Выбор интерпретируемых сверточных моделей глубокого обучения

A Preprint

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, а также определение кластеров схожих объектов. Для улучшения интерпретируемости в статье вводится модификация метода OpenBox работающего с кусочно-линейными нейронными сетями. В нём модель представляется в виде набора интерпретируемых линейных классификаторов, при этом каждый из них определен на выпуклом многограннике, что 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 позволяет значительно улучшить интепретируемость. 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 улучшения интерпретируемости модели, где под интерпретируемостью понимается простота выделения важных признаков на выборке данных и способность относить схожие объекты выборки к одним и тем же кластерам.

Проблемой является в целом высокая сложность интерпретации сверточных нейронных сетей, требующая комплексного подхода. На данный момент существует множество различных решений проблемы интерпретации. 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, являясь при этом “model-agnostic”, то есть никак не использующий информацию о строении модели изнутри. Но он весьма неустойчив к выбросам и сильно зависим от адекватности апроксимации. В статье [2] предлагается другой подход SHAP, заключающийся в рассмотрении вклада каждого признака в результат работы модели. Таким образом удается выделять даже скрытые, но значимые признаки. Однако применимость данного подхода ограничена ввиду высоких вычислительных затрат, требуется многократное обучение модели, а также 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 доказательство конкурентоспособности по сравнению с другими существующими методами интепретации CNN.

Для анализа качества предложенного метода проводится вычислительный эксперимент на выборках изображений рукописных цифр MNIST и изображений CIFAR-10.

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.

2 Headings: first level

2.1 Headings: second level

3 Examples of citations, figures, tables, references

3.1 Citations

3.2 Figures

3.3 Tables

3.4 Lists