Рецензия на статью “On Embeddings for Numerical Features in Tabular Deep Learning”

Выполнила: Шешукова Марина
Группа: БПМИ192

Суть работы:
Использование трансформеров и MLP для задач с табличными данными активно развивается. В частности в этой статье предлагается исследовать именно эмбеддинги для числовых признаков, а не сами модели. Авторы предложили два способа получения эмбеддингов. Первый – кусочно-линейное кодирование (PLE), в котором i-ый признак разбивается на T непересекающихся бинов, которые и определяют функцию кодирования. Второй – с помощью периодических функций активации, в которых они обучают параметры. В своих экспериментах авторы показали, что такие подходы к получению эмбеддингов позволяют получать более мощные модели DL и позволяют конкурировать с GBGT на большинстве данных.

Контекст:
Статья была выпущена 31 октября 2022 года исследователями из Yandex Research – команда ученых, занимающихся проблемами в ML. Среди авторов, аспирант НИУ ВШЭ (Иван Рубачев), DL-исследователь Yandex Research (Юрий Горишный) и руководитель Yandex Research (Артем Бабенко). Иван Рубачев и Юрий Горишный уже несколько лет участвуют в различных публикациях связанных с ML и DL и имеют по 100-150 цитирований каждый. Руководитель имеет 4130 цитирований и индекс Хирша 18.

Все участники статьи до появления этой работы занимались изучением и применением глубокого обучения для табличных данных. В частности, все участники статьи являются авторами предшествующей статьи “Revisiting Deep Learning Models for Tabular Data”(2), которая значительно повлияла на данную работу.

Статья достаточно логичным образом следует из проблемы применения глубинного обучения к табличным данным. Проблемой таких моделей является то, что они не всегда оказываются лучше GBDT или разница не существенна. Трансформеры имеют успех в применении к различным областям, но вот в области табличных данных не было ответа, что лучше трансформер или GBDT. В предшествующих работах было показано, что глубокие архитектуры подобные трансформерам демонстрируют неплохой результат при применении к табличным данным. То есть на каких-то задачах, они превосходят XGBoost и CatBoost. Поскольку для архитектур трансформера нам надо сначала отобразить скалярные числовые признаки в многомерные эмбеддинги, авторы решили сосредоточиться не на основной модели, а изучить влияния эмбеддингов для числовых признаков на производительность моделей для табличных данных.

Предшественники и конкуренты:

Как уже было сказано, был ряд предшествующих работ, в которых было успешное применение трансформеров к табличным данных. В частности, это работа от этих же авторов “Revisiting Deep Learning Models for Tabular Data”, NeurlPS, 2021. В этой работе сказано, что модели на основе трансформеров являются самой сильной альтернативой GBDT, кроме того модели ResNet и MLP при качественной настройке гиперпараметров являются хорошими бейзлайнами в задачах на табличных данных. Именно на эту статью авторы опирались при проведении экспериментов в статье (1). То есть сами основные модели, с которыми авторы использовали свои эмбеддинги взяты из (2) и сравниваются они тоже со статьей (2). Это в целом логично, потому что в статье сравнили много моделей для табличного DL, показавшие хороший результат к моменту написания статьи. То есть на самом деле, можно сказать что статья (1) является просто продолжением изучения области применения DL для табличных данных.

Другая статья, которой вдохновлялись авторы: Tancik, M., et al., “Fourier features let networks learn high frequency functions in low dimensional domains”, In NeurIPS, 2020. В этой статье изменение входного пространства (positional encoding и другие манипуляции с тригонометрическими функциями) улучшило возможности обучения MLP,

что побудило авторов текущей рассматриваемой статьи проверить то же самое в отношении к табличному глубинному обучению.

Есть еще одна важная статья, она не имеет отношение к табличному глубинному обучению, но зато имеет отношение к представленному методу в статье. В PLE используется разбиение на бины и алгоритмы, которые делают это, представлены в работе Kohavi, R., et al “Error-based and entropy-based discretization of continuous features”.

Не знаю можно ли их назвать конкурентами, но мне кажется есть интересная статья. Как рассматриваемая нами работа, так и предшествующая ей (2) цитируется в следующей статье Grinsztajn L., et al., “Why do tree-based models still outperform deep learning on typical tabular data?”, NeurIPS, 2022. При этом по результатам последней упомянутой статьи трансформеры и MLP ((2), без эмбеддингов) проигрывают GBDT, но сравнения со статьей (1) у них нет. Несмотря на то, что упоминание этой статьи есть, как следствие замеченного ими факта о нейросетях в табличных данных.

Также есть пример стороннего исследования, в котором трансформеры побеждают GBDT: Somepalli et al., “SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training”, ICLR, 2022. В этой статье в основном внимание уделялось именно архитектуре трансформера, а для эмбеддингов они использовали Linear-ReLU-Linear слои.

Кроме того, 13 января 2023 года на arXiv появилась статья под названием Chen J., et al., “ExcelFormer: A Neural Network Surpassing GBDTs on Tabular Data”, они пишут, что существующее модели для табличных данных, которые показывают неплохие результаты, зависят от набора данных или предметной области и маловероятно найдут применение из-за неоднородности наборов табличных данных. Поэтому в своей статье они предлагают свою модель основанную на двух модулях внимания чередующихся друг с другом. С рассмотренной статьей они не сравнились, но видимо у них цель была в количестве данных.

И наконец последнее, что хочется отметить, это также опубликована статья от этих же авторов, но которая уже вышла позже под названием “Revisiting Pretraining Objectives for Tabular Deep Learning”, ICLR, 2023. В которой свои предыдущие результаты они называют SOTA. Они используют описанные в (1) методы для построения моделей, для которых они пробовали использовать предобучение.

Сильные стороны работы:

1) В статье представлены две простые схемы получения эмбеддингов, позволяющие нейронным сетям работать на табличных данных лучше.
2) Представлены результаты демонстрирующие, что эмбеддинги для числовых признаков действительно полезны и эффективны для моделей.
3) Статья написана хорошо и легко читается.
4) Представлена ссылка на github с кодом.

Слабые стороны работы:

1) Большинство данных, на которых авторы производят сравнения небольшие, и не до конца понятно будет ли такой же хороший результат на больших данных.
2) Методов представлено два, но не пояснено какой из них эффективнее использовать в реальных приложениях и как выбирать между ними.
3) Авторы не дают никакого теоретического обоснования. Поэтому не до конца понятно, почему эти методы не зависят от проведенных экспериментов и помогают оптимизации.

Дальнейшие исследования:

1) Можно попробовать покопаться в теории и объяснить, как именно обсуждаемые методы помогают оптимизации на теоретическом уровне.
2) Попробовать применить предложенные методы к большим датасетам (по количеству признаков и по количеству данных) и посмотреть дают ли они такой же хороший результат.

3) Как предложили и сами авторы, можно применять ко всем признакам не одно и то же преобразование, а разные, что возможно приведет к лучшим результатам.

Список литературы

1) Gorishniy Y. et al., Features in Tabular Deep Learning”
2) Gorishniy Y. et al., “Revisiting Deep Learning Models for Tabular Data”
3) Tancik M. et al., “Fourier features let networks learn high frequency functions in low dimensional domains”
4) Kohavi, R. et al., “Error-based and entropy-based discretization of continuous features”.
5) Grinsztajn L. et al., “Why do tree-based models still outperform deep learning on typical tabular data?”
6) Somepalli G. et al., 2021 “SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training”
7) Chen J. et al., “ExcelFormer: A Neural Network Surpassing GBDTs on Tabular Data”
8) Rubachev I. et al., “Revisiting Pretraining Objectives for Tabular Deep Learning”