Выбор интерпретируемых рекуррентных моделей глубокого обучения

A Preprint

Гапонов Максим
gaponov.me@phystech.edu

Бахтеев Олег
bakhteev@phystech.edu

Яковлев Константин
...

Стрижов Вадим
...

Abstract

В данной работе рассматривается задача интерпретации рекуррентных нейронных сетей. Интерпретируемость нейронных сетей необходима для повышения уровня доверия к предсказаниям таких моделей. Для решения задачи предлагается обобщить метод OpenBox, предназначенный для кусочно-линейных нейронных сетей. Модель представляется в виде набора интерпретируемых линейных классификаторов. Каждый из классификаторов определён на выпуклом многограннике, поэтому интерпретации близких объектов оказываются согласованными. Для проверки работоспособности предложенного метода проводится эксперимент на выборке DBLP (данные о цитированиях).

1 Введение

В работе рассматривается метод интерпретации рекуррентных моделей глубокого обучения. Такие модели используются для обработки последовательностей данных [5]. Однако для принятия решения при помощи такой модели необходимо доверять её предсказаниям. По этой причине необходимо построить интерпретируемую модель, объясняющую предсказания исходной модели. Работа посвящена построению и выбору таких интерпретируемых моделей для произвольной рекуррентной сети.

Рекуррентная нейронная сеть представляет из себя последовательность из слоёв нейронов: входного слоя, скрытых слоёв и выходного слоя [6]. Отличительной особенностью от обыкновенных нейронных сетей являются циклические связи между нейронами. Такие связи позволяют запоминать состояния для обработки последовательностей данных.

Интерпретация должна быть точной и согласованной [3]. Интерпретация называется точной, если она согласуется с поведением модели. То есть показывает значимыми те признаки, которые значительно повлияли на предсказание модели. Интерпретация называется согласованной, если интерпретации близких объектов похожи на исходную интерпретацию. То есть метод похожим образом объясняет предсказания на близких объектах.

Анализ зависимости предсказания модели от входных данных в реккурентных нейронных сетях является сложной задачей. Методы, рассмотренные в работах [4] [8] [1] [2] [9] [7] являются либо неточными, то есть не отвечают реальному поведению модели, либо несогласованными, то есть совершенно разным образом объясняет предсказания на близких объектах.

В работах [4] [8] рассмотрены методы, основанные на интерпретации признаков, выученных скрытыми слоями. Такой подход отражает поведение скрытых слоёв, но не показывает поведение сети в целом.

В работах [1] [2] предлагаются методы подражания модели. В этом подходе строится модель с похожим на исходную поведением. Построенная модель имеет более простую структуру, поэтому проще интерпретируется. Однако из-за недостаточно сложной структуры построенная модель недостаточно хорошо приближает исходную.

Наконец, в работах [9] [7] рассмотрены методы локальной интерпретации, в которых происходит анализ поведения модели в окрестности исходного объекта. В таком подходе получается точная интерпретацию

для одного объекта, однако интерпретации для близких объектов могут существенно отличаться. Таким образом, данный подход не предоставляет согласованные интерпретации.

Рассмотренный в данной работе подход является обобщением метода OpenBox. В [3] предложен метод OpenBox для интерпретации кусочно-линейных нейронных сетей. Модель представляется в виде эквивалентного набора линейных классификаторов. Каждый из них классифицирует объекты внутри некоторого выпуклого многогранника в пространстве признаков, поэтому интерпретации получаются согласованными.

На выборке DBLP проведён вычислительный эксперимент для анализа качества метода, а также для проверки интерпретаций на точность и согласованность.

2 Постановка задачи

Рассмотрим рекуррентную нейронную сеть $\mathcal{N}$ с кусочно-линейными функциями активации. Входные данные обозначим через $x \in \mathcal{X}$, где $\mathcal{X} \subset \mathbb{R}^d$ – пространство признаков. Предсказание модели $\mathcal{N}$ обозначим через $y \in \mathcal{Y}$, где $\mathcal{Y}$ – пространство предсказаний.

Модель $\mathcal{N}$ представляет из себя функцию классификации $F : \mathcal{X} \to \mathcal{Y}$. Из-за сложной структуры сети трудно понять поведение функции F, поэтому необходимо получить интерпретацию, которую способен понять человек.

Функция F является кусочно-линейной, так как $\mathcal{N}$ – рекуррентная нейронная сеть с кусочно-линейными функциями активации. Области, на которых функция F линейна являются выпуклыми многогранниками.

Предлагается рассмотреть модель $\mathcal{M}$, представляющую из себя множество линейных классификаторов $\{F_1, F_2, \dots, F_h, \dots\}$. Пространство признаков $\mathcal{X}$ разбивается на конечное множество выпуклых многогранников $\{P_1, P_2, \dots, P_h, \dots\}$. То есть $\bigsqcup_h P_h = \mathcal{X}$. Классификатор F_h определён на многограннике P_h. Модель $\mathcal{M}$ должна быть математически эквивалентна модели $\mathcal{N}$, то есть множество функций $\{F_1, F_2, \dots, F_h, \dots\}$ определённых на $\{P_1, P_2, \dots, P_h, \dots\}$ тождественно равно функции F, определённой на $\mathcal{X}$.

Модель $\mathcal{M}$ обладает свойством точности, так как математически эквивалентна модели $\mathcal{N}$. То есть поведение двух моделей совпадает. $F|_{P_h} \equiv F_h$. Модель $\mathcal{M}$ также обладает свойством согласованности, так как близкие объекты попадают в один и тот же выпуклый многогранник P_h, а значит, классифицируются одним и тем же линейным классификатором F_h.

Задача состоит в поиске такой модели $\mathcal{M}$, построении множества классификаторов по произвольной рекуррентной нейронной сети $\mathcal{N}$.

Список литературы

[1] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.

[2] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction, 2019.

[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] Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks, 2016.

[5] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning, 2015.

[6] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena, 404:132306, Mar 2020.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[8] Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection, 2018.

[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015.