Journal of Signal Processing Systems (2023) 95:569–570
https://doi.org/10.1007/s11265-023-01854-y

Guest Editorial: Special Issue on Systems Optimizations for DSP and AI Applications

Kuan-Hsun Chen[1] · Yung-Chia Lin[2] · Jenq-Kuen Lee[3]

Published online: 8 March 2023

Embedded systems are increasingly important for signal processing, machine learning, and multimedia applications in recent years. While such systems will look to play an important role ahead for signal processing and AI application designs, many challenging problems, such as large data computations and memory footprints, remain to be resolved, especially from the efficiency aspect. Programming models, compilers, API designs, architecture designs, and software tools all need to contribute to the advance of system designs. This special issue aims to bring together researchers in the related areas to present the latest developments and technical solutions concerning various aspects of signal processing and AI applications in system optimizations.

This special issue consists of seven papers that are briefly discussed as follows:

First off, Wu et al.'s paper on accelerating OpenVX provides a compiler framework to generate efficient binary code. Specifically, given a program written in OpenVX, the framework proposed in this work translates the program to Halide for scheduling, and then converts it to MLIR. The results show that an order of magnitude speedup for kernels used in image processing and AI applications can be achieved via utilizing well-designed dialects of MLIR and Halide scheduler.

Hsieh et al. present an optimization assistant DLOOPT to make the auto-tuning module, namely AutoTVM, efficient in Apache TVM, which is a well-known open-source deep learning compiler. Although AutoTVM facilitates the tuning of optimization configurations (e.g., tiling size and loop order) for users, it takes quite a long time to search the optimum configurations. DLOOPT adopts a set of tuning strategies to simplify the tuning process, which can reduce more than 99% of time in terms of developing adequate optimizations for operators in a model, and make the optimization process much easier to improve the development of learning-based applications.

Lee et al. proposes an effective method for programmers to improve the performance of matrix multiplication layers in DNN applications through Halide scheduling primitives. With the base version of OpenCL framework, this work provides Halide scheduling primitives for sparse matrix compression and includes sparse matrix multiplication and sparse convolution method. Experimental results show that the DNN model with the proposed Halide compression scheduler can be executed in almost 2x speed-up, in comparison to the original model without any compression schemes.

Lai and Yang present a new binary translator named Rabbit, which uses the latest version of the LLVM framework (version 8.0), and introduces a novel optimization technique, i.e., platform-independent hyperchaining. The main idea is to chain static and dynamic translated code blocks together. In the past, translated code blocks can only be chained with the same type of translated blocks. Moreover, platform-dependent hyperchaining is also introduced, which can recompile source binaries to x86-64 and RISC-V and gain further performance improvement.

Through Halide scheduling, Zhao et al. investigate how to use Halide to build a framework for fast prototyping and optimization on OpenVX DAG graphs. After rewriting OpenVX kernels to Halide (i.e., over different data access modes), the auto-scheduler of Halide is utilized in a systematic way to achieve the acceleration. The usage of Halide can greatly shorten the codes, reduce the coding time, but also improve the performance. Experimental results show that the Halide scheduling can achieve great improvement for the OpenVX NNE.

Several approaches for sparse signal recovery have been developed to provide accurate recovery from a small portion of available data. Zaric et al. provides an approach to combine

✉ Kuan-Hsun Chen
k.h.chen@utwente.nl

Yung-Chia Lin
yungchia.lin@mediatek.com

Jenq-Kuen Lee
jklee@cs.nthu.edu.tw

[1] University of Twente, Hallenweg 19, Enschede 7522NH, the Netherlands

[2] MediaTek USA Inc. (Woburn), Woburn, MA 01801, USA

[3] National Tsing Hua University, Hsinchu 300, Taiwan

gradient and threshold-based approaches, by which both accurate and computationally efficienct signal reconstruction is possible. A software tool is also provided, which allows users to choose a signal from the database, select the sparsity domain, calculate its initial transformation in the selected domain and perform the reconstruction using the designed approach.

Wang et al., presents an approach to consider EfficientNets as backbone to build lightweight version and adjust the architecture toward a more hardware-friendly structure, namely Network Candidate Search. To improve the efficiency of the searching process, grouping and elimination steps are additionally introduced. Based on state-of-the-art for hardware-friendly CNN, this work focuses on the scaling down principle through three dimensions, i.e., input resolution, depths and channels, where outperformed models consuming similar hardware cost can be derived.

Overall, these papers provide frontier information related to embedded computing, embedded compilers, and embedded programming tools. The first five papers cover language and compilation techniques, and some specifically focus on AI applications. The last two papers advance the techniques of signal processing and learning model with hardware-awareness.

We thank all the authors, the reviewers, the JSPS journal administrative staff and the JSPS Editor-in-Chief for all their contributions to making the high quality of this JSPS Special Issue possible.

We hope you enjoy reading the articles.

Kuan-Hsun Chen is a tenured assistant professor in the Department of Computer Science at University of Twente in the Netherlands. From Aug. 2019 to Aug. 2021, he was a postdoc at TU Dortmund University in Germany. He earned his Ph.D. (Dr.-Ing.) in Computer Science from TU Dortmund University with a distinction in 2019. He earned his master's degree in Computer Science from National Tsing Hua University (Taiwan) in 2013. Furthermore, he received one dissertation award and one best student paper award (RTCSA'18) and best paper nominations in DATE'21. His research interests include real-time embedded systems, architecture-aware software design, and dependable computing.

Yung-Chia Lin, Ph.D., received his Ph.D. in computer science and B.S. degree in physics from National Tsing-Hua University. He is working at MediaTek, Inc. for more than 15 years, and specializes in optimizing compiler development and DSP processor architecture design. His research interests include LLVM as well as other general compiler technologies, parallel programming models, modern processor architectures, and all emerging technologies in embedded systems.

Jenq Kuen Lee received the B.S. degree in computer science from National Taiwan University in 1984. He received a Ph.D. in computer science from Indiana University in 1992, where he also received an M.S. (1991) in computer science. He was a key member of the team who developed the first version of the pC++ language and SIGMA system while at Indiana University. He was also a recipient of the most original paper award in ICPP '97 with the paper entitled "Data Distribution Analysis and Optimization for Pointer-Based Distributed Programs". In 2005, he received Taiwan MOEA funding to lead a research team to develop compilers for PAC VLIW DSP processors with distributed register files by collaborating with ITRI STC. He is also a recipient of Google Research Award (Mountain View), 2009. He was also a director for Taiwan MOE ESW (embedded system software) consortium 2008-2013. In 2010, he received a Taiwan MOEA economic contribution award (Deep Plow Award) for his contribution in embedded compiler research. From Oct. 2015 to Oct. 2018, he participated in the new version of OpenCL proposals with Khronos OpenCL Next DSP Feature Set. His current research interests include optimizing compilers, AI framework compilers, embedded multicore compilers and systems, and computer architectures.