摘要

大语言模型（LLM）在多个领域的成功应用推动了人工智能发展，但其存在计算成本高、内存需求大等缺点，优化其推理效率、降低资源消耗成核心问题。主流大模型多为 Decoder-Only 架构，其中矩阵向量乘算子（GEMV）在大模型推理中占主导地位。GEMV 因自身特点为内存瓶颈任务，GPU 等计算密集型硬件利用率低，在边端场景下常成系统瓶颈。近存计算是新兴计算范式，能解决传统计算架构内存瓶颈问题。新兴的近存计算硬件，有高带宽、高存储、高并行性、高能效等优势，也存在计算能力弱、通信开销大等局限。如何从软件角度适配硬件加速，以及如何修改设计硬件优化其应用性能的软硬协同优化方案成为待深入研究的课题。

本论文基于近存计算技术研究算子 GEMV 的软硬协同加速方案：1）在近存计算硬件上，设计基于查找表的矩阵向量乘算法，使用查找表消除复杂计算，对近存计算硬件中的多级存储结构分别做出了访存优化，包括查找表分块算法减少通过 DMA 访问内存的次数，以及矩阵的行列重排减少对缓存的访问，增强寄存器的数据局部性；2）在模拟器上，为解决真实硬件上软件算法的计算瓶颈，基于周期精确近存激素模拟器设计硬件及指令以增强其计算能力，包括融合查表和加法指令用于高效快速查询乘积并累加，以及设计基于查找表的向量指令进行向量化的查表和访存，提高计算效率。最后对上述算法进行了详细的测试，分别在 CPU 和 GPU 以及近存硬件三个硬件平台上，进行了包括对 GEMV 算子总吞吐和能效比的测试。实验结果表明，在近存计算硬件上 GEMV 算子吞吐为 355GOPS，大概达到理论性能的 93%，是 CPU 平台的 14.7 倍，能效比大概是 CPU 平台的 8.6 倍，是 GPU 平台的 1.13 倍。

关键词：近数计算　矩阵向量乘　查找表

Abstract

The successful application of Large Language Models (LLMs) in multiple domains has propelled the advancement of artificial intelligence. Nevertheless, they are plagued by drawbacks such as high computational costs and substantial memory demands. Optimizing the inference efficiency and reducing resource consumption have emerged as core issues. Mainstream LLMs adopt the Decoder-Only architecture, within which the Generalized Matrix-Vector Multiplication (GEMV) holds a dominant position in LLM inferences. Due to its inherent characteristics, GEMV constitutes a memory bounded task, with the utilization of computing-intensive hardware like GPUs being low. It has always become a system bottleneck in edge scenarios. Processing-in-Memory (PIM) represents an emerging computing paradigm capable of resolving the memory bounded problem in traditional computing architectures. Emerging PIM hardware boasts advantages such as high bandwidth, high storage capacity, high parallelism, and high energy efficiency, yet it also suffers from limitations including weak computing capabilities and significant communication overhead. How to adapt hardware acceleration from a software perspective and how to modify the design of hardware to optimize its application performance through a hardware-software co-design solution have become subjects warranting in-depth research.

This thesis investigates the hardware-software co-design solution for GEMV based on Processing-in-Memory technology: 1) On the PIM hardware, a matrix-vector multiplication algorithm based on lookup tables is devised. Lookup tables are employed to eliminate complex computations, and memory access optimizations are implemented for the multi-level storage structure within the PIM hardware. This encompasses the lookup table blocking algorithm to reduce the number of memory accesses via DMA and the rearrangement of matrix rows and columns to minimize cache access and en-

hance data locality in registers. 2) On the simulator, to address the computational bottlenecks of software algorithms on real hardware, special processing unit and instructions are designed based on the cycle-accurate PIM simulator to enhance its computing capacity. This includes the fusion of table lookup and addition instructions for efficient and rapid query of products and accumulation, as well as the design of vector instructions based on lookup tables for vectorized lookup and memory access to boost computing efficiency. Finally, detailed tests were carried out on the aforementioned algorithms on three hardware platforms, namely CPU, GPU, and PIM hardware, including tests on the total throughput and energy efficiency of the GEMV operator. The experimental results indicate that the throughput of the GEMV operator on the PIM hardware amounts to 355 GOPS, approximately reaching 93% of the theoretical performance. It is 14.7 times that of the CPU platform, and the energy efficiency is approximately 8.6 times that of the CPU platform and 1.13 times that of the GPU platform.

Key Words : Processing-in-Memory GEMV Lookup-Table

目录

图目录

表目录