



Моделирование последовательностей событий

Иван Карпухин

Руководитель направления в ЦФИ

Более 10 лет в deep learning, 4 в исследованиях

R&D:

- Computer vision
 - Face recognition & image retrieval
 - Optical Character Recognition (OCR)
 - Generative models (GANs & VAE)
- Speech (ASR, Speaker Recognition, Voice Conversion)

Research:

- Metric learning
- Classification
- Ensembles
- Event Sequences & Fintech



Основные темы

Последовательности событий – это отдельная область, которой нужны уникальные алгоритмы

Генеративная модель DeTRP

- Ограничения авторегрессии
- Решение основных проблем предсказания на длинный горизонт
- Адаптеры под типичные задачи (отток, дефолт, отклик)

Токены истории для извлечения эмбеддингов

- Почему трансформер не даёт хорошего эмбеддинга
- Как обучить эмбеддинг при помощи обычного NLP лосса

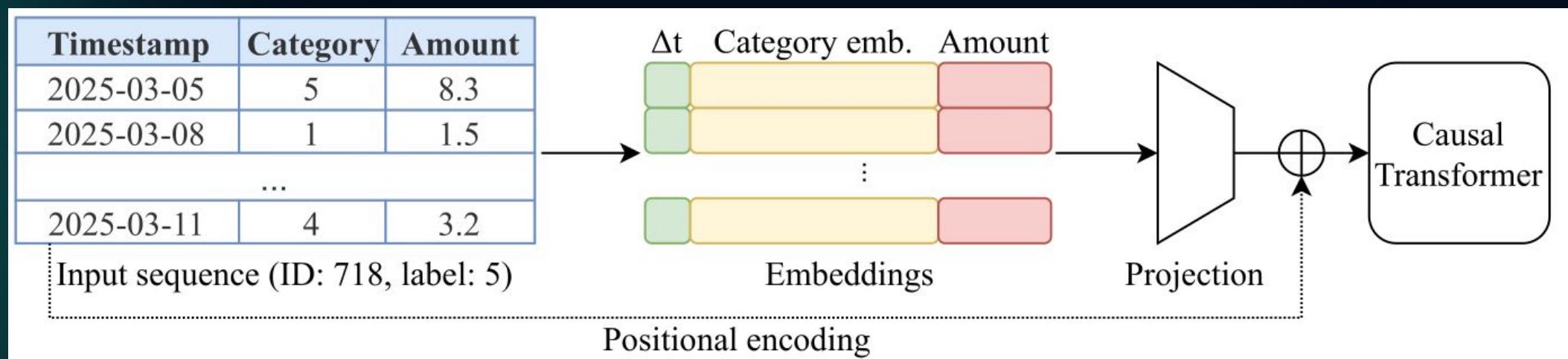
Последовательности событий

Домен

- Для каждого ID клиента имеется цепочка событий
- Каждое событие это несколько полей: дата, категория покупки, стоимость, признаки продавца и т.п.
- Иногда для цепочки есть метка класса

Задачи

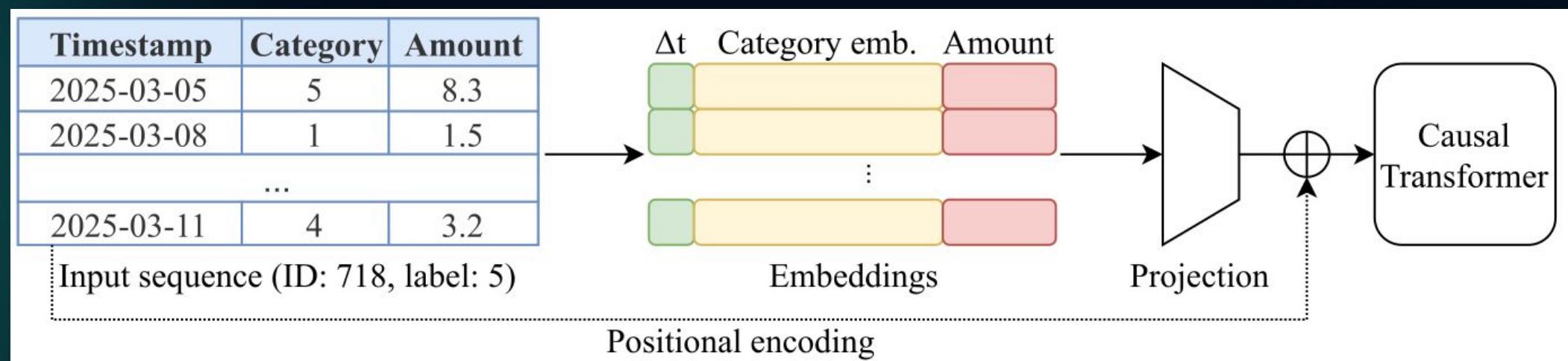
- Предсказание будущей активности (uplift, отток, CLTV)
- Предобучение без учителя и извлечение эмбеддингов при ограниченном объеме размеченных данных



Последовательности событий

Ключевые особенности:

- Нерегулярный шаг времени
- Множество разнотипных полей для каждого события
- Большое количество неразмеченных данных
- В задачах классификации нас интересуют события в будущем, а не свойства данных (в отличие от классификации текстов, таблиц и временных рядов)



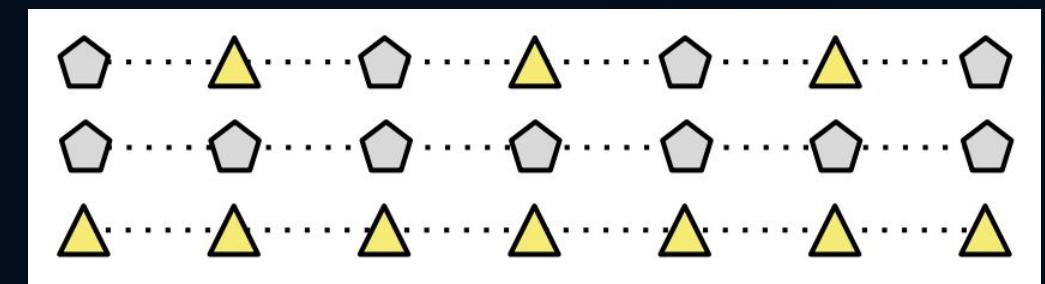
Object Detection

в задаче предсказания событий

Проблема: долгосрочные предсказания

Задача: предсказать события на неделю / месяц вперёд

- Точно предсказать невозможно
 - Что я куплю через неделю?
- Обычно применяют авторегрессию
 - Низкая скорость inference
 - Проблема однотипных предсказаний



Пример предсказаний

Идея: Object Detection

Предсказание событий похоже на поиск объектов на изображении

Object Detection

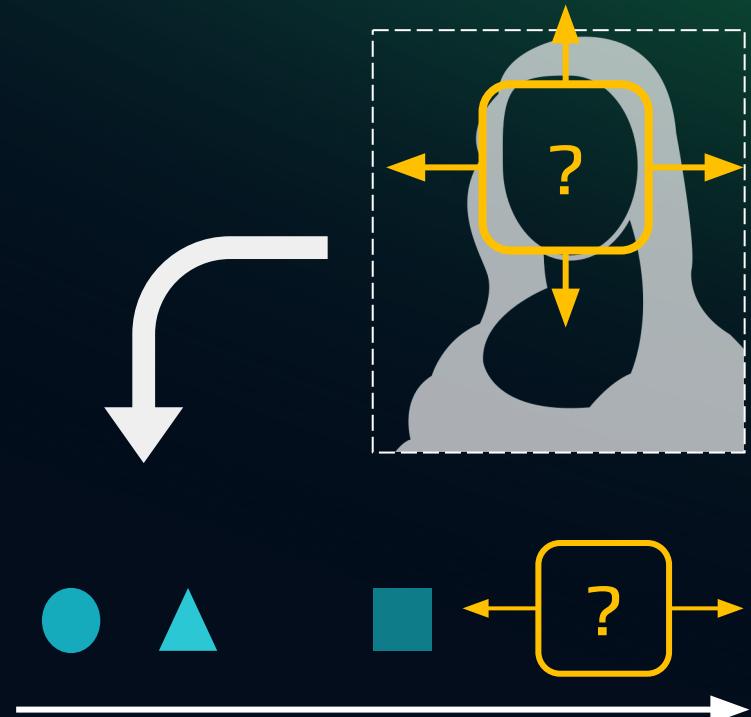
- Обвести прямоугольником объект на изображении (где?)
- Указать тип объекта (что?)

Последовательности событий

- Определить время будущего события (когда?)
- Установить тип события (что?)
 - Иногда нужно предсказать дополнительные признаки

Мотивация

- Object Detection никто не делает через авторегрессию
- Методы и метрики из CV можно перенести в нашу область



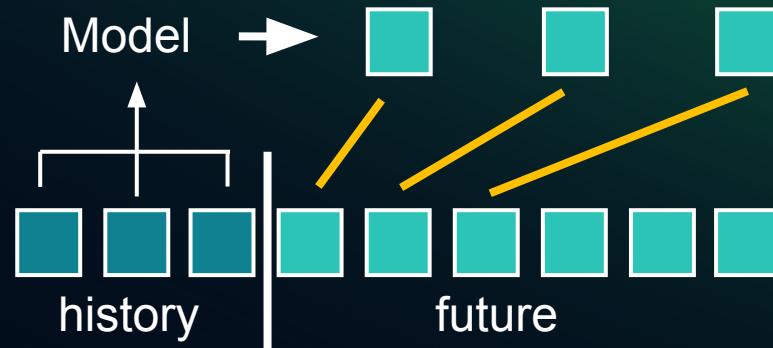
Детектор событий

Адаптация архитектуры DeTR

- Считаем лосс между ближайшими событиями
А не между одинаковыми позициями, как в NTP
- Для этого делаем матчинг предсказаний и GT
Также реализуем быстрое CUDA ядро
- Модель учится хорошо предсказывать те события, которые можно предсказать, и игнорировать шум

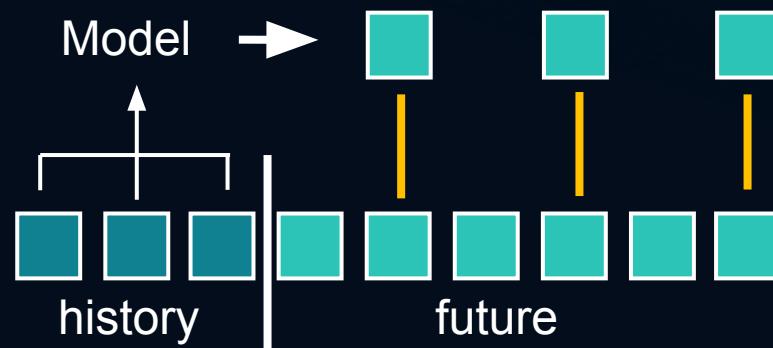
Получаем быструю и надёжную модель
предсказаний на горизонт

Обычный loss:



Большая
ошибка

Detection loss:



Средняя
ошибка

DeTPP: Результаты

Лучшая на текущий момент генеративная модель долгосрочных предсказаний

Обгоняет

- Temporal Point Processes
- Диффузии
- Статистические бейзлайны
первая DL модель, которая побила бейзлайны на группе задач

Легко добавлять новые поля для предсказания

Ряд задач можно решить в обход генерации

- Оценка распределений событий на горизонте
- Оценка суммарных трат на горизонте

Model	Metrics (OTD / T-mAP)				
	StackOverflow	Amazon	Retweet	MIMIC-IV	Transactions
IFTPP	13.64 / 8.31% ±0.05 / ±0.50%	6.52 / 22.56% ±0.05 / ±0.52%	172.7 / 31.75% ±4.4 / ±4.44%	11.53 / 21.67% ±0.01 / ±0.21%	6.90 / 5.88% ±0.01 / ±0.13%
RMTPP	13.17 / 12.72% ±0.05 / ±0.16%	6.57 / 20.06% ±0.03 / ±0.33%	166.7 / 44.74% ±3.3 / ±0.94%	13.71 / 21.08% ±0.03 / ±0.29%	6.88 / 6.69% ±0.01 / ±0.12%
NHP	13.24 / 11.96% ±0.02 / ±0.40%	9.02 / 26.29% ±0.35 / ±0.55%	165.8 / 45.07% ±1.6 / ±0.34%	18.60 / 7.32% ±0.19 / ±1.33%	6.98 / 5.61% ±0.01 / ±0.05%
AttNHP	13.30 / 11.13% ±0.02 / ±0.32%	7.30 / 14.62% ±0.06 / ±0.80%	171.6 / 25.85% ±1.0 / ±1.08%	14.68 / 22.46% ±0.08 / ±0.40%	7.50 / 1.48% N/A / N/A
ODE	13.27 / 10.52% ±0.03 / ±0.23%	9.46 / 22.96% ±0.08 / ±0.61%	165.3 / 44.81% ±0.5 / ±0.69%	14.74 / 15.18% ±0.34 / ±0.15%	6.97 / 5.52% ±0.01 / ±0.13%
HYPERO	13.26 / 14.69% N/A / N/A	6.61 / 20.53% N/A / N/A	170.7 / 46.99% N/A / N/A	14.87 / 16.77% N/A / N/A	7.05 / 7.05% N/A / N/A
DeTPP	12.06 / 23.11% ±0.04 / ±0.08%	5.98 / 37.18% ±0.01 / ±0.07%	134.4 / 57.37% ±0.8 / ±0.67%	12.85 / 30.63% ±0.26 / ±0.14%	6.66 / 9.17% ±0.03 / ±0.11%

(a) Качество



(b) Скорость (RPS)

HT-Transformer

Токены истории для извлечения эмбеддингов

Проблема

RNN на многих задачах превосходят трансформеры

- Даже лучшие трансформерные архитектуры проигрывают RNN на малых и средних наборах данных
- На больших наборах данных трансформеры могут давать преимущество, но в разы более затратные

Гипотеза: в трансформерах нет качественного эмбеддинга

- В RNN скрытое состояние сохраняет всю необходимую информацию о префикссе
- В Transformer, механизм внимания использует активации всех токенов в истории
 - Последний эмбеддинг содержит мало информации
 - Усреднение всех эмбеддингов по истории приводит к потере информации



Почему не contrastive learning

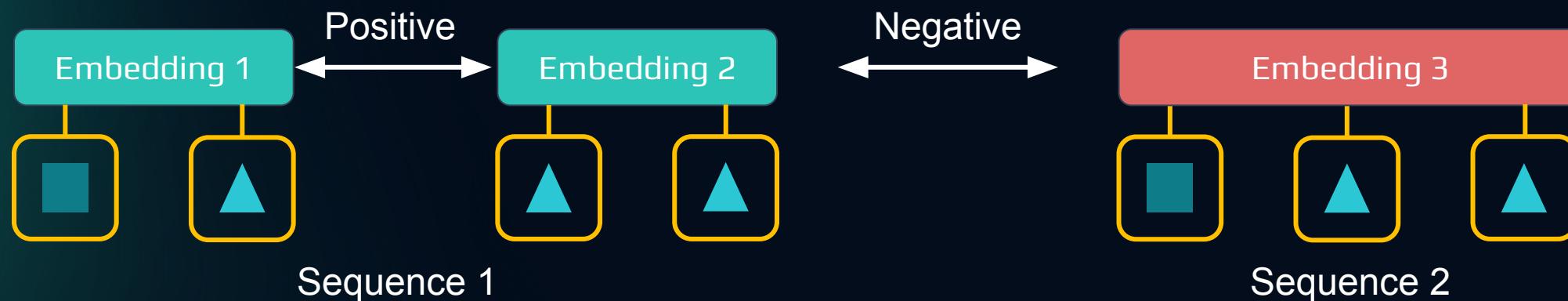
Часто эмбеддинги предобучают контрастивным подходом

- Сближаются эмбеддинги подцепочек одного клиента
- Отдаляются эмбеддинги разных клиентов

Проблемы контрастивных эмбеддингов

- Извлекается информация о цепочке целиком, а не то, что нужно для предсказания будущих событий
- Чувствительность к “easy features”

Хотим извлечь эмбеддинг после NTP pretrain!



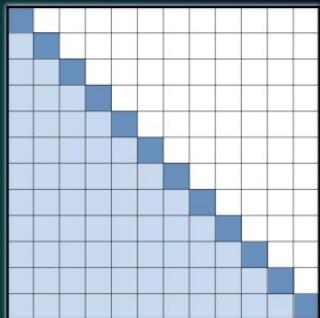
Токены Истории

Цель: сформировать эмбеддинг во время NTP pretrain

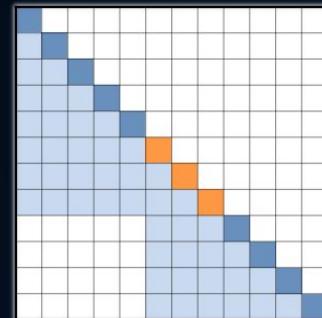
- Добавить <history-token> в несколько мест цепочки
- Иногда использовать их вместо предыдущих событий
- В итоге <history-token> будет содержать эмбеддинг контекста
- Его можно использовать для дообучения на downstream

Близкие идеи есть в работах, вроде Recurrent Memory Transforemer*

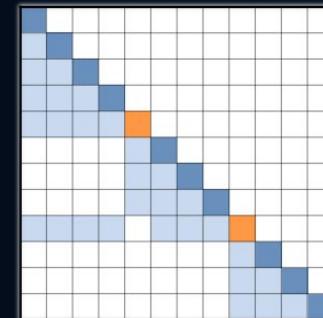
- В них фокус на long context, а не на embedding extraction



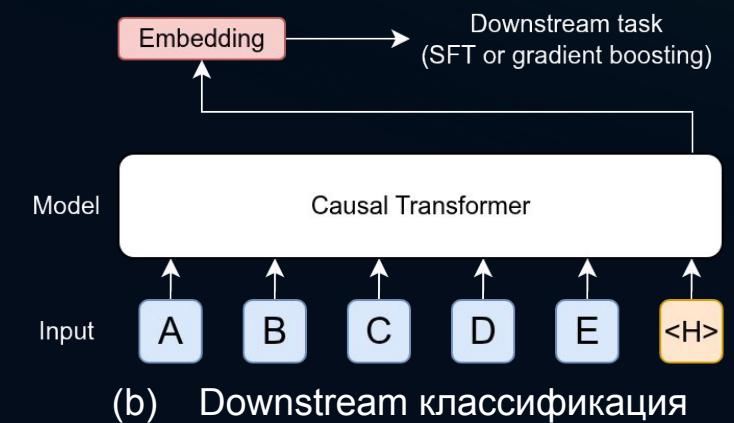
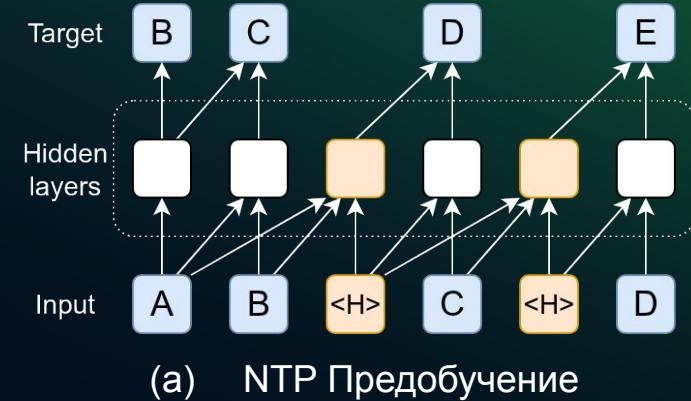
Causal mask



Recurrent Memory



HT-Transformer



Результаты на реальных данных

- Получили значимый прирост относительно обычного трансформера
 - Прогнозирование оттока
 - Прогнозирование дефолта
 - Предсказание смертности
- SOTA на 3-х датасетах
 - AgePred – это global задача

Отправили статью на конференцию*

Method	Churn	AgePred	Alfabattle	MIMIC-III	Taobao
Supervised RNN	79.10 ± 0.80	61.18 ± 0.49	76.47 ± 1.13	91.46 ± 0.10	84.91 ± 1.17
Supervised Transformer	80.92 ± 0.66	54.88 ± 2.37	74.90 ± 0.08	77.48 ± 1.22	79.71 ± 1.68
NTP RNN	81.56 ± 0.59	60.05 ± 0.29	79.83 ± 0.05	90.68 ± 0.07	83.28 ± 1.42
NTP Transformer	80.92 ± 0.66	56.16 ± 0.51	78.63 ± 0.12	91.28 ± 0.10	83.39 ± 1.43
NTP Rec. Mem. Transf.	80.23 ± 0.21	58.43 ± 0.39	80.25 ± 0.05	91.82 ± 0.04	80.54 ± 0.76
NTP Longformer	81.48 ± 0.66	57.64 ± 0.29	65.91 ± 0.34	89.26 ± 0.18	84.76 ± 1.67
CoLES RNN	82.82 ± 0.28	62.42 ± 0.33	79.30 ± 0.08	87.44 ± 0.20	85.56 ± 1.14
CoLES Transformer	78.92 ± 0.49	59.92 ± 0.30	78.40 ± 0.00	87.06 ± 0.38	82.03 ± 0.98
HT-Transformer	83.34 ± 0.42	60.10 ± 0.39	80.42 ± 0.12	92.00 ± 0.09	84.65 ± 1.07
<i>Impr. over NTP Transf.</i>	+2.42	+3.94	+1.79	+0.72	+1.26

* Karpukhin I., Savchenko A. HT-Transformer: Event Sequences Classification by Accumulating Prefix Information with History Tokens //arXiv preprint arXiv:2508.01474. – 2025.

Спасибо!



НоTPР: Бенчмарк и метрика



ДеTPР: Генерация



HT-Transformer: Эмбеддинги