2024 IEEE 32nd Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) | 979-8-3503-7243-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/FCCM60383.2024.00040

High-Performance Reconfigurable Accelerator for Knowledge Graph Reasoning

Hanning Chen*, Ali Zakeri*, Yang Ni*, Fei Wen†, Behnam Khaleghi‡, Hugo Latapie§, and Mohsen Imani*
*University of California, Irvine, †Texas A&M University, ‡Qualcomm, §Cisco Systems
Email: *{hanningc, m.imani}@uci.edu

***Abstract*—In recent times, a plethora of hardware accelerators has emerged, catering to graph learning applications. However, the focus has primarily been on accelerating graph analysis, graph clustering, and graph mining, with a lack of attention to knowledge graph reasoning. Graph reasoning requires a more complex model to handle the complicated knowledge graph compared to other graph learning tasks. A primary knowledge graph reasoning task is to find the implicit relations between entities of a given knowledge graph, which demands a significantly longer training time than traditional graph learning algorithms due to the model complexity. Therefore, it is essential to develop an acceleration method to mitigate the training cost for the practical deployment of this task. Prior work in this field has solely considered using a single GPU or distributed GPU cluster to accelerate translational embedding models. However, as demonstrated in this paper, such general-purpose GPUs don't provide satisfactory results for more complex reinforcement learning-based models. Hence, it becomes necessary to design customized domain-specific accelerators. This work proposes GraFlex, the first domain-specific accelerator for reinforcement learning-based knowledge graph reasoning, implemented on FPGA. We first develop a compression method for knowledge graphs. Then, we explore FPGAs of different sizes, analyze their on-chip resources, and suggest a mechanism to achieve high-speed training on devices with insufficient resources using the aforementioned compression method.**

I. Introduction and GraFlex Architecture

Knowledge graphs (KGs) have found wide applications in assorted artificial intelligence (AI) tasks, facilitating the organization and storage of large amounts of existing knowledge. KGs are multi-relational graphs, comprising entities and relations represented as nodes and edges in the graph, respectively. Figure 1 is an architecture overview of GraFlex implemented with Xilinx Alveo platform. The input knowledge graph is first processed on host CPU. After the preprocessing step, each entity will be represented by an embedding vector of dimension M, with elements quantized to 16-bit fixed-point numbers. Embedding vectors are then loaded into FPGA and stored in either the off-chip DRAM or the off-chip high bandwidth memory (HBM). To maximize memory bandwidth utilization, we partition the graph and distribute entity word vectors over multiple HBM pseudo channels (PCs), with each channel keeping L entity words. Beside loading entity embedding vectors into FPGA, we also store the KG triples in the on-chip storage such as UltraRAM.

II. Experimental Result

We implement and run previous KGR work, DeepPath [1], on both CPU (Intel Xeon Silver 4114) and GPU (NVIDIA GTX 1660 Ti and RTX 3090) as hardware acceleration baseline. We synthesize and compile GraFlex on two different classes of Xilinx FPGAs: Alveo U280 (U280) and Zynq UltraScale+ ZCU104 (ZCU104). To test our agent's reasoning capability, we choose the two most common knowledge graph datasets, FB15K-237 [2] and NELL-995 [3]. Our evaluations on extensive reasoning tasks show that GraFlex on Xilinx Alveo U280 achieves a 65× speedup over CPU (Intel Xeon 4114) and an approximately 8× speedup compared to GPU (RTX 3090). Regarding energy efficiency, our design on Alveo U280 shows over 30× improvement in comparison to CPU and GPU. On the Xilinx ZCU 104, GraFlex achieves over 3× speedup and 14× energy efficiency improvement over over GTX 1660 Ti.

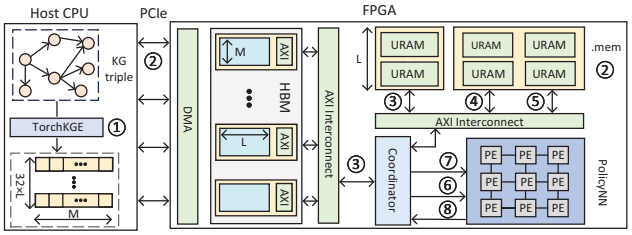

Fig. 1. GraFlex acceleration architecture. 1. Entity and relation word embedding. 2. Transfer of embedding vector from CPU to FPGA via PCIe and Xilinx DMA. 3. Entity embedding vectors preload. 4. Address access. 5. Outgoing relations access. 6. Teacher's guide for agent 7. Provision of state, reward, and loss to RL agent. 8. Agent's action probability distribution.

Acknowledgment

This work was supported in part by the DARPA Young Faculty Award, the National Science Foundation (NSF) under Grants #2127780, #2319198, #2321840, #2312517, and #2235472, the Semiconductor Research Corporation (SRC), the Office of Naval Research through the Young Investigator Program Award, and Grants #N00014-21-1-2225 and N00014-22-1-2067. Additionally, support was provided by the Air Force Office of Scientific Research under Award #FA9550-22-1-0253, along with generous gifts from Xilinx and Cisco.

References

[1] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *AAAI 2014*, vol. 28, 2014.
[2] K. Toutanova *et al.*, "Representing text for joint embedding of text and knowledge bases," in *EMNLP 2015*, pp. 1499–1509, 2015.
[3] H. Wang *et al.*, "Incorporating graph attention mechanism into knowledge graph reasoning based on deep reinforcement learning," in *EMNLP-IJCNLP 2019*, pp. 2623–2631, 2019.

2576-2621/24/$31.00 ©2024 IEEE
DOI 10.1109/FCCM60383.2024.00040