

ПРОГНОЗИРОВАНИЕ ПОПУЛЯРНОСТИ ДАННЫХ В ФВЭ: ПОДХОД С ИСПОЛЬЗОВАНИЕМ ТРАНСФОРМЕРОВ

М. А. Григорьева^{1, *}, *Н. Н. Попова*^{1, **}, *М. В. Шубин*^{1, ***}

¹ Московский государственный университет им. М. В. Ломоносова, Москва, Россия

Современные эксперименты в области физики высоких энергий (ФВЭ) генерируют и хранят огромные объемы данных, к которым обращаются пользователи. Характер обращений пользователей — несистематический и нерегулярный. Эффективное управление данными требует точного прогнозирования популярности датасетов (datasets) для оптимизации хранения, кэширования и стратегий распределения данных. В этой работе мы предлагаем подход для прогнозирования будущих шаблонов доступа к датасетам с использованием моделей глубокого обучения на основе трансформеров. На основе исторических журналов взаимодействий пользователей с датасетами ФВЭ наш метод находит временные зависимости и контекстуальные изменения для прогнозирования как краткосрочного, так и среднесрочного спроса на данные.

Modern high-energy physics (HEP) experiments generate and store vast volumes of data, which users access through complex and irregular patterns. Efficient data management in such environments requires accurate forecasting of dataset popularity to optimize storage, caching, and data distribution strategies. In this work, we propose an approach for predicting future dataset access patterns using transformer-based deep learning models. By leveraging historical logs of user interactions with HEP datasets, our method captures temporal dependencies and contextual signals to forecast both short- and medium-term data demand.

PACS: 07.05.Kf

ВВЕДЕНИЕ

На сегодняшний день количество данных, которые обрабатываются и хранятся в распределенных системах больших научных экспериментов класса мегасайенс, чрезвычайно велико. В ходе работы научных установок физики высоких энергий непрерывно генерируется интенсивный поток данных, который затем обрабатывается на вычислительных

* E-mail: magsend@gmail.com

** E-mail: popova@cs.msu.ru

*** E-mail: mih.shub@gmail.com

ресурсах. Обработка может подразумевать несколько различных этапов, таких как реконструкция, вывод (derivation), моделирование. Пользователи распределенной системы, как правило, работают с уже подготовленными данными, прошедшими ряд определенных этапов. Такие данные представлены в виде наборов логически сгруппированных файлов, именуемых датасетами. Для хранения всех датасетов используются различные носители информации, включая диски и магнитные ленты. Огромная часть датасетов после одного-двух обращений к ним остаются невостребованными в дальнейшем; такие датасеты подлежат архивированию на ленты. Однако некоторые датасеты подвергаются повторным обращениям со стороны пользователей, при этом характер обращений может быть хаотичный и нерегулярный. При обращениях к датасетам их приходится копировать либо кэшировать на дисках. Прогнозирование востребованности (популярности) данных среди пользователей является важным и полезным для принятия решений по размещению и кэшированию данных. Управление репликами датасетов является ключевым фактором, обеспечивающим эффективное функционирование распределенной системы класса мегасайнс.

Задача прогнозирования популярности датасетов для эффективного управления репликами датасетов или кэширования активно исследуется в ряде работ. Например, в [1] была разработана нейросетевая модель, которая успешно предсказывает количество обращений на следующий день, но только для 100 наиболее популярных наборов данных.

В данной работе предлагается подход на основе трансформеров для прогнозирования популярности групп датасетов.

ПРОГНОЗИРОВАНИЕ ПОПУЛЯРНОСТИ ДЛЯ ГРУПП ДАТАСЕТОВ

Под *популярностью датасетов* будем понимать количество обращений к ним за определенный промежуток времени (день, неделя, месяц). Предлагаемый подход подразумевает построение прогноза не для конкретных датасетов, а для групп датасетов, объединяющих датасеты по общему признаку (например, формат данных, проект, год проведения эксперимента). Такой прогноз позволит выделить наиболее востребованные группы данных и решать задачу реплицирования датасетов уже для этой группы.

Популярность группы датасетов определим как суммарное количество обращений к датасетам данной группы за определенный промежуток времени.

На сегодняшний день современные трансформерные архитектуры широко используются для прогнозирования временных рядов. Infig-
met [2] — модель для прогнозирования длинных последовательностей

данных с использованием разреженного механизма внимания. Autoformer [3] — модель, сочетающая идеи автокодировщиков и механизмов внимания. Учитывает сезонные и трендовые компоненты. FEDformer (Frequency Enhanced Decomposition Transformer) [4] — модель, использующая частотное разложение для выделения различных компонентов временных рядов.

В качестве нейросетевых моделей на основе трансформерной архитектуры в данной работе рассматриваются модели:

- CATS (Cross-Attention-only Time Series transformer) [5] — легковесная модель, в которой оставлен только механизм внимания, связывающий энкодер и декодер.

- Chronos T5 [6] — предобученное семейство моделей от Amazon, основанных на языковой архитектуре T5. Модель может использоваться в формате zero-shot learning, не требующем предварительного обучения.

Такой выбор моделей обосновывается тем, что для обучения не требуется большого количества данных: CATS является легковесной моделью, в то время как Chronos T5 не требует обязательного обучения.

РЕЗУЛЬТАТЫ ВЫЧИСЛИТЕЛЬНЫХ ЭКСПЕРИМЕНТОВ

Для оценки качества прогнозирования рассматриваемыми моделями проводилась блочная кросс-валидация, которая заключается в выделении во временном ряде блоков фиксированного размера и разбиении каждого блока на обучающий и тестовый временные отрезки. Исследование проведем в соответствии с ранее представленной работой [7], в которой получены результаты для модели случайный лес (Random Forest). Качество оценивается с помощью метрики SMAPE (Symmetric Mean Absolute

Percentage Error), которая вычисляется как $\frac{Y_{\text{forecast}} - Y_{\text{actual}}}{(Y_{\text{forecast}} + Y_{\text{actual}})/2}$ и зависит от длины обучающего временного ряда (в исследовании рассматривались: 3, 6, 9 мес., 1 год) и от горизонта предсказания (от 1 до 15 дней).

На рис. 1 приведены результаты исследования зависимости ошибки предсказания от горизонта предсказания для различных длин временного отрезка обучения.

Видно, что модель Chronos T5 превосходит наивный метод при любой длине входного временного ряда. На горизонтах предсказания 1–7 дней при длине входного временного ряда 6–9 мес. трансформерная модель показывает себя немного лучше модели случайного леса. Модель CATS превосходит наивный метод только для случая с длиной временного ряда 6 мес. и на малых горизонтах предсказания, показывая себя хуже модели случайного леса.

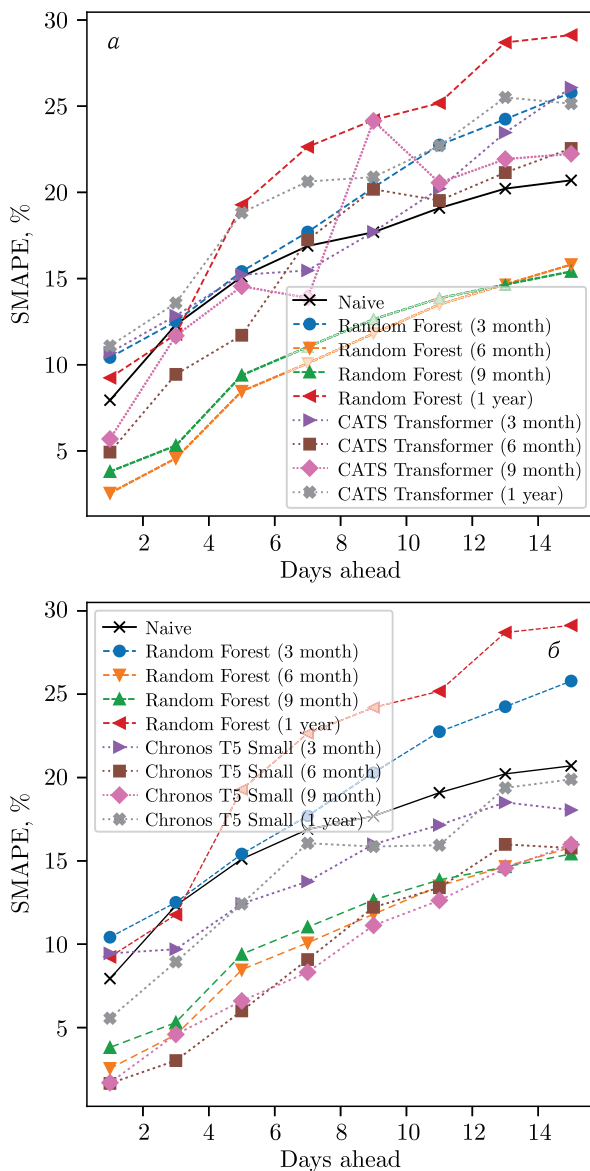


Рис. 1. Результаты кросс-валидации модели CATS (а) и модели Chronos T5 Small (б) в сравнении с моделью случайного леса. Штриховой линией отмечены результаты для случайного леса. Результаты трансформеров отмечены пунктирной линией

ЗАКЛЮЧЕНИЕ

В работе апробирован подход на основе трансформеров для прогнозирования популярности групп датасетов при различных параметрах длины отрезка обучения и горизонта предсказания. Эксперименты демонстрируют преимущество модели Chronos T5 по сравнению с моделью случайного леса на малых горизонтах прогноза (1–7 дней). Модель CATS при параметрах обучения по умолчанию показала превосходство по сравнению с наивным методом только для случая длины обучающего отрезка 6 мес. на малых горизонтах прогноза. Для обучения трансформерных моделей с нуля в этом подходе требуется настройка гиперпараметров архитектуры и проведение обучения с учетом специфики временных рядов. Направлением дальнейшего развития является автоматизация выбора модели, настройка ее гиперпараметров и проведение обучения для укладываемых пользователем данных.

Работа выполнена в Московском государственном университете им. М. В. Ломоносова при поддержке Российского научного фонда (соглашение № 25-11-00181).

СПИСОК ЛИТЕРАТУРЫ

1. Bellavita J., Sim C., Wu K., Sim A., Yoo S., Ito H., Garonne V., Lancon E. Understanding Data Access Patterns for dCache System // Eur. Phys. J. Web Conf. 2024. V. 295. P. 01053; DOI: 10.1051/epjconf/202429501053.
2. Zhou H., Zhang S., Peng J., Zhang S., Li J., Zhang J. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. arXiv:2012.07436. 2020.
3. Wu H., Xu J., Wang J., Long M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. arXiv:2106.13008. 2021.
4. Zhou T., Ma Z., Wen Q., Wang X., Sun L., Jin R. FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting // Proc. 39th Intern. Conf. on Machine Learning (ICML), Baltimore, Maryland, July 17–23, 2022. arXiv:2201.12740. 2022.
5. Kim D., Park J., Lee J., Kim H. Are Self-Attentions Effective for Time Series Forecasting? arXiv:2405.16877. 2024.
6. Ansari A. F., Stella L., Turkmen C., Zhang X., Mercado P., Shen H., Shchur O., Rangapuram S. S., Arango S. P., Kapoor S., Zschiegner J., Maddix D. C., Wang H., Mahoney M. W., Torkkola K., Wilson A. G., Bohlke-Schneider M., Wang Y. Chronos: Learning the Language of Time Series. arXiv:2403.07815. 2024.
7. Grigorieva M., Schulz M., Klimentov A., Popova N., Shubin M. Operational Analytics Studies for ATLAS Distributed Computing: Data Popularity Forecast and Ranking of the WLCG Centers // CHEP 2023 Proc. Eur. Phys. J. Web Conf. EDP Sci. 2024. 04033.