پژوهشگاه دانشهای بنیادی
پژوهشکده علوم کامپیوتر

IPM

"سخنرانی‌های علمی"

Leveraging DNN Approximation in ML Cloud Services

سیدمرتضی نبوی نژاد، پژوهشگاه دانش‌های بنیادی (IPM)

Abstract

Deep Neural Networks (DNNs) provide state-of-the-art results in various areas such as computer vision, speech Recognition, and natural language processing. Coupled with efficient hardware accelerators and large datasets, these modern neural networks use very large and deep architectures to achieve state-of-the-art classification accuracy results. These modern DNNs require enormous computational resources for both learning and inference. One of the promising solutions to address the computational demand of DNNs is employing GPU accelerators. To improve the performance of DNN inference on GPU accelerators , hardware manufactures such as Nvidia and AMD offer reduced-precision arithmetic, which requires less resources (memory, computation and power), together with software libraries to convert trained models using higher precision arithmetic while trying to preserve the accuracy of the model. This approach is suitable for ML cloud service providers who are interested in faster and more resource-efficient approaches to improve their resource utilization and customer satisfaction.

In this talk, we first briefly introduce several control knobs that we have investigated in system level to manage the performance of DNN inference on GPUs. Then we focus on the reduce-precision arithmetic and its advantages and disadvantages. Finally, we present our proposed approach for incentivizing the employment of reduced-precision instructions of GPUs to maximize the profit of ML cloud service providers.

Biography

Seyed Morteza Nabavinejad is a Post-Doctoral Research Fellow at the School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. He received his Ph.D. degree from Sharif University of Technology in 2018. He has published several peer reviewed papers in venues such as IEEE Transactions on Cloud Computing, IEEE Transactions on Big Data, IEEE JETCAS, IEEE CAL, DATE, and CCGrid. His research interests include DNN accelerators, approximate computing, big data processing, and cloud computing.

زمان : چهارشنبه 1399/9/12 - ساعت ۱۵:۰۰
ارائه به صورت مجازی انجام خواهد شد.

***** شرکت برای عموم علاقه مندان آزاد است*****