Выбор интерпретируемых рекуррентных моделей глубокого обучения

Гапонов Максим, Бахтеев Олег, Яковлев Константин, Стрижов Вадим

Модели глубокого обучения, в частности рекуррентные нейронные сети, всё чаще используются для принятия решений в социально важных сферах. Чтобы эксперты могли доверять решениям моделей, необходимо предоставлять некоторые объяснения того, как модель получила такой результат. В данной работе мы опишем метод интерпретации свёрточных моделей глубокого обучения. Основная идея метода заключается в анализе влияния входных данных на принятое моделью решение. Мы сравним наш метод с уже существующими методами, проведём вычислительные эксперименты для анализа производительности и точности решения.

1 Введение

В работе рассматривается метод интерпретации рекуррентных моделей глубокого обучения. Такие модели используются экспертами во многих сферах для обработки последовательностей данных. Однако для принятия решения при помощи такой модели необходимо доверять её предсказаниям. По этой причине необходимо помимо самого предсказания модели предоставлять интерпретируемые данные о том, почему модель сделала такое предсказание. Работа посвящена генерации таких интерпретируемых данных (интерпретаций) для произвольной рекуррентной сети.

Рекуррентная нейронная сеть представляет из себя последовательность из слоёв нейронов: входного слоя, скрытых слоёв и выходного слоя. 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** [?]. 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. Существующие методы являются либо неточными, то есть несогласуются с реальным поведением модели, либо несогласованными, то есть совершенно разным образом объясняет предсказания на близких объектах.

Уже существующие методы можно разделить на три основные группы.

1. Методы анализа скрытых слоёв

Происходит интерпретация признаков выученных скрытыми слоями. Такой подход отражает поведение скрытых слоёв, но не показывает поведение сети в целом.

2. Методы подражания модели

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.

3. Методы локальной интерпретации

Анализируется поведение модели в окрестности исходного объекта. В таком подходе получаем точную интерпретацию для одного объекта, однако интерпретации для близких объектов могут существенно отличаться. Таким образом, данный подход не предоставляет согласованные интерпретации.

Рассмотренный в работе подход является обобщением метода OpenBox. В [?] рассмотрено применение метода OpenBox для 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.

Проведён вычислительный эксперимент для анализа производительности метода, а также для проверки интерпретаций на точность и согласованность.