南京航空航天大學
Nanjing University of Aeronautics and Astronautics

CT-BERT: Learning Better Tabular Representations Through Cross-Table Pre-training

Chao Ye*
Zhejiang University
Hangzhou, China
ye.chao@zju.edu.cn

Guoshan Lu*
Zhejiang University
Hangzhou, China
luguoshan@zju.edu.cn

Haobo Wang
Zhejiang University
Hangzhou, China
wanghaobo@zju.edu.cn

Liyao Li
Zhejiang University
Hangzhou, China
liliyao@zju.edu.cn

Sai Wu
Zhejiang University
Hangzhou, China
wusai@zju.edu.cn

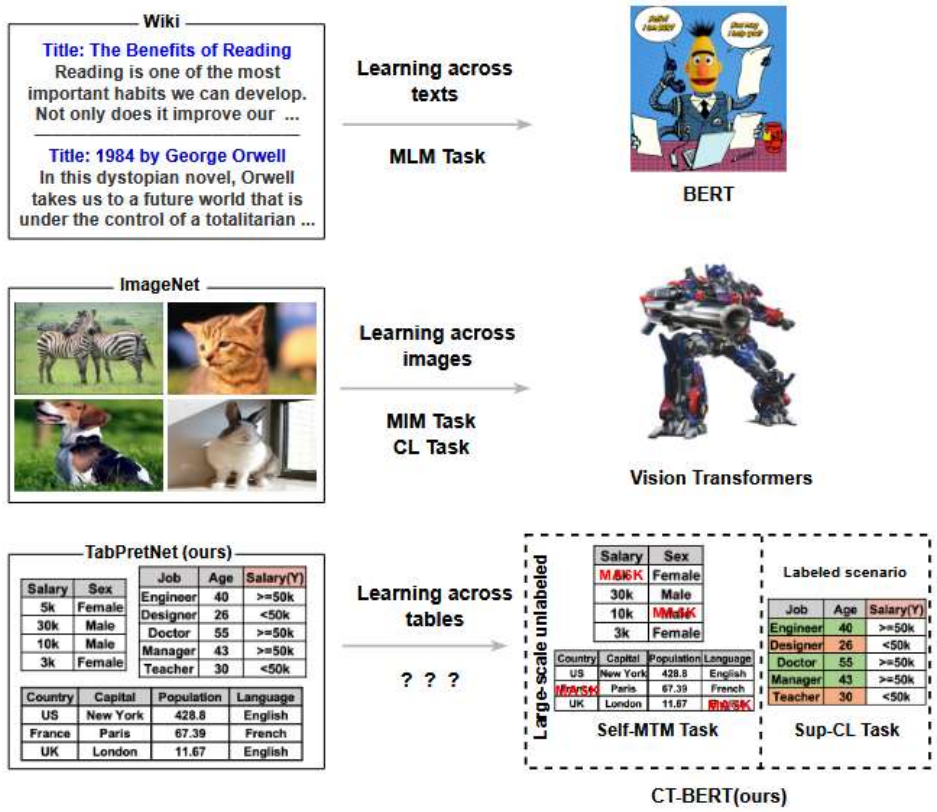
Gang Chen
Zhejiang University
Hangzhou, China
cg@zju.edu.cn

Junbo Zhao†
Zhejiang University
Hangzhou, China
j.zhao@zju.edu.cn

Background

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

- Traditional pre-training methods for tabular data are often limited to a single table or within a fixed-format architecture.
- How to extract common feature representations from multiple tabular data through cross-table pre-training has become an important research issue.



Challenges

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

C1. How can pre-training models accept inputs from heterogeneous tables as there are significant differences between different tables? For instance, the feature value "apple" appears under the column names "fruit" and "My_Laptop" in two different tables, conveying completely different meanings.

heterogeneity

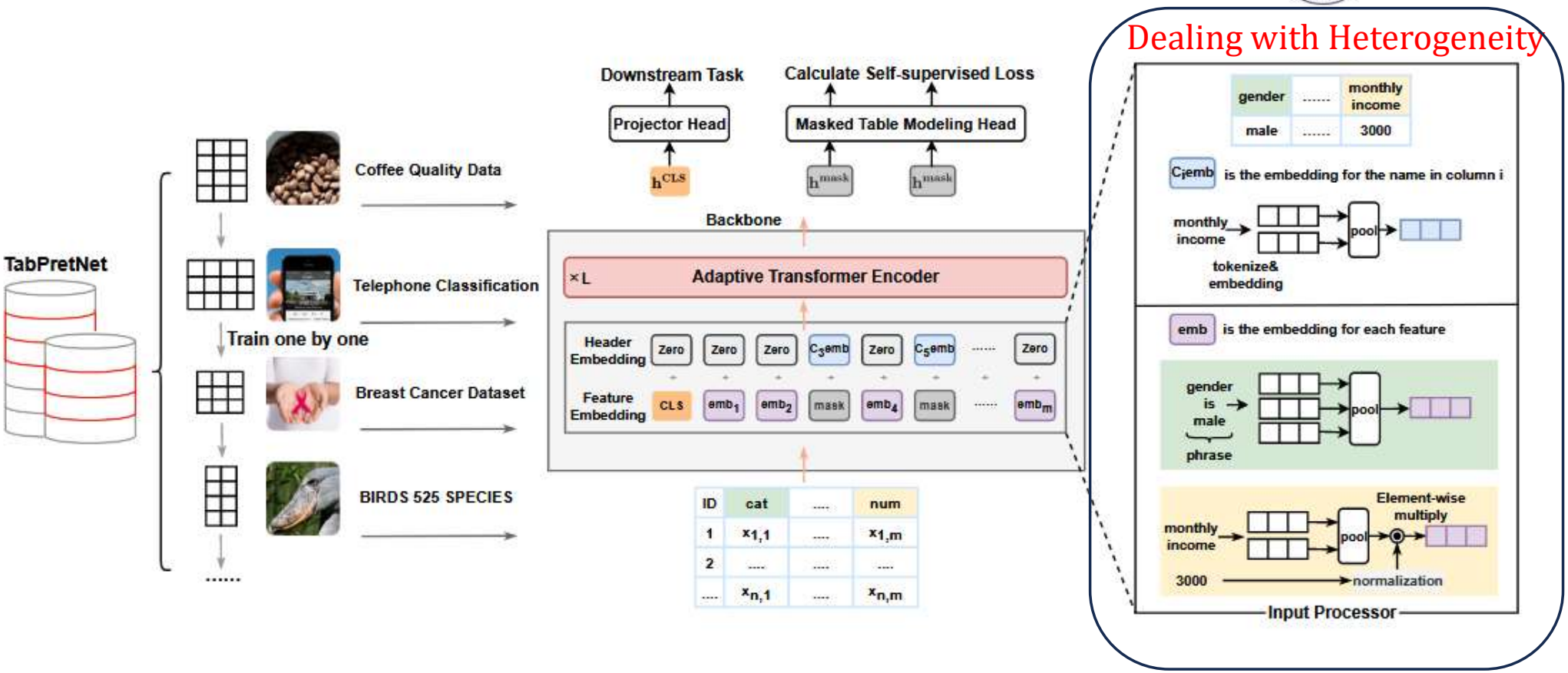
C2. Unlike image or text data where the pixels and word/character tokens are ordered, arbitrarily permuting any tables' rows or columns does not change its semantic meaning. We dub this property as *permutation invariance* uniquely to tabular data. Thus, how can the pre-training mechanism be compatible with this nature of tabular data?

permutation invariance

C3. Still driven by the difference against common vision or text data, how to design a suitable cross-table pre-training task objective because there is no obvious context or spatial structure in the tabular data?

Lack of apparent context or spatial structure

Methods — Overall Architecture



Methods — Dealing With Permutation Invariance

for NLP: cat loves Tom ≠ Tom loves cat

for Tabular Learning:

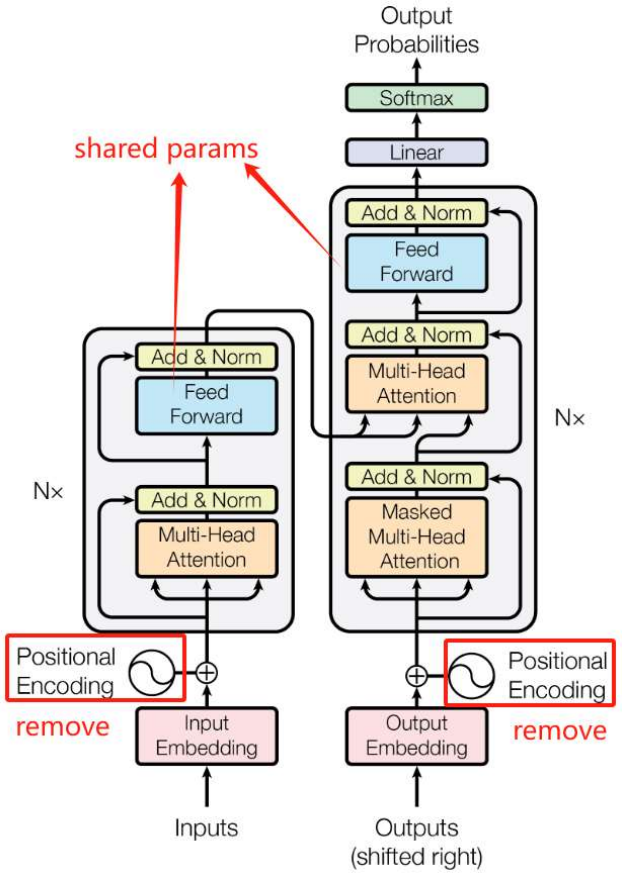
name	age	gender
Tom	23	male
Jenny	31	female

=

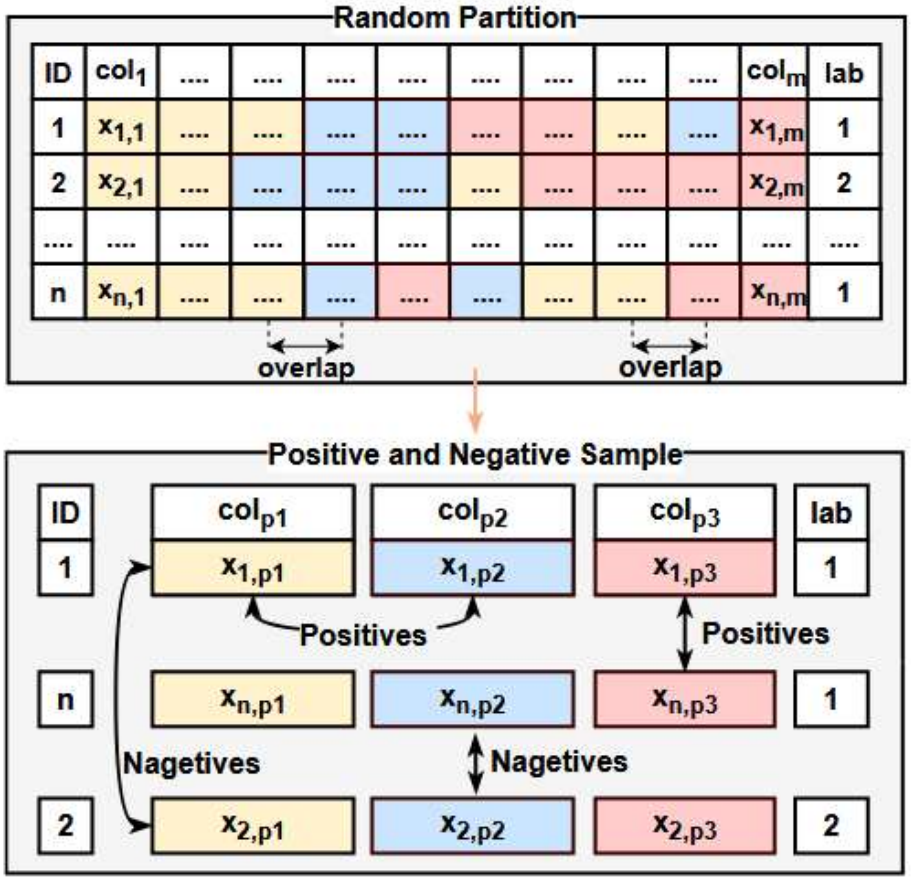
name	gender	age
Tom	male	23
Jenny	female	31

Due to the shared parameters, each feature column is treated equally and the model does not have a preference for columns in a particular position, thus maintaining the invariance of the arrangement of feature columns.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



Methods — Randomly Subsampled Supervised Contrastive Learning



$$\Psi(\mathbf{z}_i^{CLS}, \mathbf{z}_p^{CLS}) = -\log\left(\frac{\exp(sim(\mathbf{z}_i^{CLS}, \mathbf{z}_p^{CLS})/\tau)}{\sum_{i' \in B} \exp(sim(\mathbf{z}_i^{CLS}, \mathbf{z}_{i'}^{CLS})/\tau)}\right)$$

$$\mathcal{L}_{pretrain}^{CL}(\mathbf{X}, \mathbf{y}) = \frac{1}{|B|} \sum_{i \in B} \frac{1}{|P(i)|} \sum_{p \in P(i)} \Psi(\mathbf{z}_i^{CLS}, \mathbf{z}_p^{CLS})$$

Methods — Self-supervised MTM

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

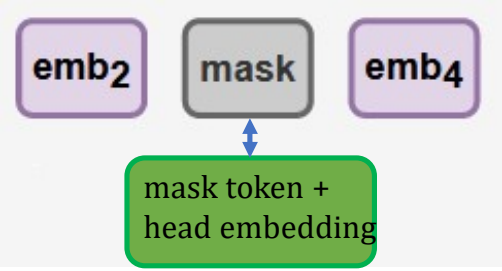
Step 1:Mask features

generate a binary mask vector: p^{mask}

$$\mathbf{m} = [m^1, m^2, \ldots, m^{a+b}] \in \{0,1\}^{a+b}$$

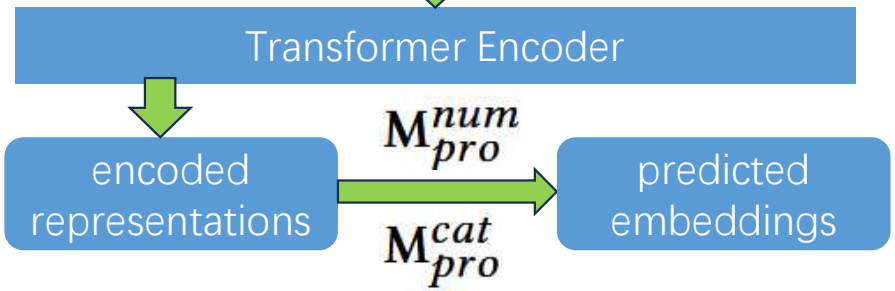
0 -> original feature | 1 -> masked feature

Step 2:Replace masked features



Step 3:Reconstruct masked features

$$\mathbf{x} = \{\mathbf{e}^j | m^j = 0\} \cup \{\mathbf{e}^{mask} + \mathbf{c}^j | m^j = 1\}$$



$$\Phi(\mathbf{x_i}, \mathbf{e_i}, \mathbf{z_i}) = \frac{1}{N^{num}} \sum_{j=1}^{N^{num}} (x_i^j - z_i^j)^2 + \frac{1}{N^{cat}} \sum_{j'=1}^{N^{cat}} (1 - sim(\mathbf{e}_i^{j'}, \mathbf{z}_i^{j'}))$$

$$\mathcal{L}_{pretrain}^{mask}(\mathbf{X}) = \frac{1}{|B|} \sum_{i \in B} \Phi(\mathbf{x_i}, \mathbf{e_i}, \mathbf{z_i})$$

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Dataset Setup:

1000 labeled datasets
1000 unlabeled datasets
} Large-scale cross-table pre-training dataset (TabPretNet)

15 common and high-quality tabular datasets from OpenML-CC18 → Downstream tabular tasks dataset

Baselines:

Shallow baselines: Logistic Regression, XGBoost, LightGBM
Neural network-based baselines: MLP, TransTab, TabNet, eg.

Metrics:

We use **AUC** as the main evaluation metric and improve on it using **5-fold cross-validation** as the final result.

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Overall Performance

Dataset	Shallow Methods			NN-Based Methods								Our Methods		
	LR	XGB	LightGBM	DCN-v2	AutoInt	MLP	FT-Trans	Saint	TabNet	VIME	TransTab	CT-BERT_NoPT	CT-BERT_P_M	CT-BERT_P_S
pc4	0.8621	0.7264	0.7938	0.8532	0.8741	0.8031	0.8990	0.8957	0.8449	0.8442	0.8882	0.8816	0.8837	**0.9030**
kc1	0.8004	0.6488	0.6623	0.7161	0.7710	0.7894	0.7957	**0.8459**	0.7917	0.7901	0.7945	0.7950	0.7846	0.7880
car	0.7393	0.9950	0.9134	0.9896	0.9664	0.9961	0.9981	0.9588	0.9713	0.9921	0.9039	0.9997	0.9996	**0.9998**
wilt	0.7098	0.8883	0.9249	0.9894	0.9601	0.6406	0.6978	0.9744	0.9934	0.9134	0.9850	0.9930	**0.9946**	0.9937
higgs	0.6346	0.6730	0.6935	0.6435	0.6237	0.6430	0.7063	0.7324	0.5474	0.6354	0.7284	0.6610	0.7002	**0.7348**
adult	0.8360	0.7894	0.8314	0.8923	0.8879	0.9023	0.9161	0.9152	0.9003	0.9128	0.9134	0.9150	**0.9155**	0.9143
climate	0.9449	0.7217	0.7014	0.8549	0.9097	0.4048	**0.9584**	0.8145	0.7951	0.8647	0.9345	0.9164	0.9204	0.9376
credit-g	0.7251	0.6755	0.7152	0.6912	0.7253	0.7370	0.7675	0.7817	0.6630	0.7659	0.7600	0.7701	0.7703	**0.7867**
vehicle	0.8912	0.9286	0.9305	0.9125	0.8883	0.9277	0.9231	0.8053	0.7877	0.7752	0.9178	0.9291	**0.9306**	0.9197
segment	0.9703	0.9929	0.9923	0.9746	0.9881	0.9858	0.9913	0.9809	0.9633	0.9752	0.9922	0.9908	0.9919	**0.9930**
amazon	0.5315	0.5231	0.6012	0.5564	0.5372	0.5461	0.5099	0.5550	0.5190	0.5081	0.5551	0.5698	**0.6092**	0.5313
satimage	0.9722	0.9889	0.9501	0.8023	0.9530	0.9863	0.9867	0.9838	0.9831	0.9126	0.9868	0.9856	**0.9897**	0.9888
phishing	0.9786	0.9669	0.9810	0.9389	0.9789	0.9943	0.9936	0.9923	0.9911	0.9913	0.8296	**0.9949**	**0.9949**	0.9942
mice-protein	0.9973	0.9993	0.9989	0.8894	0.9112	0.9997	0.9987	0.9973	0.9477	0.9579	0.9998	0.9981	**0.9999**	0.9998
cylinder-bands	0.7498	0.8197	0.7706	0.7465	0.7203	0.7070	0.8303	0.7415	0.5640	0.6916	0.8537	0.7629	0.8581	**0.8715**
mean	0.8229	0.8225	0.8307	0.8301	0.8463	0.8042	0.8648	0.8650	0.8175	0.8354	0.8695	**0.8775**	**0.8895**	**0.8904**

Experiments

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

Few-shot 5-fold AUC (%) on 6 datasets from the OpenML-CC18

Our Methods	Datasets						Mean
	vehicle	pc4	adult	phishing	cylinder	car	
shot=5							
CT-BERT_NoPT	0.6771	0.7459	0.7561	0.7389	0.6315	0.7527	0.7170
CT-BERT_P_M	0.7020	**0.7825**	0.8131	0.8435	0.6498	0.7260	0.7528
CT-BERT_P_S	**0.7126**	0.7684	**0.8778**	**0.8596**	**0.7246**	**0.8629**	**0.8010**
shot=10							
CT-BERT_NoPT	0.7440	**0.7721**	0.8082	0.8471	0.6694	0.8681	0.7848
CT-BERT_P_M	0.7514	0.7607	0.8499	0.9023	0.6652	0.8651	0.7991
CT-BERT_P_S	**0.7690**	0.7598	**0.8797**	**0.9424**	**0.7202**	**0.9186**	**0.8316**
shot=20							
CT-BERT_NoPT	**0.8312**	0.7695	0.8312	0.9253	0.6875	0.9591	0.8341
CT-BERT_P_M	0.8028	**0.7826**	0.8601	0.9512	0.6825	0.9464	0.8376
CT-BERT_P_S	0.8227	0.7601	**0.8805**	**0.9546**	**0.7734**	**0.9743**	**0.8609**

In the case of few samples, cross-table pre-training can significantly improve model performance. As the number of samples increases, the performance improvement gradually decreases.

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Ablation studies of different pooling strategies (mean AUC %)

Pooling Strategy		CT-BERT_NoPT	CT-BERT_P_M	CT-BERT_P_S
No-Pooing	token-level	0.8556	0.8630	0.8633
Max	feature-level	0.8695	0.8883	0.8774
Average	feature-level	**0.8775**	**0.8895**	**0.8904**
Self-Attention	feature-level	0.8681	0.8733	0.8710

Conclusion:

1. In tabular data, feature-level modeling is significantly better than token-level modeling
2. Average pooling gives the best results

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Thanks