Yao Jianyu

Residence
Beijing

Email
yaojianyu89@gmail.com

My name is Yaojianyu.

Education

Bachelors degree in Software Engineering. Shandong University [2015 - 2019]

Masters degree. Institute of Computing Technology, Chinese Academy of Sciences [2019 - now]

Professional Experience

Small GEMM

November 2020 - May 2021

GEMM with the small size of input matrices is becoming widely used in many fields like HPC and machine learning. Although many famous BLAS libraries already supported small GEMM, they cannot achieve near-optimal performance. This is because the costs of pack operations are high and frequent boundary processing cannot be neglected. This paper proposes an input-aware adaptive tuning framework(IAAT) for small GEMM to overcome the performance bottlenecks in state-of-the-art implementations. IAAT consists of two stages, the install-time stage and the run-time stage. In the run-time stage, IAAT tiles matrices into blocks to alleviate boundary processing. This stage utilizes an input-aware adaptive tile algorithm and plays the role of runtime tuning. In the install-time stage, IAAT auto-generates hundreds of kernels of different sizes to remove pack operations. Finally, IAAT finishes the computation of small GEMM by invoking different kernels, which corresponds to the size of blocks. The experimental results show that IAAT gains better performance than other BLAS libraries on ARMv8 platform.

C Assembly Makefile

Small GEMM for cross platform

June 2021 - now

GEMM with the small size of input matrices is becoming widely used in many fields like HPC and machine learning. Although many famous BLAS libraries already supported small GEMM, they cannot achieve near-optimal performance. This is because the costs of pack operations are high and frequent boundary processing cannot be neglected. This paper proposes an input-aware adaptive tuning framework(IAAT) for small GEMM to overcome the performance bottlenecks in state-of-the-art implementations. IAAT consists of two stages, the install-time stage and the run-time stage. In the run-time stage, IAAT tiles matrices into blocks to alleviate boundary processing. This stage utilizes an input-aware adaptive tile algorithm and plays the role of runtime tuning. In the install-time stage, IAAT auto-generates hundreds of kernels of different sizes to remove pack operations. Finally, IAAT finishes the computation of small GEMM by invoking different kernels, which corresponds to the size of blocks. The experimental results show that IAAT gains better performance than other BLAS libraries on ARMv8 platform.

C Assembly

OpenVML:

August 2020 - April 2021

Vector Math Library

C Assembly SIMD

C	++++	C++	++++	Assembly	+++	SIMD	++++	CUDA	+++	MPI	+++
pthread	+++	Makefile	+++	CMake	+++	Latex	++++	Linux	++++	Git	++++
Python	+++	Java	+++	JavaFX	++++	Android	++++	SQL	++		