پژوهشگاه دانش‌های بنیادی
پژوهشکده علوم کامپیوتر

سخنرانی علمی

Challenges in Accelerating DNNs

هاجر فلاحتی

پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)

Abstract

Machine learning (ML) algorithms, with a special focus on deep neural networks (DNNs), have become integral to a myriad of applications. However, the execution of DNNs poses challenges such as handling a vast number of parameters, leading to increased computation, memory accesses, on-chip storage demands, and requirements for off-chip and on-chip memory bandwidth. A promising platform for accelerating ML algorithms is Graphics Processing Units (GPUs). However, executing DNNs on GPUs presents notable challenges. Despite benefiting from memory hierarchy and high thread-level parallelism (TLP), GPUs still face challenges related to memory inefficiency, resulting in suboptimal performance.

Prior research shows that prefetching is an effective technique for addressing memory inefficiency in GPUs. However, existing methods only rely on fixed strides. In this talk, I elaborate on challenges in DNN execution. Then, I introduce a novel prefetcher, called Snake, which is built upon chains of variable strides, using throttling and memory decoupling strategies. Snake prefetches 80% of memory requests with 90% accurately prefetched requests, improves GPU performance by 17%, and reduces energy consumption by 17% in memory-bound GPGPU applications.

Biography

Hajar Falahati is a Senior Postdoctoral Researcher at the School of Computer Science, Institute for Research in Fundamental Sciences (IPM). She received her BS.c from Isfahan University of Technology in 2009. She received her PhD and MS.c from the Sharif University of Technology in 2016 and 2011, respectively. She had work experience in driver development, data analysis, and ML model development. Her research interests include machine learning, AI accelerators, hardware accelerators, GPU architecture and applications, near-data processing, and bioinformatics.

زمان : سه‌شنبه ۱۴۰۲/۱۱/۲۴ – ساعت ۱۵:۰۰

ارائه به صورت مجازی انجام خواهد شد.

https://vmeeting2.ipm.ir/b/com-hh1-n07-vil

***** شرکت برای عموم علاقه‌مندان آزاد است*****