
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 выборках изображений рукописных цифр MNIST и изображений CIFAR-10.

2 Problem Definition

Задана выборка изображений для классификации.

Для CNN \mathcal{N} содержащей L слоев назовем l слой \mathcal{L}_l . Так \mathcal{L}_1 - входной слой, а \mathcal{L}_L - выходной слой. Слои со 2 по $L - 1$ - это скрытые слои. На вход сети \mathcal{N} подается $x \in \mathcal{X}$, где $X \subseteq \mathbb{R}^d$, из пространства размерности d .

Слои с 2 по $L - 1$ свёрточной сети \mathcal{N} представляют собой свёртки, пулинги и нормализации. Каждый из которых можно представить в виде линейных операций. Между слоями применяются кусочно линейные функции активации.

На выходе сети \mathcal{N} получается вектор $a \in \mathcal{Y}$, где $Y \subseteq \mathbb{R}^{n_L}$, пространство размерности n_L . Выходной слой сети \mathcal{L}_L применяет функцию softmax для получения выходного вектора.

CNN работает как классификатор $F : \mathcal{X} \rightarrow \mathcal{Y}$, сопоставляющий элементам из множества \mathcal{X} элементы из множества \mathcal{Y} . При этом F представляет собой систему тяжело интерпретируемую человеком.

Требуется построить интерпретацию CNN \mathcal{N} в виде линейной модели \mathcal{M} , хорошо поддающейся интерпретации. Таким образом для \mathcal{M} выдвигаются два требования:

- Точность: Модель \mathcal{M} математически эквивалента модели \mathcal{N} .
- Консистентность: Модель \mathcal{M} даёт близкие интерпретации для близких объектов выборки.

3 Our setup

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

4 Computational experiment

Эксперимент заключается в получении baseline качества интерпретаций для дальнейшего исследования.

4.1 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].

4.2 Configuration of algorithm run

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

4.3 Preliminary report



Рис. 1: Accuracy and cosine similarity between decision features

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.