# Выбор интерпретируемых сверточных моделей глубокого обучения

Тимур Мурадов
МФТИ

Олег Бахтеев
МФТИ

Константин Яковлев
МФТИ

Вадим Стрижов
МФТИ

## Abstract

В статье рассматривается задача построения интерпретируемой сверточной нейронной сети. Под интерпретируемостью модели понимается выделение наиболее важных признаков и определение кластеров схожих объектов. Для повышения интерпретируемоси в статье вводится модификация метода OpenBox работающего с кусочно-линейными нейронными сетями. В нём модель представляется в виде набора интерпретируемых линейных классификаторов. Каждый из них определен на выпуклом многограннике. Это позволяет классифицировать схожие объекты одним и тем же классификатором. Метод обобщается на работу с более широким классом нейронных сетей: сверточными нейронными сетями. Предлагается математически эквивалентная замена слоев свёрточной сети на линейные модели. Что значительно повышает интепретируемость. Вычислительный эксперимент проводится на выборках изображений рукописных цифр MNIST и изображений CIFAR-10.

Keywords Model interpretability · Deep Learning · OpenBox · Convolutional neural networks

## 1 Вступление

В данном исследовании стоит задача повышения интерпретируемости модели, где под интерпретируемостью понимается простота выделения важных признаков на выборке данных и классификация близких объектов одним и тем же классификатором.

Проблемой является в целом высокая сложность интерпретации сверточных нейронных сетей, требующая комплексного подхода. На данный момент существует множество различных решений проблемы интерпретации [1, 2, 3] . В статье [1] описан метод LIME, предлагающий линейную апроксимацию предсказаний модели в некоторой небольшой окрестности вокруг объектов из тестовой выборки. Такой подход позволяет получить простую для интерпретации модель без использования информации о строении модели изнутри “model-agnostic”. Но он весьма неустойчив к выбросам и сильно зависим от точности апроксимации. В статье [2] предлагается подход SHAP, заключающийся в рассмотрении вклада каждого признака в работу модели. Таким образом удается выделять даже скрытые, но значимые признаки. Однако применимость данного подхода ограничена ввиду высоких вычислительных затрат: требуется многократное обучение модели, и он весьма зависим от выборки данных. Ещё один подход к интерпретации OpenBox, описываемый в статье [3] предлагает построение математически эквивалентных линейных моделей для линейных нейронных сетей. Он показал более высокую эффективность по сравнению с LIME и весьма перспективен для дальнейшей работы.

В данной работе предлагается адаптация метода OpenBox для работы со свёрточными нейронными сетями: математически эквивалентно представить в виде линейных моделей такие слои как свёртка, пулинг и нормализация. И сравнение с альтернативными методами интепретации CNN.

Для анализа качества предложенного метода проводится вычислительный эксперимент на выборке изображений Fashion-MNIST [4].

## 2 Постановка проблема интерпретируемости сверточных нейронных сетей

Задана выборка $\mathbf{x} \in \mathbf{X}$, где $\mathbf{X} = \mathbb{R}^2$, то есть двумерные изображения. $\mathbf{y} \in \{1, 2, ...k\}$, заданное конечное множество классов.

Задача построенить модель глубокого обучения для задачи классификации.

Модель $\mathbf{f}(\mathbf{X}, \mathbf{w})$ — сверточная нейронная сеть, для краткости CNN, это суперпозиция подмоделей $\mathbf{f_1} \circ \mathbf{f_2} \dots \mathbf{f}_n$.

Функции $\mathbf{f_i}$ — слои нейронной сети, это одни из функций: линейные $\mathbf{f_i} = \mathbf{w_0} + \Sigma \mathbf{w_i} * \mathbf{x_i}$, свертки $S(i,j) = (K * I)(i,j) = \Sigma_m \Sigma_n I(i+m, j+n) K(m,n)$, батч-нормы $\hat{x_i^{(k)}} = \frac{x_i^{(k)} - E(x_i^{(k)})}{\sqrt{D(x_i^{(k)})}}$ или пулинги.

В модели $\mathbf{f}(\mathbf{X}, \mathbf{w})$ оптимизируется функция кросс энтропии $\mathcal{L}(\mathbf{g}, \mathbf{p})$, $\mathbf{g}$ — функция softmax $\mathbf{g}$: $\mathbf{X} \to \mathbf{y}$ на выходе предсказанное распределение вероятности соответствия объектов классам, функция $\mathbf{p}$: $\mathbf{X} \to \mathbf{y}$ на выходе истинное распределение вероятности соответствия объектов классам.

$$\mathbf{g}(\mathbf{x})_i = \frac{exp(\mathbf{x_i})}{\Sigma_j exp(\mathbf{x_j})},$$
$$\mathcal{L} = -\Sigma_i \mathbf{p}(\mathbf{x})_i \log \mathbf{g}(\mathbf{x})_i \to \max$$

Кроме задачи оптимизации модель должна удовлетворять двум требованиям к итерпретируемости: точность и консистентность.

Точность: Математическая эквивалентность.

$$\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{p}(\mathbf{x})$$

Консистентность: Близкие интерпретации для близких объектов выборки.

$$\mathbf{x_i} \in U_\epsilon(\mathbf{x_j}) \Longrightarrow \mathbf{f}(\mathbf{x_i}, \mathbf{w}) \in U_{\mathbf{f}(\epsilon, \mathbf{w})} \mathbf{f}(\mathbf{x_j}, \mathbf{w})$$

## 3 Линейность сверточных нейронных сетей

Theorem 1 Слои сверточной нейронной сети: линейные, свертки, батч-нормализации, пулинги — это линейные операции.

Свертка представима как линейная операция, если расписать её как произведение матрицы входного изображения на матрицу с весами фильтра. Пулинг на максимум представим в виде взаимодействия фильтра на изображение как на политоп. Батч нормализация представима как скалярное произведение, применённое поэлементно к каждому изображению.

## 4 Вычислительный эксперимент

Цель эксперимента: сравнить качество базового метода LIME [1] с предлагаемой альтернативой OpenBox [3]. Критерием качества рассматривается точность предсказания класса объектов. Косинусное расстояние между распределением вероятностей классов $\mathbf{q}(\mathbf{x})$ полученные методом и распределением вероятностей $\mathbf{g}(\mathbf{x})$, полученным классификатором.

## 5 Базовый эксперимент

Строим CNN и при помощи метода LIME [1] получаем интерпретации признаков модели. Результаты работы LIME по выделению важных признаков показан на рисунке 1.

### 5.1 Исходные данные

Fashion-MNIST датасет содержащий 60000 изображений в train и 10000 изображений в test из 10 различных классов. Каждое изображение имеет разрешение 28*28 пикселей [4].

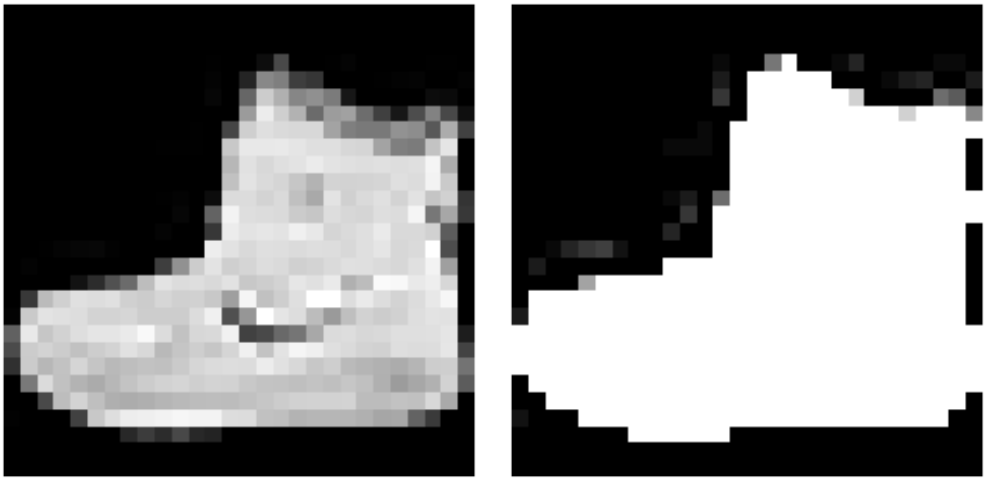

Рис. 1: Lime features decision

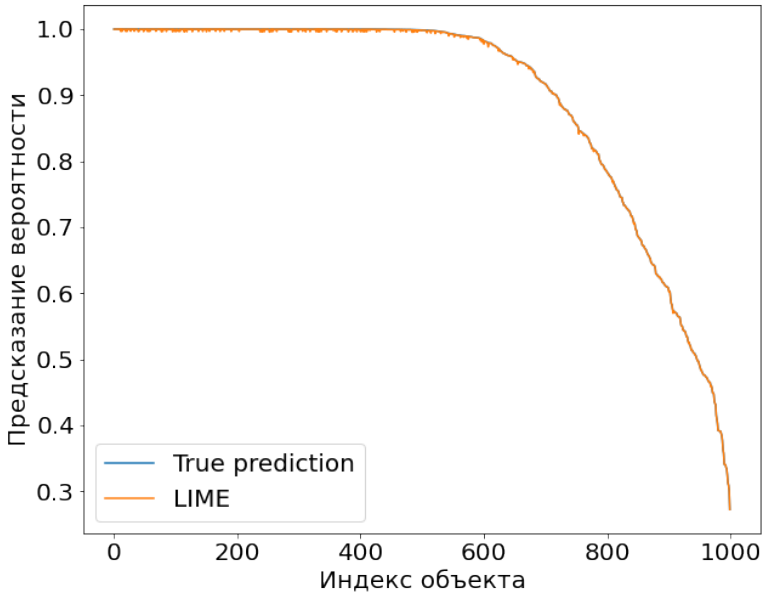


Рис. 2: Lime accuracy

### 5.2 Конфигурация запуска алгоритма

Считаем точность предсказаний и расстояние между признаками, полученные при помощи алгоритма LIME [1] (рисунок 2).

### 5.3 Предварительный отчет

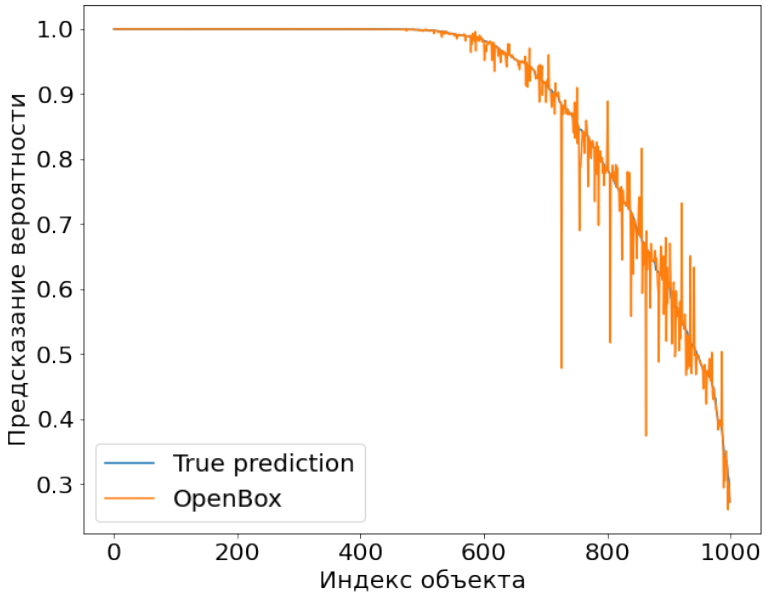


Рис. 3: OpenBox accuracy

### 5.4 Анализ ошибки

Рассмотриваем простейшую модель сверточной нейронной сети и адаптируем методу OpenBox [3], далее сравниваем полученные результаты с приминением базового метода LIME [1].

Эксперимент заключается в анализе влияния изменения картинок на значения градиентов при backward проходе. Что позволяет выделить на картинках наиболее важные признаки и степень их влияения на результат работы сверточной нейронной сети. График точности метода OpenBox указан на рисунке 3.

### 5.5 Заключение

Результаты исследовательской работы:

1. Предложена адаптация метода OpenBox в применении к работе со сверточными нейронными сетями.
2. Доказана гипотеза о линейности слоев сверточных нейронных сетей.

## Список литературы

[1] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier, 2016.

[2] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[3] Lingyang Chu, Xia Hu, Juhua Hu, Lanjun Wang, and Jian Pei. Exact and consistent interpretation for piecewise linear neural networks: A closed form solution, 2019.

[4] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.