обобщение (MeGe) и относительная согласованность (ReCo).

Рис. 4. Интерпретация LIME для модели классификации, обученной на данных MNIST

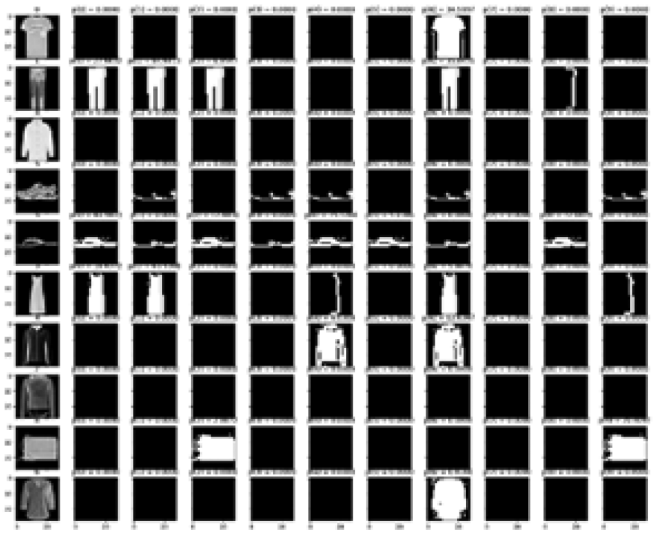


Рис. 5. Интерпретация LIME для модели классификации, обученной на данных FASHION MNIST

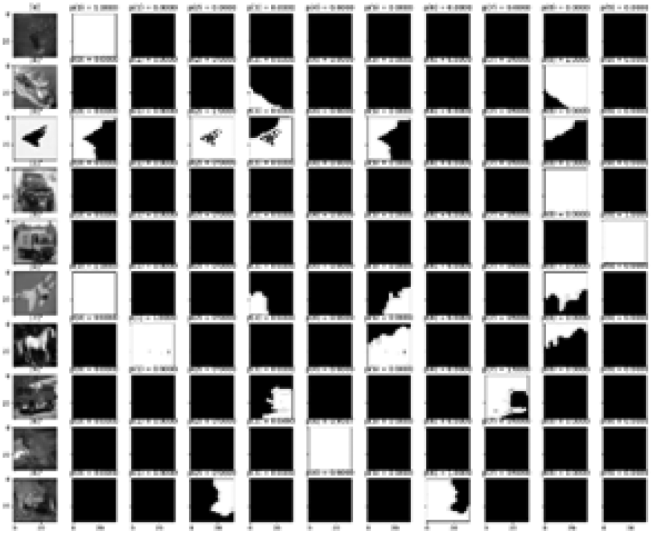


Рис. 6. Интерпретация LIME для модели классификации, обученной на данных CIFAR10

MeGe оценивает среднюю обобщаемость объяснений внутри одной и той же категории. Он пытается определить, насколько хорошо объяснения работают для разных примеров внутри одной категории. MeGe близкий к 1 означает, что объяснения хорошо работают для разных примеров внутри одной категории.

Мы модифицировали исходную метрику MeGe, используемую в статье [1], так как в нашем случае мы используем одну модель. Поэтому вместо сравнений объяснений моделей на объектах, мы сравниваем объяснения внутри каждой категории, считаем среднюю попарную разность объяснений в каждой из категорий. Далее мы вычитаем нормированное среднее отклонение из 1, таким образом, максимальное значение данной метрике равно 1

$$MeGe = 1 - \sum \frac{cons_i}{max(cons)},$$

где $cons_i$ — среднее попарное отклонение внутри *i*-ой категории,

max(*cons*) — максимальное из отклонений.

Помимо метрики MeGe мы также применяем метрику ReCo, измеряющая согласованность интерпретаций между разными классами в задачах классификации, определяя, насколько сильно объяснения различных категорий различаются друг от друга. Высокое значение ReCo означает, что объяснения разных категорий сильно различаются.

Данную метрику мы так же модифицировали по сравнению со статьей [1]. Мы считаем попарные расстояния между элементами внутри объяснений в одной и той же группе, и в разных. Далее мы вводим классификатор, определяющий по разности объяснений, находятся ли два изображения в одном классе или нет (если разница между объяснениями меньше отсечки δ, то в одном классе, иначе в разных). С введением данного классификатора γ, наша метрика приобретает вид

$$ReCo = max\left(\frac{TPR(\gamma) + TNR(\gamma)}{2}\right),$$

где $TPR(\gamma)$ — соотношение числа верно определенных положительных результатов к общему числу истинных положительных результатов;

$TNR(\gamma)$ — соотношение числа верно определенных отрицательных результатов к общему числу истинных отрицательных результатов.

Описание эксперимента

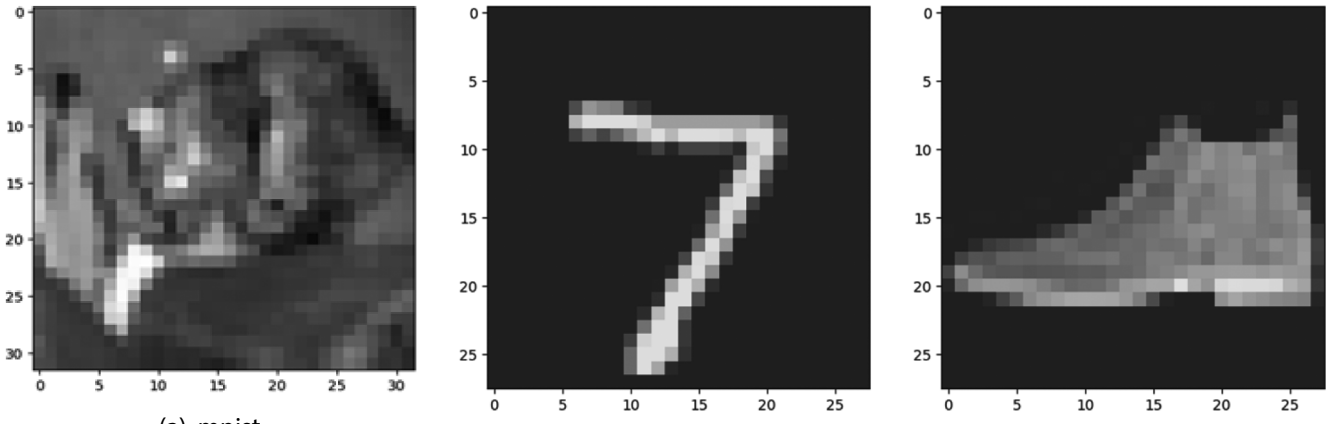
В работе используется 3 наборы данных для обучения моделей классификации изображений: mnist, fashion-mnist и cifar10. Наборы данных mnist и fashion-mnist содержат 60000 обучающих и 10000 тестовых данных, а cifar10 содержит 50000 обучающих и 10000 тестовых данных. Каждый из 3 наборов данных предназначен для мультиклассовой классификации и содержит 10 меток класса. На рисунке 7 приведены примеры данных из каждого набора данных.

Наше исследование направлено на анализ взаимосвязи между качеством модели и качеством интерпретации модели в контексте серии из 20 итераций, охватывающих обучение, предсказание и интерпретацию. В качестве базовой модели была выбрана полносвязная нейронная сеть с одним скрытым слоем, состоящим из 50 нейронов. Этот выбор обусловлен ограниченностью вычислительных ресурсов и достаточной простотой исходных датасетов. На рисунке 8 приведена подробная конфигурация модели.

На каждой итерации мы произвели:

— Обучение модели. Нейросеть обучалась 5 эпох. Установлено, что после пяти эпох точность модели, не существенно увеличивается дальше.

— Предсказания модели. Полученная модель была использована для предсказаний на тестовых данных.



(a). mnist (b). fashion mnist (c). cifar10

Рис. 7. Пример данных mnist, fashion mnist и cfar10

— Для каждого из методов мы получили интерпретацию полученных результатов на тестовой выборке при помощи LIME и SHAP, рассмотренных ранее.
— Для каждого из методов оценки мы вычислили качество интерпретации с помощью метрик MeGe и ReCo.

Layer (type)	Output Shape	Param #
flatten_7 (Flatten)	(None, 2352)	0
dense_13 (Dense)	(None, 50)	117650
dense_14 (Dense)	(None, 10)	510

Total params: 118160 (461.56 KB)
Trainable params: 118160 (461.56 KB)
Non-trainable params: 0 (0.00 Byte)

Рис. 8. Конфигурация моделей

Ввиду вычислительной сложности интерпретации моделей мы использовали лишь 100 объектов из каждой тестовой выборки для вычисления интерпретации. Такой объем гарантирует ненулевое значение каждого из классов среди классифицируемых объектов, в то же время позволяя произвести вычисления за разумное время.

Результаты эксперимента

Мы собрали значения метрик ReCo и MeGe на каждой эпохе для каждой модели (MNIST, FASHION MNIST, CIFAR10) с применением кросс-валидации на 5 фалдах. Далее мы применили к полученным данным коэффициент корреляции Пирсона и получили тепловые карты изображенные на рисунке 9 и 10.

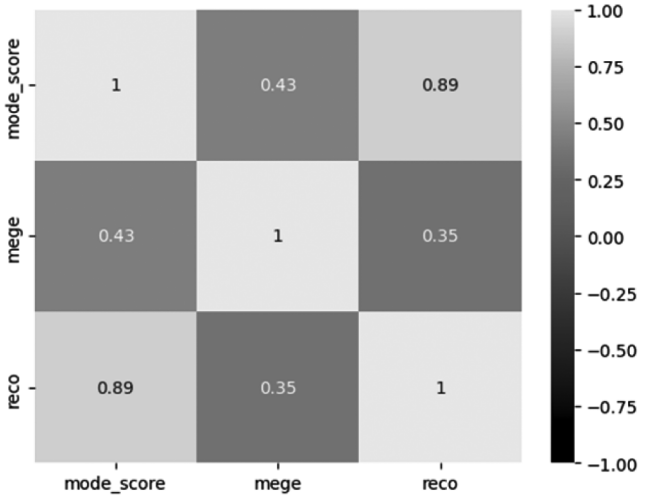


Рис. 9. Корреляция между качеством модели классификации и интерпретации данной модели при помощи LIME

Наиболее выраженная корреляция прослеживается между метрикой точности (accuracy) и ReCo. Эта явная связь указывает на согласованность между тем, насколько точно модель предсказывает данные, и степенью, с которой метрика ReCo отражает уровень интерпретируемости модели.

Стоит так же отметить, что попарная корреляция метрик ReCo и MeGe не столь значима, в то время как обе они сильно коррелируют с точностью модели. Это подтверждает тот факт, что, отражая различные аспекты качества интерпретации, обе эти метрики зависимы от точности модели.

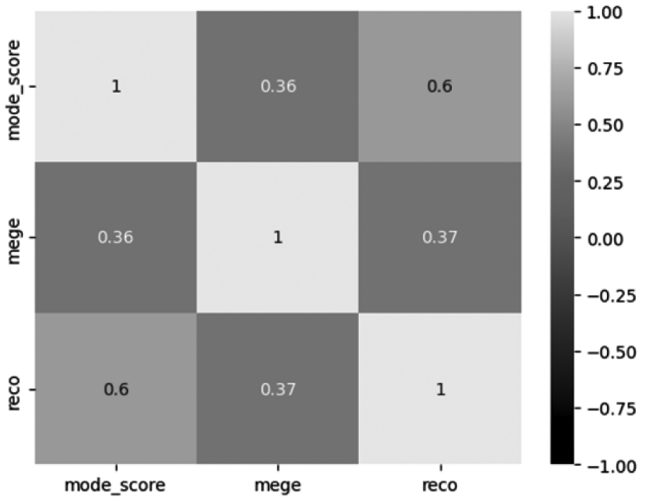


Рис. 10. Корреляция между качеством модели классификации и интерпретации данной модели при помощи SHAP

Для метода SHAP мы так же видим строго положительную корреляцию с accuracy. На основании результатов эксперимента мы приходим к Теореме 1:

Теорема 1. Пусть задана полносвязная нейросетевая модель φ для классификации изображений, и пусть ξ — случайная величина, равная качеству (accuracy) модели, обученной на некоторой случайной выборке данных. Пусть также μ_1 и μ_2 — случайные величины, равные метрикам MeGe и ReCo, посчитанным при помощи LIME метода для некоторой случайной валидационной подвыборки. Пусть η_1 и η_2 — аналогичные случайные величины для SHAP метода. Тогда

$$\mathrm{corr}(\xi, \mu_1) > 0,\ \mathrm{corr}(\xi, \mu_2) > 0,\ \mathrm{corr}(\xi, \eta_1) > 0,\ \mathrm{corr}(\xi, \eta_2) > 0.$$

Согласно Теореме 1 качество интерпретаций связано с качеством работы моделей. Исходя из этого нами была разработана метрика, учитывающая качество интерпретации с учетом качества исходной модели.

Метрика InterpretationAccuracy предназначена для оценки качества интерпретации модели компьютерного зрения. Она позволяет учитывать не только качество интерпретаций, но и качество интерпретируемой модели

$$Interpretation\ Accuracy = \alpha\frac{N_{Np}}{N} + \beta\frac{A \cup B}{A \cap B},$$

где N_{Np} — кол-во правильных интерпретаций с истинными значениями;

N — общее кол-во интерпретаций;

$\frac{A \cup B}{A \cap B}$ — сходство Жаккара;

коэффициенты α и β в сумме дают всегда единицу и задают приоритет (вес) для оценки интерпретации или качества модели.

В качестве метрики для оценки интерпретаций моделей компьютерного зрения мы используем сходство Жаккара. Во многих работах [5, 6] сходство Жаккара применяется в качестве метрики сходства признаков для изображений. Мы объединили её с метрикой, которую мы определили как отношение количества правильно объясненных примеров к общему числу примеров.

Ниже приведем реализацию InterpretationAccuracy на языке Python.

Листинг 1.

Interpretation accuracy на языке программирования Python

```
class InterpretationAccuracy:
 def init(self, predictions, x_test, true_labels, masks):
 self.predictions = predictions
 self.total_count = x_test.shape[0]
 self.true_labels = true_labels
 self.img = x_test
 self.masks = masks

 def model_accuracy(self):
 correct_count = 0

 for i in range(self.total_count):
 predicted_class = np.argmax(self.predictions[i])
 true_class = self.true_labels[i]

 if predicted_class == true_class:
 correct_count += 1

 accuracy = correct_count / self.total_count
 return accuracy

 def jaccard_index(self):
 img, masks = np.array(self.img), np.array(self.masks)
 intersection = np.logical_and(img, masks)
 union = np.logical_or(img, masks)

 return (np.sum(intersection) / np.sum(union))

 def index_IA(self):
 return (0.5 * self.model_accuracy() + 0.5 * self.jaccard_index())
```

Заключение

Нами была выявлена значительная положительная корреляция между предсказаниями модели и качеством интерпретации. Изучение взаимосвязи между этими двумя аспектами позволило обнаружить наличие устойчивой зависимости, что говорит о тесной связи между точностью модели и метриками интерпретации.

Также были рассмотрены метрики оценки интерпретаций моделей машинного обучения. Важным аспектом нашего подхода стало модифицированное применение этих методов: вместо сравнения нескольких моделей, мы адаптировали их для оценки одной конкретной модели. Это позволило углубить понимание того, как именно данная модель работает и какие признаки она учитывает при принятии решений.

В результате эксперимента была доказана Теорема 1. Данная теорема доказывает непосредственную взаимосвязь между точностью модели и качеством ее интерпретации. Доказательство данной теоремы стало фундаментальным шагом в понимании того, как взаимодействуют качество модели и уровень ее интерпретации.

Исходя из доказанной теоремы, нами была разработана метрика, учитывающая точность модели при оценке ее интерпретации. Это значительный шаг вперед, особенно в контексте моделей компьютерного зрения, где важны не только точность предсказаний, но и способность понять, как именно модель принимает свои решения на основе визуальных данных.

ЛИТЕРАТУРА

1. Fel T., Vigouroux D., Cad R., Serre T. How Good is your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks, 2021 — URL: https://arxiv.org/pdf/2009.04521.pdf (дата обращения: 19.10.2023).
2. Ермакова Т.Н., Ромашкова О.Н. Математическая модель оценки финансовых показателей средней общеобразовательной организации // В книге: Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Материалы Всероссийской конференции с международным участием. Российский университет дружбы народов. 2016. С. 93–95.
3. Ribeiro M.T., Singh S., Guestrin C. Why should i trust you?: Explaining the predictions of any classifier, 2016 — URL: https://arxiv.org/pdf/1602.04938.pdf (дата обращения: 02.10.2023).
4. Lundberg S., Lee S. A unified approach to interpreting model predictions», 2017 — URL: https://arxiv.org/pdf/1705.07874.pdf (дата обращения: 02.10.2023).
5. Mahbod A., Dorffner G., Ellinger I., Woitek R., Hatamikia S. Improving Generalization Capability of Deep Learning-Based Nuclei Instance Segmentation by Non-deterministic Train Time and Deterministic Test Time Stain Normalization, 2023 — URL: https://arxiv.org/pdf/2309.06143.pdf (дата обращения: 05.10.2023).
6. Benatti A., Costa L. Multilyaer Multiset Neuronal Networks — MMNNs, 2023 — URL: https://arxiv.org/pdf/2308.14541v1.pdf (дата обращения: 05.10.2023).

Журнал «Современная наука: актуальные проблемы теории и практики»