ARIA
APPLIED RESEARCH IN ACTION

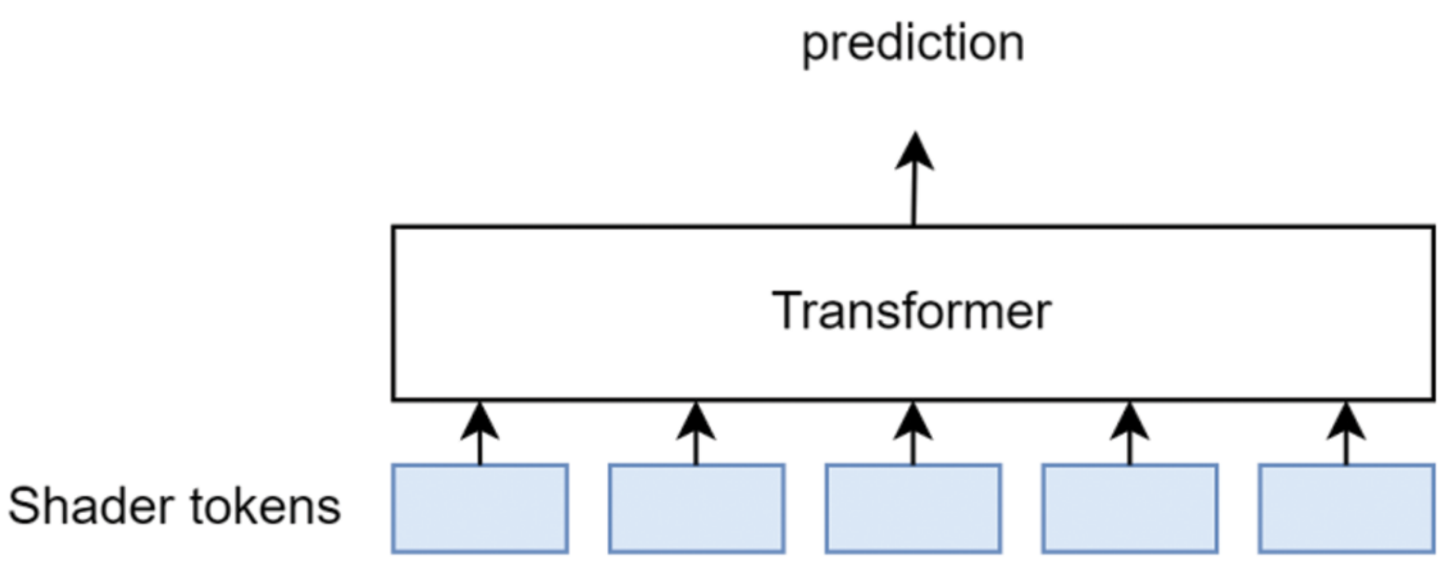
Machine Learning for Pattern-Matching Optimization in GPU Compilers

Advancing GPU performance through AI-Driven compiler optimizations

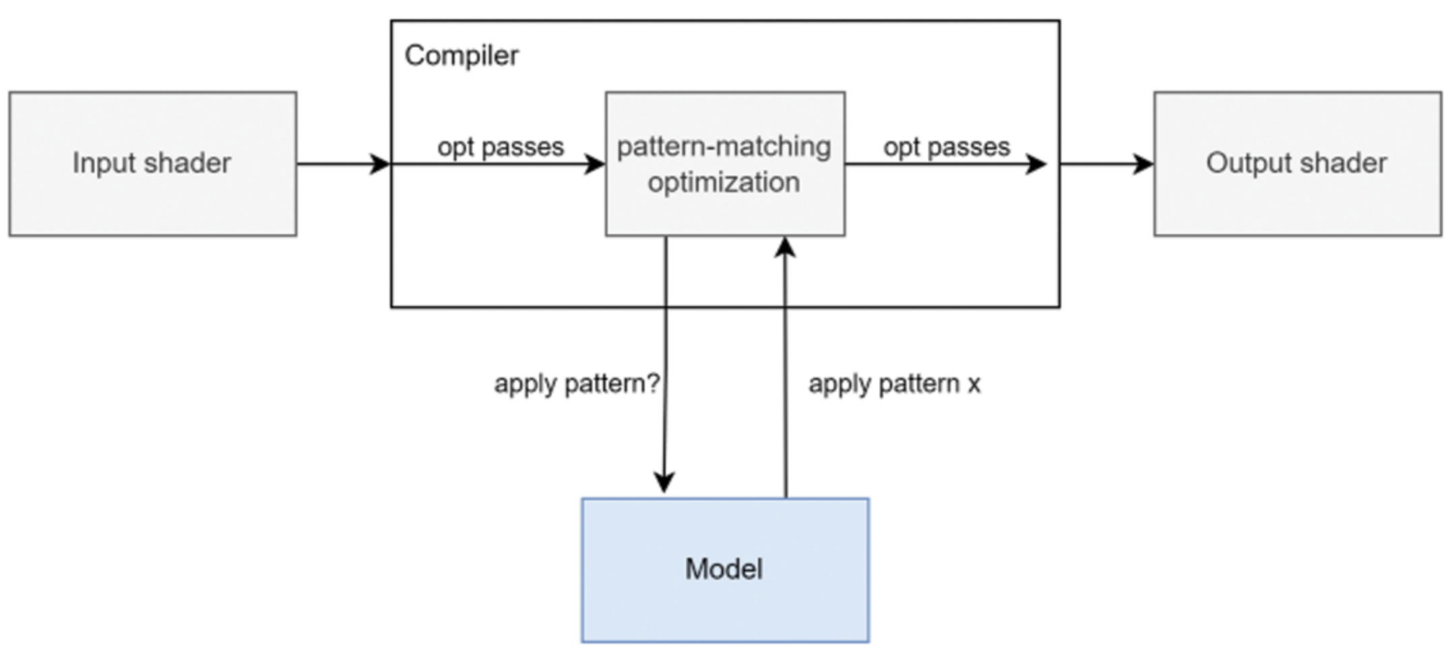
Zichuan Guan

Nandita Vijaykumar
ACADEMIC SUPERVISOR

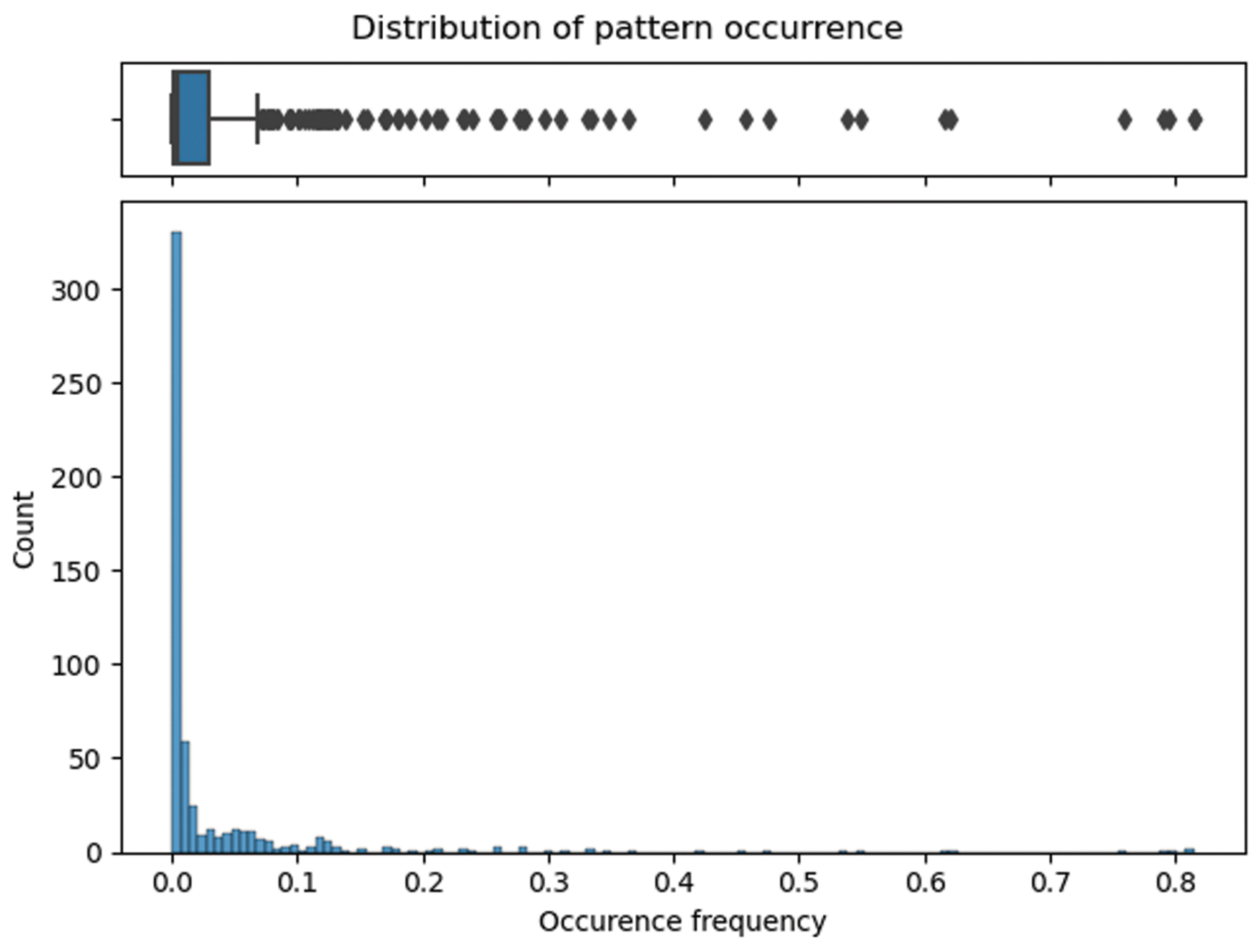
Olga Golovanevsky
INDUSTRY SUPERVISOR



(a) NLP approach



(b) RL approach



PROJECT SUMMARY

Compiler optimizations play a pivotal role in improving program efficiency and reducing code size, particularly in the domain of GPU computing, where harnessing parallel processing capabilities is critical. While machine learning techniques have been widely applied to various optimization problems, GPU compilers, which employ a range of proprietary optimizations targeted to specific hardware architectures, continue to rely heavily on expert-engineered heuristics. In this study, we focus on enhancing an instruction-based pattern-matching optimization technique designed for AMD platforms. We explore two distinct machine learning approaches and analyze the usage and dynamics of different hand-crafted optimization patterns. First, we adopt a novel approach from a natural language processing (NLP) perspective using an end-to-end transformer-based model. Then, we also explore a reinforcement learning (RL) approach by reframing the problem as a Markov decision problem (MDP) and conduct a comparative analysis of the two approaches. Furthermore, we investigate the challenges and limitations associated with integrating machine learning into a rapidly evolving production GPU compiler, highlighting challenges such as data availability, model generalization, and real-time compilation constraints.

AMD

Computer Science
UNIVERSITY OF TORONTO

Master of Science in
Applied Computing