Выбор интерпретируемых сверточных моделей глубокого обучения

Тимур Русланович Мурадов

Московский физико-технический институт

Курс: Моя первая научная статья
(В. В. Стрижов)/Группа Б05-907б

Консультанты: О. Бахтеев, К. Яковлев

2022

Цель исследования

Задача

Выбор интерпретируемой сверточной нейронной сети.

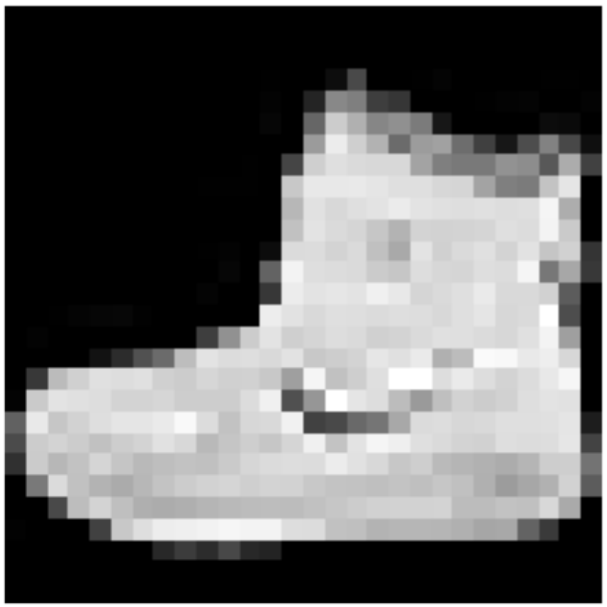
Проблема

Высокая сложность интерпретации сверточных нейронных сетей.

Решение

Адаптация метода OpenBox для работы со сверточными нейронными сетями.

Интерпретируемость изображений



Выделение наиболее важных признаков для изображений подразумевает собой подсветку формы объекта соответствующего класса.

Литература

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.

Постановка задачи: модель и задача оптимизации

Задана выборка $x \in X$, где $X \in \mathbb{R}^m$. Вектор меток классов $y \in \{1, 2, ...k\}$ — заданное конечное множество классов.
Модель $f(X, w)$ — сверточная нейронная сеть, это суперпозиция подмоделей $f_1 \circ f_2 \dots f_n$.
Функции f_i — слои нейронной сети, это одни из функций: линейные, свертки, операции побатчевой нормализации или пулинги.
В модели $f(X, w)$ оптимизируется функция кросс-энтропии $\mathcal{L}(g, y)$, g — функция softmax, g: $\mathbb{R}^m \to \{1, \dots, k\}$, на выходе предсказанное распределение вероятности соответствия объектов классам.

$$g(x)_i = \frac{\exp(x_i)}{\Sigma_j \exp(x_j)}.$$

$$\mathcal{L} = -\Sigma_i y_i \log g(x)_i \to \max.$$

Постановка задачи: требования к модели

Модель должна удовлетворять двум требованиям к интерпретируемости: **точность** и **консистентность**.

1. Точность: Близость предсказаний построенной модели к исходной.

$$\|f(X, w) - f_{method}(X, w)\| \to \min_{f_{method}}$$

   Где f - исходная модель, f_{method} - построенная модель.

2. Консистентность: Близкие интерпретации для близких объектов выборки.

$$x_i \in U_\epsilon(x_j) \Longrightarrow f_{method}(x_i, w) \in U_{f_{method}(\epsilon, w)} f_{methood}(x_j, w).$$

Постановка задачи: критерий качества

Критерием качества рассматривается точность предсказания класса объектов. Отличие оценок предсказаний для метода от истинных предсказаний, полученных из классификатора.

Решение задачи интерпретации CNN

Предлагается адаптация 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. Метод обобщается на работу с более широким классом нейронных сетей: сверточными нейронными сетями.

Линейность сверточных нейронных сетей

Теорема

Слои сверточной нейронной сети: линейные, свертки, операции побатчевой нормализации, пулинги — это линейные операции.

Описание вычислительного эксперимента

Фиксируется подвыборка размера 1000 из датасета **Fashion-MNIST**. Для каждого объекта генерируется похожий объект изменением случайного пикселя картинки.

Строится интерпретация для каждого объекта из подвыборки. Считается предсказание вероятности класса объекта, на самом объекте при помощи аппроксимации для базового метода **LIME** и на близком объекте для предложенного метода **OpenBox**.

Для методов строятся графики предсказания наиболее вероятного класса относительно предсказания классификатора исходной модели.

Базовый эксперимент LIME

На графике представлена работа базового метода **LIME** по предсказанию класса объекта.

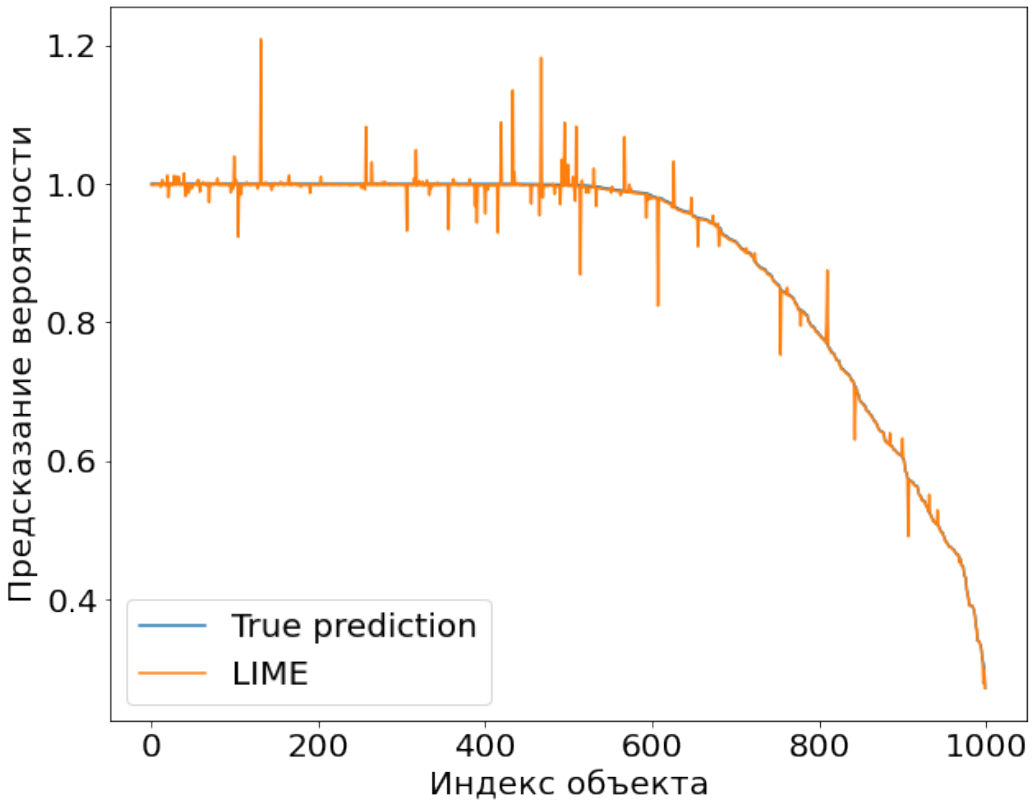
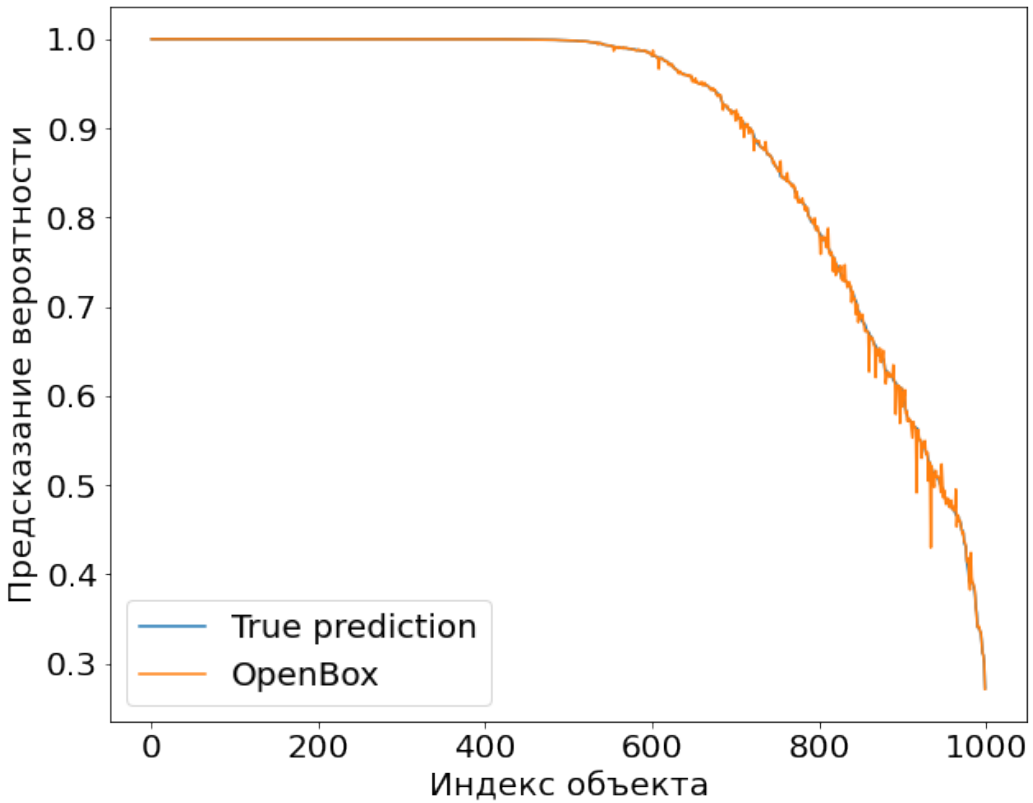


График говорит о наличии значимых отклонений в работе базового пакета.

Эксперимент по адаптации метода OpenBox

На графике представлена работа метода **OpenBox** по предсказанию класса на близких объектах.



Заметим отсутствие больших скачков в предсказаниях на графике.

Заключение

Результаты

- Предложена адаптация метода **OpenBox** в применении к работе со сверточными нейронными сетями.
- Доказана теорема о линейности слоев сверточных нейронных сетей.
- Проведен вычислительный эксперимент, по результатам которого показана более высокая точность полученного метода **OpenBox** по сравнению с базовым методом **LIME**