Выбор интерпретируемых сверточных моделей глубокого обучения

Тимур Русланович Мурадов

Московский физико-технический институт

Курс: Моя первая научная статья
(В. В. Стрижов)/Группа Б05-907б

Консультант: О. Бахтеев

Консультант: К. Яковлев

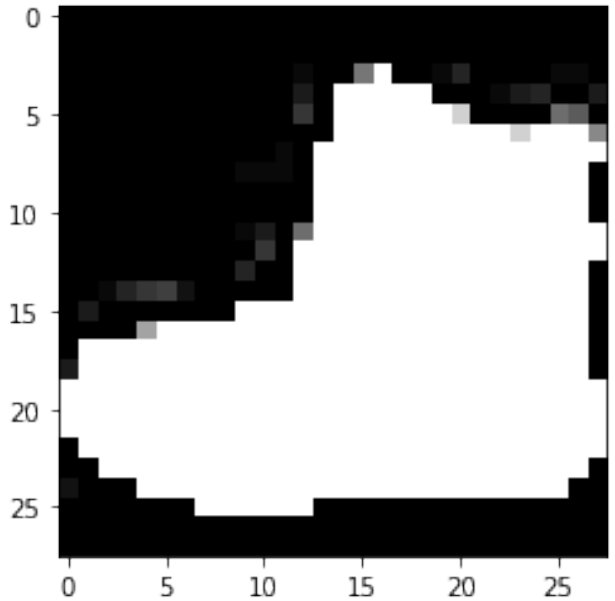
2022

Цель исследования

Решается задача построения интерпретируемой сверточной нейронной сети

Под интерпретируемостью понимается простота выделения важных признаков на выборке данных и классификация близких объектов одним и тем же классификатором.

Интерпретируемость изображений



Выделение наиболее важных признаков для изображений подразумевает собой подсветку формы объекта соответствующего класса.

Литература

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.

Постановка задачи: модель и задача оптимизации

Задана выборка $x \in X$ двумерные трехканальные изображения.
$X \in \{1, 2, ...k\}$, 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 $f_i = w_0 + \Sigma w_i * x_i$, свертки
$S(i,j) = (K * I)(i,j) = \Sigma_m \Sigma_n I(i+m, j+n) K(m,n)$, батч-нормы $\hat{x_i}^{(k)} = \frac{x_i^{(k)} - E(x_i^{(k)})}{\sqrt{D(x_i^{(k)})}}$ и пулинги.
$f(X, w)$ оптимизирует функцию кросс энтропии $\mathcal{L}$, g — функция softmax.

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

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

Постановка задачи: требования и критерии качества

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

- **Точность**: Математическая эквивалентность.

$$f(X, w) = f_{true}(X, w)$$

- **Консистентность**: Близкие интерпретации для близких объектов выборки.

$$x_i \approx x_j \Longrightarrow f(x_i, w) \approx f(x_j, w)$$

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.

Решение задачи интерпретации CNN

Предлагается адаптация 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. Метод обобщается на работу с более широким классом нейронных сетей: сверточными нейронными сетями.

Теоретическая часть

Гипотеза

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

Доказательство.

- Линейный слой линеен по определению.
- 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 в виде взаимодействия фильтра на изображение как на политоп.
- Батч нормализация представима как скалярное произведение, применённое поэлементно к каждому изображению.

□

Базовый эксперимент

На графике представлена работа базового метода **LIME** по предсказанию класса объекта.

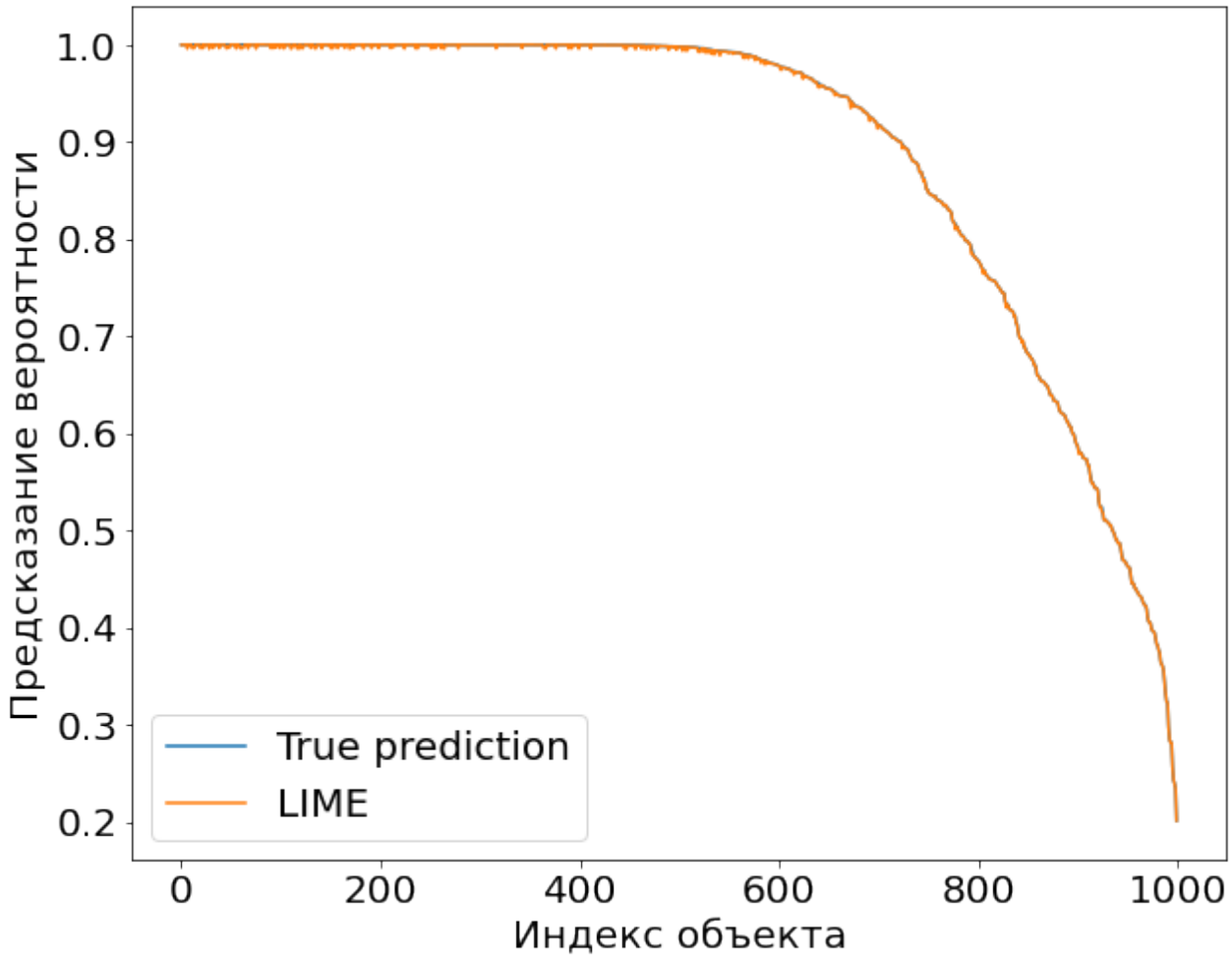


График говорит о хорошей аппроксимации базового метода.

Эксперимент по адаптации метода OpenBox

Рассматриваем простейшую модель сверточной нейронной сети и адаптируем метод **OpenBox**, далее сравниваем полученные результаты с приминением базового метода **LIME**.
Эксперимент заключается в анализе влияния изменения картинок из датасета **Fashion MNIST** на значения градиентов при backward проходе. Что позволяет выделить на картинках наиболее важные признаки и степень их влияения на результат работы сверточной нейронной сети.

Заключение

Результаты

- Предложена адаптация метода OpenBox в применении к работе со сверточными нейронными сетями.
- Доказана гипотеза о линейности слоев сверточных нейронных сетей.