Industrial Forum

Yanzhi Wang

Northeastern University, Chairman and former CEO of CoCoPIE Inc., USA

"Towards Best Possible Deep Learning Acceleration on the Edge – A Compression-Compilation Co-Design Framework"

Abstract:

Mobile and embedded computing devices have become key carriers of deep learning to facilitate the widespread of machine intelligence. However, there is a widely recognized challenge to achieve real-time DNN inference on edge devices, due to the limited computation/storage resources on such devices. Model compression of DNNs, including weight pruning and weight quantization, has been investigated to overcome this challenge. However, current work on DNN compression suffer from the limitation that accuracy and hardware performance are somewhat conflicting goals difficult to satisfy simultaneously.

We present our recent work CoCoPIE, representing Compression-Compilation Codesign, to overcome this limitation towards the best possible DNN acceleration on edge devices. We propose novel fine-grained structured pruning schemes, including pattern-based pruning, block-based pruning, etc. They can simultaneously achieve high hardware performance (similar to filter/channel pruning) while maintaining zero accuracy loss, with the help of compiler, which is beyond the capability of prior work. Similarly, we present novel quantization scheme that achieves ultra-high hardware performance close to 2-bit weight quantization, with almost no accuracy loss. Through the CoCoPIE framework, we are able to achieve real-time on-device execution of a number of DNN tasks, including object detection, pose estimation, activity detection, speech recognition, just using an off-the-shelf mobile device, with up to 180X speedup compared with prior work. Our comprehensive demonstrations are at :

https://www.youtube.com/channel/UCCKVDtg2eheRTEuqIJ5cD8A

Biography:

Yanzhi Wang is currently an associate professor and faculty fellow at Dept. of ECE at Northeastern University, Boston, MA. He received the B.S. degree from Tsinghua University in 2009, and Ph.D. degree from University of Southern California in 2014. His research interests focus on model compression and platform-specific acceleration of deep learning applications. His work has been published broadly in top conference and journal venues (e.g., DAC, ICCAD, ASPLOS, ISCA, MICRO, HPCA, PLDI, ICS, PACT, ISSCC, AAAI, ICML, NeurIPS, CVPR, ICLR, IJCAI, ECCV, ICDM, ACM MM, FPGA, LCTES, CCS, VLDB, PACT, ICDCS, RTAS, Infocom, C-ACM, JSSC, TComputer, TCAS-I, TCAD, TCAS-I, JSAC, TNNLS, etc.), and has been cited above 12,000 times. He has received six Best Paper and Top Paper Awards, and one Communications of the ACM cover featured article. He has another 12 Best Paper Nominations and four Popular Paper Awards. He has received the U.S. Army Young Investigator Program Award (YIP), IEEE TC-SDM Early Career Award, Massachusetts Acorn Innovation Award, Martin Essigmann Excellence in Teaching Award, Massachusetts Acorn Innovation Award,

Ming Hsieh Scholar Award, and other research awards from Google, MathWorks, etc. He has received 22 federal grants from NSF, DARPA, IARPA, ARO, ARFL/AFOSR, etc.. He has participated in a total of $40M funds with personal share $8.5M. Six of his former Ph.D./postdoc students become tenure track faculty at Univ. of Connecticut, Clemson University, Chongqing University, Beijing University of Technology, Texas A&M University, Corpse Christi, and Cleveland State University.

Shuhao Wang, Ph.D.

Co-founder and CTO of Thorough Future

"Empowering future pathology with artificial intelligence"

Abstract:

The shift from traditional microscopy to digital pathology has paved the way for the use of AI-assisted diagnostic systems in pathology diagnosis. Through nearly five years of technology exploration and clinical practice, we have successfully built the AI-assisted pathological diagnostic platform. The deep learning system has a sensitivity close to 100% and a specificity over 80% for the recognition of malignant tumors in stomach, intestine, lung, and prostate, and is able to

complete the diagnosis of tumor subtypes. In the platform deployed at PLAGH, the diagnostic models of all organs are embedded. The platform supports all digital scanners on the market. We have also connected the platform with the information system in the hospital, so that we can obtain information about the samples and export the diagnostic report. Thus, the platform can be seamlessly embedded into the diagnostic process for pathologists, improving their working efficiency. Every day, all slides supported by the intelligent diagnostic platform are scanned and uploaded, and pathologists can use digital slides and artificial intelligence in the interface of the information system in their daily diagnoses, and issue reports with a single click. In this report, we will introduce the large-scale application of the AI pathology diagnosis platform in the real world.

Biography:

Doctor Shuhao Wang, the co-founder and CTO of Thorough Future, has a Ph.D. from Tsinghua University, was a postdoctoral fellow at the Institute for Interdisciplinary Information Sciences, Tsinghua University, and an assistant researcher at Baidu, NovuMind, and JD, and has more than 20 national patents, and has published many academic papers in top journals/conferences such as Nature Communications, Modern Pathology, ICCV, etc. He received the Elite Award of “30 New Generation Digital Economy Talents” at the World Internet Conference 2019. Dr. Shuhao Wang has extensive experience in the implementation of cutting-edge AI techniques and has a background in medical AI research for many years.