4784

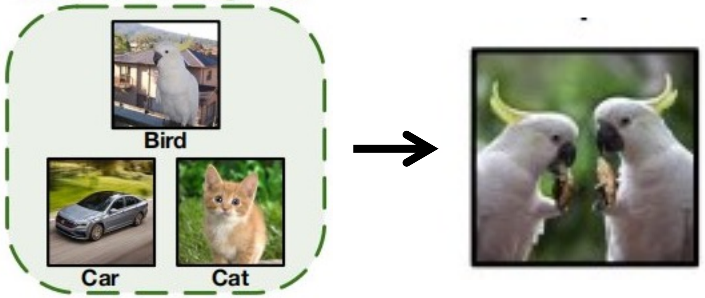
Making Pre-trained Language Models Great on Tabular Prediction

Jiahuan Yan[1], Bo Zheng[1], Hongxia Xu[1], Yiheng Zhu[1], Danny Z Chen[2], Jimeng Sun[3], Jian Wu[1,†], Jintai Chen [3,†]

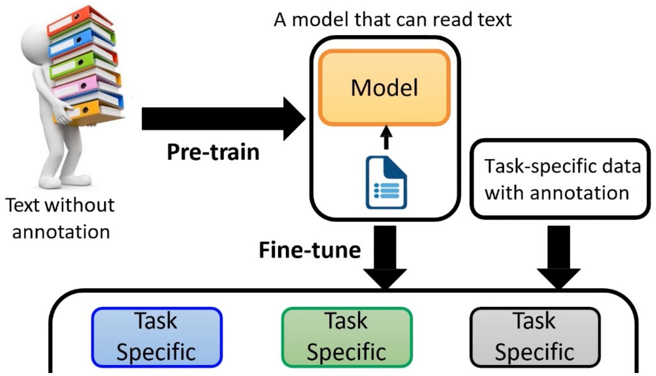
[1]Zhejiang University, Hangzhou, China
[2]University of Notre Dame, Notre Dame, USA
[3]University of Illinois Urbana-Champaign

Background

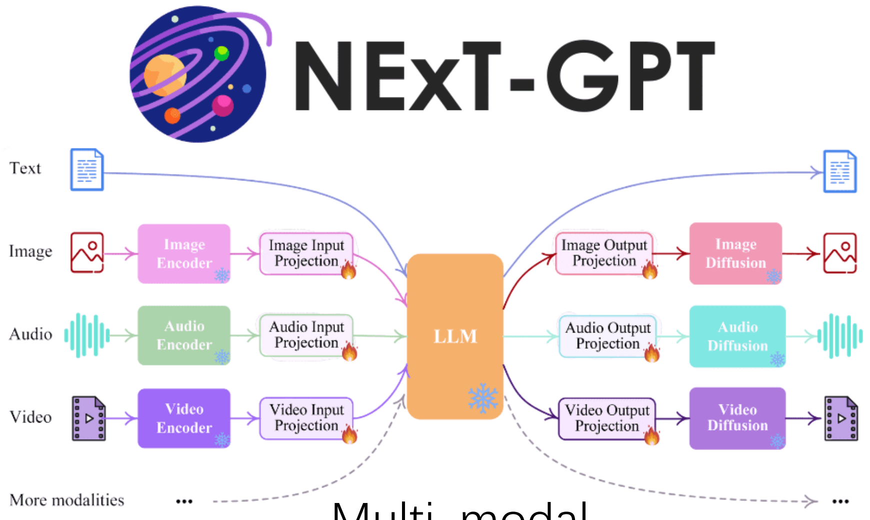
1. Universal success of DNN transfer learning on unstructured data



Vision transferability



Textual transferability

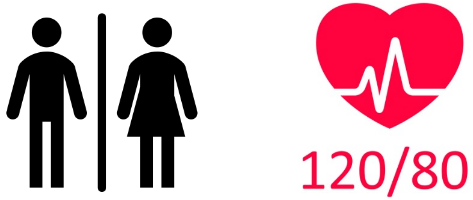


Multi-modal transferability

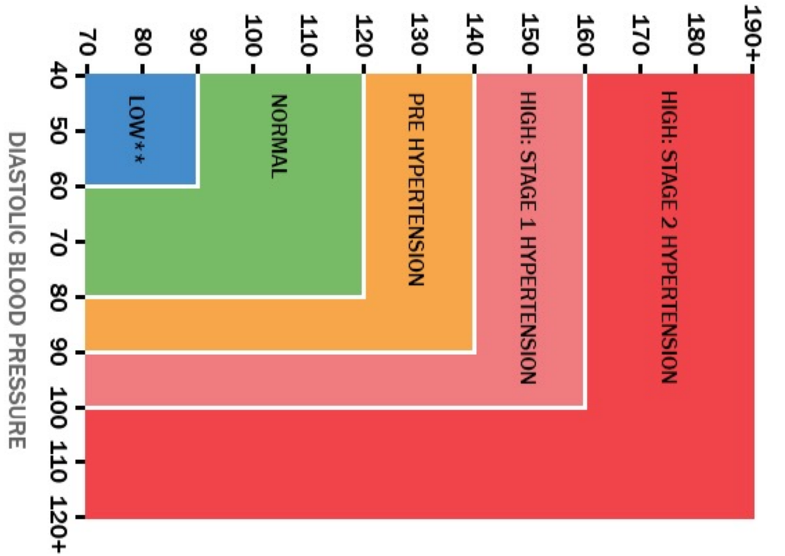
2. How to transfer on structured tabular data? (heterogeneity problem & numerical insensitivity)

Tabular Features

Gender	female
Blood Pressure	163 (mmHg)
Name	Value



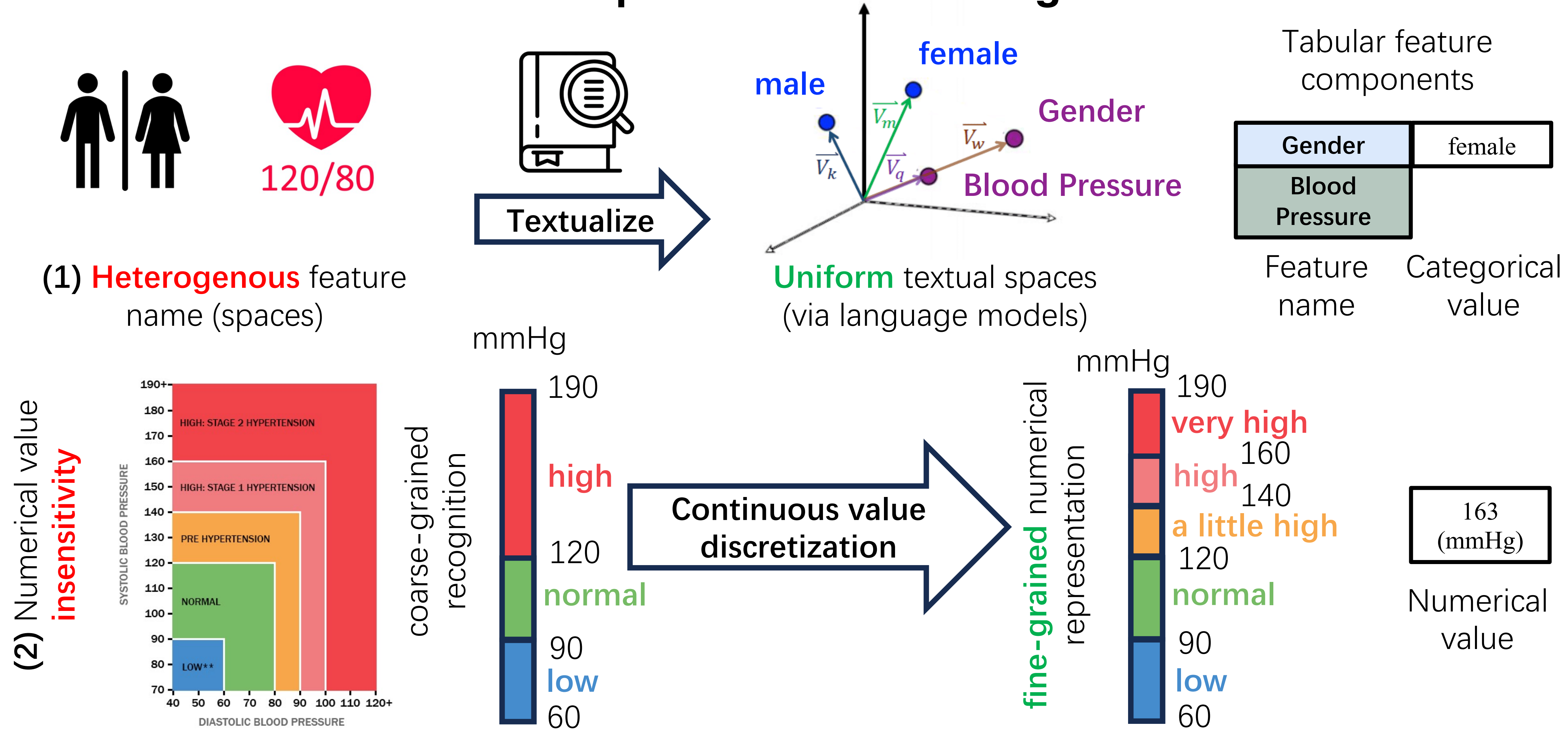
distinct feature spaces



numerical value representation

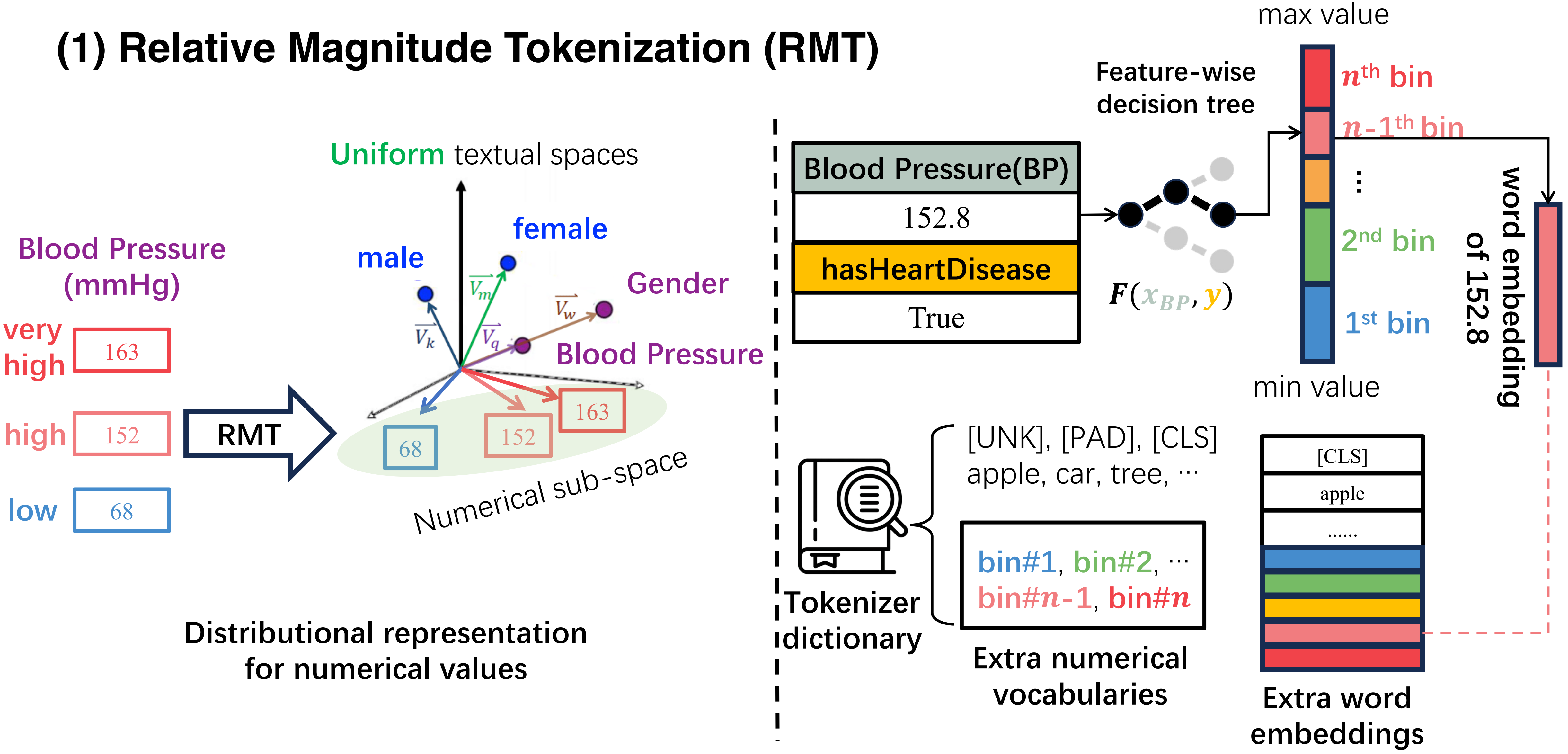
Inspiration

3. How to address the two representation challenges?



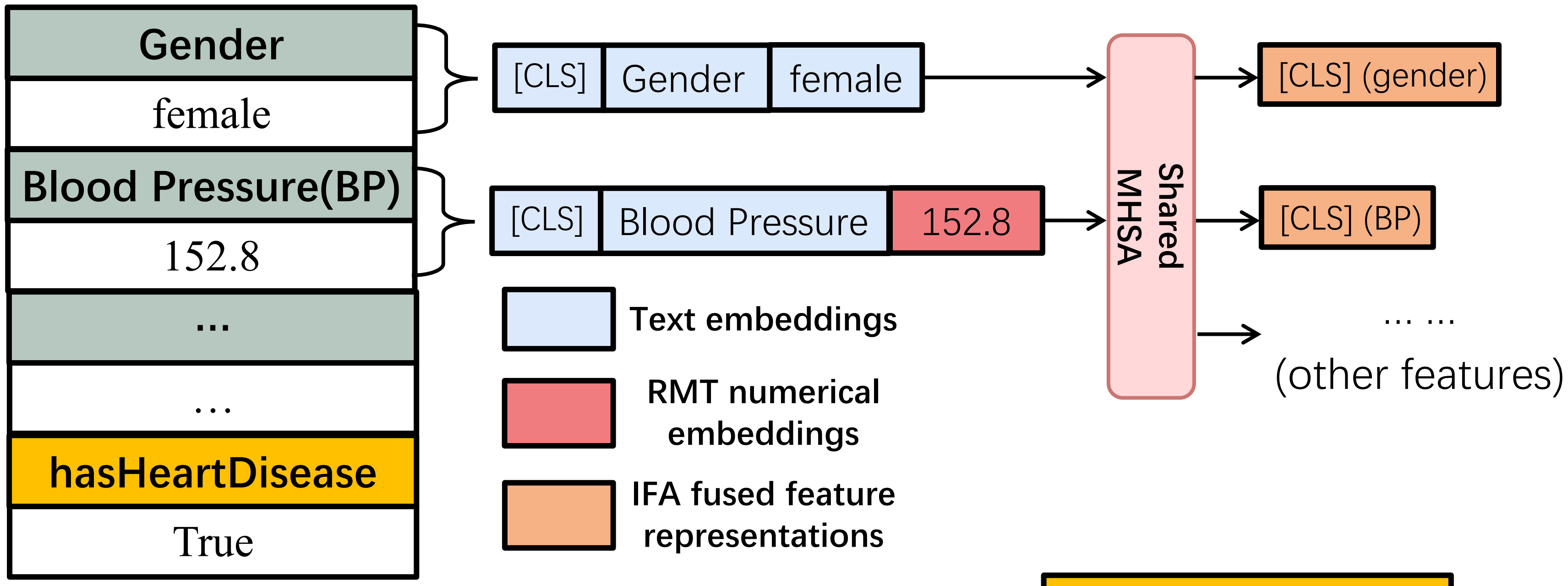
Method: Numerical Tokenization + LLM

(1) Relative Magnitude Tokenization (RMT)

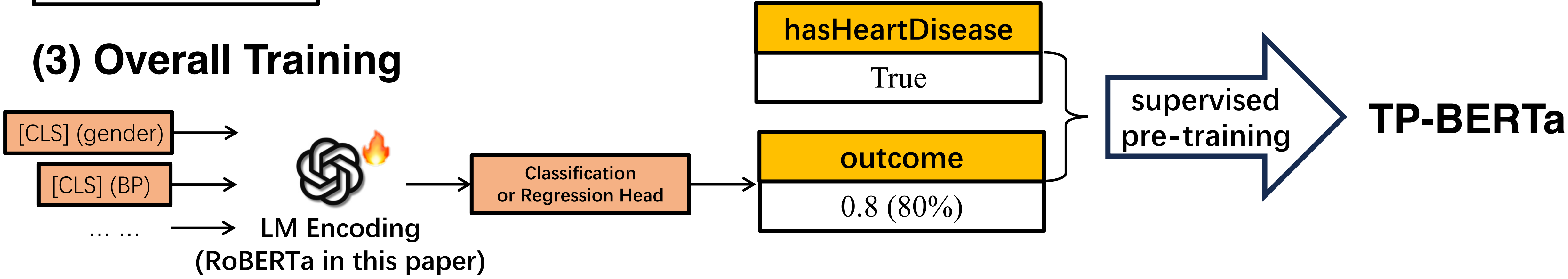


Method: Numerical Tokenization + LLM

(2) Intra-feature Attention (IFA)



(3) Overall Training



Main Experiments

Pre-training (supervised): 101 binary classification & 101 regression datasets
Downstream (supervised): 80 binary classification & 65 regression datasets

Table 1: The average values (standard deviations) of ranks

Baselines	80 downstream binary classification tasks						65 downstream regression tasks					
	All	$\alpha > 0$	$\alpha \geq 1$	$\alpha = 0$	$\beta > 0$	$\beta > 0.5$	All	$\alpha > 0$	$\alpha \geq 1$	$\alpha = 0$	$\beta > 0$	$\beta > 0.5$
XGBoost(d)	7.7(4.0)	7.8(4.1)	9.2(4.0)	6.8(3.5)	8.2(4.1)	8.3(3.9)	7.7(4.4)	7.7(4.6)	7.3(4.1)	7.8(4.0)	8.0(4.7)	9.2(4.3)
CatBoost(d)	6.7(4.1)	6.8(4.0)	7.4(4.0)	6.0(4.6)	7.0(4.1)	6.8(4.2)	5.5(2.7)	5.5(2.6)	5.5(2.7)	5.6(3.0)	5.5(2.7)	5.8(3.2)
FTT(d)	7.1(3.5)	7.0(3.5)	6.6(3.5)	6.9(3.6)	6.9(3.6)	7.2(3.6)	7.8(2.7)	7.8(2.5)	8.2(3.0)	7.6(3.2)	8.0(2.6)	8.3(1.3)
TransTab(d)	11.0(4.5)	11.2(4.5)	11.2(4.1)	10.2(4.6)	11.6(4.3)	11.7(4.2)	12.1(4.0)	12.1(3.8)	13.3(2.2)	12.4(4.5)	12.0(4.0)	13.6(1.2)
XGBoost(t)	6.2(4.1)	6.3(4.1)	6.5(4.3)	5.9(4.2)	6.5(4.2)	6.7(4.5)	4.5(3.7)	4.3(3.8)	**3.3(3.3)**	5.0(3.5)	4.7(3.9)	4.1(3.2)
CatBoost(t)	5.9(3.8)	6.3(3.9)	7.1(4.1)	**4.9(3.1)**	6.4(3.9)	6.4(4.1)	5.5(3.6)	5.7(3.6)	5.8(3.5)	4.9(3.7)	5.7(3.7)	6.1(3.8)
MLP(t)	8.6(4.0)	8.9(3.9)	8.7(4.1)	8.5(4.1)	8.5(3.9)	8.3(4.1)	8.5(3.6)	8.8(3.4)	9.3(3.2)	7.6(4.1)	9.0(3.4)	7.5(3.8)
AutoInt(t)	8.0(3.5)	7.8(3.3)	7.4(3.4)	8.6(4.0)	7.7(3.4)	7.7(3.2)	8.3(3.0)	8.6(3.0)	8.5(2.7)	7.4(3.1)	8.3(3.0)	8.2(3.2)
DCNv2(t)	7.9(3.9)	8.0(3.9)	8.4(3.8)	7.9(4.0)	7.7(3.9)	8.8(3.3)	8.4(3.4)	8.4(3.5)	8.5(3.1)	8.5(3.2)	8.4(3.5)	7.2(3.5)
TabNet(t)	12.1(3.5)	12.4(3.3)	12.7(2.7)	11.5(4.2)	12.3(3.4)	12.3(3.8)	12.6(3.6)	13.2(2.6)	13.1(2.4)	10.5(5.1)	13.5(1.9)	14.1(1.4)
SAINT(t)	8.2(3.8)	8.0(3.7)	8.1(4.1)	8.7(4.2)	7.9(3.8)	7.5(3.9)	7.6(3.8)	7.3(3.9)	7.7(3.3)	8.4(3.7)	6.6(3.6)	7.2(3.0)
FTT(t)	6.8(3.5)	6.8(3.6)	6.5(3.4)	6.2(3.3)	6.9(3.6)	6.9(3.9)	7.9(3.4)	7.6(3.3)	7.7(3.1)	9.0(3.4)	7.2(3.0)	6.8(3.2)
XTab(t)	9.8(4.0)	9.7(4.0)	8.9(3.8)	10.5(4.1)	9.4(4.0)	9.9(3.7)	12.4(2.8)	12.5(2.8)	13.3(1.6)	12.0(3.0)	12.4(2.9)	13.1(1.8)
Ours_j(d)	8.4(4.5)	7.7(4.5)	7.0(5.0)	9.9(4.1)	7.9(4.6)	7.0(4.7)	6.9(4.6)	6.3(4.4)	4.8(3.9)	8.5(5.0)	6.5(4.5)	5.2(3.9)
Ours_s(d)	**5.8(4.0)**	**5.1(3.9)**	**4.4(3.3)**	7.5(3.7)	**5.2(4.1)**	**4.5(3.4)**	**4.3(2.8)**	**4.1(2.6)**	3.9(2.4)	**4.8(3.4)**	**4.3(2.7)**	**3.6(2.8)**

Pre-training
Name: gender
Value: female

transfer

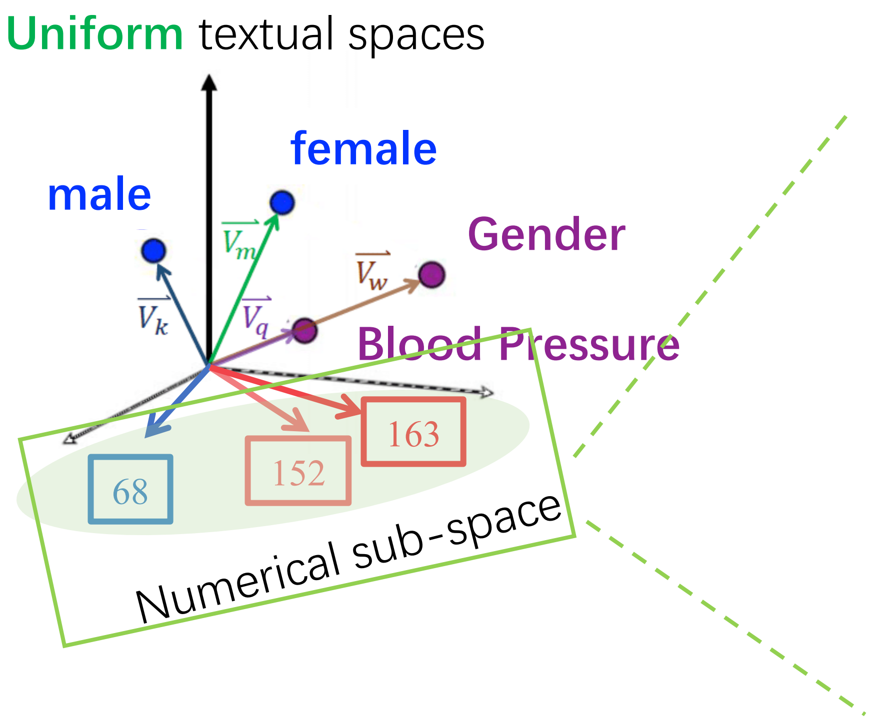
Downstream
Name: sex
Value: girl

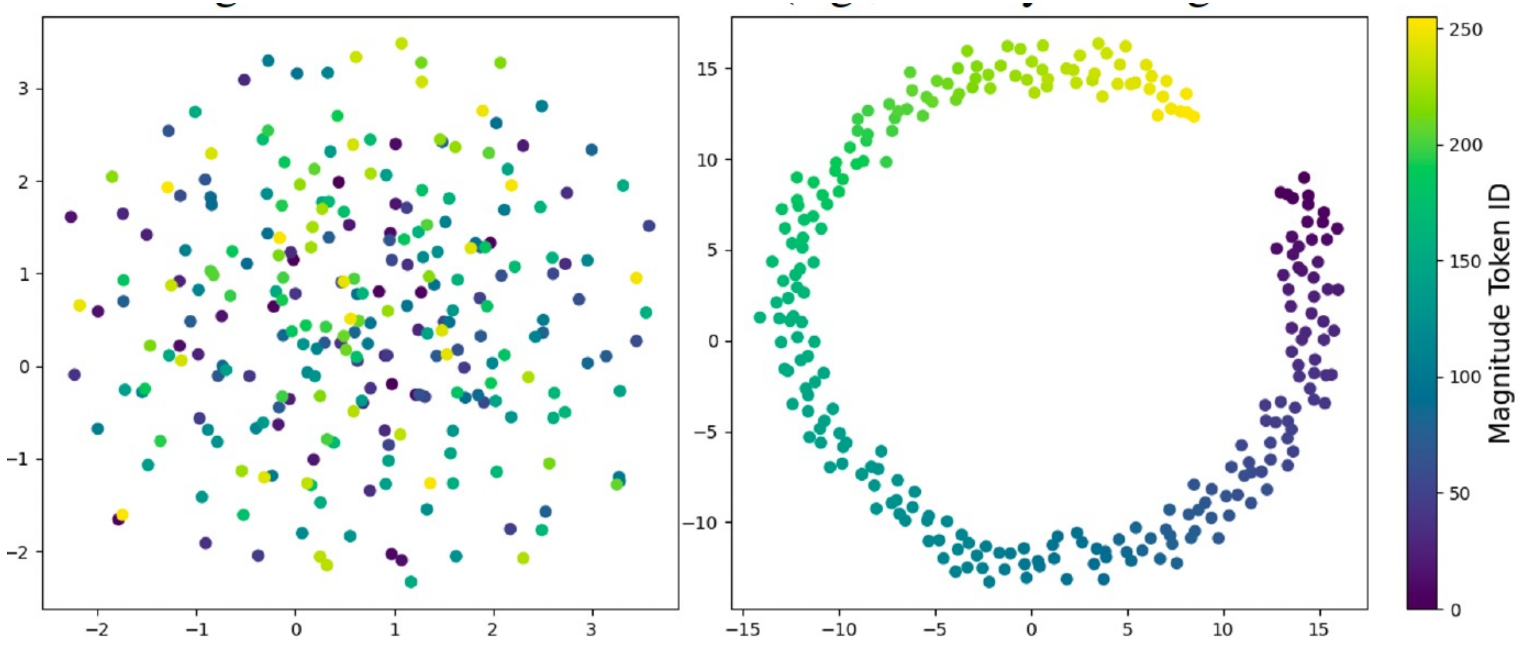
Denotation α: the amount ratio of categorical features, β: the significance of discrete features

[1] Database: OPENTABS, *Towards Cross-Table Masked Pretraining for Web Data Mining,* (WWW 24')

RMT Interpretability

Figure: The t-SNE visualization of 256 magnitude token embeddings before and after pre-training





Conclusion

- Firstly deal with **fundamental difficulties in LM adaptation to tabular data** (numeric feature handling and tabular feature organization) and propose general adaptation techniques RMT & IFA.

- Develop LM-based tabular DNNs and pre-train a tabular-data-tailored LM called **TP-BERTa**.

- Comparisons on extensive downstream datasets demonstrate that **pre-trained LMs can be superior to non-LM tabular DNNs and competitive with GBDTs in typical tabular regime**.

Project repo https://github.com/jyansir/tp-berta
Personal homepage https://jyansir.github.io

WeChat