NEURAL OBLIVIOUS DECISION ENSEMBLES FOR DEEP LEARNING ON TABULAR DATA

Sergei Popov Yandex
Stanislav Morozov Lomonosov Moscow State University
Artem Babenko National Research University Higher School of Economics

RELATED WORK

- The state-of-the-art for tabular data
  - XGBoost (Chen Guestrin, 2016)
  - CatBoost (Prokhorenkova et al., 2018)
  - LightGBM (Ke et al., 2017)
- Oblivious Decision Trees
- Differentiable trees
- Entmax
- Multi-layer non-differentiable architectures
- Specific DNN for tabular data

DIFFERENTIABLE OBLIVIOUS DECISION TREES

$$h(x) = R[\mathcal{H}(f_1(x) - b_1), ..., \mathcal{H}(f_d(x) - b_d)]$$

$$\hat{f}_i(x) = \sum_{j=1}^{n} x_j + entmax_\alpha(F_{ij})$$

$$\sigma_\alpha(x) = entmax_\alpha([x, 0])$$

$$c_i(x) = \sigma_\alpha(\frac{f_i(x) - b_i}{\tau_i})$$

$$C(x) = \begin{bmatrix} c_1(x) \\ 1 - c_1(x) \end{bmatrix} \otimes \begin{bmatrix} c_2(x) \\ 1 - c_2(x) \end{bmatrix} \otimes ... \otimes \begin{bmatrix} c_d(x) \\ 1 - c_d(x) \end{bmatrix}$$

$$\hat{h}_i(x) = \sum_{i_1,...,i_d} R_{i_1,...,i_d} * C_{i_1,...,i_d}(x)$$

ODT Layer

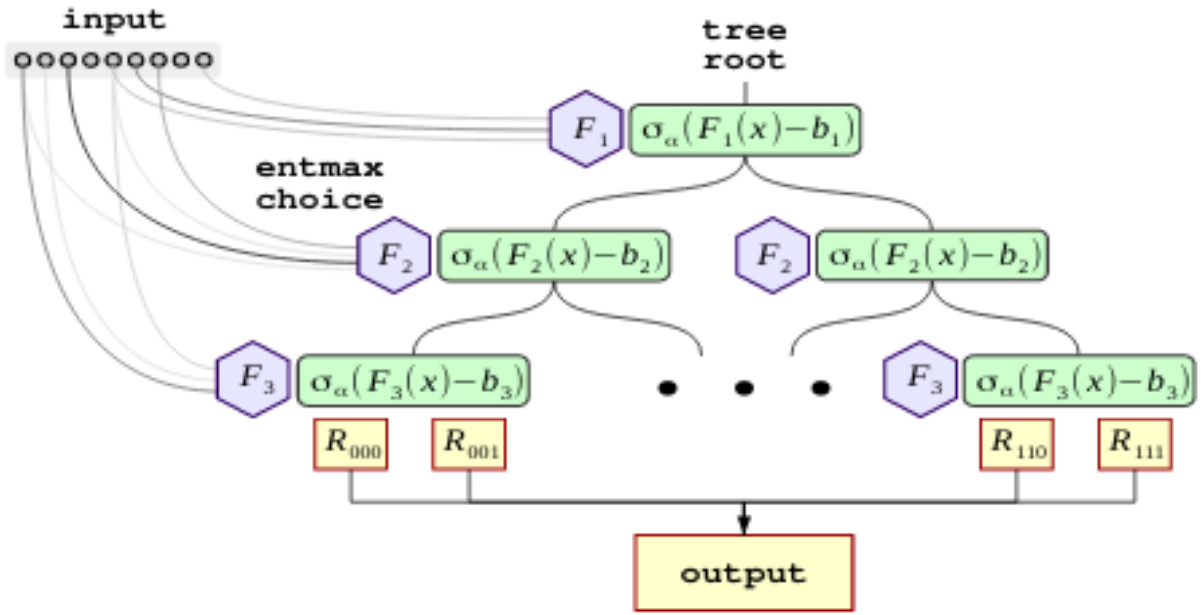

Figure 1: The single ODT inside the NODE layer. The splitting features and the splitting thresholds are shared across all the internal nodes of the same depth. The output is a sum of leaf responses scaled by the choice weights.

Architecture

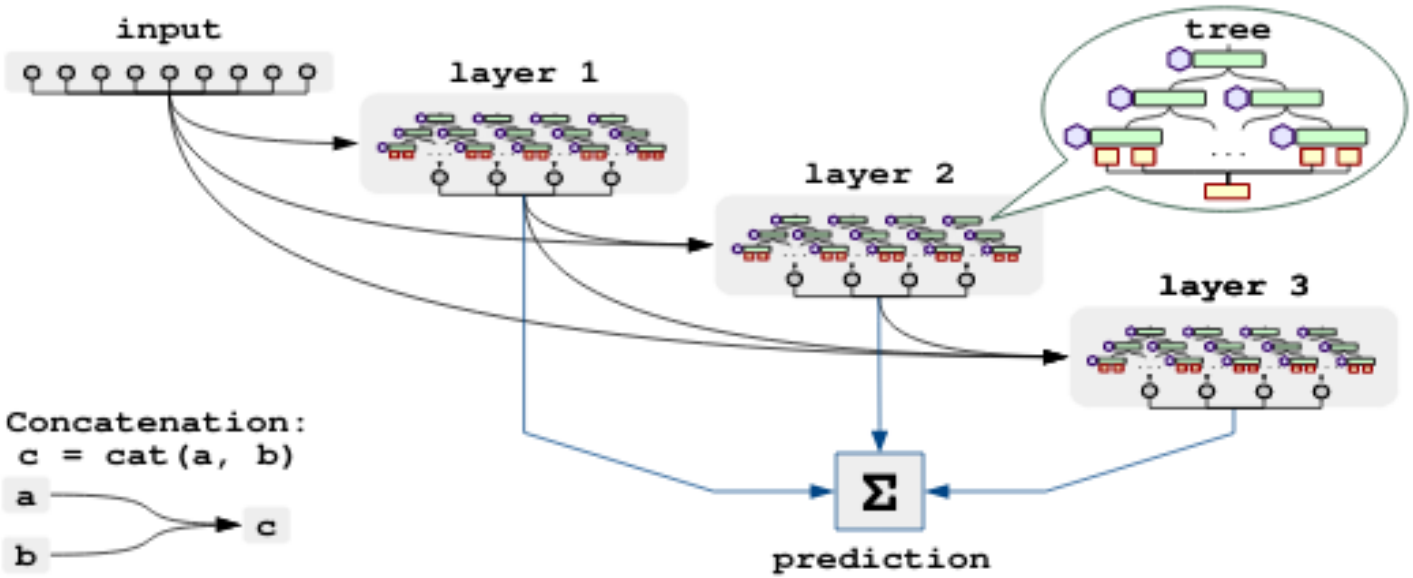

Figure 2: The NODE architecture, consisting of densely connected NODE layers. Each layer contains several trees whose outputs are concatenated and serve as input for the subsequent layer. The final prediction is obtained by averaging the outputs of all trees from all the layers.

Training

	Train	Test	Features	Task	Metric	Description
Epsilon[4]	400K	100K	2000	Classification	Error	PASCAL Challenge 2008
YearPrediction[5]	463K	51.6K	90	Regression	MSE	Million Song Dataset
Higgs[6]	10.5M	500K	28	Classification	Error	UCI ML Higgs
Microsoft[7]	723K	241K	136	Regression	MSE	MSLR-WEB10K
Yahoo[8]	544K	165K	699	Regression	MSE	Yahoo LETOR dataset
Click[9]	800K	200K	11	Classification	Error	2012 KDD Cup

Table 5: The datasets used in our experiments.

Comparison without fitting hyperparameters

	Epsilon	YearPrediction	Higgs	Microsoft	Yahoo	Click
Default hyperparameters						
CatBoost	$0.1119 \pm 2e{-4}$	80.68 ± 0.04	$0.2434 \pm 2e{-4}$	$0.5587 \pm 2e{-4}$	$0.5781 \pm 3e{-4}$	$0.3438 \pm 1e{-3}$
XGBoost	0.1144	81.11	0.2600	0.5637	0.5756	0.3461
NODE	$\mathbf{0.1043 \pm 4e{-4}}$	$\mathbf{77.43 \pm 0.09}$	$\mathbf{0.2412 \pm 5e{-4}}$	$\mathbf{0.5584 \pm 3e{-4}}$	$\mathbf{0.5666 \pm 5e{-4}}$	$\mathbf{0.3309 \pm 3e{-4}}$

Table 1: The comparison of NODE with the shallow state-of-the-art counterparts with default hyperparameters. The results are computed over ten runs with different random seeds.

Comparison with fitting hyperparameters

	Epsilon	YearPrediction	Higgs	Microsoft	Yahoo	Click
Tuned hyperparameters						
CatBoost	$0.1113 \pm 4e{-4}$	79.67 ± 0.12	$0.2378 \pm 1e{-4}$	$0.5565 \pm 2e{-4}$	$0.5632 \pm 3e{-4}$	$0.3401 \pm 2e{-3}$
XGBoost	$0.1112 \pm 6e{-4}$	78.53 ± 0.09	$0.2328 \pm 3e{-4}$	$\mathbf{0.5544 \pm 1e{-4}}$	$\mathbf{0.5420 \pm 4e{-4}}$	$0.3334 \pm 2e{-3}$
FCNN	$0.1041 \pm 2e{-4}$	79.99 ± 0.47	$0.2140 \pm 2e{-4}$	$0.5608 \pm 4e{-4}$	$0.5773 \pm 1e{-3}$	$0.3325 \pm 2e{-3}$
NODE	$\mathbf{0.1034 \pm 3e{-4}}$	$\mathbf{76.21 \pm 0.12}$	$\mathbf{0.2101 \pm 5e{-4}}$	$0.5570 \pm 2e{-4}$	$0.5692 \pm 2e{-4}$	$\mathbf{0.3312 \pm 2e{-3}}$
mGBDT	OOM	80.67	OOM	OOM	OOM	OOM
DeepForest	0.1179	—	0.2391	—	—	0.3333

Table 2: The comparison of NODE with both shallow and deep counterparts with hyperparameters tuned for optimal performance. The results are computed over ten runs with different random seeds.

ODT Layer

Dataset	YearPrediction				Epsilon			
Function	softmax	Gumbel	sparsemax	entmax	softmax	Gumbel	sparsemax	entmax
1 layer	78.41	79.39	78.13	77.43	0.1045	0.1979	0.1083	0.1043
2 layers	77.61	79.31	76.81	77.05	0.1041	0.2884	0.1052	0.1031
4 layers	77.58	79.69	76.60	76.21	0.1034	0.2908	0.1058	0.1033
8 layers	77.47	80.49	76.31	76.17	0.1036	0.3081	0.1058	0.1036

Table 3: The experimental comparison of various choice functions and architecture depths. The values represent mean squared error for YearPrediction and classification error rate for Epsilon.

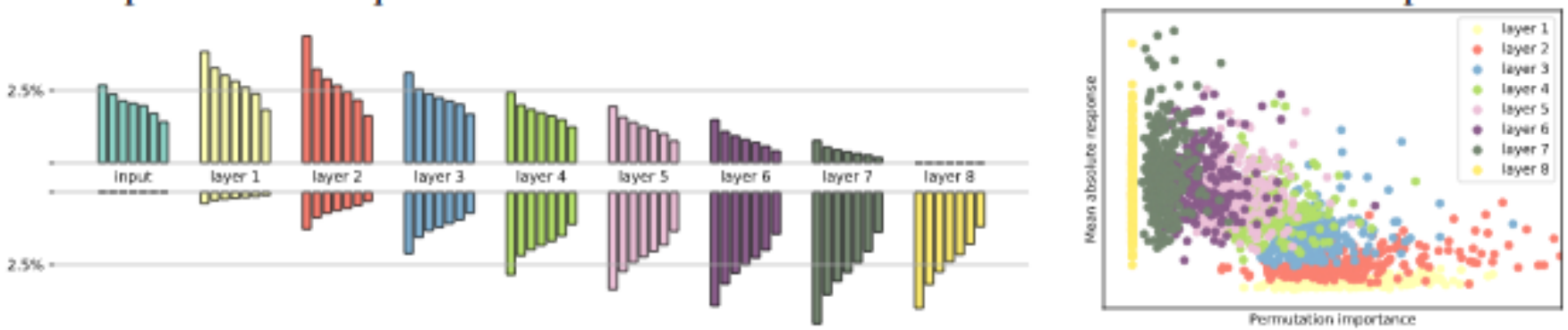

Figure 3: NODE on UCI Higgs dataset: **Left-Top:** individual feature importance distributions for both original and learned features. **Left-Bottom:** mean absolute contribution of individual trees to the final response. **Right:** responses dependence on feature importances. See details in the text.