Scanning the Issue

Software-Defined Imaging: A Survey
by S. Jayasuriya, O. Iqbal, V. Kodukula, V. Torres, R. LiKamWa, and A. Spanias
Image sensing has become ubiquitous in modern society, ranging from industrial uses in the workplace all the way to personal entertainment through the sharing of photographs and videos via social media. Driven by the development of CMOS image sensors in the 1990s and 2000s, image sensing has become cheap, affordable, and when integrated with smartphones, extremely portable, and easy to use. Indeed, billions of photographs are taken and uploaded on the internet each day. In accordance with the development of image sensor technology, the rise of image processing, computer vision, and computational photography has been similarly meteoric. Advances in algorithms, including state-of-the-art machine learning using deep neural networks, have enabled high-fidelity visual computing. However, while both image-sensing hardware and the software that implements and realizes the algorithms have grown exponentially, there is still relatively little co-design between the two domains. This is due to several factors which include the technical challenges of interfacing analog sensing components with digital computation, but also social and cultural challenges of different communities of researchers and industries communicating with one another across the stack.

This month's regular papers cover a broad range of topics including megahertz wireless power transfer and resistive neural hardware accelerators.

For example, the traditional image sensor faces major hurdles to allow for flexibility and reconfiguration in its operating modes. A recent study showed that changing image sensor resolution costs hundreds of milliseconds in latency, and this was mostly due to software operating system (OS) bottlenecks. Most image sensors have been slow to expose knobs to the software developer such as resolution, exposure, and quantization; in contrast, most imaging and vision algorithms do not exploit these sensing mechanisms efficiently and adaptively at runtime. This leads to lost opportunities for improved reconfigurability of imaging systems in practice. To address these issues, an interdisciplinary community of researchers and practitioners has embraced software-defined imaging (SDI) to improve the technology of visual computing systems. A software-defined image sensor offers several dimensions of sensing configurability along with system support and programming abstractions to support application-specific needs. Hardware parameters such as exposure, resolution, and frame rate are programmable in image sensors, and software-defined image sensors exploit this programmability alongside software algorithms to optimize certain imaging task metrics such as energy, latency, and task performance. This field has a vertical-oriented mindset connecting knowledge of sensor physics and electronics, analog and mixed-signal circuits, digital systems and architectures, Oss and programming languages, and end applications of computer vision and computational photography.

This article surveys this emerging area of SDI and highlights key works across the hardware and software stack in the literature. It points out how this area intersects with other popular areas of research including embedded computer vision, deep learning, computational photography, and digital hardware acceleration for imaging applications. The unique nature of SDI research, cross-cutting traditional boundaries to allow synergy among the stack is stressed.

Model-Based Deep Learning
by N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis
Traditional signal processing is dominated by algorithms that are based on simple mathematical models which are hand-designed from domain knowledge. These domain-knowledge-based processing algorithms, or model-based methods, carry out inference based on knowledge of the underlying model by relating the observations at hand and the desired information. Model-based methods do not rely on data to learn their mapping, though data is often used to estimate a small number of parameters. Classical statistical models rely on simplifying assumptions (e.g., linear systems,

Digital Object Identifier 10.1109/JPROC.2023.3267661

Gaussian and independent noise, etc.) that make inference tractable, understandable, and computationally efficient. On the other hand, simple models frequently fail to represent nuances of high-dimensional complex data and dynamic variations.

The incredible success of deep learning in applications has initiated a general data-driven mindset. The current trend is to replace simple principled models with purely data-driven pipelines, trained with massive, labeled datasets. The benefits of data-driven methods over model-based approaches are twofold. First, purely data-driven techniques do not rely on analytical approximations and thus can operate in scenarios where analytical models are not known. Second, for complex systems, data-driven algorithms can recover features from observed data which are needed to carry out inference. This is sometimes difficult to achieve analytically, even when complex models are perfectly known.

While popular, the computational burden of training and utilizing highly parametrized deep neural networks (DNNs), as well as the fact that massive datasets are typically required to train such DNNs to learn a desirable mapping, may constitute major drawbacks in various signal processing, communications, and control applications. This is particularly relevant for hardware-limited devices, such as mobile phones, unmanned aerial vehicles, and Internet-of-Things systems, which are often limited in their ability to utilize highly parametrized DNNs and require adapting to dynamic conditions. Furthermore, DNNs are commonly utilized as black boxes; understanding how their predictions are obtained and characterizing confidence intervals tends to be quite challenging. As a result, deep learning does not yet offer the interpretability, flexibility, versatility, and reliability of model-based methods.

The limitations associated with model-based methods and black-box deep learning systems have given rise to a multitude of techniques for combining signal processing and machine learning to benefit from both approaches. These methods are application-driven and are thus designed and studied in light of a specific task. For example, the combination of DNNs and model-based compressed sensing (CS) recovery algorithms was shown to facilitate sparse recovery as well as enable CS beyond the domain of sparse signals. The proliferation of hybrid model-based/data-driven systems, each designed for a unique task, motivates establishing a concrete systematic framework for combining domain knowledge in the form of model-based methods and deep learning, which is the focus of this article. This article reviews leading strategies for designing systems whose operation combines domain knowledge and data via model-based deep learning in a tutorial fashion.

Resistive Neural Hardware Accelerators

by K. Smagulova, M. E. Fouda, F. Kurdahi, K. N. Salama, and A. Eltawil

Neural networks (NN) have found widespread use due to their ability to generalize and learn adaptively, making them a potent tool for tackling abstract and complex problems. They have found their way into numerous fields such as agriculture, robotics, medicine, renewable energy, ecology, and climate change, among others. However, the advancement in state-of-the-art machine learning algorithms and deep neural networks is tightly coupled to intensive computational resources that consume high power and memory resources. Furthermore, the constant increase in data volume compounds this issue. Proposed solutions aim to counter these limitations by increasing the storage capacity of the main memory and enhancing interconnection bandwidth.

Traditional memory devices are struggling to keep up with NN architecture requirements giving rise to active research in the field of emerging nonvolatile memory (NVM). These new devices operate based on a change in resistance instead of charge, where the resistance-switching phenomenon has been observed in various materials. The most common NVM devices include phase-change memory (PCM), spin-transfer torque random-access memory (STT-RAM), and resistive random-access memory (ReRAM). Emerging technology devices have the potential to not only store more data but also serve as efficient compute-in-memory (CIM) architectures. Among different NVM technologies, ReRAM has better characteristics than PCM, including a higher density and scalability than STT-RAM and better write and read performance.

ReRAM devices are widely considered a promising technology for the implementation of multiply-accumulate (MAC)—the key operation in ML and NNs. Efficient ReRAM-based hardware accelerators could significantly speed up the progress of intelligent hardware architectures, allowing them to process larger amounts of data while utilizing less resources. Towards that end, various application-specific ReRAM-based accelerator designs have been demonstrated, including standalone macros, co-processors, and many-core processors. This article reviews several state-of-the-art many-core and multi-node ReRAM-based CIM neural network accelerators, examines their limitations and proposes potential directions for improvement.

Overview of Megahertz Wireless Power Transfer

by Y. Wang, Z. Sun, Y. Guan, and D. Xu

Following the wireless transmission of information, wireless power transfer (WPT) will profoundly impact all aspects of production and daily life in human society. Due to its advantages of convenience, safety, reliability, and noncontact, it provides new solutions for some special applications and further promotes many innovative developments that urgently need multidisciplinary crossover and integration. For instance, when WPT technology is combined with biology and medicine,

it can provide better conditions for the sustainable charging of implantable medical devices, such as retinal prostheses, pacemakers, and cortical implants, significantly reducing the risk caused by surgical battery replacement. In the power electronics (PE) field, WPT has dramatically improved the user's charging experience. Wireless and multistandard compatibility make charging consumer electronic devices easier and tidier. In addition, the simultaneous charging technology of multiple devices also reduces the number of chargers that need to be carried. The impact of WPT is also considerable in areas such as the charging of electric vehicles, mobile robots, and more. In the Internet of Things (IoT), WPT is deeply integrated with wireless information technology and radio frequency (RF) microwave, which is considered a tremendous potential technology. With the continuous iteration of wide bandgap devices, it provides solid conditions for the system's high-frequency design. Furthermore, in most practical applications, the power transfer often needs to be accompanied by information feedback, which has led to optimal conditions for developing the synchronous wireless information and power transfer (SWIPT) technology.

Nowadays, there are two major WPT standards. One is Qi, another is AirFuel. For Qi, the operating frequency is around 80 to 300 kHz. For WPT systems operating in such a frequency range, it is more worthy of priority to be considered in high-power applications (e.g., several kW power levels). It performs well in terms of system robustness, power transfer level, and efficiency. But relatively, its volume is usually large, the freedom degree of space transfer is comparatively low and it is relatively sensitive to space misalignment. For AirFuel, the operating frequency is within the industrial scientific medical band, such as 6.78, 13.56, and 27.12 MHz. At this point, the increase in frequency significantly improves the coupler's ability to transmit power, compensating for the degraded coupling performance in the case of misalignment. In other words, the spatial flexibility of MHz systems is higher, further providing feasibility for multiple loads to be charged simultaneously. But on the other hand, substantial switching losses, extreme eddy current effects, significant proximity effects, sensitive parasitic parameters, and nonlinear capacitive reactance of the rectifier and other problems will arise.

MHz power conversion is a very interesting research area, where such a system can be designed from both the PE and RF point of view. When reviewing MHz systems (usually above 10 MHz as RF) or even higher gigahertz (GHz) systems (i.e., microwave techniques) from an RF perspective, extreme space transmission capability, efficient energy harvesting, beamforming, and low-reflection rectenna design are the main concerns and research hotspots. In general, for a GHz system, power can be transmitted over a longer distance, from a few meters to several kilometers, but its transmission efficiency is relatively low. This article mainly focuses on the research and analysis of MHz WPT systems from a PE perspective. The article discusses the inverter, rectifier, impedance compression, dynamic impedance matching, coupling structure design, and multiload modeling while providing comparisons with different methods. It also provides insights into existing challenges and future research hotspots in this exciting area of research. ■