Self-Supervised Representation Learning for Relational Multimodal Data

I. Introduction

Deep Learning models can optimize **pre-training objectives** on an unlabelled dataset to learn a representation of the data which can be followed by fine-tuning for downstream ML tasks. Pre-training has proven significant benefits in other fields, but has not been applied with multi-task learning to relational multimodal tabular data.

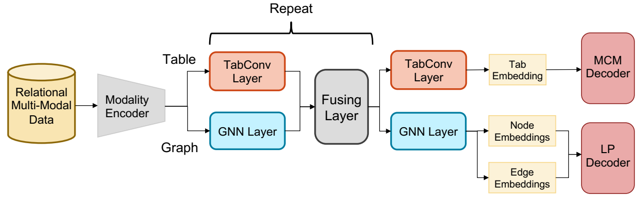

***Fig 1**. The relational multimodal framework provided by the project supervisors used in the experiments.*

II. Research Questions

Main RQ: Can a combination of pre-training objectives improve self-supervised metrics?

SQ1: What is the best data masking strategy?
SQ2: How does combining pre-training objectives affect self-supervised metrics?
SQ3: Can the multi-task algorithm MoCo [1] improve self-supervised metrics?

References

[1] Fernando, H., Shen, H., Liu, M., Chaudhury, S., Murugesan, K., & Chen, T. (2023, May). Mitigating gradient bias in multi-objective learning: A provably convergent approach. International Conference on Learning Representations

[2] I. Rubachev, A. Alekberov, Y. Gorishniy, and A. Babenko, "Revisiting Pretraining Objectives for Tabular Deep Learning." arXiv, Jul. 12, 2022. doi: 10.48550/arXiv.2207.03208.

[3] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 11033–11043. Accessed: May 03, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/hash/7d97667a3e056acab9aaf653807b4a03-Abstract.html

[4] M. Yasunaga *et al.*, "Deep Bidirectional Language-Knowledge Graph Pretraining," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37309–37323, Dec. 2022.

[5] W. Hu *et al.*, "Strategies for Pre-training Graph Neural Networks." arXiv, Feb. 18, 2020. doi: 10.48550/arXiv.1905.12265.

III. Method

Experimental setup

1 — Select pre-training objectives

MCM: Mask Cell Modelling **[2]**.
MV: Mask Vector prediction **[3]**.
LP: Link Prediction **[4]**.
AM: Attribute Masking **[5]**.

2 — Select data mask types

replace: mask by sampling from the column's empirical distribution.
remove: mask with fixed token.
bert: a combination of the two.

3 — Select subsets of objectives

Combine the 4 objectives in 9 different subsets.

Experiments

4 — Experiment 1

Find best mask type by running {MCM,MV,AM} with each of the three mask types on two datasets. The self-supervised metrics of accuracy (ACC) and Root Mean Squared Error (RMSE) (both for MCM and AM) are compared.

5 — Experiment 2

Quantify the effect of combining objectives by comparing the self-supervised metrics of ACC and RMSE (for MCM) and Mean Reciprocal Rank (for LP) on each of the subsets chosen.

6 — Experiment 3

Quantify the effect of using MoCo to combine the loss functions of the objectives. Evaluate by comparing the same self-supervised metrics as experiment 2.

IV. Results

Mask Type	IBM AML					Amazon Fashion				
	ACC↑		RMSE↓		MV ACC↑	ACC↑		RMSE↓		MV ACC↑
	MCM	AM	MCM	AM		MCM	AM	MCM	AM	
replace	**0.8517**	**0.8545**	0.0925	0.0908	0.8341	**1**	**0.9999**	1.5046	1.5042	0.9671
bert	0.8207	0.8240	0.0933	0.0892	0.9065	0.9620	0.9616	**1.5687**	1.5179	0.9295
remove	0.8136	0.8170	0.0927	0.0908	**0.9992**	0.9608	0.9607	1.5203	1.5019	**0.9976**

***Table 1**. Experiment 1 results. The replace mask type dominates the other two for ACC, but not RMSE. Remove reaches the highest MV accuracy as it's trivial to predict the masked value.*

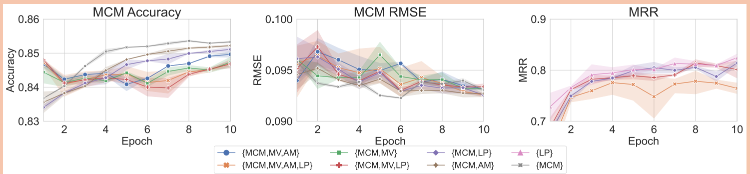

***Fig 2**. Experiment 2 results. The difference in converged metric values is small between subsets but larger subsets tend to perform worse on ACC and MRR.*

Pretext tasks	IBM AML					
	ACC↑		RMSE↓		MRR↑	
	Sum	MoCo	Sum	MoCo	Sum	MoCo
{MCM,AM}	0.8522	**0.8531**	**0.0927**	0.0929	n/a	n/a
{MCM,MV}	0.8474	**0.8530**	**0.0931**	0.0932	n/a	n/a
{MCM,LP}	0.8511	**0.8520**	**0.0926**	0.0937	**0.8139**	0.8032
{MCM,MV,AM}	0.8496	**0.8522**	0.0932	**0.0928**	n/a	n/a
{MCM,MV,LP}	0.8468	**0.8488**	0.0934	**0.0932**	0.8012	**0.8153**
{MCM,MV,AM,LP}	0.8469	**0.8510**	**0.0926**	0.0931	0.7645	**0.7966**
σ	0.0023	0.0016	0.0004	0.0003	0.0256	0.0095
% Change in σ	-30.4%		-25%		-62.9%	

***Table 2**. Experiment 3 results. MoCo improves ACC in all subsets, but not RMSE and MRR. The standard deviation of the converged values between subsets is reduced with MoCo.*

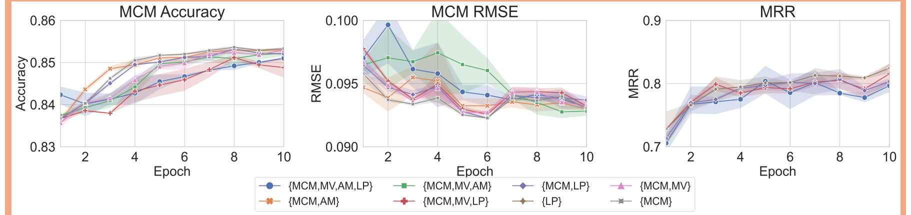

***Fig 3**. Experiment 3 results. The difference in performance is visibly smaller than in Fig 2 for ACC and MRR, indicating MoCo optimizes each objective at little cost to other objectives.*

V. Limitations & Future work

- The quality of the Amazon Fashion dataset needs imporovement and the experiments should be conducted on more datasets.
- Different variations of the "replace" mask type should be investigated.
- More runs are required to establish statistical certainty.
- The combination of objectives should also be evaluatated based on downstream task performance after pre-training.

VI. Conclusions

- **SQ1**: The "replace" mask type performs the best in ACC, but not in RMSE.
- **SQ2**: Combining objectives leads to marginal differences in self-supervised metrics, with larger subsets performing slightly worse.
- **SQ3**: Using MoCo offers two benefits: slightly improved self-supervised metrics and significantly reduced variance between subsets.
- MoCo allows for a more diverse representation to be learned as more pre-training objectives can be used.

Author: **Ilias Mc Auliffe** (i.mcauliffe@student.tudelft.nl)   Responsible Professor: **Kubilay Atasu**   Supervisor: **Atahan Akyıldız**