

Relational Table Learning

Teacher: Zheng Wang Social Network Analysis (NIS8023) Shanghai Jiao Tong University



Outline

Relational Table Learning (RTL)
Generalized RTL

Discussion

Relational/Graph Data are everywhere



Recall that "relational links" exist in Relational DBs



Relational databases

Relational databases domain the world data



Ranking scores per category in percent, February 2025



We need to pay great attention to **Relational Table Learning (RTL)**.



RTL is a field of AI that focuses on enabling computers to understand the content across multiple tables linked by primary and foreign keys.



1. Li, W., Huang, X., Zheng, J., Wang, Z., Wang, C., Pan, L., & Li, J. (2024). rLLM: Relational table learning with LLMs. *arXiv preprint arXiv:2407.20157*.





- How to jointly model multiple tables and their relationships?
 - Table data are the data stored in each table.
 - Non-table data the relationships by primary and foreign keys.

Can machine learning methods do this totally automatically?

Observation on RTL



- A relational database is a graph
 - Nodes: rows in various tables
 - Links: relationships presented by primary and foreign keys



Relational databases

A very simple example: movie rate prediction

IMDb IMDb IMC IMC

- On IMDb, users rate movies, with data stored in three key tables:
 - User Table: Stores user details (e.g., age, gender).
 - Movie Table: Contains movie details (e.g., actors, directors).
 - Rating Table: Records user ratings as (userid, movieid, score) tuples.



How to conduct RTL



- The existing solution is very straightforward:
 - Adopting Table Neural Networks to model table data
 - Adopting Graph Neural Networks to model their relationships

We think more **seamless** RTL type methods will be proposed soon.



rLLM: Relational Table Learning with LLMs

Weichen Li¹, Xiaotong Huang¹, Jianwu Zheng¹, Zheng Wang¹, Chaokun Wang², Li Pan¹, Jianhua Li¹ ¹ Shanghai Jiao Tong University, China ² Tsinghua University, China {wzheng,panli,lijh888}@sjtu.edu.cn,chaokun@tsinghua.edu.cn

Overview of rLLM





The rLLM Overall Architecture

Project: https://github.com/rllm-project/rllm

The Key of rLLM (relationLLM)





rLLM (relationLLM) is an easy-to-use Pytorch library for Relational Table Learning with LLMs, by performing two key functions:

- Breaks down state-of-the-art GNNs, LLMs, and TNNs as standardized modules.
- Facilitates novel RTL model building in a "combine, align, and co-train" way.

Project page: https://github.com/rllm-project/rllm

Three new relational table datasets with standard classification tasks

- TMLIM is derived from the classical MovieLensIM dataset.
- TLF2K is derived from the classical LastFM2K dataset.
- TACM12K is derived from the ACM heterogeneous graph dataset.

Dataset	Tables [#row/#col]	Relation Tables	Label	Classes	#Train/#Val/#Test
TML1M	users [6,040/5] movies [3,883/11] ratings [1,000,209/4]	ratings: user-movie	Age range of user	7	[140/500/1000]
TLF2K	artists [9,047/10] user_artists [80,009/3] user_friends [12,717/3]	user_artists: user-artist user_friends: user-user	Genre of artist	11	[220/500/1000]
TACM12K	papers [12,499/5] authors [17,431/3] citations [30,789/2] writings [37,055/2]	citations: paper-paper writings: paper-author	Conference of paper	14	[280/500/1000]

An illustration RTL method - BRIDGE





Relational Tabular-data

Table Encoder

Graph Encoder

Figure 4: The architecture of BRIDGE

BRIDGE utilizes TNNs to process table data and leverages the "foreign keys" in relational tables to construct relationships between table samples, which are then analyzed using GNNs.

Pseudo-code of BRIDGE



from rllm.nn.conv.graph_conv import GCNConv
from rllm.nn.conv.table_conv import TabTransformerConv

```
# Define the encoders
```

```
g_encoder = GraphEncoder(GCNConv, ...)
t_encoder = TableEncoder(TabTransformerConv, ...)
```

```
# Define the Bridge class
class Bridge:
    def __init__(self, t_encoder, g_encoder):
        self.t_encoder = t_encoder
        self.g_encoder = g_encoder
    def forward(self, table, non_table, adj):
        t_embeds = self.t_encoder(table)
        node_feats = COMBINE(t_embeds, non_table)
```

```
return self.g_encoder(node_feats, adj)
```

In practical implementations, the code lines are around 40. Without rLLM, more than 400+ lines are needed!

Table 2: Classification accuracy.

Methods\Datasets	TML1M	TLF2K	TACM12K
Random	0.144 ± 0.01	0.091 ± 0.03	0.075 ± 0.00
TabTransformer	0.347 ± 0.02	$0.137 {\pm} 0.08$	$0.142 {\pm} 0.01$
TabNet	0.259 ± 0.08	0.135 ± 0.03	0.120 ± 0.02
FT-Transformer	0.352 ± 0.02	0.132 ± 0.01	0.128 ± 0.01
BRIDGE	$0.428 {\pm} 0.02$	$0.454{\pm}0.01$	$0.309{\pm}0.02$

How to try



cd ./examples

set parameters if necessary

python bridge/bridge_tml1m.py
python bridge/bridge_tlf2k.py
python bridge/bridge_tacm12k.py

Project page: https://github.com/rllm-project/rllm

s relationIIm.readthedocs.io/en/latest/

rLLM

Q Search

INTRODUCTION Graph Data Handle Table Data Handle LLM Data Handle TUTORIALS Understanding Transform Understanding Convolution Design of GNNs Design of TNNs Design of RTLs Design of LLM Methods

rLLM Documentation

rLLM (relationLLM) is an easy-to-use Pytorch library for Relational Table Learning with LLMs, by performing two key functions:

Breaks down state-of-the-art GNNs, LLMs, and TNNs as standardized modules.
 Facilitates novel model building in a "combine, align, and co-train" way.

		Models		
Cor	mbine	Align	Co-Train	
1		Ť	Ť	
		Modules		
G	NNs	LLMs	TNNs	
Grap	hConv	Prediction	TableConv	
Graph	[ransform]	Enhancement	TableTransform	
		†	t t	
		Data Engine		
Data Lo	oader	Graph Builder	Table Marker	

The rLLM Overall Architecture



Outline

Relational Table Learning (RTL)
 Generalized RTL

Discussion

Data Lake(house): a central hub for machine learning





Question: how to conduct machine learning (maybe RTL) in Data Lake(house)?

Image from https://helicaltech.com/services/data-lake-services/

Preliminary: the data stores in data lake(house)



Open-source file formats, like Parquet and ORC



Question: can we convert columnar data to table data? Answer: Yes, but only obtain very sparse tables.

Apache Parquet is a binary, efficient columnar data format.

How to prepare data in data lake(house)



Data discover (i.e., find their relationships) is one of the most important tasks, including Not Joinable

- Joinable Tables
- Unionable Tables
- Subsetable Tables



Table discovery in data lake(house)

Deng, Yuhao, et al. "LakeBench: A Benchmark for Discovering Joinable and Unionable Tables in Data Lakes." *Proceedings of the VLDB Endowment* 17.8 (2024): 1925-1938. Srinivas, Kavitha, et al. "Lakebench: Benchmarks for data discovery over data lakes." arXiv preprint arXiv:2307.04217 (2023).

Machine Learning in a Data Lakehouse: Possible Steps

- Machine learning in data lake(house), possible steps:
 - Define a Task: Identify the target table and define the goal.
 - Discover Related Tables: Find other task related tables.
 - Model Learning: Use the combined data for AI training and get insights.

Generalized Relational Table Learning (GRTL)

- Compared to RTL, GRTL further extends the model's ability to capture more general relationships between tables, including:
 - Joins on different keys (e.g., <pri, for>, <pri, pri>, <anycol, anycol>)
 - Union (combining result sets from two or more SELECT statements)
 - Subset (some tables are subsets of others)

LakeMLB (Data Lake Machine Learning Benchmark) Github: <u>https://github.com/zhengwang100/LakeMLB</u>

Outline

Relational Table Learning (RTL)
Generalized RTL
Discussion

Applications of RTL methods: Self-driving car

- Sensors: cameras, LiDAR, radar, audio, and ultrasound, GPS, and inertial measurement.
- Data source: single/multiple sources
- Data platform: RDB, Data Lake(house), ...
- Method: Deep Learning -> Large Model

Question:

I. Should the processed sensor data be stored in tables?2. Can RTL aid in developing self-driving models?

Berkeley Deep Drive dataset

Applications of RTL methods: robots

Modalities:	unitreerobotics eo	/G1_CameraF	ackaging_l	Dataset 🗅	♥ like 3 Follow U	Unitree I
🖗 Dataset card	Wiewer →I≣	Files and versions	<i>(</i> Communit	ÿ		
 Dataset Previe 	ew (j)				🖀 Embed 🖪 Full Scree	n Viewer
Split (1) train			\checkmark			
▶ The full datase	et viewer is not ava	ailable (click to	read why). Only	showing a previ	iew of the rows.	
_data_files list	_fingerprint string	_format_columns null	_format_kwargs dict	_format_type null	_output_all_columns	_split null
[{ "filename": "data-00000-of-	4a577e130d30e2a9	null	{}	null	false	null

Game Developer Conference (GDC) 2024

A dataset of Unitree Robotics

Question:

I. Should the processed sensor data be stored in tables?2. Can RTL aid in developing robot models?

Thanks for your time. QA.