
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 интерпретируемости 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 интерпретируемость. Вычислительный эксперимент проводится на выборках изображений рукописных цифр MNIST и изображений CIFAR-10.

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

1 Introduction

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 аппроксимацию 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 зависит от точности аппроксимации. 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 зависит 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 интерпретации CNN.

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

2 Problem Definition

Задана выборка $\mathbf{x} \in \mathbf{X}$ двумерные трехканальные изображения. $\mathbf{X} \in \{1, 2, \dots, k\}$, заданное конечное множество классов.

Рассматривается задача построения модели 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 $x_i^{(k)} = \frac{x_i^{(k)} - E(x_i^{(k)})}{\sqrt{D(x_i^{(k)})}}$ или пулинги.

$\mathbf{f}(\mathbf{X}, \mathbf{w})$ оптимизирует функцию кросс энтропии \mathcal{L} , \mathbf{g} — функция softmax.

$$\mathbf{g}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\Sigma_j e^{\mathbf{x}_j}}$$

$$\mathcal{L} = -\Sigma_i \log \mathbf{g}(\mathbf{x})_i \rightarrow \max$$

Кроме задачи оптимизации модель также должна удовлетворять следующим требованиям к интерпретируемости: Точность и Консистентность.

- Точность: Математическая эквивалентность.

$$\mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{f}_{\text{true}}(\mathbf{X}, \mathbf{w})$$

- Консистентность: Близкие интерпретации для близких объектов выборки.

$$\mathbf{x}_i \approx \mathbf{x}_j \implies \mathbf{f}(\mathbf{x}_i, \mathbf{w}) \approx \mathbf{f}(\mathbf{x}_j, \mathbf{w})$$

3 Theory

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

- Линейный слой линеен по определению.

$$\mathbf{f}_i = \mathbf{w}_0 + \Sigma \mathbf{w}_i * \mathbf{x}_i$$

- Свертка представима как линейная операция, если расписать её в специфичном виде как произведение матрицы входного изображения на матрицу с весами фильтра.
- Пулинг на максимум представим как взаимодействие фильтра на изображение как на политоп.
- Батч нормализация представима в виде скалярного произведения, примененная поэлементно к каждому изображению.

4 Computational experiment

Цель эксперимента: проверить работоспособность базового метода LIME [1] и сравнить качество работы с предлагаемой альтернативой OpenBox [3]. Критерием качества рассматривается точность предсказания класса объектов. Отличие $P(i)$ для метода от истинной вероятности, полученной из классификатора, где P - вероятность принадлежности к классу, i - индекс рассматриваемого объекта.

5 Our setup

Строим CNN и при помощи метода LIME [1] получаем интерпретации признаков модели.

5.1 Data

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].

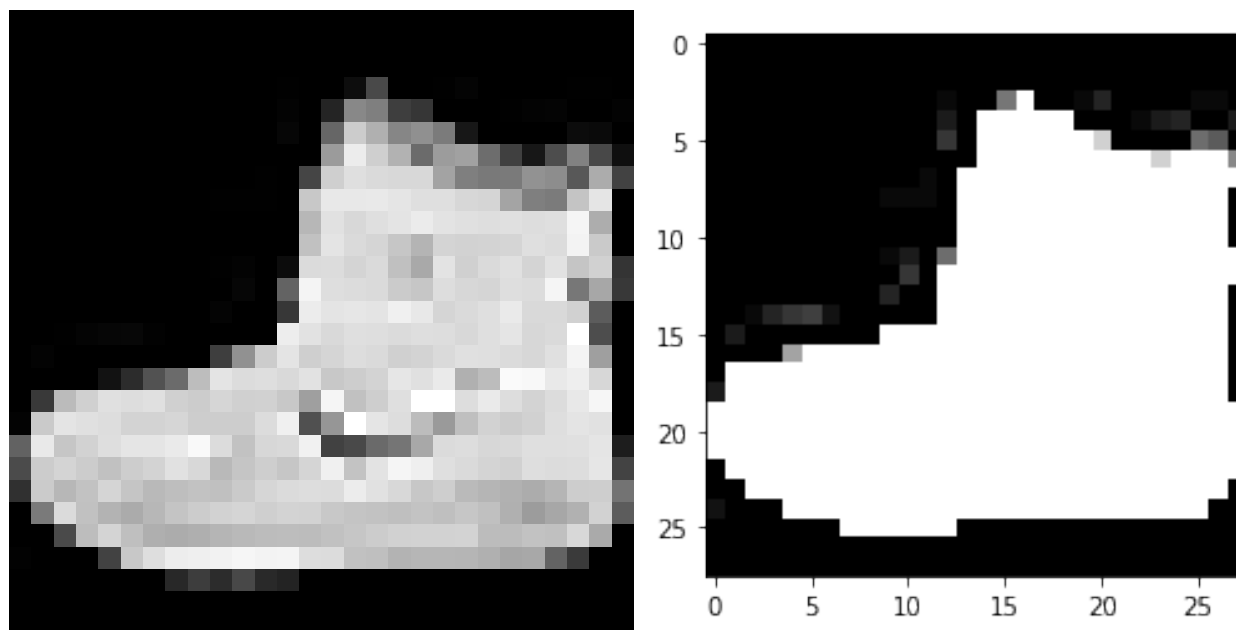


Рис. 1: Lime features decision

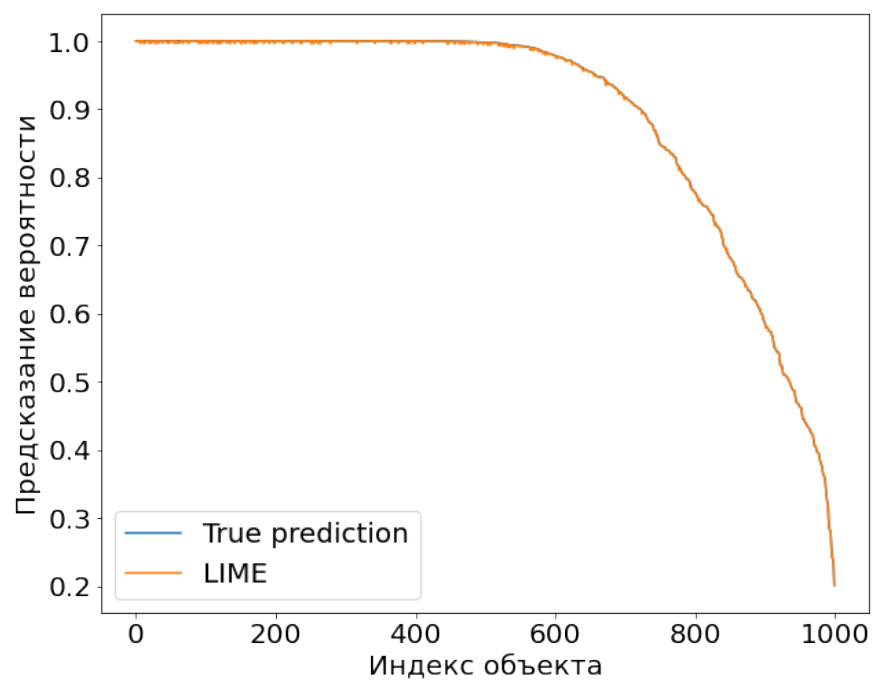


Рис. 2: Lime accuracy

5.2 Configuration of algorithm run

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

5.3 Preliminary report

5.4 Error analysis

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

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.