Выбор интерпретируемых рекуррентных моделей глубокого обучения

Гапонов Максим

Московский физико-технический институт

Эксперт: Бахтеев Олег

Консультант: Яковлев Константин

2022

Цель исследования

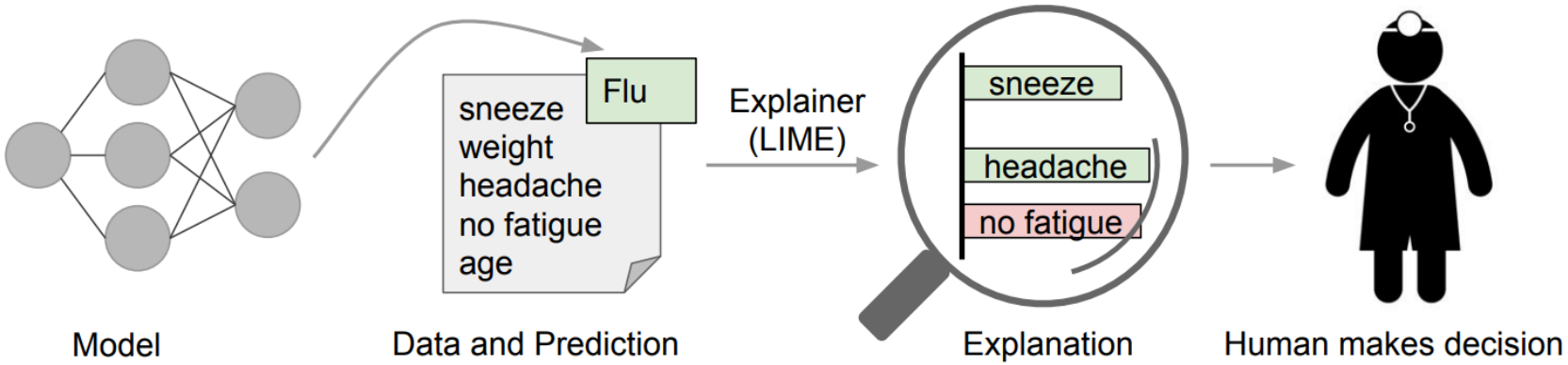
Решается задача выбора интерпретируемых рекуррентных моделей.

Требуется найти метод получения интерпретаций предсказаний произвольной рекуррентной модели глубокого обучения с кусочно-линейными функциями активации.

Предлагается обобщить ранее известный метод OpenBox, который предназначен для обыкновенных нейронных сетей с кусочно-линейными функциями активации.

Выбор интерпретируемых рекуррентных моделей

Получение интерпретации предсказаний модели глубокого обучения необходимо для того, чтобы эксперты в различных областях могли "доверять"модели и принимать решения на основе её предсказаний.



Интерпретация должна быть **точной**, то есть поведение интерпретируемой модели похоже на исходную модель, и **согласованной**, то есть интерпретации близких объектов должны быть похожи.

Основная литература

- 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.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.
- Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning, 2015.

Обозначения

L — число слоёв нейронной сети
$\mathbf{x}_0$ — исходный объект
$\mathbf{a}_i$ — вход слоя с номером i
f_i — функция активации после слоя с номером i
$\mathbf{W}_i$ — матрица линейного преобразования слоя с номером i
$\mathbf{b}_i$ — вектор сдвига

$$a_i = f_i\left(\mathbf{W}_i\mathbf{a}_i + \mathbf{b}_i\right)$$

$$\mathbf{a}_1 = \mathbf{x}_0$$

Рассматриваются кусочно-линейные функции активации, то есть

$$f(x) = \begin{cases} r_1 x + t_1 & \text{if } x \in I_1 \\ r_2 x + t_2 & \text{if } x \in I_2 \\ \dots & \dots \\ r_u x + t_u & \text{if } x \in I_u \end{cases}$$

Описание метода

Исходная модель представляется в виде математически эквивалентного множества линейных классификаторов $F_1, \ldots, F_h$, заданных на выпуклых многогранниках $P_1, \ldots, P_h \subset \mathcal{X}$, где $\mathcal{X}$ – пространство признаков.

Линейный классификатор с небольшим числом ненулевых коэффициентов считается интерпретируемой моделью.

Свойство точности интерпретаций выполнено автоматически, так как построенная модель математически эквивалентна исходной.

Также интерпретации будут согласованными, так как близкие объекты в пространстве $\mathcal{X}$ будут принадлежать одному выпуклому многограннику P_i, а значит, будут классифицироваться одним и тем же линейным классификатором F_i.

Вычислительный эксперимент

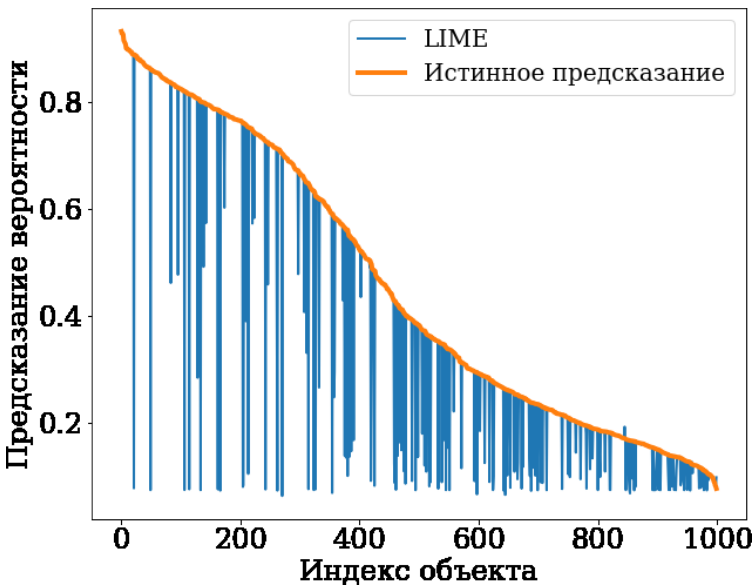
Зафиксируем подвыборку размера 1000. Для каждого объекта найдём ближайший в смысле коэффициента Жаккара.

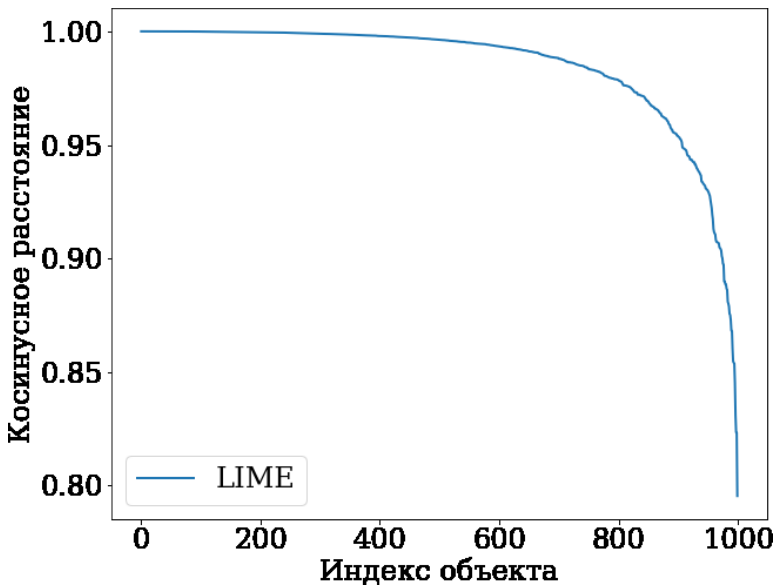
Построим интерпретацию каждого объекта по отдельности. Далее будем оценивать предсказания исходной модели либо на текущем объекте (для метода LIME), либо на соседнем объекте (для обобщения метода OpenBox).

Построим график истинных предсказаний модели и оценок предсказаний, полученных данными методами. Также построим графики косинусного расстояния между истинными предсказаниями и их оценками. Помимо этого, измерим точность предсказаний метрикой RMSE (Root mean squared error).

Вычислительный эксперимент

Существующий метод **LIME** не обладает достаточной точностью.

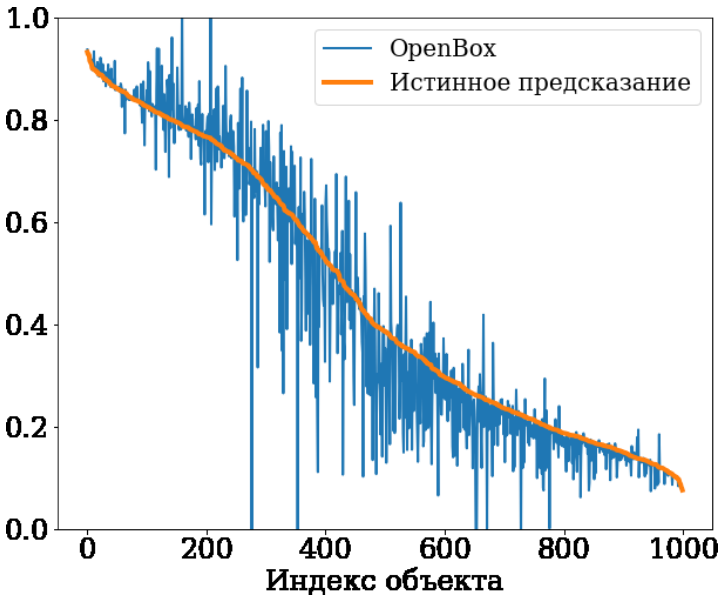


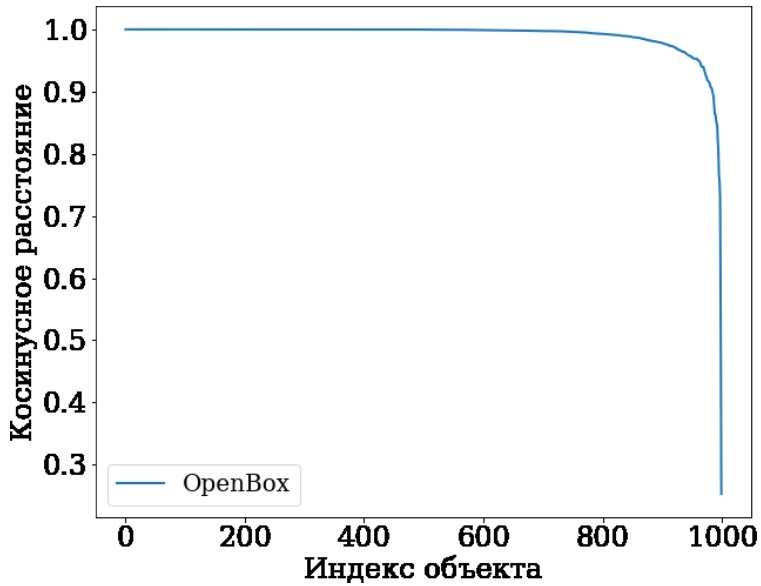


Как видно из графиков, оценки предсказаний, полученные методом LIME значительно отклоняются от истинных для большой доли объектов.

Вычислительный эксперимент

Построим аналогичные графики для метода OpenBox. Напомним, что в данном случае оценивается предсказание модели не в исходном объекте, а в соседнем.





RMSE

- OpenBox: 0.083
- LIME: 0.118

Заключение

- Предложен метод выбора интерпретируемых рекуррентных моделей с кусочно-линейными функциями активации
- Доказана теорема о представлении произвольной рекуррентной нейронной сети в виде математически эквивалентного набора линейных классификаторов
- Проведён вычислительный эксперимент, который показал, что предложенный в работе метод обладает более высокой точностью и согласованностью предсказаний, чем ранее известный метод.