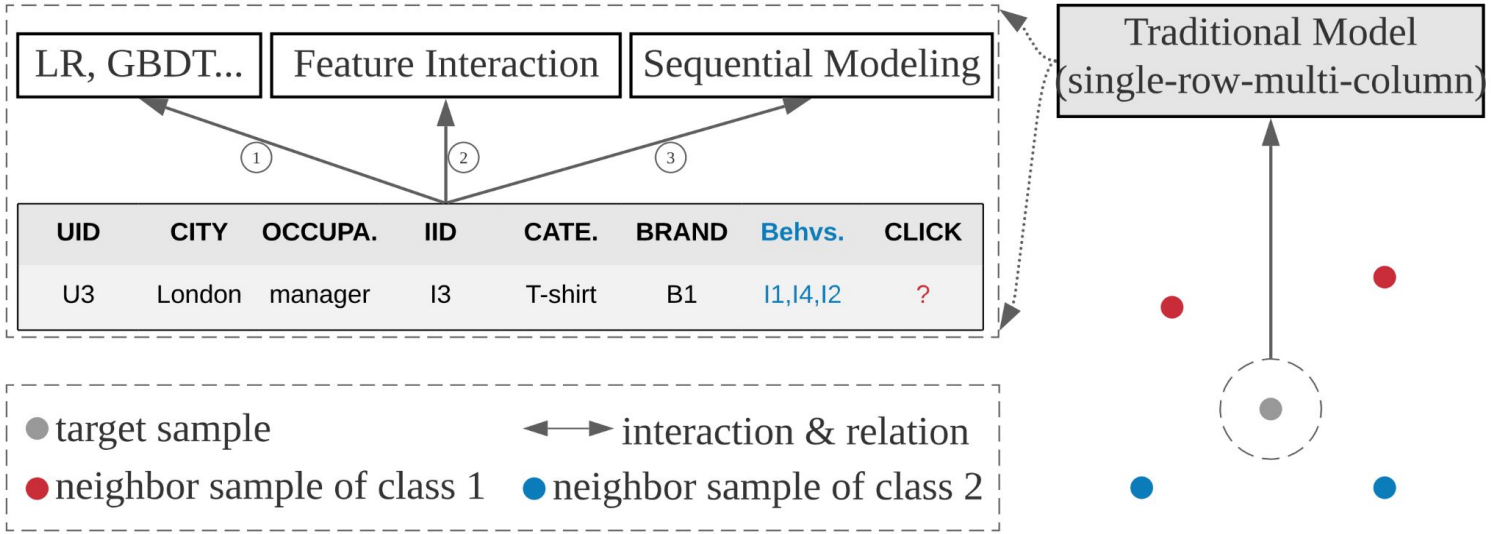
# Tabular Retrieval

Sedov Sergey

# Что происходит в области?

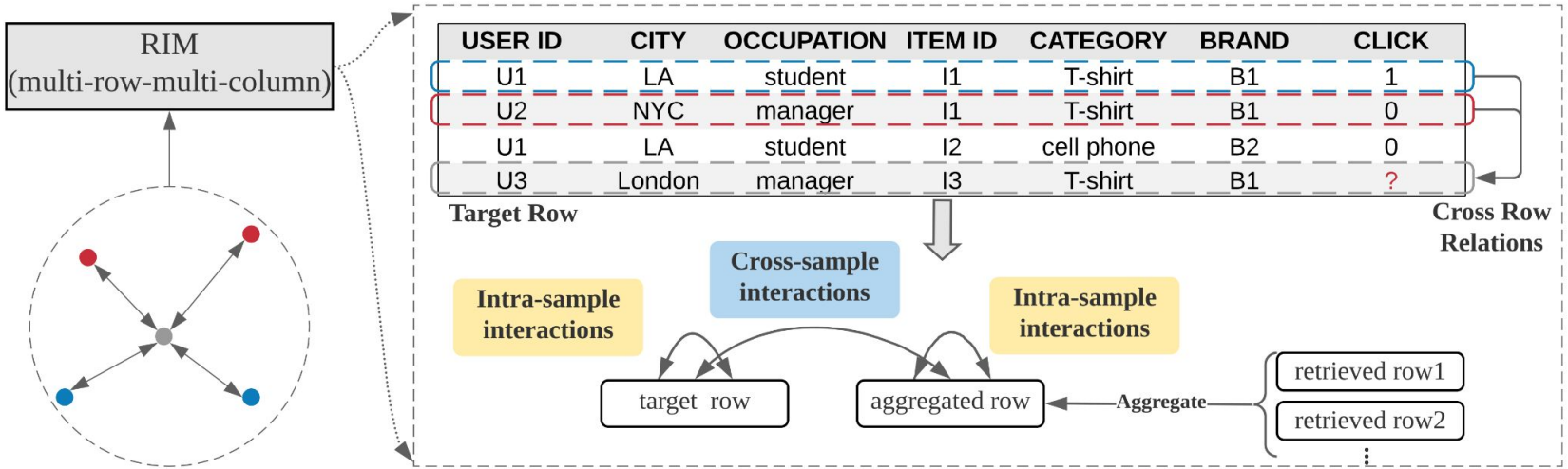
- Все бьются с бустингами, стараясь их окончательно победить
- С точки зрения практических применений, бустинги как правило удобнее
- С одной стороны, активно ресёрчат трансформеры на табличных данных, с другой - показывают, что вкачанный MLP так-то хорош
- Это наталкивает на мысль, что наши модели всё ещё плохо улавливают зависимости в данных, в частности между разными объектами
- Так мы сталкиваемся с необходимостью модернизировать аттеншн и retrieval-механизмы для эффективной работы на табличных данных

# Retrieval in Tabular Deep Learning



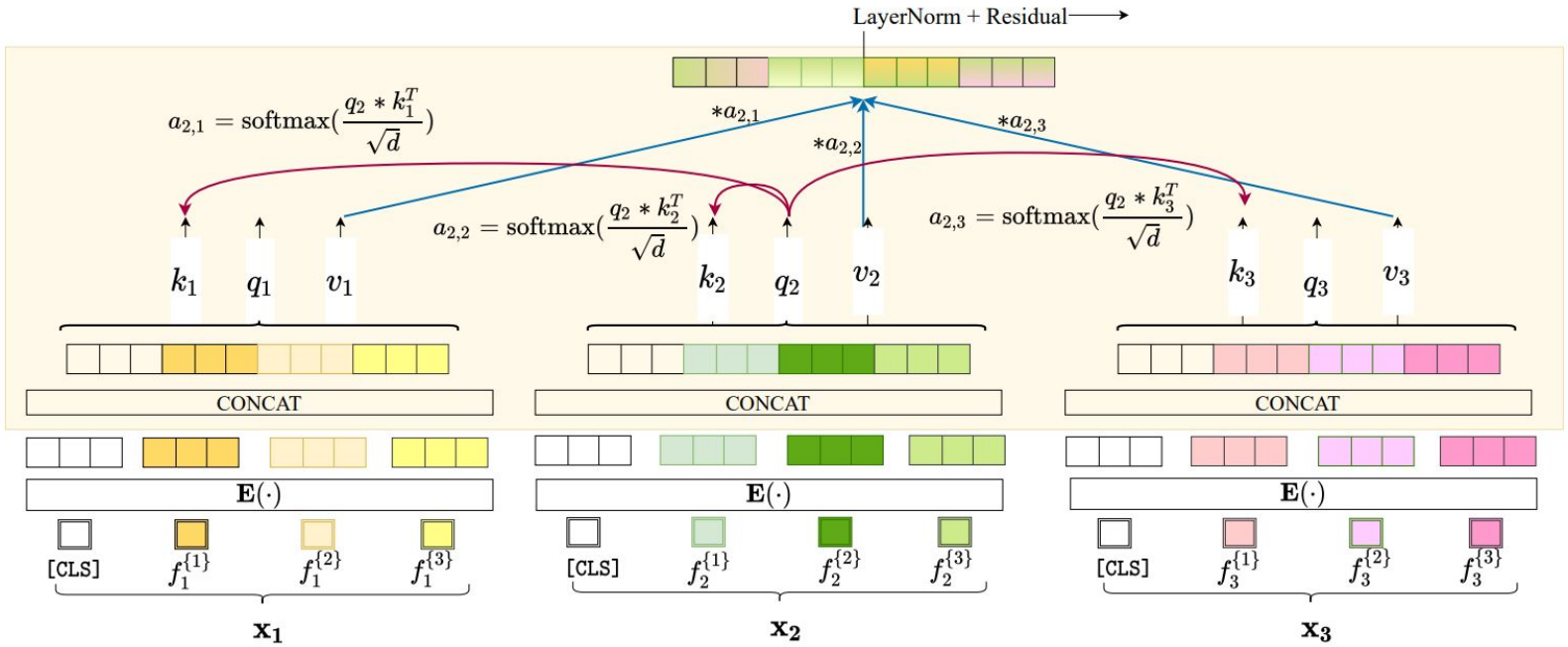

Qin et al. 2021

# Retrieval in Tabular Deep Learning




Qin et al. 2021

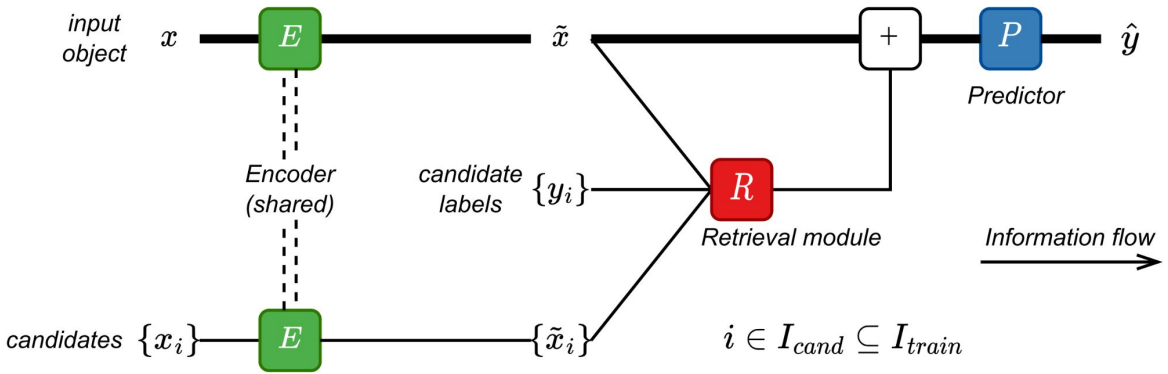
# SAINT - intersample attention on batches




Somepalli et al. 2021

# Общая постановка

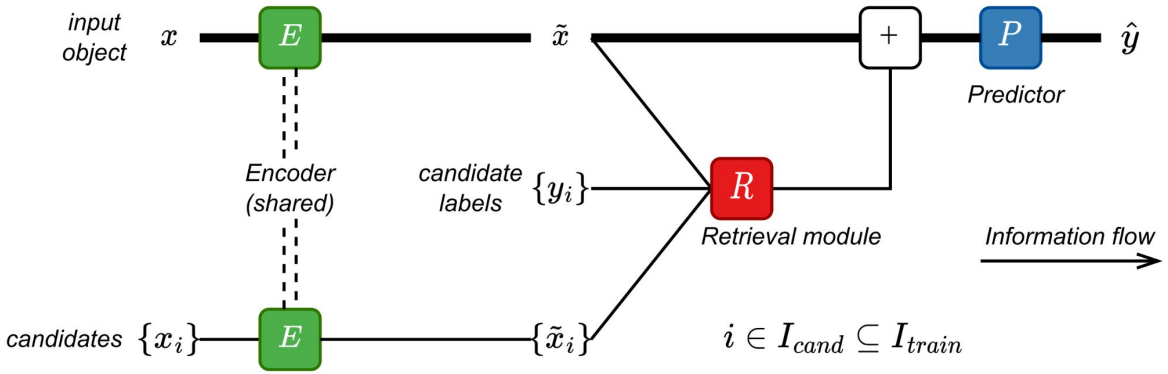
- Отбираем пул кандидатов для сравнения $\{x_i\}$ из всего датасета
- Преобразуем все выбранные строки таблицы с shared Encoder
- Используя таргетные значения $\{y_i\}$, прогоняем через Retrieval module
- Добавляем информацию о соседях к эмбеддингу входного объекта
- Отправляем всё в предиктор



# Общая постановка

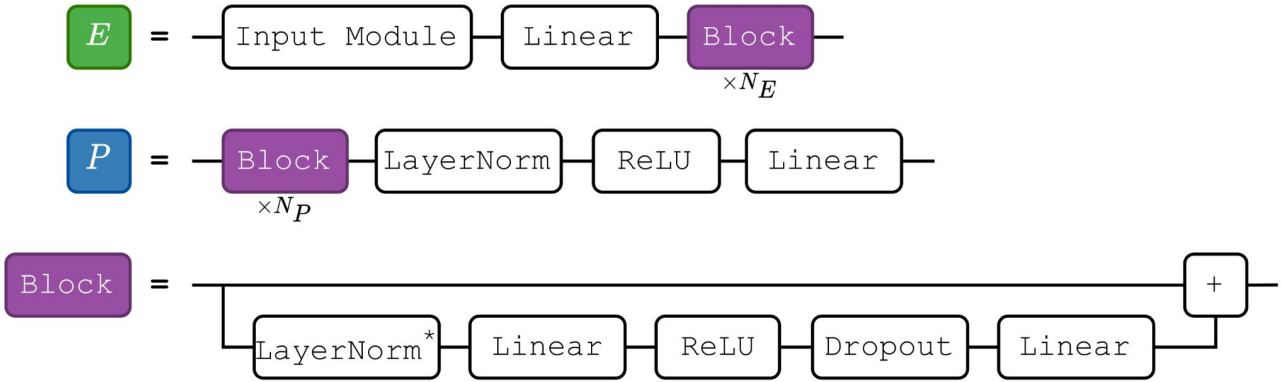
не то что бы

- Отбираем пул кандидатов для сравнения $\{x_i\}$ из всего датасета
- Преобразуем все выбранные строки таблицы с shared Encoder
- Используя таргетные значения $\{y_i\}$, прогоняем через Retrieval module
- Добавляем информацию о соседях к эмбеддингу входного объекта
- Отправляем всё в предиктор



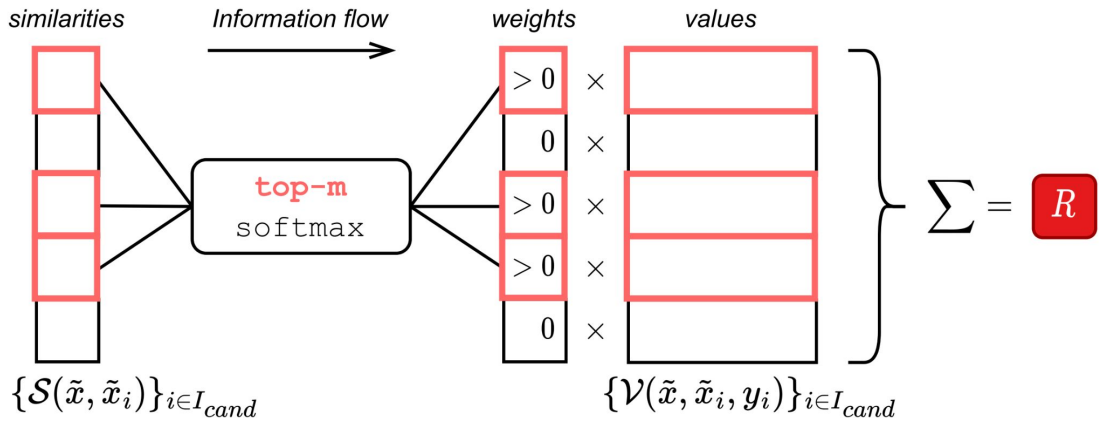
# Общая постановка: Encoder & Predictor

Мы будем фокусироваться на Retrieval модуле, так что архитектуры энкодера и предиктора оставим достаточно простыми. Вообще говоря, в процессе подбора эффективного Retrieval: $N_E = 0$, $N_P = 1$



# Общая постановка: Retrieval

- Вычисление Similarity-Score между целевым объектом и остальными
- Отбор top-m после softmax
- Умножение полученных значений на Values
- Values могут учитывать лейблы остальных объектов



# Улучшаем пайплайн

1. Запишем обычный аттеншн:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = W_Q(\tilde{x})^T W_K(\tilde{x}_i) \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_V(\tilde{x}_i)$$

# Улучшаем пайплайн

1. Запишем обычный аттеншн:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = W_Q(\tilde{x})^T W_K(\tilde{x}_i) \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_V(\tilde{x}_i)$$

2. Добавим лейблы в Values:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = W_Q(\tilde{x})^T W_K(\tilde{x}_i) \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = \underline{W_Y(y_i) +} W_V(\tilde{x}_i)$$

# Улучшаем пайплайн

1. Запишем обычный аттеншн:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = W_Q(\tilde{x})^T W_K(\tilde{x}_i) \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_V(\tilde{x}_i)$$

2. Добавим лейблы в Values:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = W_Q(\tilde{x})^T W_K(\tilde{x}_i) \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = \underline{W_Y(y_i) +} W_V(\tilde{x}_i)$$

3. Откажемся от Query-Key подхода:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = \underline{-\|W_K(\tilde{x}) - W_K(\tilde{x}_i)\|^2} \cdot d^{-1/2} \qquad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_Y(y_i) + W_V(\tilde{x}_i)$$

# Улучшаем пайплайн

4. Перейдем к расстоянию между объектами в Values:

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = -\|W_K(\tilde{x}) - W_K(\tilde{x}_i)\|^2 \cdot d^{-1/2} \quad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_Y(y_i) + \underline{T(W_K(\tilde{x}) - W_K(\tilde{x}_i))}$$

$$T(\cdot) = \texttt{LinearWithoutBias}(\texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(\cdot))))$$

В чём мотивация такого подхода?

- Мы стараемся извлечь абсолютную информацию о контекстных объектах только из их лейблов - $W_Y(y_i)$
- В то же время преобразованные эмбеддинги самих объектов лучше использовать для измерения смещения между ними
- Вдохновлён обобщением kNN для задачи регрессии - Nader et al. 2022

# А откуда здесь kNN?

Рассмотрим обычный kNN для регрессии: $\eta_{\text{KNN}}(x) = \frac{1}{k} \sum_{X_m \in B_{x,\#k}} Y_m.$

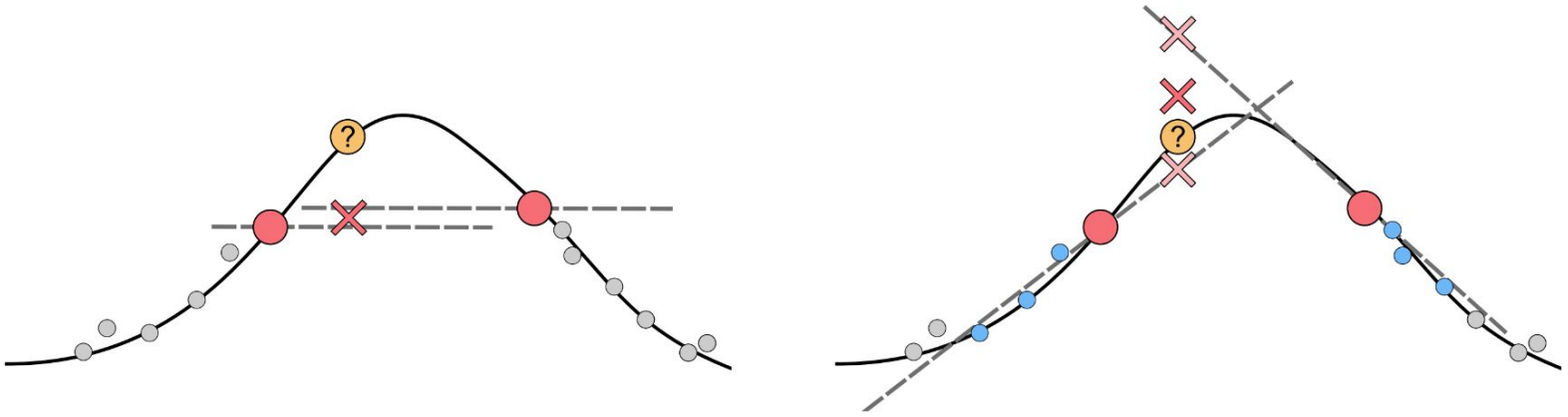
Так как это функция от данных, рассмотрим её приближение через Тейлора:

$$\eta_{\text{known}\nabla}(x) = \frac{1}{k} \sum_{X_m \in B_{x,\#k}} (Y_m + \nabla\eta(X_m)(x - X_m))$$

Градиент функции по Y мы очевидно не знаем, но мы можем оценить его, детали в статье Differential Nearest Neighbors Regression - Nader et al. 2022.

Ключевой вывод для нас - таргет в задаче регрессии можно приблизить через лейблы других объектов и функцию на расстояниях между этими объектами, что и делают авторы нашей статьи.

# Визуализация DNNR




(a) KNN (b) DNNR

*Figure 1.* **(a)** An illustration of KNN regression. To predict a value for a query (circle with question mark), the target values of the nearest points (red circles) are averaged. KNN's prediction is marked by the red cross. The other data points (gray circles) are not used for prediction. **(b)** Similar illustration of DNNR. The local gradient (gray dashed line) is estimated for each neighbor and a target value is interpolated linearly (light red crosses). The final prediction (red cross) is the average of these interpolated values.

# Последние улучшения

$$\mathcal{S}(\tilde{x}, \tilde{x}_i) = -\|W_K(\tilde{x}) - W_K(\tilde{x}_i)\|^2 \cdot d^{-1/2} \quad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_Y(y_i) + \underline{T(W_K(\tilde{x}) - W_K(\tilde{x}_i))}$$

$$T(\cdot) = \texttt{LinearWithoutBias}(\texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(\cdot))))$$

Эмпирически заметим, что нормировка Similarity Score на размерность эмбеддинга нам только вредит, так и получим TabR:

$$k = W_K(\tilde{x}),\ k_i = W_K(\tilde{x}_i) \quad \mathcal{S}(\tilde{x}, \tilde{x}_i) = -\|k - k_i\|^2 \quad \mathcal{V}(\tilde{x}, \tilde{x}_i, y_i) = W_Y(y_i) + T(k - k_i)$$

В ablation study авторы отмечают, что на задачах классификации предложенная версия Values не даёт таких очевидных результатов, так что вероятно модуль нужно дизайнить по-другому.

# Сравнения

| | CH ↑ | CA ↓ | HO ↓ | AD ↑ | DI ↓ | OT ↑ | HI ↑ | BL ↓ | WE ↓ | CO ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| MLP | 0.854 | 0.499 | 3.112 | 0.853 | 0.140 | 0.816 | 0.719 | 0.697 | 1.905 | 0.963 |
| (Step-0) The vanilla attention baseline | 0.855 | 0.484 | 3.234 | 0.857 | 0.142 | 0.814 | 0.719 | 0.699 | 1.903 | 0.957 |
| (Step-1) + Context labels | 0.855 | 0.489 | 3.205 | 0.857 | 0.142 | 0.814 | 0.719 | 0.698 | 1.906 | 0.960 |
| (Step-2) + New similarity module $\mathcal{S}$ | 0.860 | 0.418 | 3.153 | 0.858 | 0.140 | 0.813 | 0.720 | 0.692 | 1.804 | 0.972 |
| (Step-3) + New value module $\mathcal{V}$ | 0.859 | 0.408 | 3.158 | 0.863 | 0.135 | 0.810 | 0.722 | 0.692 | 1.814 | 0.975 |
| (Step-4) + Technical tweaks = **TabR** | 0.860 | 0.403 | 3.067 | 0.865 | 0.133 | 0.818 | 0.722 | 0.690 | 1.747 | 0.973 |

Здесь выделены улучшения > std относительно предидущего этапа.

# Сравнения

| | CH ↑ | CA ↓ | HO ↓ | AD ↑ | DI ↓ | OT ↑ | HI ↑ | BL ↓ | WE ↓ | CO ↑ | MI ↓ | Avg. Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| kNN | 0.837 | 0.588 | 3.744 | 0.834 | 0.256 | 0.774 | 0.665 | 0.712 | 2.296 | 0.927 | 0.764 | $6.0 \pm 1.7$ |
| DNNR (Nader et al., 2022) | – | 0.430 | 3.210 | – | 0.145 | – | – | 0.704 | 1.913 | – | 0.765 | $4.8 \pm 1.9$ |
| DKL (Wilson et al., 2016) | – | 0.521 | 3.423 | – | 0.147 | – | – | 0.699 | – | – | – | $6.2 \pm 0.5$ |
| ANP (Kim et al., 2019) | – | 0.472 | 3.162 | – | 0.140 | – | – | 0.705 | 1.902 | – | – | $4.6 \pm 2.5$ |
| SAINT (Somepalli et al., 2021) | **0.860** | 0.468 | 3.242 | 0.860 | 0.137 | 0.812 | 0.724 | 0.693 | 1.933 | 0.964 | 0.763 | $3.8 \pm 1.5$ |
| NPT (Kossen et al., 2021) | 0.858 | 0.474 | 3.175 | 0.853 | 0.138 | 0.815 | 0.721 | 0.692 | 1.947 | 0.966 | 0.753 | $3.6 \pm 1.0$ |
| MLP | 0.854 | 0.499 | 3.112 | 0.853 | 0.140 | 0.816 | 0.719 | 0.697 | 1.905 | 0.963 | 0.748 | $3.7 \pm 1.3$ |
| MLP-PLR | 0.860 | 0.476 | **3.056** | **0.870** | 0.134 | 0.819 | **0.729** | 0.687 | 1.860 | 0.970 | **0.744** | $2.0 \pm 1.0$ |
| TabR-S | 0.860 | **0.403** | **3.067** | 0.865 | **0.133** | 0.818 | 0.722 | 0.690 | 1.747 | 0.973 | 0.750 | $1.9 \pm 0.7$ |
| TabR | **0.862** | **0.400** | 3.105 | **0.870** | **0.133** | **0.825** | **0.729** | **0.676** | **1.690** | **0.976** | 0.750 | $1.3 \pm 0.6$ |

Различие между TabR и TabR-S заключается в используемых эмбеддингах для действительнозначных признаков. В TabR используются PLR (Piecewise-Linear) эмбеддинги из Gorishny et al. 2022, “сближающие” представления категориальных и действительнозначных признаков. Отрыв от MLP-PLR всё же невелик, но мы скорее компенсируем его.

# Сравнения

| | CH ↑ | CA ↓ | HO ↓ | AD ↑ | DI ↓ | OT ↑ | HI ↑ | BL ↓ | WE ↓ | CO ↑ | MI ↓ | Avg. Rank |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tuned hyperparameters | | | | | | | | | | | | |
| XGBoost | 0.861 | 0.432 | 3.164 | **0.872** | 0.136 | **0.832** | 0.726 | 0.680 | 1.769 | 0.971 | 0.741 | $2.5 \pm 0.9$ |
| CatBoost | 0.859 | 0.426 | 3.106 | **0.872** | 0.133 | 0.827 | 0.727 | 0.681 | 1.773 | 0.969 | **0.741** | $2.5 \pm 1.1$ |
| LightGBM | 0.860 | 0.434 | 3.167 | **0.872** | 0.136 | **0.832** | 0.726 | 0.679 | 1.761 | 0.971 | 0.741 | $2.4 \pm 0.9$ |
| TabR | **0.865** | **0.391** | **3.025** | **0.872** | **0.131** | **0.831** | **0.733** | **0.674** | **1.661** | **0.977** | 0.748 | $1.3 \pm 0.9$ |
| Default hyperparameters | | | | | | | | | | | | |
| XGBoost | 0.856 | 0.471 | 3.368 | 0.871 | 0.143 | 0.817 | 0.716 | 0.683 | 1.920 | 0.966 | 0.750 | $3.4 \pm 0.9$ |
| CatBoost | 0.861 | 0.432 | 3.108 | **0.874** | 0.132 | 0.822 | **0.726** | 0.684 | 1.886 | 0.924 | **0.744** | $2.1 \pm 0.8$ |
| LightGBM | 0.856 | 0.449 | 3.222 | 0.869 | 0.137 | **0.826** | 0.720 | 0.681 | 1.817 | 0.899 | 0.744 | $2.5 \pm 0.9$ |
| TabR-S | **0.864** | **0.398** | **2.971** | 0.859 | **0.131** | 0.824 | 0.724 | 0.688 | **1.721** | **0.974** | 0.752 | $2.0 \pm 1.3$ |

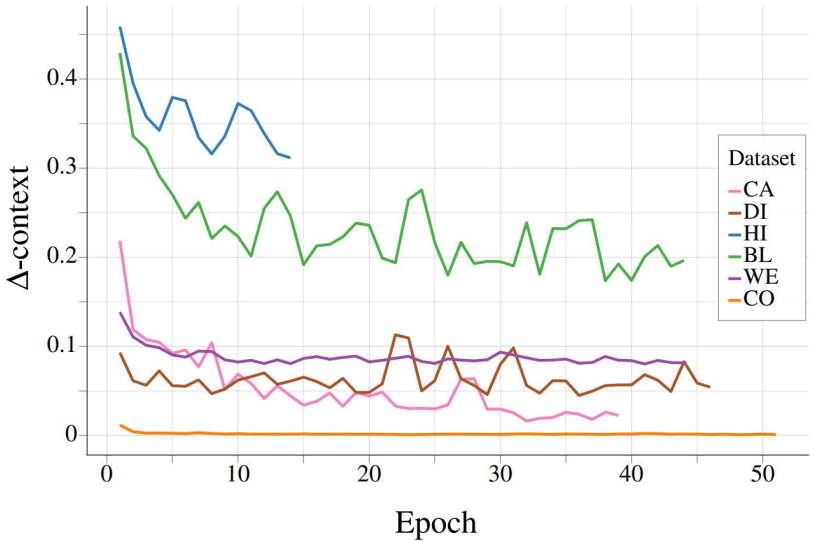
В основном качество сравнимо с бустингами, но на отдельных датасетах новая архитектура позволяет получить значительное улучшение.

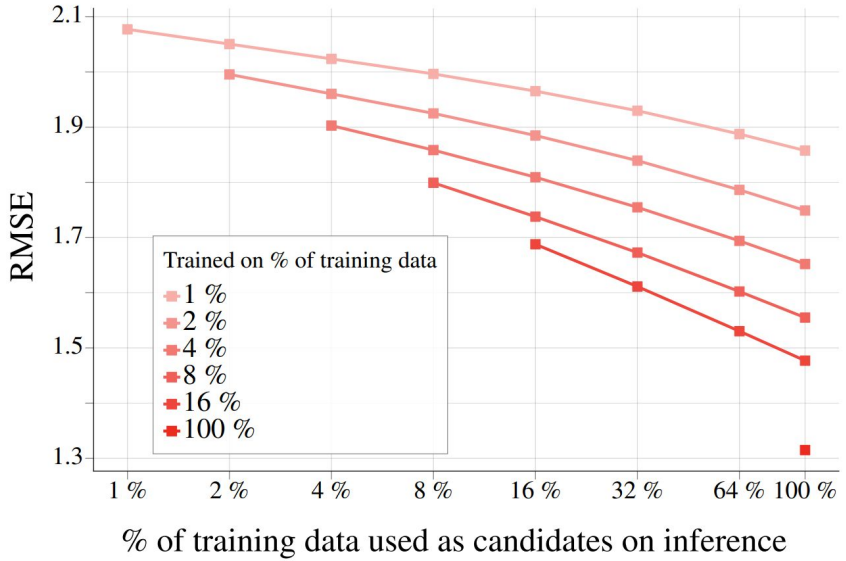
# Ограничения в использовании

- Проблемы с интерпретируемостью
- По-хорошему, нужно отбирать контекст, а не рассматривать весь датасет
- При этом отбор должен работать так же для новых объектов:
  - например, в рекомендательных системах мы не можем отбирать документы для каждого запроса в обучающей выборке, так как в реальности информации о документах у нас нет
- TabR скорее всего сложно скейлить на большие датасеты, хоть он и работает много быстрее других retrieval-моделей

# Ограничения в использовании

Сможем ли мы как-то уменьшить контекст?



# Ограничения в использовании

- TabR скорее всего сложно скейлить на большие датасеты, хоть он и работает много быстрее других retrieval-моделей:
  - Авторы замечают, что достаточно быстро контексты большинства объектов стабилизируются, и предлагают версию TabR с заморозкой контекста.
  - Это даёт значительное улучшение в скорости работы (вплоть до x7) с малыми потерями в качестве, но вместе с отбором кандидатов проблема всё ещё актуальна

| | CA ↓ | DI ↓ | HI ↑ | BL ↓ | WE ↓ | CO ↑ | WE (full) ↓ |
|---|---|---|---|---|---|---|---|
| TabR-S (CF-1) | 0.414 (0.72) | 0.137 (0.47) | **0.718** (0.80) | 0.692 (0.61) | 1.770 (0.57) | **0.973** (0.49) | 1.325 (0.13) |
| TabR-S (CF-4) | **0.409** (0.71) | 0.136 (0.51) | 0.717 (0.73) | **0.691** (0.62) | 1.763 (0.56) | **0.973** (0.59) | – |
| TabR-S | **0.406** (1.00) | **0.133** (1.00) | **0.719** (1.00) | **0.691** (1.00) | **1.755** (1.00) | **0.973** (1.00) | **1.315** (1.00) |

# Ограничения в использовании

- TabR скорее всего сложно скейлить на большие датасеты, хоть он и работает много быстрее других retrieval-моделей:


| | CH | CA | HO | AD | DI | OT | HI | BL | WE | CO | MI |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XGBoost | 0:00:01 | 0:00:20 | 0:00:05 | 0:00:05 | 0:00:02 | 0:00:35 | 0:00:15 | 0:00:08 | 0:02:02 | 0:01:55 | 0:03:43 |
| LightGBM | 0:00:00 | 0:00:04 | 0:00:01 | 0:00:01 | 0:00:03 | 0:00:34 | 0:00:10 | 0:00:07 | 0:06:40 | 0:06:22 | 0:06:45 |
| MLP | 0:00:02 | 0:00:18 | 0:00:09 | 0:00:17 | 0:00:15 | 0:00:31 | 0:00:24 | 0:01:38 | 0:00:29 | 0:04:01 | 0:02:09 |
| MLP-PLR | 0:00:03 | 0:00:43 | 0:00:14 | 0:00:24 | 0:00:25 | 0:02:09 | 0:00:17 | 0:00:52 | 0:20:01 | 0:03:32 | 0:30:30 |
| Retrieval-augmented models | | | | | | | | | | | |
| TabR-S (CF-4) | 0:00:08 | 0:00:25 | 0:00:30 | 0:00:34 | 0:00:43 | 0:00:57 | 0:01:02 | 0:03:08 | 0:09:08 | 0:23:13 | – |
| TabR-S | 0:00:20 | 0:01:20 | 0:01:23 | 0:03:04 | 0:01:44 | 0:01:17 | 0:02:09 | 0:11:22 | 0:12:11 | 0:49:59 | 0:55:04 |
| TabR | 0:00:16 | 0:00:40 | 0:00:55 | 0:01:30 | 0:01:24 | 0:01:47 | 0:06:22 | 0:04:14 | 1:03:18 | 0:37:03 | 1:46:07 |
| DKL | – | 0:06:15 | 0:03:55 | – | 0:21:59 | – | – | 1:04:10 | – | – | – |
| ANP | – | 0:37:40 | 0:42:16 | – | 2:14:38 | – | – | 1:32:27 | 6:00:11 | – | – |
| SAINT | 0:00:23 | 0:06:04 | 0:01:44 | 0:00:58 | 0:01:55 | 0:05:37 | 0:03:47 | 0:06:22 | 2:55:51 | 6:17:20 | 5:39:37 |
| NPT | 0:08:44 | 0:06:58 | 0:12:21 | 0:11:22 | 0:54:55 | 10:45:42 | 3:26:47 | 0:55:04 | 5:28:56 | 12:05:28 | 8:07:36 |

# Время для выводов

- Сделали важные шаги в развитии retrieval модулей в tabular dl
- Оценили необходимость каждого из улучшений
- Оставили много открытых вопросов, упомянув про них
- Практические ограничения всё ещё достаточно суровы
- В общем, до обучения формул на трансформерах ещё далеко (так ли?)

спасибо за внимание!