

Energy Efficient Accelerator for Machine Learning

هاجر فلاحتی

پژوهشگاه دانش‌های بنیادی

Abstract

Machine Learning algorithms, especially neural networks (NNs) as the most promising ones, are widely used in a myriad of applications and moving towards larger and deeper structures (DNNs). Executing these overparameterized DNN architectures imposes high computations and memory demands (i.e., frequent data movements between the memory and compute units, high memory storage, and off-chip/on-chip memory bandwidth), which make performing DNN tasks on resource-constrained edge devices challenging. Although ML algorithms are usually executed on general purpose platforms, such as CPUs and GPUs, these platforms are expensive in terms of consumed energy. Recently, there have been several attempts in both academic and industry to design special-purpose hardware accelerators. To reduce both memory and computation demands in DNN accelerators, some prior research develops compression techniques. However, the compression techniques make DNNs irregular.

In this talk, I will briefly introduce the challenges in processing DNNs and compression techniques, and present in detail a compression technique which exploits input similarity, and paves the way to exploit from potential behind similarity and sparsity in both weights and inputs effectively, while mitigating the irregularity overheads.

Biography

Hajar Falahati received her B.Sc. degree in Computer Engineering from Isfahan University of Technology, in 2009, and her M.Sc. and Ph.D. in Computer Architecture from Sharif University of Technology, in 2011 and 2016, respectively. Now, she is a postdoctoral researcher with the Institute for Research in Fundamental Sciences (IPM), working at Computer Architecture, Hardware Accelerators, Energy Efficient GPU, Machine Learning.

زمان: چهارشنبه ۱۴۰۰/۰۶/۰۳ - ساعت ۱۵:۰۰

ارائه به صورت مجازی انجام خواهد شد.

<https://conf.ipm.ir/b/lot-0ed-uys-360>

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