Efficient Computer Vision for Embedded Systems

George K. Thiruvathukal, Loyola University Chicago
Yung-Hsiang Lu, Purdue University

The winners, as well as the organizers and sponsors of the IEEE Low-Power Computer Vision Challenge, share their insights into making computer vision (CV) more efficient for running on mobile or embedded systems. As CV (and more generally, artificial intelligence) is deployed widely on the Internet of Things, efficiency will become increasingly important.

Computer vision (CV) is at the center of the past decade's impressive improvements of learning-based artificial intelligence (AI). It also remains one of the grand challenges in AI, where significant R&D effort is required to achieve CV's fullest potential. One of the driving forces assessing progress in CV is competitions that compare different solutions using the same data sets. Most CV competitions focus on accuracy, without the consideration of efficiency on hardware with limited resources. As a result, researchers use increasingly deeper neural networks (NNs), running on fast computers (sometimes supercomputers) with one or more GPUs. Since 2015, the IEEE Low-Power Computer Vision Challenge (LPCVC) has compared CV solutions running on battery-powered devices such as mobile phones and miniature autonomous robots. Over the years, 108 teams from around the world have submitted more than 500 solutions for CV problems including object detection, image classification, moving-object tracking, and character recognition. This virtual roundtable collects the opinions from experts in efficient CV about the status of technologies and directions for future improvements. LPCVC was called the *Low-Power Image Recognition Challenge* (*LPIRC*) in 2015–2019. It was renamed *LPCVC* in 2020 when a video track was added.

Digital Object Identifier 10.1109/MC.2022.3145677
Date of current version: 8 April 2022

ROUNDTABLE PANELISTS

Yiran Chen is a professor in the Department of Electrical and Computer Engineering at Duke University and serves as the director of the National Science Foundation (NSF) AI Institute for Edge Computing Leveraging the Next-Generation Networks and the NSF Industry–University Cooperative Research Center for Alternative Sustainable and Intelligent Computing, and the codirector of Duke Center for Computational Evolutionary Intelligence. His group focuses on the research of new memory and storage systems, machine learning and neuromorphic computing, and mobile computing systems. He was a co-organizer of the 2018–2021 IEEE Low-Power Computer Vision Challenge (LPCVC). He serves as chair of ACM's Special Interest Group on Design Automation. He is a Fellow of IEEE and ACM.

Soonhoi Ha is a professor of computer science and engineering at Seoul National University. Ha received a Ph.D. in electrical engineering and computer science from the University of California at Berkeley and was the winner of the 2017 LPCVC. His research interests include hardware-software co-design methodology of embedded systems, system simulation and performance estimation, embedded machine learning, and the Internet of Things. He is a Fellow of IEEE.

Song Han is an assistant professor at the Massachusetts Institute of Technology. His research focuses on efficient deep learning computing, model compression, neural architecture searches, and hardware acceleration. He received Best Paper awards at the 2016 International Conference on Learning Representations and the 2017 International Symposium on Field-Programmable Gate Arrays, and multiple faculty awards from Amazon, Sony, Facebook, Nvidia, and Samsung. He was named one of the "35 Innovators Under 35" by *MIT Technology Review*. He received the NSF CAREER Award for "Efficient Algorithms and Hardware for Accelerated Machine Learning" and IEEE's "AIs 10 to Watch: The Future of AI" award. He was the winner of the 2019–2021 LPCVC.

Naveen Purushotham received an M.S. in electrical and computer engineering from the University of New Mexico before joining Lattice Semiconductors as a product engineer. Currently, he is a staff applications engineer with Xilinx, which was a financial sponsor of LPCVC from 2019 to 2021. He is a part of the Xilinx University Program team, a technical group inside the Xilinx CTO's office. He is actively engaged in the Xilinx open source project PYNQ. He enjoys working in the machine intelligence area.

Tao Sheng is the director of computer vision services at Oracle Cloud. He has extensive industrial R&D experience spanning several leading companies, including Oracle, Amazon, Qualcomm, and Intel. He has launched multiple products used by millions of customers. He has obtained more than 10 U.S. and international patents and published more than 10 research papers. Since 2018, he has co-organized the EMC^2 Workshop on Efficient ML and AI. He is the winner of the 2018 and 2019 LPCVC and co-organizer of the 2021 LPCVC.

Michelle Tubb, a Certified Association Executive, is director of marketing and sales of the IEEE Computer Society. The Computer Society is the premier source for information, inspiration, and collaboration in computer science and engineering. Connecting members worldwide, the Computer Society empowers the people who advance technology by delivering tools for individuals at all stages of their professional careers. Our trusted resources include international conferences, peer-reviewed publications, a robust digital library, globally recognized standards, and continuous learning opportunities. The Computer Society is the financial and administrative sponsor of the 2020–2021 LPCVC.

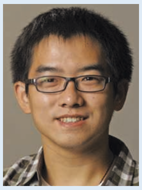
Ying Wang is an associate professor at the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include chip design automation, energy-efficient computer architecture, and memory systems. Ying received a Ph.D. in computer architecture from the Chinese Academy of Sciences. He has published more than 100 papers in refereed journals and conferences, and received Best Paper awards at the IEEE International Conference on Computer Design, Great Lakes Symposium on Very Large Scale Integration, Asian Test

(*Continued*)

ROUNDTABLE PANELISTS (*Continued*)

Symposium, and International Test Conference in Asia. He is also the recipient of the 2022 Under-40 Innovators Award and was the winner of the 2016 LPCVC.

Yu Wang is a tenured professor in the Department of Electronic Engineering, Tsinghua University. His research interests include domain-specific, hardware-software co-design and multiagent systems. He has authored and coauthored more than 200 papers in refereed journals and conferences. He has received four Best Paper awards at the Asia and South Pacific Design Automation Conference, International Symposium on Field-Programmable Gate Arrays, Non Volatile Memory Systems and Applications Symposium, and IEEE Computer Society Annual Symposium on Very Large Scale Integration, and 10 Best Paper nominations. He cofounded DeePhi Tech (acquired by Xilinx in 2018), a leading deep learning computing platform provider. He is the winner of the 2015 LPCVC. He is a Fellow of IEEE.



Zhangyang "Atlas" Wang is the Jack Kilby/Texas Instruments Endowed Assistant Professor in the Department of Electrical and Computer Engineering at The University of Texas at Austin. He also holds a visiting researcher position at Amazon. He was an assistant professor of computer science and engineering at Texas A&M University from 2017 to 2020. His research interests include machine learning, computer vision, and optimization and their interdisciplinary applications. Most recently, he has studied automated machine learning, learning to optimize, robust learning, efficient learning, and graph neural networks. He is the winner of the 2020 and 2021 LPCVC.

LPCVC

***COMPUTER*:** Why did you participate in LPCVC?

YU WANG: Our team started studying energy-efficient hardware design for NNs in 2012. By 2015, we had already formalized our software-hardware co-design methodology for efficient NN inference. We chose to participate in LPCVC to validate our workflow.

YING WANG: The challenge is unique because it concerns the realistic implementation of cutting-edge embedded CV systems, while most of the other competitions overemphasize a single aspect or design goal of the algorithms. It is very worthwhile and exciting to push the efficiency limit of solutions in real hardware devices and under the performance and power constraints.

SOONHOI HA: My main research area is the design methodology for embedded systems. As machine learning (ML) applications are becoming more popular in embedded systems, we were naturally interested in how to support ML applications in an embedded system that has tight constraints on real-time performance and energy consumption. When we were introduced to LPIRC early 2017, we were just starting to study deep learning. I thought it was a good way to participate in a challenge to understand the problem and find the related research issues, such as which hardware platform and algorithm to use, how to optimize the software with resource constraints, how to optimize multiple design objectives, and so on.

ATLAS WANG: Low-power CV (LPCV) is a very important topic and one of my main interests. LPCV provides an influential and authoritative platform to pursue real use cases and benchmark achievable performances.

Since 2015, the IEEE Low-Power Computer Vision Challenge has compared CV solutions running on battery-powered devices such as mobile phones and miniature autonomous robots.

SONG HAN: LPCVC pushes the frontier of LPCV on edge devices; new TinyML techniques always emerge during the competition.

TAO SHENG: LPCVC builds a strong community of LPCV from both academia and industry. It's a great experience to know cutting-edge solutions used to solve the efficiency problem of CV from all the participants.

***COMPUTER*:** How is LPCVC relevant to activities in the IEEE Computer Society?

MICHELLE TUBB: Supporting the development of CV technologies is an important part of the Computer Society's mission. The Computer Society has the Technical Community on Pattern Analysis and Machine Intelligence, which addresses pattern recognition, artificial intelligence, expert systems, natural language understanding, image processing, and CV. This community counts among its conference portfolio the Computer Vision and Pattern Recognition (CVPR) conference. The Guide2Research placed CVPR at the top of all conferences in computer science in terms of its influence. CV technologies, with applications that impact security and biometrics, image and video synthesis, 3D CV, representation learning, and improving model efficiency, are enabling the creation of opportunities for the future. The Computer Society fosters this type of growth and activity, and as a professional Society, also delivers a forum for discussing the implications of the technology—its potential, policies, trends, and ethics. With the growth potential of CV, the Computer Society will continue to influence R&D efforts and foster new areas of focus.

LPCV research may have important implications for alleviating global warming and achieving carbon neutrality.

LPCV RESEARCH

***COMPUTER*:** Why is research in LPCV important?

WANG (YU): Our team considered that AI of Things (AIoT) applications would become ubiquitous in the future. Due to the limited power of and resource budgets for AIoT devices, industrial deployment tools at that time could not fill the gap between complex NN models and their constrained resources. Only critical, breakthrough research in academia can effectively close this gap. Thus, LPCV is a good start for bridging the gap between academia and industry. LPCV research may have important implications for alleviating global warming and achieving carbon neutrality.

WANG (YING): LPCV needs the cooperation of both the software and hardware communities, and to solve one of the most important technical issues that prevents the powerful CV technology from being adopted in IoT and edge devices. My team has been looking into the system and architectural aspects of edge AI computing. I am very interested to see the result of our research and how it works in a realistic setup for AIoT devices. So I think LPCV is a perfect playground for my students and me to try all-layer solutions to approach the limit of LPCV tasks.

HA: At the application level, new ML algorithms are being developed that use less computing power while maintaining a similar level of accuracy. With a given algorithm and a hardware platform, various software-optimization techniques such as quantization, pruning, and low-rank approximation have been developed to reduce energy consumption. At the hardware level, deep learning accelerators, called *neural processing units* (*NPU*), and power-efficient hardware platforms that equip both GPUs and NPU, are being developed. Reducing energy consumption needs a comprehensive technology that harmonizes hardware, software, and algorithms. LPCV encourages researchers and engineers to understand the interaction of various areas in computer science and engineering. Vision has a huge impact on human cognition and behavior. It would be of great benefit if we could obtain visual information through our mobile devices in a situation where visibility is poor. For example, it will give blind people the opportunity to indirectly restore their sight. Infrared sensors can be used to recognize objects even at night, increasing safety. If it is applied to an application that visualizes various types of data with images and obtains necessary information from the images, human information-recognition ability will be greatly improved. Furthermore, when it is applied to the metaverse in which virtual space will be augmented with real space captured by CV, our living space will be able to expand without boundaries.

SHENG: CV is widely used in industry, for example, manufacturing, robotics, and so on. CV can help the aging population.

WANG (ATLAS): Training AI models, especially deep networks, includes significant energy consumption, financial costs, and environmental impacts. For instance, the carbon footprint of training one deep NN (DNN) can be as high as five American cars' lifetime emissions. More efficient models are also crucial for bringing AI-powered features to more resource-constrained devices, such as mobile phones and wearables. As one of the most important AI applications, CV is especially heavy in energy costs, and therefore urgently demands low-power solutions.

HAN: Today's AI is too big. DNNs demand extraordinary levels of data and computation, and therefore power, for training and inference. This severely limits the practical deployment of AI in edge devices. TinyML techniques can make AI greener, faster, more efficient, and more accessible.

COMPUTER: Can you describe one (or several) "grand challenges" using CV; the solutions will significantly change the world, but are they far beyond today's technologies?

HAN: Smart home, smart retail, smart factory, smart transportation, smart health care, smart agriculture, and more.

WANG (YING): Predicting large, worldwide issues like famine, epidemics, and warfare using nongovernment-managed satellite vision data and other multimodal data sources from connected public sensor sources so that we may have a slight chance of avoiding the occurrence of human crises by making the evidence available to everyone and take timely measures.

SHENG: Humanoid robots.

WANG (ATLAS): Making CV runnable on the smallest-possible chips/sensors; in particular, further equipping them with incremental/lifelong learning capability to adapt to/interact with changing environments.

COMPUTER: If you have unlimited resources, what would you like to see in the area of LPCV?

CHEN: Building a demonstration of a swarm of intelligent LPCV devices that are connected and function as a single, highly integrated system.

SHENG: New competitions and new data sets.

HAN: New data sets that are closer to real-world applications.

HA: A new competition to solve real complex problems without limitations on hardware, software, and algorithms.

WANG (YING): More interesting data sets and easy-to-use frameworks are important to democratize the research of LPCV.

COMPUTER: Would you like to add a concluding thought?

WANG (YING): What are the ultimate solutions to LPCV when the development of deep learning vision technologies stop making further progresses? Are the solutions specialized hardware platforms or algorithms? Will general programmable hardware be the solution?

TinyML techniques can make AI greener, faster, more efficient, and more accessible.

HA: In image classification, a neural architecture search (NAS) that automates the design of a new DNN structure becomes a de facto technique. It needs to be extended to other vision applications that will be run on a hardware platform under a set of given constraints on real-time performance and energy consumption. Even though existing NAS techniques can find algorithms for a given hardware platform, the algorithm-hardware co-design cooperative will also be an interesting topic. In addition, it would be good if software-optimization techniques are considered in the design space exploration.

HAN: Could LPCV use solutions that are not based on deep learning and achieve competitive or even better accuracy and efficiency than deep learning solutions? My guess is yes, and this new thinking leaves much open room to explore new competitions.

NAVEEN PURUSHOTHAM: We should encourage more participation. Make the LPCVC tracks more organic to each other. Maybe have a common state-of-the-art LPCV ML problem, a common hardware (say, Xilinx), and common ML framework, say, PyTorch. This way, the organizers and sponsors can collaborate more, and the common track will have a larger pool of contestants. Unifying the contest in this way may have a larger impact and may also attract more contestants. Another suggestion is about publicity worldwide: we need a more uniform distribution of contestants across the world, within reason. Currently, it seems that most of contestants are from the United States and Asia. We need a more solid publicity plan.

SHENG: Could LPCV discuss the research direction on how to bridge the gap from cloud-trained CV models to be effectively deployed to low-power edge devices? The challenges are 1) cloud training may not know the hardware architecture of edge devices, 2) training data are not sufficiently captured from edge devices, and 3) image quality is very dynamic on handheld devices.

GEORGE K. THIRUVATHUKAL is a professor of computer science and the department chairperson at Loyola University Chicago, Chicago, Illinois, 60626, USA. He is also a visiting computer scientist at Argonne National Laboratory in the Leadership Computing Facility. Contact him at gthiruvathukal@luc.edu.

YUNG-HSIANG LU is a professor of electrical and computer engineering and a university faculty scholar at Purdue University, West Lafayette, Indiana, 47907-2035, USA. Contact him at yunglu@purdue.edu.