Выбор интерпретируемых рекуррентных моделей глубокого обучения

Гапонов Максим, Бахтеев Олег, Яковлев Константин, Стрижов Вадим
Московский физико-технический институт
{gaponov.me, bakhteev, strijov}@phystech.edu

Аннотация

Рассматривается задача интерпретации рекуррентных нейронных сетей. Под интерпретируемостью модели понимается возможность получить линейную зависимость выходных данных от входных. Предлагается обобщить метод OpenBox, предназначенный для интерпретации нейронных сетей с кусочно-линейными функциями активации. В данном методе нейронная сеть представляется в виде ансамбля интерпретируемых линейных классификаторов, каждый из которых определён на выпуклом многограннике, поэтому интерпретации близких объектов согласованны. Для анализа качества предложенного метода проводится эксперимент на выборке IMDB Review dataset.

Abstract

Consider the problem of interpretation of recurrent neural networks. We understood the interpretability of the model as the ability to obtain a linear dependence of the output data on the input data. We propose to generalize the OpenBox method designed for the interpretation of neural networks with piecewise linear activation functions. In this method, the neural network is represented as an ensemble of interpreted linear classifiers, each of which is defined on a convex polyhedron. Therefore, the interpretations of close objects are consistent. To test the operability of the proposed method, we conducted an experiment on a sample of the IMDB Review dataset.

Ключевые слова: Интерпретация моделей · Рекуррентные нейронные сети · Кусочно-линейные модели

1 Введение

В работе рассматривается метод интерпретации рекуррентных моделей глубокого обучения. Такие модели используются для обработки последовательностей данных [1]. Однако

из-за большого числа слоёв в нейронной сети она работает по принципу "чёрного ящика". То есть нет явной зависимости выходных данных от входных. Работа посвящена построению и выбору интерпретируемых моделей для рекуррентных нейронных сетей с кусочно-линейными функциями активации.

Рекуррентная нейронная сеть представляет последовательность слоёв нейронов: входного слоя, скрытых слоёв и выходного слоя [2]. Отличительной особенностью от 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 предсказания интерпретируемой модели похожи на предсказания исходной. Интерпретация называется согласованной, если интерпретации близких объектов похожи, то есть значимости признаков близки.

Анализ зависимости предсказания модели от входных данных в рекуррентных нейронных сетях — открытая проблема. Методы, рассмотренные в работах [4, 5, 6, 7, 8, 9] являются либо неточными, то есть построенные интерпретируемые модели не отвечают поведению исходной модели, либо несогласованными, то есть значимости признаков для близких объектов значительно отличаются.

В работах [4, 5] рассмотрены методы, основанные на интерпретации признаков, выученных скрытыми слоями. Такой подход отражает поведение скрытых слоёв, но не показывает поведение сети в целом. В работах [6, 7] 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 плохо приближает исходную. Наконец, в работах [8, 9] рассмотрены методы локальной интерпретации, в которых происходит анализ поведения модели в окрестности исходного объекта. В таком подходе получается точная интерпретация одного объекта, однако интерпретации близких объектов могут существенно отличаться. Таким образом, данный подход не предоставляет согласованные интерпретации.

Предложенный в данной работе подход является обобщением метода OpenBox, предназначенного для интерпретации нейронных сетей с кусочно-линейными функциями активации [3]. Нейронная сеть представляется в виде ансамбля линейных классификаторов. Каждый из них классифицирует объекты внутри выпуклого многогранника в пространстве признаков, поэтому интерпретации получаются согласованными.

На выборке IMDB Movie Review Dataset [10] проведён вычислительный эксперимент для анализа качества метода, а также для проверки интерпретаций на точность и согласованность.

2 Задача выбора интерпретируемых моделей

Рассмотрим рекуррентную нейронную сеть $\mathbf{f}_{\mathbf{w}}(\mathbf{x})$ с кусочно-линейными функциями активации. Входные данные обозначим через $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbf{X}$, где $\mathbf{X} \subset \mathbb{R}^d$ — простран-

ство признаков. Выходные данные обозначим через $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n \in \mathbf{Y}$, где $\mathbf{Y} \subset \mathbb{R}^k$ — пространство предсказаний.

Модель является функцией $\mathbf{f_w} : \mathbf{X} \to \mathbf{Y}$. Из-за большого числа параметров $\mathbf{w}$ невозможно получить явную зависимость значения функции $\mathbf{f_w}$ от аргумента $\mathbf{x}$.

Требуется построить модель, соответствующую функции $\mathbf{g}$, которая обладает следующими свойствами.

Во-первых модель должна быть интерпретируемой. То есть можно построить линейную зависимость выходных данных от входных.

Также модель должна быть точной, то есть минимизировать

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{f_w}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}_i)\right)^2 \to \min_{\mathbf{g}} \tag{1}$$

Модель должна быть согласованной, то есть должно выполняться

$$\exists \varepsilon > 0 \; \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbf{X} \; ||\mathbf{x}_1 - \mathbf{x}_2|| < \varepsilon \implies \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_1) = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}_2) \tag{2}$$

3 Интерпретация рекуррентных нейронных сетей

Далее описаны методы интерпретации рекуррентных нейронных сетей.

3.1 Метод OpenBox

Рассмотрим рекуррентную нейронную сеть $\mathbf{f}(\mathbf{x})$. Для последовательностей фиксированной длины она представима в виде нейронной сети, если заменить рекуррентные слои на последовательность обыкновенных слоёв нейронов. А именно, пусть дан слой рекуррентной нейронной сети состоящий из n нейронов с кусочно-линейной функцией активации f_i. Обозначим число рекуррентных запусков данного слоя через k. Заменим в архитектуре нейронной сети данный слой на k слоёв, состоящих из n нейронов, с функциями активации f_i. Таким образом, нейронная сеть представляется в виде обыкновенной сети.

Для слоя с номером i введём следующие обозначения:

- $\mathbf{a}_i$ — вход слоя
- $\mathbf{h}_i$ — скрытое состояние
- f_i — функция активации
- $\mathbf{W}_i$ — матрица линейного преобразования
- $\mathbf{b}_i$ — вектор сдвига

Рекуррентная нейронная сеть задаётся соотношением (3)

$$[\mathbf{a}_{i+1}, \mathbf{h}_{i+1}] = f_i\left(\mathbf{W}_i[\mathbf{a}_i, \mathbf{h}_i] + \mathbf{b}_i\right) \tag{3}$$

Рассматриваются кусочно-линейные функции активации, то есть

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

Где $I_1, \dots, I_u$ – разбиение $\mathbb{R}$.

Для фиксированных входных данных $\mathbf{x}$ каждая кусочно-линейная функция f_i является линейной в некоторой окрестности своего аргумента.

Также условие принадлежности аргумента x интервалу I_j – это неравенство вида $p \leq x \leq q$.

Теорема 1. Пусть дана нейронная сеть $\mathbf{f}(\mathbf{x})$ с кусочно-линейными функциями активации f_i. Тогда она представима в виде набора линейных функций $F_1, \dots, F_n$, каждая из которых определена на выпуклом многограннике $P_1, \dots, P_n \subset \mathbf{X}$.

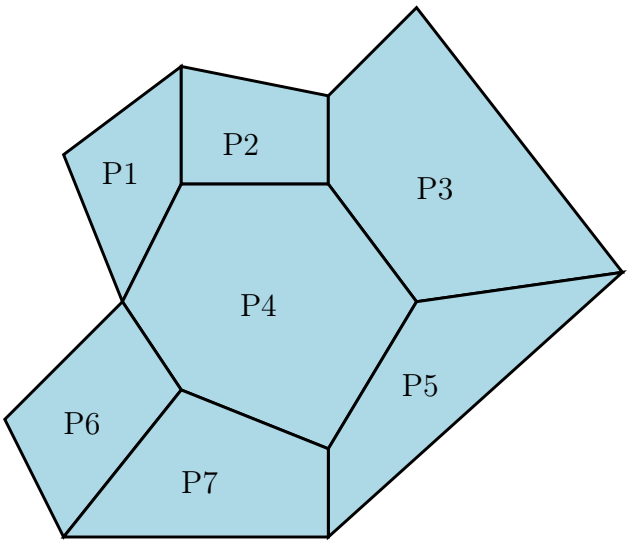


Рис. 1: R^2 разбивается на выпуклые многоульньники $P_1, P_2, P_3 \dots$

Доказательство. Рассмотрим произвольный $\mathbf{x}_0 \in \mathbf{X}$. Кусочно-линейные функции активации f_i являются линейными в некоторой окрестности своего аргумента. Причём условие принадлежности аргумента функции f_i к интервалу I_j – это пересечение двух линейных неравенств, каждое из которых задаёт полупространство в пространстве $\mathbf{X}$. Пересечением такого семейства полупространств является выпуклый многогранник. Помимо этого, композиция линейных функций является линейной функцией.

Таким образом, найден выпуклый многогранник P_h (пересечение семейства полупространств) и линейная функция F_h (композиция линейных функций) такая, что выполнено следующее

$$\forall \mathbf{x} \in P_h \; \mathbf{f}(\mathbf{x}) = F_h(\mathbf{x})$$

В силу произвольности выбора $\mathbf{x}_0$ получаем, что всё пространство $\mathbf{X}$ разбивается на выпуклые многогранники P_h, на которых заданы линейные функции F_h.

Нейронная сеть $\mathbf{f}$ имеет конечное число нейронов, значит, имеется конечное число неравенств, задающих полупространтва в пространстве $\mathbf{X}$. А конечное число неравенств задает лишь конечное число выпуклых многогранников. □

Для решения задачи интерпретации рекуррентных нейронных сетей с кусочно-линейными функциями активации предлагается следующим алгоритм.

1. Вычислить коэффициенты линейного классификатора F_h, (h такое, что $\mathbf{x}_0 \in P_h$) по формуле $\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x}_0)$

2. Выделить линейные неравенства на $\mathbf{x}_0$, соответствующие выпуклому многограннику P_h.

3. Интерпретацией модели на входном объекте $\mathbf{x}_0$ является набор коэффициентов линейного классификатора F_h, соответствующие им признаки, а также набор условий задающих P_h.

Полученная интерпретируемая модель оказывается математически эквивалентна исходной, то есть выполнено свойство точности (1).

Также модель обладает свойством согласованности (2), 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.

3.2 Метод LIME

Рассмотрим метод LIME [11] для выбора интерпретируемых моделей.

Алгоритм построения интерпретируемой модели $\mathbf{g}(\mathbf{x})$

1. Фиксируются константы n и K.

2. Порождаются объекты $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x_n}$ в окрестности $\mathbf{x}$.

3. Вычисляются предсказания $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ модели $\mathbf{f}$ на объектах $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. То есть $\mathbf{y}_i = \mathbf{f}(\mathbf{x}_i)$.

4. Построение линейной модели $\mathbf{g}(\mathbf{x})$ с K параметрами происходит при помощи метода наименьших квадратов для объектов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ и соответствующих им меток $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$.

Таким образом, поведение модели приближается линейной моделью в окрестности исходного объекта.

4 Вычислительный эксперимент

В рамках эксперимента рассматривается задача классификации отзывов на ресурсе IMDB. Используется выборка IMDB Movie Review Dataset [10] состоящяя из 50 000 отзывов, а также типов отзывов: положительный, отрицательный. Требуется по тексту отзыва определить вероятность принадлежности отзыва к классу положительных и отрицательных.

Для решения задачи использовалась рекуррентная нейронная сеть $\mathbf{f}$ (3), аргументом которой являлись последовательности предобученных векторов эмбедингов слов отзыва. Рассматриваемая модель имеет 2 592 105 параметров, она не является интерпретируемой.

Выборка разделена на две части: тренировочную и тестовую, в соотношении 3:1. Модель обучается на тренировочной выборке. После обучения измеряется качество модели на тестовой выборке. Была получена точность классификации (доля верных ответов) 65.55%.

Предсказание полученной методом LIME выглядит следующим образом.

`I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting the theater I not only saw many women in tears, but many full grown men as well, trying desperately not to let anyone see them crying. This movie was great, and I suggest that you go see it before you judge.`

Подсвеченные слова сильнее всего влияют на результат. Модель отнесла отзыв к положительному классу, основываясь на наличии слов `and, able, men, last`. Очевидно, что данные слова не влияют на характер отзыва. Таким образом, даже не учитывая значение метрики accuracy, понятно, что модель непригодна для использования, так как совершает предсказания на основе нерелевантных признаков.

`I went and saw this movie last night after being coaxed to by a few friends of mine. I'll admit that I was reluctant to see it because from what I knew of Ashton Kutcher he was only able to do comedy. I was wrong. Kutcher played the character of Jake Fischer very well, and Kevin Costner played Ben Randall with such professionalism. The sign of a good movie is that it can toy with our emotions. This one did exactly that. The entire theater (which was sold out) was overcome by laughter during the first half of the movie, and were moved to tears during the second half. While exiting`

```
the theater I not only saw many women in tears, but many full grown men as
well, trying desperately not to let anyone see them crying. This movie was
great, and I suggest that you go see it before you judge.
```

Второй пример показывает, что интерпретации несогласованны. А именно, значимые слова отличаются. Такой эффект связан с высокой сложностью рассматриваемой модели.

Вышеперечисленных недостатков мы постараемся избежать в новом подходе.

Для оценки качества интерпретации построим следующие графики:

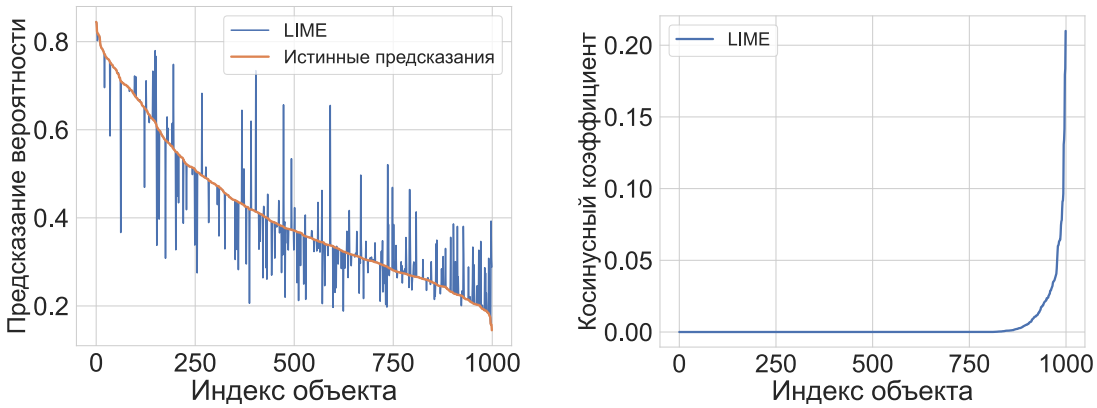
- Предсказаний интерпретируемой модели и исходной модели.
- Косинусный коэффициент между предсказаниями исходной модели и построенной интерпретируемой модели.

$$\frac{(p_{\text{source}}, p_{\text{interpretable}})}{||p_{\text{source}}|| \cdot ||p_{\text{interpretable}}||}$$

5 Анализ ошибки

Зафиксируем выборку размера 1000. Для каждого объекта сгенерируем похожий объект, заменяя некоторые слова на синонимы. Построим интерпретацию на текущем объекте и предскажем вероятность принадлежности к положительному классу для похожего объекта. Чем меньше полученная вероятность отличается от вероятности, которую предсказывает исходная модель, тем лучше.

Для метода LIME будем строить предсказывать вероятности не на похожем объекте, а на исходном.



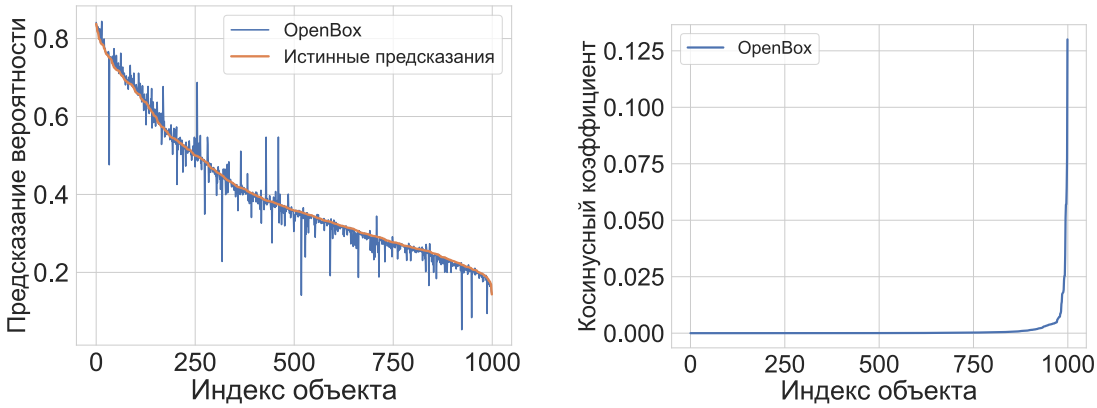


Рис. 2: Точность предсказаний, полученных методами LIME и OpenBox. Слева — графики истинных предсказаний и предсказаний метода, справа — косинусные расстояния между истинными предсказаниями и предсказаниями метода.

Метод LIME недостаточно точно предсказывает вероятности. График получился шумным, предсказанные вероятность существенно отличаются от истинных значений.

Предсказания, полученные методом LIME, для значительной доли объектов отличается по косинусному расстоянию.

Рассмотрим теперь метод OpenBox.

Как видно из графика, есть улучшение по сравнению с методом LIME. Предсказанные вероятности меньше отличаются от исходных.

График косинусного расстояния также показывает, что интерпретации, полученные методом OpenBox являются согласованными. Есть лишь небольшая доля объектов, на которых наблюдается существенное различие между истинным предсказанием и предсказанием метода.

Вычислим метрики качества.

Метод	RMSE	MAE	MAPE
OpenBox	0.01541	0.00849	0.01923
LIME	0.04085	0.01479	0.03923

Ошибки метода OpenBox меньше, даже несмотря на то, что измерения для данного метода проводились не на исходных объектах, а на похожих.

6 Заключение

В работе был предложен метод выбора интерпретируемой рекуррентной модели глубокого обучения, обладающий свойствами точности и согласованности интерпретаций.

За основу был взят метод OpenBox, предназначенный для интерпретации нейронных сетей с кусочно-линейными функциями активации.

Было проведено сравнение предложенного метода с методом LIME. Точность и согласованность предложенного метода оказалась выше.

В качестве продолжения данной темы можно рассмотреть нелинейные функции активации (такие как гиперболический тангенс, логистическая сигмоида и т.д.). Если обобщить метод на популярные нелинейные функции активации, то появится способ получать интерпретации для многих известных рекуррентных нейронных сетей.

Список литературы

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

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

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

[6] 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.

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

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

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

[10] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.