Why do tree-based models still outperform deep learning on typical tabular data?

Статья 2022 года

Публиковалась в NIPS

Авторы статьи: Leo Grinsztajn, Edouard Oyallon, Gael Varoquaux

Leo Grinsztajn

Всего 3 публикации, схожих работ нет

- Interpreting Neural Networks through the Polytope Lens (2022)
- Bayesian workflow for disease transmission modeling in Stan (2021)

Edouard Oyallon

Схожих статей все так же нет, публикуется с 2014

- On Lazy Training in Differentiable Programming (573)
  (L Chizat, E Oyallon, F Bach)
- i-RevNet: Deep Invertible Networks (277)
  (JH Jacobsen, A Smeulders, E Oyallon)
- Deep Roto-translation Scattering for Object Classification (222)
  (E Oyallon, S Mallat)

Gael Varoquaux

Очень много статей, есть уже 2 статьи этого года (но цитирований мало)

- Machine learning for medical imaging: methodological failures and recommendations for the future
  (G Varoquaux, V Cheplygina)
- International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium
  (GA Brat, GM Weber, N Gehlenborg, ...)

Revisiting Deep Learning Models for Tabular Data

Yury Gorishniy[*†‡] **Ivan Rubachev**[†♣] **Valentin Khrulkov**[†] **Artem Babenko**[†♣]

† Yandex, Russia
‡ Moscow Institute of Physics and Technology, Russia
♣ National Research University Higher School of Economics, Russia

Abstract

The existing literature on deep learning for tabular data proposes a wide range of novel architectures and reports competitive results on various datasets. However, the proposed models are usually not properly compared to each other and existing works often use different benchmarks and experiment protocols. As a result, it is unclear for both researchers and practitioners what models perform best. Additionally, the field still lacks effective baselines, that is, the easy-to-use models that provide competitive performance across different problems.

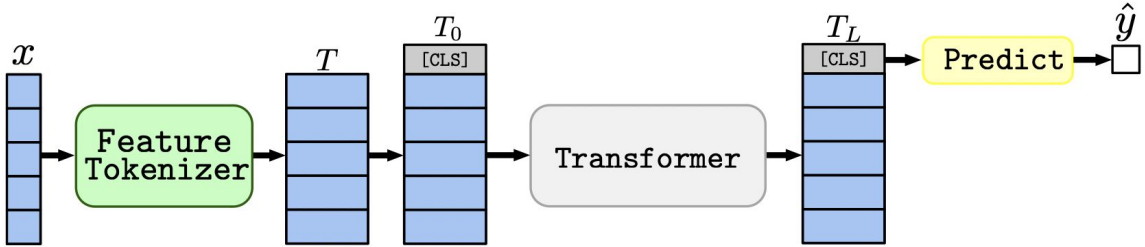
In this work, we perform an overview of the main families of DL architectures for tabular data and raise the bar of baselines in tabular DL by identifying two simple and powerful deep architectures. The first one is a ResNet-like architecture which turns out to be a strong baseline that is often missing in prior works. The second model is our simple adaptation of the Transformer architecture for tabular data, which outperforms other solutions on most tasks. Both models are compared to many existing architectures on a diverse set of tasks under the same training and tuning protocols. We also compare the best DL models with Gradient Boosted Decision Trees and conclude that there is still no universally superior solution. The source code is available at https://github.com/yandex-research/rtdl.

Revisiting Deep Learning Models for Tabular Data

- 11 datasets, comparing DL models
- ResNet

$$\begin{aligned}\texttt{ResNet}(x) &= \texttt{Prediction}\left(\texttt{ResNetBlock}\left(\ldots\left(\texttt{ResNetBlock}\left(\texttt{Linear}(x)\right)\right)\right)\right)\\ \texttt{ResNetBlock}(x) &= x + \texttt{Dropout}(\texttt{Linear}(\texttt{Dropout}(\texttt{ReLU}(\texttt{Linear}(\texttt{BatchNorm}(x))))))\\ \texttt{Prediction}(x) &= \texttt{Linear}\left(\texttt{ReLU}\left(\texttt{BatchNorm}\left(x\right)\right)\right)\end{aligned}$$

- FT-transformer



Сильные стороны

- Интересный подход с изучением влияния особенностей табличных данных
- Много важных результатов, которые могут пригодиться для поиска новых моделей для табличных данных
- Сформулированы открытые вопросы

Слабые стороны

- Качество моделей растет с ростом итераций в random search, было бы как минимум интересно посмотреть на сравнение моделей, гиперпараметры которых уже не улучшаются
- Анализ ограничен численными переменными и задачами классификации на наборах данных среднего размера
- Дифференцируемые деревья существуют (https://arxiv.org/abs/2002.07772)