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 интерпретируемоси 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 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 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 апроксимацию 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 зависим от точности апроксимации. 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 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 сравнение с альтернативными методами интепретации CNN.

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

2 Problem Definition

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

Рассматривается задача построения модели глубокого обучения для задачи классификации.

Модель $f(X, w)$ - сверточная нейронная сеть, для краткости CNN, это суперпозиция подмоделей $f_1 \circ f_2 \dots f_n$.

Функции f_i - слои нейронной сети, это одни из функций: линейные $y_i = w_0 + \Sigma w_i * x_i$, свертки, батч-нормы или пулинги.

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

$$g(x)_i = \frac{e^{x_i}}{\Sigma_j e^{x_j}}$$

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

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

- Точность: Математическая эквивалентность.
- Консистентность: Близкие интерпретации для близких объектов выборки.

3 Our setup

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

4 Computational experiment

Эксперимент заключается в получении baseline качества интрепретаций 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 [1].

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.