“On Embeddings for Numerical Features in Tabular Deep Learning” - Yury Gorishniy, Ivan Rubachev, Artem Babenko

Автор обзора-рецензии: Руденко Анастасия, БПМИ 193

Авторы изучают эмбеддинги для табличных данных. В работе описаны два метода: первый - Piecewise Linear Encoding (PLE), в котором i-й численный признак разбивается на T непересекающихся бинов, которые определяют функцию кодирования. Таким образом, PLE создает альтернативные начальные представления числовых признаков и может быть основным методом предобработки. Полученные представления вычисляются один раз, а затем используются вместо исходных скалярных значений во время основной оптимизации. Второй описанный метод - периодическая функция активации - авторы обучают параметры активации, а не оставляют их фиксированными. В эксперементах авторы показывают, что эти подходы

Основные выводы авторов: построение эмбеддингов числовых признаков - довольно неисследованная тема в табличном DL, хотя корректное построение может значительно увеличить качество предыдущих моделей. Также в экспериментах показано, что правильное построение не только улучшает качество трансформера, но и качество классических моделей MLP.

Работа написана 15 марта 2022 года. Авторы Юрий Горишний (МФТИ), Иван Рубачев (ВШЭ), Артема Бабенко (ВШЭ)

У авторов есть еще 2 работы на тему табличных данных: первая “Revisiting Deep Learning Models for Tabular Data” опубликована на 35-й конференции Neural Information Processing Systems (NeurIPS 2021), а вторая “Revisiting Pretraining Objectives for Tabular Deep Learning” в июле 22. Авторы ссылаются на собственные работы в каждой следующей, что логично, так как они продолжают изучение данной области. Первая работа о том, что модели на основе трансформаторов являются самой сильной альтернативой GBDT, а модели ResNet и MLP в сочетании с сильной настройкой гиперпараметров также являются конкурентоспособными бэйзлайнами в задачах с табличными данными. В последней работе авторы ссылаются на результаты текущей как на “*state-of-the-art solution for tabular DL”,* и используют описанные методы для построения одной из моделей.

Данная работа (судя по Google Scholar) цитировалась в еще в 4 работах, не считая работ авторов.

Можно выделить три основных статьи на которые ссылаются авторы. Первая - их предыдущая работа, связь с которой описана выше. Другие две оказали влияние на методы, используемые в статье. В PLE используется разбиение на бины (интервалы). Методы бинаризации описаны в работе “Supervised and unsupervised discretization of continuous features”. Эта работа является отправной точкой, хотя авторы и изменили оригинальный подход.

При описании второго метода, а именно, периодической функции активации авторы ссылаются на статьи, в которых данный метод используется для других задач: NLP (Vaswani et al., “Attention is all you need” 2017), CV (Li et al., “Learnable fourier features for multi-dimensional spatial positional encoding”, 2021).

Авторы описывают конкурентноспособную модель к GBDT и достигают strate-of-art для табличного DL, что является прекрасным результатом, который используется не только в дальнейших работах авторов, но и в других работаю по теме. В работе очень подробно расписаны все эксперименты, приведены сравнения для различных подходов, явно прописаны преимущества рассматриваемых методов и результаты отдельных экспериментов. Структура работы позволяет быстро выявить из нее необходимую информацию.

Не уверена, что можно это считать недостатком, но оба метода имеют довольно простую структуру. Да, на экспериментах они показывают хороший результат, но думаю усложнение этих методов может быть одним из направлений улучшения работы.