# TransTab: Learning Transferable Tabular Transformers Across Tables

Zifeng Wang, Jimeng Sun

University of Illinois Urbana-Champaign

## Flexibility in tabular learning

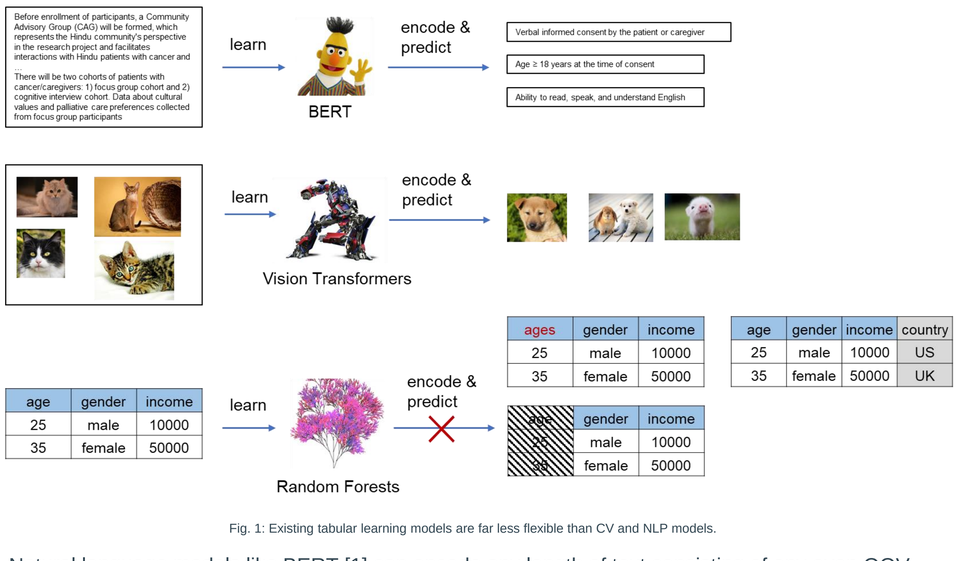


Fig. 1: Existing tabular learning models are far less flexible than CV and NLP models.

Natural language models like BERT [1] can encode any length of text consisting of any even OOV words; Computer vision models like ViT [2] can encode any images consisting of pixels ranging in [0,255].

However, existing tabular learning models are much less flexible. For example, if the model is trained on a table with three columns: age, gender, income, it can only deal with future tables that have the strictly same columns. Otherwise, the model will feel confused and not work at all.

## How flexible tabular learning can be with TransTab?

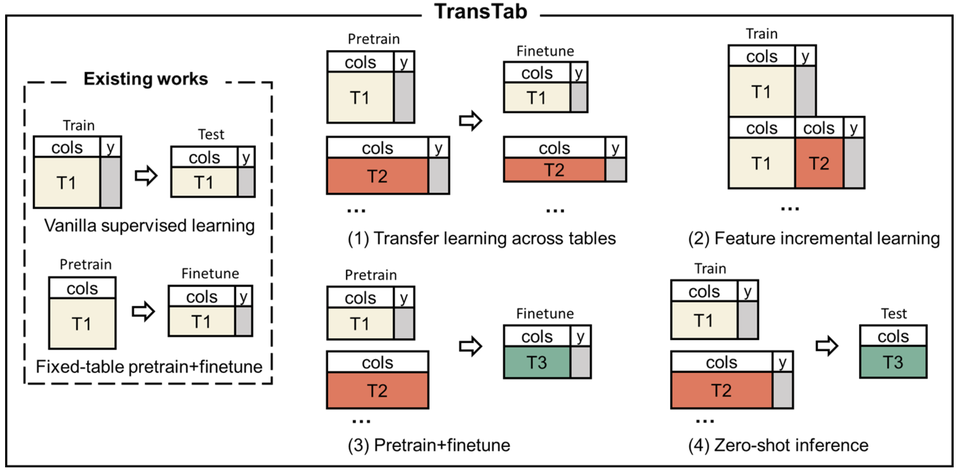


Fig. 2: The new tasks that TransTab can handle but other existing models cannot.

As the figure shows, TransTab can deal with many new tasks credited to its flexibility in encoding **any** input tables.

Specifically, the basic elements of the TransTab's inputs are no longer *features*; instead, there are *tokens*. For instance, in the table in Fig. 1, TransTab first linearizes the input row as

$$\text{age: 25 [SEP] gender: male [SEP] income: 10000}$$

which incorporates the column semantics into the encoding. It is drawn from the motivation that the feature's meaning is dependent on column semantics. We know it is 25 years old instead of 25 KG because it is under the column age.

With the above formulation, TransTab can encode any input tables by merging column semantics and feature values, which gives rise to four novel tasks:

- Transfer learning across tables.
- Feature incremental learning.
- Pretrain+finetuning
- Zero-shot inference

Readers are encouraged to read our paper for more details about these tasks. Here, we want to highlight that our work is the first to achieve **zero-shot prediction** on the tabular domain.

## Experiments

| Name | Datapoints | Categorical | Binary | Numerical | Positive ratio |
|---|---|---|---|---|---|
| NCT00041119 | 3871 | 5 | 8 | 2 | 0.07 |
| NCT00174655 | 994 | 3 | 31 | 15 | 0.02 |
| NCT00312208 | 1651 | 5 | 12 | 6 | 0.19 |
| NCT00079274 | 2968 | 5 | 8 | 3 | 0.12 |
| NCT00694382 | 1604 | 1 | 29 | 11 | 0.45 |

Table. 1:Five clinical trial patient outcome prediction datasets. Statistics of open data are in the paper.

| Methods | N00041119 | N00174655 | N00312208 | N00079274 | N00694382 | Rank(Std) |
|---|---|---|---|---|---|---|
| LR | 0.6364 | 0.8543 | 0.7382 | 0.7067 | 0.7360 | 5.40(1.14) |
| XGBoost | 0.5937 | 0.5000 | 0.6911 | 0.6784 | 0.7440 | 9.60(3.71) |
| MLP | 0.6340 | 0.6189 | 0.7427 | 0.6967 | 0.7063 | 8.00(2.83) |
| SNN | 0.6335 | 0.9130 | 0.7469 | 0.6948 | 0.7246 | 5.80(2.39) |
| TabNet | 0.5856 | 0.5401 | 0.6910 | 0.6031 | 0.7113 | 11.40(0.89) |
| DCN | 0.6349 | 0.7577 | 0.7431 | 0.6952 | 0.7458 | 5.60(2.51) |
| AutoInt | 0.6327 | 0.7502 | 0.7479 | 0.6958 | 0.7411 | 6.20(2.59) |
| TabTrans | 0.6187 | 0.9035 | 0.7069 | 0.7178 | 0.7229 | 7.20(3.56) |
| FT-Trans | 0.6372 | 0.9073 | 0.7586 | 0.7090 | 0.7231 | 4.20(2.28) |
| VIME | 0.6397 | 0.8533 | 0.7227 | 0.6790 | 0.7232 | 7.00(3.08) |
| SCARF | 0.6248 | 0.9310 | 0.7267 | 0.7176 | 0.7103 | 6.60(3.91) |
| TransTab | **0.6408** | **0.9428** | **0.7770** | **0.7281** | **0.7648** | **1.00(0.00)** |

Table. 2:Performance of supervised learning (the most naïve setting).

| Methods | N00041119 set1 | N00041119 set2 | N00174655 set1 | N00174655 set2 | N00312208 set1 | N00312208 set2 | N00079274 set1 | N00079274 set2 | N00694382 set1 | N00694382 set2 | Rank(Std) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LR | 0.625 | 0.647 | 0.789 | 0.819 | 0.701 | 0.735 | 0.635 | 0.685 | 0.675 | 0.763 | 5.33(1.73) |
| XGBoost | 0.638 | 0.575 | 0.574 | **0.886** | 0.690 | 0.700 | 0.596 | 0.647 | 0.592 | 0.677 | 7.56(3.75) |
| MLP | 0.639 | 0.621 | 0.314 | 0.857 | 0.683 | 0.744 | 0.620 | 0.675 | 0.648 | 0.765 | 6.56(3.32) |
| SNN | 0.627 | 0.634 | 0.215 | 0.754 | 0.687 | 0.732 | 0.631 | 0.683 | 0.651 | 0.759 | 7.44(2.07) |
| TabNet | 0.564 | 0.558 | 0.856 | 0.592 | 0.671 | 0.657 | 0.443 | 0.605 | 0.581 | 0.677 | 10.67(2.96) |
| DCN | 0.636 | 0.625 | 0.767 | 0.790 | 0.711 | 0.698 | 0.682 | 0.664 | 0.658 | 0.737 | 6.33(2.45) |
| AutoInt | 0.629 | 0.630 | 0.843 | 0.730 | 0.725 | 0.698 | 0.679 | 0.665 | 0.686 | 0.661 | 5.89(2.89) |
| TabTrans | 0.616 | 0.647 | 0.866 | 0.822 | 0.675 | 0.677 | 0.618 | 0.702 | 0.652 | 0.718 | 6.22(3.38) |
| FT-Trans | 0.627 | 0.641 | 0.836 | 0.858 | 0.720 | 0.741 | **0.692** | 0.692 | 0.652 | 0.740 | 4.22(2.28) |
| VIME | 0.603 | 0.625 | 0.312 | 0.726 | 0.601 | 0.642 | 0.477 | 0.668 | 0.614 | 0.715 | 10.44(1.51) |
| SCARF | 0.635 | **0.657** | 0.651 | 0.814 | 0.653 | 0.686 | 0.682 | 0.701 | 0.671 | **0.776** | 5.56(3.40) |
| TransTab | **0.653** | 0.653 | **0.904** | 0.846 | **0.730** | **0.756** | 0.680 | **0.711** | **0.747** | 0.774 | **1.78(1.30)** |

Table. 2:Performance of transfer learning within the same data source.

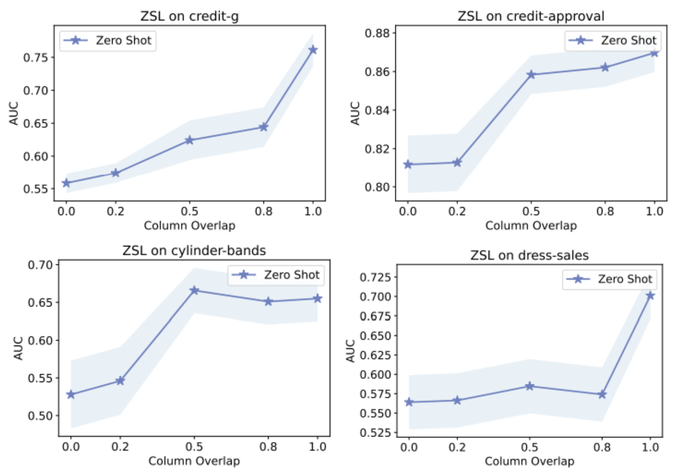


Fig. 3: Performance of zero-shot prediction when column overlapping ratio varies.

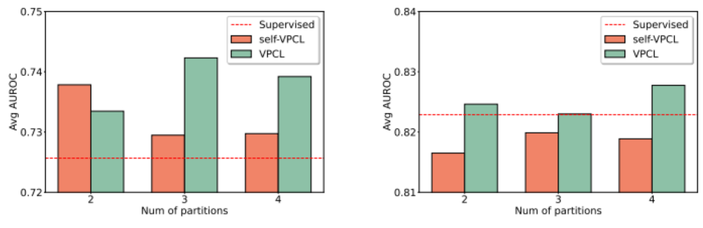


Fig. 4: Performance of the average gain by pretraining.

## References

[1] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171-4186).

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020, September). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.