Sharc Lab @ GT

HW-SW Co-design of Efficient and Scalable Training for Relational Graph Representation Learning

Lakshmi Sathidevi, Cong (Callie) Hao
School of Electrical and Computer Engineering

Georgia Tech | Center for Research into Novel Computing Hierarchies
Georgia Tech | School of Electrical and Computer Engineering, College of Engineering

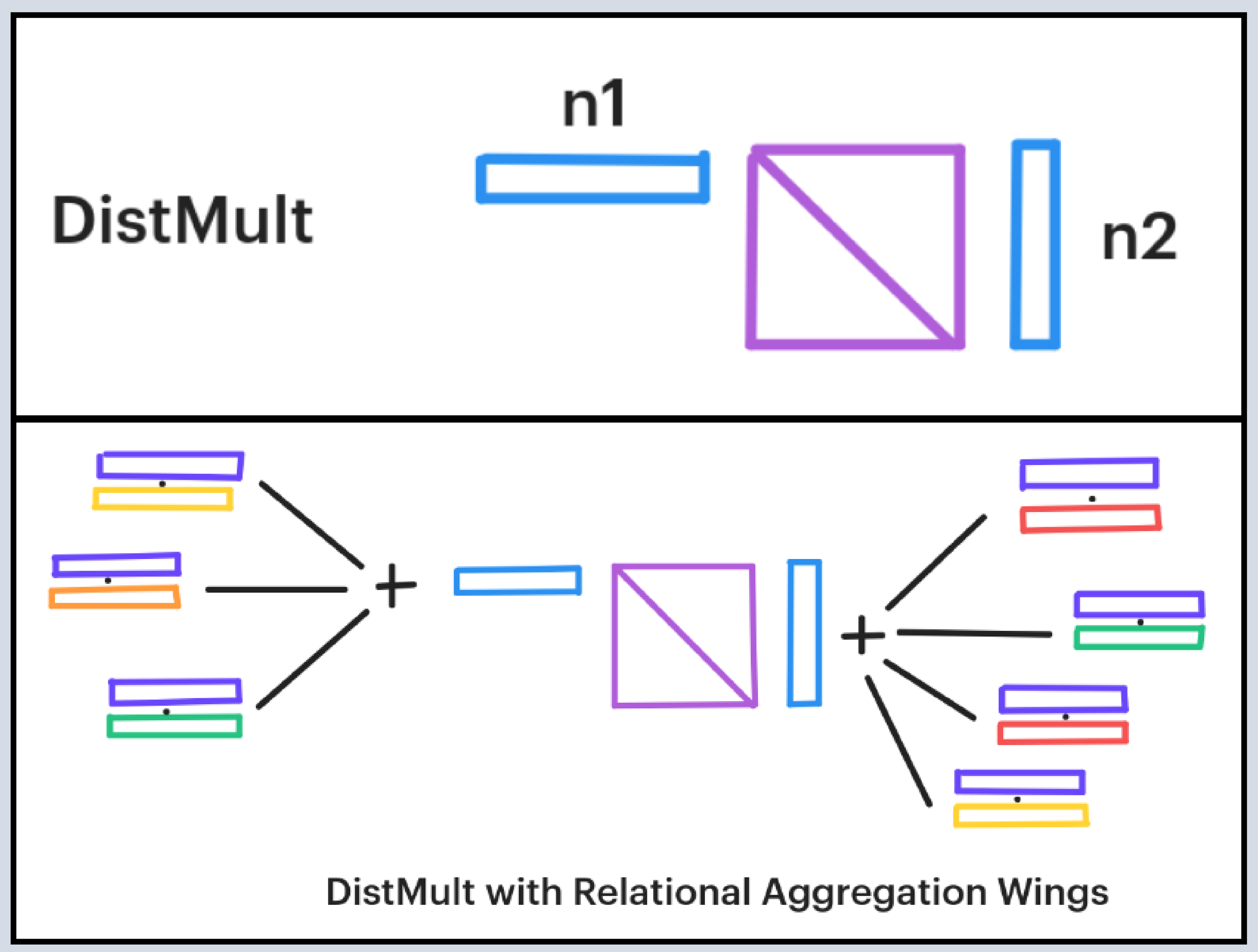
1 Knowledge Representation

Being able to easily store and retrieve knowledge, as well as predict missing information using existing knowledge, is key to enabling a faster pace of R&D across a variety of high-impact domains.

Towards this, **large-scale relational graph representation learning on knowledge graphs** holds great promise.

2 Co-design of a Scalable Model

We propose **DMRAW** (DistMult with Relational Aggregation Wings) — a SW model that **inductively learns on knowledge graphs** via self-supervised relational link prediction.



DistMult with Relational Aggregation Wings

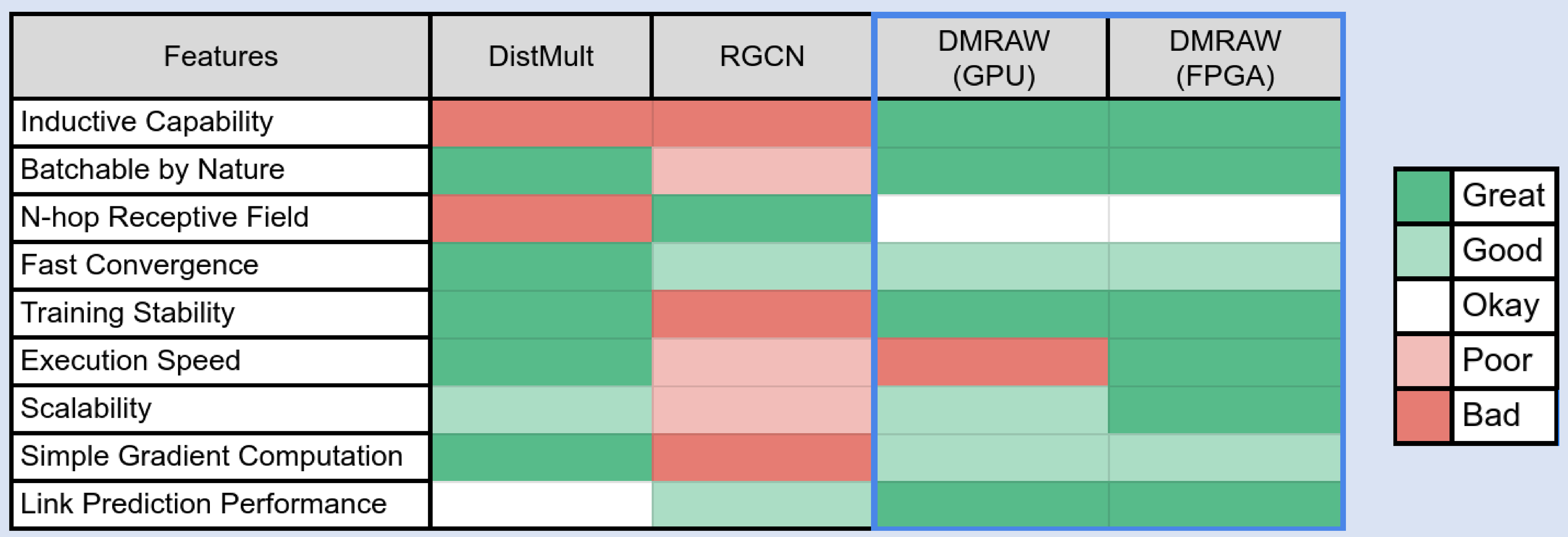
DMRAW is co-designed at the SW level considering HW level opportunities and limitations.

3 DistMult with Relational Aggregation Wings

DMRAW applies 1-hop relational aggregation and learns a linear transformation to generate node embeddings from the aggregate.

Key Features

- **DMRAW is inductive,** unlike **DistMult** and **RGCN** models.
- **DMRAW is stable** during training, unlike **RGCN.**
- **DMRAW** has **1-hop** receptive field, unlike **DistMult**.
- **DMRAW** is by default **batchable and scalable**, unlike **RGCN**.
- **DMRAW is linear**, unlike **RGCN**.
- **DMRAW** is **parameter efficient** than **RGCN**.



DMRAW's unique design makes it **great for custom HW acceleration** on FPGA.

4 Relational Link Prediction Performance

Test MRR Score	tiny	small	small-moderate	moderate	large
	Nations	UMLS	codex-s	WN18RR	FB15k-237
DistMult	0.773	0.685	0.305	0.393	0.181
RGCN	0.769	0.755	0.327	**0.502**	**0.233**
DMRAW	**0.803**	**0.778**	**0.488**	0.434	0.228

DMRAW performed on average 8.2% better than **RGCN.**
DMRAW performed on average 22% better than **DistMult.**
Despite being **inductive, linear, and having just 1-hop receptive field.**

5 Why Accelerate Relational Graph Representation Learning?

Works on very large datasets in domains like genomics and drug discovery, are seen to fall back to using older, simpler heuristics, to avoid the computational limitations of current relational representation learning schemes. This trend extends to various other high-impact domains.

6 HW Acceleration Opportunity

Execution Time (s) (160 Epochs)	small-moderate	moderate
	codex-s	WN18RR
GPU	1600	1600
FPGA	383	924
Minimum Expected Speed-Up	4.2x	1.7x

HW architecture latency simulation indicates even a sparingly optimized custom HW architecture for DMRAW can give **significant speed-up on FPGA over GPU**, thanks to DMRAW's suitability for custom acceleration.

7 Future Scope

Our results for our co-designed SW model show promise for efficient, stable, inductive and scalable training on large-scale knowledge graphs by coupling with an optimized custom hardware acceleration architecture.

Lakshmi S: lsathidevi3@gatech.edu