PORTAL: Scalable Tabular Foundation Models via Content-Specific Tokenization

Marco Spinaci, Marek Polewczyk, Johannes Hoffart, Markus C. Kohler, Sam Thelin, Tassilo Klein

SAP SE

Introduction

PORTAL (Pretraining One-Row-at-a-Time for All tabLes) is a framework to predict values in table cells that handles various data modalities without the need for cleaning or preprocessing. This approach can be effectively pre-trained on multiple datasets and fine-tuned to match state-of-the-art methods on complex classification and regression task.

Encoding

Depending on the type of input data, distinct encoding mechanisms are employed:

- Date (Day, Month, Year) → Learnable embedding layer to d dimension
- Numeric $x = \pm\alpha \cdot 2^{\beta}$ → Learnable embedding layer for sign ($\pm$), α, β
- Text Column name → all-MiniLM-L6-v2 + linear layer to d dimension

At the end of the encoding process, each cell is represented by a single vector.

Decoding

Decoding transforms the tokens that have gone through the backbone encoder into features similar to those in the encoding.

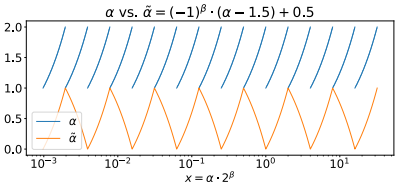
Loss functions:

- **Cross-entropy** loss for discrete fields: day, month, year, sign, exponent
- **Binary cross-entropy** loss for the fraction "$\tilde{\alpha}$" of numbers
- **Huber** loss for text embeddings

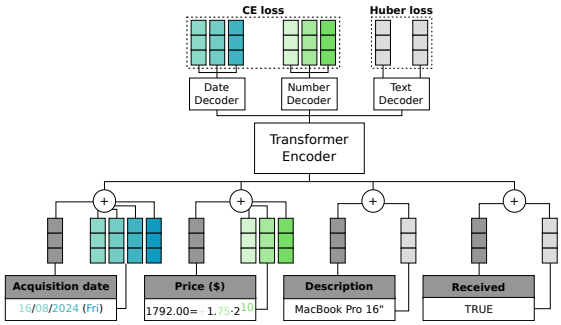
Numerical decoding improvement for scientific notation: $x = \pm\alpha \cdot 2^{\beta}$

$$\tilde{\alpha} = (-1)^{\beta}\left(\alpha - \frac{3}{2}\right) + \frac{1}{2}$$

($\alpha - 1$ if β is even; $2 - \alpha$ if β is odd)



Model Architecture



Encoder-only transformer as our primary architecture, optimally adapted for generating embeddings from heterogeneous data types.

Pre-training

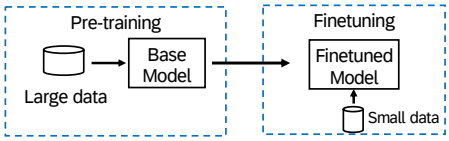
Our model is pre-trained using tabular data from English Wikipedia:

- infoboxes (treated as single-row tables),
- inline tables (multi-row tables found in article texts).

Each cell is picked with 30% probability:

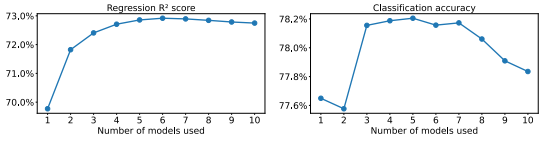
- zeroed out with 80% probability
- replaced with a random value from the same column with 10% probability
- left unchanged with 10% probability

Header names are not masked.



Model Ensemble

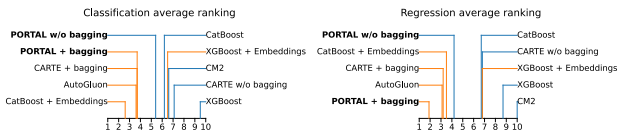
State-of-the-art performance was achieved for the ensemble of 10 PORTAL models using bagging.



Finetuning Results

Finetuning results on 51 predominantly textual datasets from (Kim et al. 2024) to perform classification and regression on the target column:

Method	Acc. (%)	R^2
CARTE w/o bagging (Kim et al., 2024)	75.2	67.6
CatBoost (Prokhorenkova et al., 2018)	75.4	66.7
XGBoost (Chen and Guestrin, 2016)	71.8	59.0
CM2 (Ye et al., 2024)	76.3	4.9
PORTAL w/o bagging	**77.0**	**71.4**
CARTE 10 models bagging	78.3	72.3
CatBoost + Embeddings	**78.4**	72.3
XGBoost + Embeddings	76.5	67.5
AutoGluon (Erickson et al., 2020)	**78.4**	72.6
PORTAL 10 models bagging	77.8	**73.8**



Regression Target Encoding Comparison

Method	Targets	Binned?	Loss	Normalization	R^2 score (%)	# Failures
PORTAL L^2	y	No	L^2	Standard	**70.9**	0
PORTAL $\tilde{\alpha}$	$\pm, \tilde{\alpha}, \beta$	No	XE	None	67.3	0
Raw L^2	y	No	L^2	None	58.1	0
Percentile	y	Yes	XE	None	63.8	0
Not continuous	$\pm, \alpha - 1, \beta$	No	XE	None	57.8	4
Binned $\tilde{\alpha}$	$\pm, \tilde{\alpha}, \beta$	Yes	XE	None	64.6	1
Continuous L^2	$\pm, \tilde{\alpha}, \beta$	No	L^2	None	63.8	1
Standard $\tilde{\alpha}$	$\pm, \tilde{\alpha}, \beta$	No	XE	Standard	64.1	0
Power L^2	y	No	L^2	Power	67.2	2
Power $\tilde{\alpha}$	$\pm, \tilde{\alpha}, \beta$	No	XE	Power	58.0	4

References

Myung Jun Kim, Leo Grinsztajn, and Gael Varoquaux (2024). "CARTE: Pretraining and transfer for tabular learning." In: Forty-first International Conference on Machine Learning.

Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin (2018). "Catboost: unbiased boosting with categorical features." In: NeurIPS, pages 6639–6649.

Tianqi Chen and Carlos Guestrin (2016). "XGBoost: A scalable tree boosting system." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794

Chao Ye, Guoshan Lu, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao (2024). "Towards cross-table masked pretraining for web data mining." In: Proceedings of the ACM on Web Conference 2024, pages 4449–4459.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola (2020). "Autogluon-tabular: Robust and accurate automl for structured data." arXiv preprint arXiv:2003.06505.