SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training

Gowthami S., Micah G., Avi S.,
Bayan C.B., & Tom G.

Code and more materials available at
https://go.umd.edu/saint

SCAN ME

What is SAINT?

A transformer for tabular data.

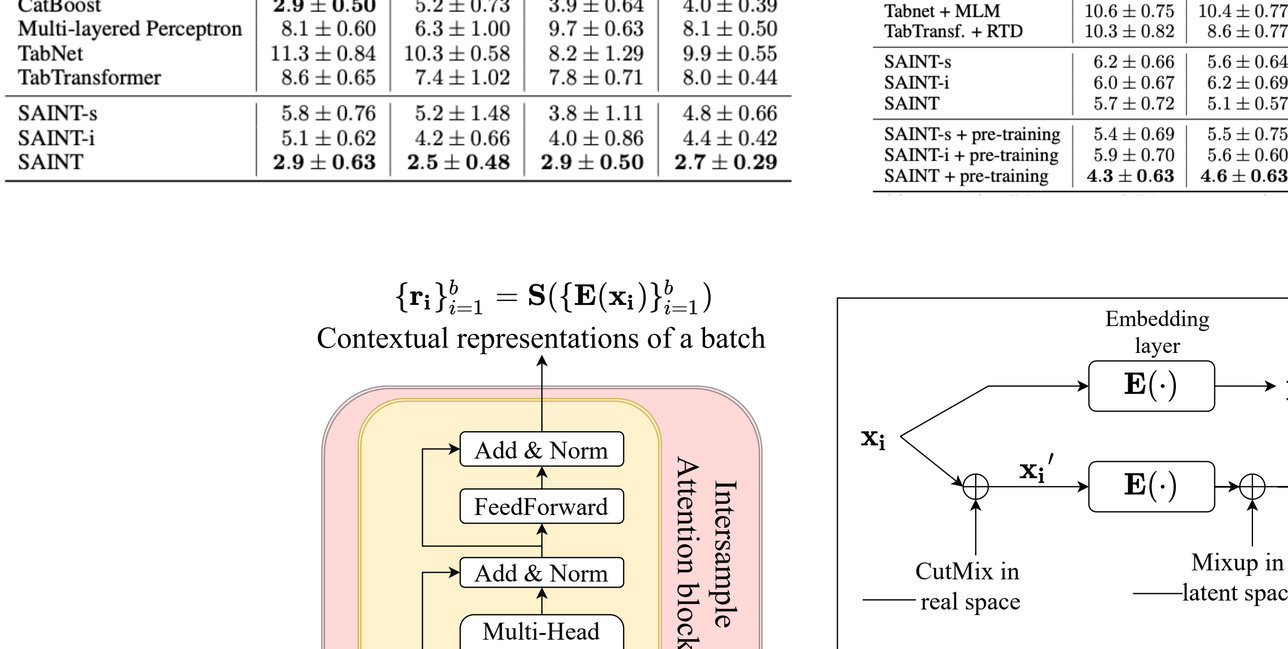
- Works for both classification and regression.
- Can handle categorical or numerical features.
- Allows task-agnostic pre-training

Our contributions

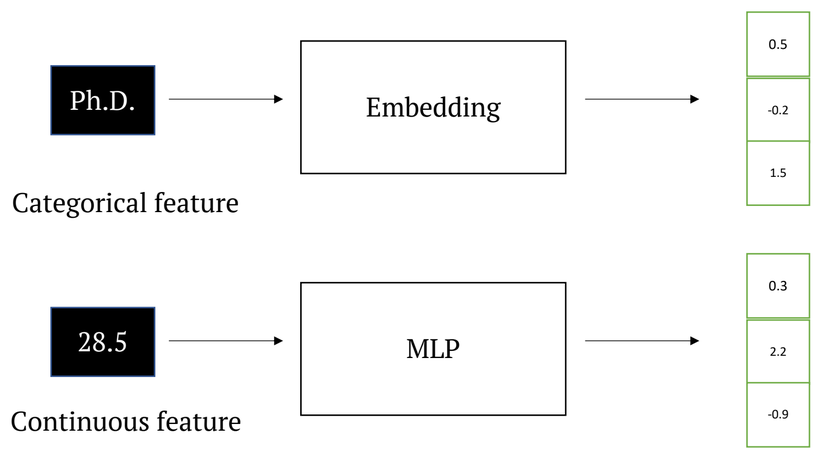
Transformer +

- ✓ Simultaneous embedding of numerical & categorical features.
- ✓ Intersample attention.
- ✓ New augmentation strategy for tabular data (Cutmix in real, mixup in latent space).
- ✓ Pre-training pipeline.

Intersample attention



Embedding all data types



$$\mathcal{L}_{\text{pre-training}} = \underbrace{-\sum_{i=1}^{m} \log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k=1}^{m} \exp(z_i \cdot z_k'/\tau)}}_{\text{Contrastive Loss}} + \lambda_{\text{pt}} \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n} [\mathcal{L}_j(\text{MLP}_j(\mathbf{r_i'}), \mathbf{x_i})]}_{\text{Denoising Loss}}$$

Model \ Task	Binary	Multiclass	Regression	Overall
RandomForest	5.9 ± 0.75	5.5 ± 0.56	7.3 ± 1.22	6.2 ± 0.52
ExtraTrees	5.8 ± 0.95	7.2 ± 0.63	6.7 ± 1.11	6.6 ± 0.55
KNeighborsDist	11.5 ± 0.43	8.2 ± 0.98	10.3 ± 0.63	10.0 ± 0.46
KNeighborsUnif	12.2 ± 0.47	8.5 ± 1.23	11.4 ± 0.54	10.7 ± 0.53
LightGBM	4.3 ± 0.60	3.2 ± 0.65	4.8 ± 0.93	4.3 ± 0.45
XGBoost	3.1 ± 0.67	4.5 ± 0.65	5.4 ± 0.65	5.0 ± 0.42
CatBoost	$\mathbf{2.9 \pm 0.50}$	5.2 ± 0.73	3.9 ± 0.64	4.0 ± 0.39
Multi-layered Perceptron	8.1 ± 0.60	6.3 ± 1.00	9.7 ± 0.63	8.1 ± 0.50
TabNet	11.3 ± 0.84	10.3 ± 0.58	8.2 ± 1.29	9.9 ± 0.55
TabTransformer	8.6 ± 0.65	7.4 ± 1.02	7.8 ± 0.71	8.0 ± 0.44
SAINT-s	5.8 ± 0.76	5.2 ± 1.48	3.8 ± 1.11	4.8 ± 0.66
SAINT-i	5.1 ± 0.62	4.2 ± 0.66	4.0 ± 0.86	4.4 ± 0.42
SAINT	$\mathbf{2.9 \pm 0.63}$	$\mathbf{2.5 \pm 0.48}$	$\mathbf{2.9 \pm 0.50}$	$\mathbf{2.7 \pm 0.29}$

Model \ # Labeled	50	200	All
RandomForest	9.9 ± 0.79	8.5 ± 0.67	8.1 ± 0.58
ExtraTrees	10.2 ± 0.75	8.0 ± 0.78	8.4 ± 0.63
KNeighborsDist	11.7 ± 0.67	13.2 ± 0.42	12.4 ± 0.58
KNeighborsUnif	12.3 ± 0.77	13.8 ± 0.48	13.1 ± 0.67
LightGBM	6.3 ± 0.58	7.1 ± 0.69	6.0 ± 0.54
XGBoost	7.5 ± 0.73	7.2 ± 0.62	6.7 ± 0.50
CatBoost	7.2 ± 0.61	6.0 ± 0.64	5.7 ± 0.46
MLP	10.9 ± 0.66	11.7 ± 0.57	11.1 ± 0.95
Tabnet + MLM	10.6 ± 0.75	10.4 ± 0.77	9.9 ± 0.61
TabTransf. + RTD	10.3 ± 0.82	8.6 ± 0.77	8.8 ± 0.83
SAINT-s	6.2 ± 0.66	5.6 ± 0.64	7.5 ± 0.95
SAINT-i	6.0 ± 0.67	6.2 ± 0.69	6.1 ± 0.43
SAINT	5.7 ± 0.72	5.1 ± 0.57	$\mathbf{4.2 \pm 0.36}$
SAINT-s + pre-training	5.4 ± 0.69	5.5 ± 0.75	6.6 ± 0.72
SAINT-i + pre-training	5.9 ± 0.70	5.6 ± 0.60	6.1 ± 0.87
SAINT + pre-training	$\mathbf{4.3 \pm 0.63}$	$\mathbf{4.6 \pm 0.63}$	4.3 ± 0.77

- Embedding the continuous data is important and can boost performance significantly.
- SAINT-i is likely to outperform other variants whenever the number of features is large.
- Intersample attention makes the model robust to noise in features.
- Pre-training does not improve performance when all the data is labeled.