عنوان پایاننامه

استفاده از سخت افزارهای شتاب دهنده برای بهبود کارآیی معماری در پردازش در حافظه

چکیده پایان نامه

The performance ofmodern microprocessors is mainly limited by power and off-chip bandwidth walls. The emerging process-in-memory architectures present a unique opportunity to reduce both power usage and data movement overheads by moving computation closer to memory. This is achieved by stacking a logic layer on top of one or multiple memory layers in a ٣D fashion. Current process-in-memory proposals leverage the logic layer to build varying processing units, ranging from application-specific accelerators to general-purpose cores. In this paper, we propose a new processing-in-memory architecture that uses a neural network as the memory-side general-purpose accelerator. This architecture is mainly motivated by the observation that in many real-world applications, some program regions, or even the entire program, can be replaced by a neural network that is learned to mimic the function of the region. This architecture benefits from both flexibility of general-purpose processors and superior performance of application-specific accelerators. Experimental results show that this method provides up to ٢.٣x speedup over a processor-side neural network accelerator and up to ١.٤X speedup over a memory-side general-purpose core.

کلمات کلیدی

پردازش در حافظه
شبکه عصبی

کلمات کلیدی انگلیسی

Processing-in-memory
Neural network
Hardware acceleration