УДК 004.89

ИНТЕРПРЕТАЦИЯ ВРЕМЕННЫХ ЗАВИСИМОСТЕЙ В МОДЕЛЯХ АНАЛИЗА ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Тюльпанов А.Д., студент гр. ИУ12-41М, II курс
МГТУ им. Н. Э. Баумана, г. Москва

Модели анализа последовательностей, в особенности применяемые к временным рядам, потокам событий, текстам или мультимодальным данным, требуют особого подхода к интерпретации, поскольку их поведение определяется не просто статическим состоянием признаков, а динамикой, контекстом и порядком следования информации во времени. Такие модели – будь то рекуррентные нейросети (RNN, LSTM, GRU), трансформеры, сверточные архитектуры с временными свёртками или модели с attention-механизмами – способны захватывать долгосрочные зависимости, нелинейные паттерны и сложные взаимодействия во временной структуре данных. Однако именно это усложняет задачу объяснения их решений, поскольку «значимость» отдельных элементов неразрывно связана с их положением в контексте, историей и перспективой предсказания. В этой связи интерпретация временных зависимостей становится необходимым компонентом построения доверенных, верифицируемых и управляемых ИИ-систем, работающих с последовательной информацией [1].

В отличие от классических моделей, работающих с фиксированными признаками, временные модели принимают на вход последовательность состояний, каждое из которых может само по себе иметь высокую размерность. Например, в задачах предсказания отказа оборудования временной ряд может включать десятки сенсорных сигналов, записанных с высокой частотой; в текстовых задачах – это последовательности токенов, а в финансовом анализе – цепочки транзакций или цен. Объяснение предсказания в такой модели должно отвечать не только на вопрос «какие признаки?», но и «на каком шаге времени?», а также «в каком взаимодействии с другими шагами?».

Одним из базовых подходов к интерпретации таких моделей является анализ важности временных окон. Это может реализовываться через обнуление (occlusion) отдельных фрагментов входной последовательности и анализ изменений в предсказании модели. Например, если удаление 5–10 шагов назад вызывает радикальное изменение выхода модели, можно сделать вывод, что соответствующее окно было критичным. Такой метод прост в реализации и интуитивно понятен, однако страдает от низкой точности, поскольку не учитывает нелинейные и контекстные зависимости. Более развитые версии включают

smooth occlusion, где влияние определённых сегментов ослабляется, а не полностью исключается, что позволяет избежать скачков и выделить регионы постепенного влияния.

В случае рекуррентных нейросетей используются методы визуализации скрытых состояний (hidden states). Накапливаемые в памяти модели скрытые векторы отражают, какие аспекты предыдущего контекста важны на каждом временном шаге. Их анализ позволяет обнаружить, на какие сигналы сеть «реагирует» во времени. Например, при анализе последовательностей событий в медицинских записях можно визуализировать, как определённые диагнозы, назначения или симптомы изменяют динамику внутреннего состояния модели и влияют на вероятность наступления события (например, рецидива). Однако прямой анализ скрытых векторов сложен, поскольку они неинтерпретируемы напрямую. Здесь помогают проекции, кластеризация или attention-механизмы [2].

Важнейшим инструментом интерпретации временных зависимостей стали механизмы внимания (attention). В трансформерных моделях они позволяют явно зафиксировать, какие временные точки влияют на текущее предсказание. Attention-матрицы можно визуализировать в виде тепловых карт, отражающих силу связи между позициями последовательности. Это даёт мощный и интуитивный способ локализации зависимости: например, можно показать, что предсказание на текущем шаге опирается на события, произошедшие 3 и 7 шагов назад, с определённой значимостью. При этом, в отличие от обнуления, attention показывает не влияние конкретных значений признаков, а связность шагов – то есть структуру временной памяти модели. В задачах перевода, анализа поведения пользователя, предсказания спроса такие карты становятся неотъемлемой частью аудита модели [3].

Однако следует отметить, что сами по себе веса attention не всегда являются достоверным объяснением – они не обязательно коррелируют с причинным вкладом соответствующих шагов. Поэтому появляются гибридные методы, совмещающие attention с градиентным анализом или методами Шепли (SHAP values). В таких подходах вычисляется не только, на что модель «смотрит», но и как изменение этих входов влияет на выход. В результате можно получить более стабильные и устойчивые объяснения, учитывающие как временную структуру, так и функциональные зависимости.

В задачах с нерегулярными временными рядами (например, в здравоохранении или телеметрии), где промежутки между событиями варьируются, используются темпорально-сознательные интерпретаторы, которые учитывают не только порядок, но и интервалы между шагами. Это позволяет уточнить, влияет ли событие из прошлого из-за своей важности, или из-за его близости к моменту предсказания. Применение темпоральных масок, анализ временных

дельт и визуализация временной плотности значимости позволяет сделать объяснение более согласованным с человеческим восприятием причинно-следственных связей.

В последнее время набирают популярность методы контрфактической интерпретации временных моделей, где моделируется альтернатива – «что было бы, если бы событие X произошло раньше/позже/иначе». Такие методы позволяют не только объяснить текущее поведение модели, но и выявить возможности для вмешательства, например, предложить оптимальный момент для принятия решения в системах предиктивного обслуживания или медицинского реагирования.

Для поддержки интерпретации временных зависимостей активно развиваются специализированные инструменты:
– TimeSHAP, расширяющий SHAP на модели с временной структурой;
– TSInterpret, предлагающий набор объяснителей для временных рядов;
– Captum Temporal, расширение PyTorch Captum для работы с последовательностями;
– ExplainX и Truera, предоставляющие визуализацию временных attention и динамики вкладов [4].

Интерпретация временных зависимостей особенно важна в производственных условиях. Например, в предиктивной аналитике оборудования визуализация вкладов температурных или вибрационных показателей за последние часы помогает инженеру принять решение о плановом обслуживании. В логистике анализ последовательностей маршрутов с указанием ключевых событий помогает понять, что повлияло на задержку поставки. В медицине визуализация траектории пациента с указанием, какие наблюдения повлияли на диагноз или прогноз, позволяет повысить уверенность врача в выводах модели [5].

Таким образом, интерпретация временных зависимостей в моделях анализа последовательностей – это неотъемлемый элемент инженерии доверенного ИИ. Она требует комплексного подхода, сочетающего анализ вклада отдельных шагов, визуализацию внимания, моделирование альтернатив и семантическую привязку объяснений к контексту задачи [6].

Только при наличии таких инструментов становится возможным не только объяснять, но и контролировать, проверять и улучшать поведение интеллектуальных систем, работающих с последовательными и временными данными.

Список литературы:

1. Hochreiter, S. Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – Vol. 9, № 8. – P. 1735–1780. – DOI 10.1162/neco.1997.9.8.1735.

2. Cho, K. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation / K. Cho, B. van Merriënboer, C. Gulcehre [et al.] //

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. – Doha : ACL, 2014. – P. 1724–1734. – DOI 10.3115/v1/D14-1179.

3. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

4. Lim, B. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting / B. Lim, S. Ö. Arık, N. Loeff, T. Pfister // International Journal of Forecasting. – 2021. – Vol. 37, № 4. – P. 1748–1764. – DOI 10.1016/j.ijforecast.2021.03.012.

5. Jain, S. Attention Is Not Explanation / S. Jain, B. C. Wallace // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. – Minneapolis : ACL, 2019. – P. 3543–3556. – DOI 10.18653/v1/N19-1357.

6. Пылов, П. А. Экстракция признаков в моделях последовательного глубокого обучения / П. А. Пылов, А. В. Протодьяконов // Россия молодая : сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года / редкол. : К. С. Костиков (отв. ред.) [и др.]. – Кемерово : Кузбасский государственный технический университет имени Т. Ф. Горбачева, 2022. – С. 31525.1–31525.3. – EDN MZEHGM.