Выбор интерпретируемых сверточных моделей глубокого обучения

A Preprint

Тимур Мурадов
МФТИ

Олег Бахтеев
МФТИ

Константин Яковлев
МФТИ

Abstract

В статье рассматривается проблема слабой интерпретируемости сверточных нейронных сетей, то есть затруднённого выделения наиболее важных признаков, а также определения кластеров схожих объектов. Для улучшения интерпретируемоси в статье ведётся модификация доказавшего свою эффективность метода OpenBox работающего с кусочно-линеными нейронными сетями. Метод обобщается на работу с более широким классом нейронных сетей: сверточными нейронными сетями. Предлагается математически эквивалентная замена слоев: свертка, пулинг, нормализация на линейные, что позволяет значительно улучшить интепретируемость.

Keywords Machine Learning · CNN · OpenBox · Explicit

1 Introduction

При работе со сверточными нейронными сетями частой проблемой является повышение качества, а также анализ получаемых результатов. 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 [citeribeiro2016why]описан метод LIME, предлагащий 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, однако он весьма неустойчив к выбросам и сильно зависим от адекватности апроксимации. В статье [?] предлагается другой подход SHAP, заключающийся в рассмотрении вклада каждого признака в результат работы модели, таким образом удается выделять даже скрытые, но значимые признаки, но применимость подхода ограничена ввиду высоких вычислительных затрат, так как требует многократного обучения модели, а также весьма зависим от выборки данных. Ещё один подход к интерпретации OpenBox, описываемый в статье [?] редлагает построение математически эквивалентных линейных моделей для линейных нейронных сетей, он показал более высокую эффективность по сравнению с LIME и весьма перспективен для дальнейшей работы.

Задачей проекта является 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.

В качестве валидационного датасета предложены: выборка изображений рукописных цифр MNIST и выборка изображений CIFAR-10. Эксперимент выполнялся локально в среде Jupiter Notebook, а также удаленно задействуя графический ускоритель в среде Google Collab. Для построения моделей был выбран язык Python по средствам библиотек с открытым кодом Pytorch и Scikit-Learn.