ARIA
APPLIED RESEARCH IN ACTION

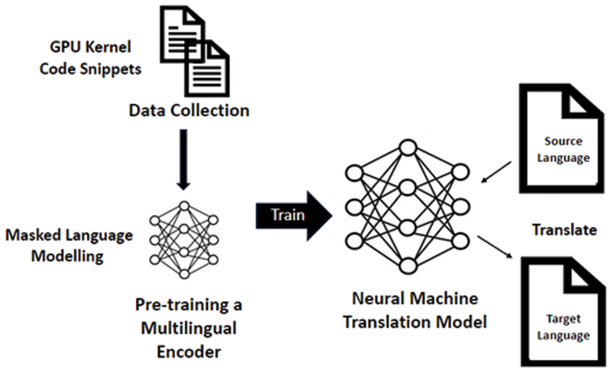
Learning GPU Code Structure Using Large Language Models

Revolutionizing GPU Kernel Optimization: Harnessing Transformer-Based Models to Improve Performance and Efficiency

Sanjukta De

Maryam Mehri Dehnavi
ACADEMIC SUPERVISOR

Abhinav Vishnu
INDUSTRY SUPERVISOR



PROJECT SUMMARY

In the dynamic landscape of modern machine learning and computational efficiency, optimizing GPU kernels presents a complex challenge. These kernels constitute the foundational elements of deep learning frameworks, driving groundbreaking advancements across diverse domains. This project delves into the intricate realm of GPU kernel optimization, exploring strategies to address their complexity and resource-intensive nature.

Previous research has largely utilized deep learning models to decipher the performance attributes and underlying structures of GPU kernels. The research done in this project builds upon these foundations by employing transformer-based large language models to comprehend the intricate architecture of GPU kernels. Leveraging the successes of these models in various applications, a multilingual encoder is trained on a large corpus of GPU kernel code using masked language modelling. The final model uses this encoder and finetunes it to perform downstream tasks like GPU code generation and translation.

AMD

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing