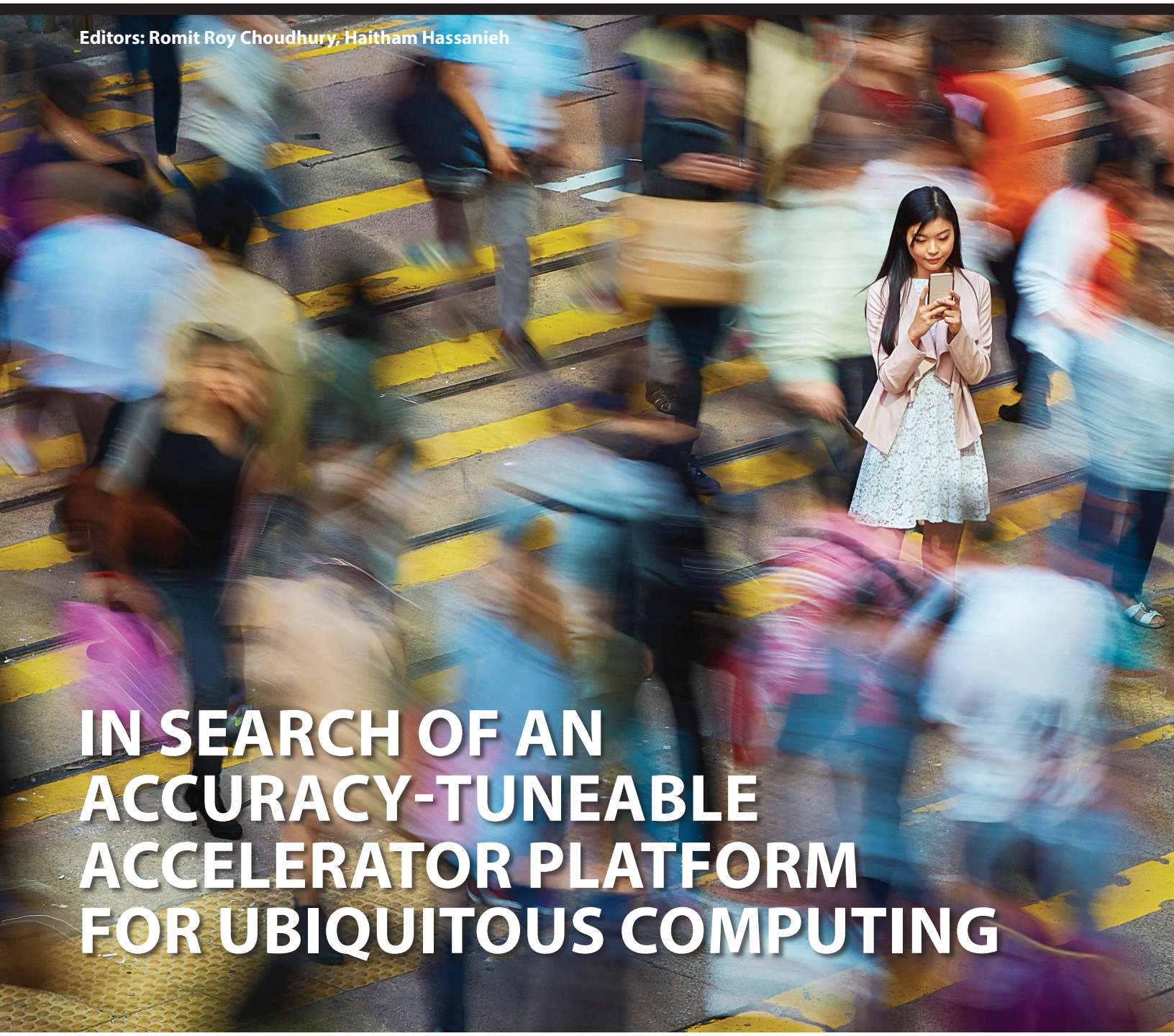
Ratko Pilipović, Veljko Pejović, Octavian Machidon
Faculty of Computer and Information Science, University of Ljubljana, Slovenia

Editors: Romit Roy Choudhury, Haitham Hassanieh



IN SEARCH OF AN ACCURACY-TUNEABLE ACCELERATOR PLATFORM FOR UBIQUITOUS COMPUTING

Paving the way towards the realization of Mark Weiser's vision of ubiquitous computing [1], the research community has made incredible advancements on several fronts. When it comes to interacting with humans, for example, computers can already use pretty much anything as a touch-pad [2]. Similarly, when it comes to sensing the environment, computers can unobtrusively detect anything from a driver fatigue [3] to the presence of the queen bee in a hive [4]. When compared with these, advancements on the "core" front — the computing itself — appear to be rather orthodox and limited.

Photo, istockphoto.com

The last decade instilled a particularly strong driver that reshapes the way we reason about the computing ecosystem. The name of this force is deep learning (DL), and it demonstrated that with sufficiently capable computers we can unleash the power of data collected by omnipresent sensors. Successful integration of the sensing-learning-interaction pipeline has already transformed the way we track our exercise regime, our means of authentication, and our approach towards getting rid of pests in agriculture.

Deep learning, however, often represents an insurmountable obstacle to ubiquitous computing. Its computational appetite is not poised to be satisfied with small edge devices that cannot afford to host high-performance CPUs and GPUs. Energy is especially critical, as deep learning not only incurs high power consumption, but also, due to the computational complexity, reduces the time a device can spend in a low-power mode. Instead, a nowadays common solution is to transfer the data to the cloud. Here, heavy processing can take place without incurring any cost to the edge devices' resources. In lieu, the price is paid through a potential loss of privacy and data confidentiality, additional communication delay, and the loss of autonomy.

In their 2019 Turing award lecture, Hennessey and Patterson predict the move away from general purpose architectures towards specialized computing hardware [5]. Indeed, equipping edge devices with special accelerators represents a viable means of making local deep learning feasible, while avoiding cloud computing and keeping the data private. These accelerators are based on field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs) and are the subject of active research. Prototypes optimized for various deep learning models and applications have been demonstrated. Yet most of these accelerators fail to embrace the key property of ubiquitous computing — its dynamic nature.

In our preliminary work, we examined how an on-device deep learning model for a spoken keyword detection performs on a mobile device that transitions through environments characterized by different levels of noise [6]. We "slimmed" the model to different widths to observe how different levels of approximation impact the model performance and resource consumption. We demonstrated that successful keyword detection in noisy environments requires the full model, whereas detection in quiet environments can be completed with a model approximated to 25% of its parameters. Modern hardware accelerators that should bring deep learning to a wide gamut of devices, however, are either suited for a particular model or the other, and cannot adapt model parameters, their precision, or employ any other approximation to reduce the amount of computation on the fly.

Context-driven dynamic adaptation of deep learning is the guiding principle of our research vision. In Figure 1 (next page), we depict the idea: different contextual situations, such as varying environmental noise levels, make the inference problem more or less challenging (upper row); tuneable accelerators adapt the level of approximation, for instance, parameter quantization, according to the context (middle row), which in turn leads to computation and energy savings in periods when approximation is tolerated (bottom row).

BEYOND STATIC ACCELERATION

To identify the opportunities for dynamic hardware accelerator adaptation, we first need to dig into the functioning of these chips that enable efficient handling of massive computation required by deep neural networks. These accelerators harness parallelization and consist of a network-on-chip

(NoC) of processing elements (PEs) that perform multiple multiply-and-accumulate (MAC) operations at a time. Initial accelerator implementations had fixed size PE sets regardless of the size of the network layers, which significantly limited their throughput.

Adaptability is a way to address the above issue, and more recent implementations use multiple PE sets of different sizes, use packet-switched NoCs, or exploit reconfigurability for changing to match the size of the PE sets and match the layer size. Such an approach is evident in the reconfigurable ASIC accelerator for convolutional neural networks (CNNs) described by Zhao et al [7, 1]. Each of its 24 reconfigurable PE supports nine 16-×16-bit MAC operations in parallel, and by reconfiguring the PEs, the accelerator can handle different sizes of convolution operations such as 1×1, 3×3, 5×5, 7×7 and 11×11. Another reconfigurable accelerator that can adapt its structure to the size and dataflow pattern of a CNN layer is RC-CNN [8, 2]. It relies on a reconfigurable on-chip interconnection fabric that can organize a subset of the accelerator's PEs as a PE set with the same size/dimension of the target CNN layer and customize the inter-PE connections for the layer's dataflow pattern. RC-CNN reportedly increases the accelerator's throughput due to improved PE utilization while reducing the network latency and energy consumption.

In the mobile realm, however, opportunities arise dynamically due to the varying needs for highly accurate computation. For instance, a network for inferring a user's physical activity has to rely on more accurate computation once a user is performing a set of complex exercises, while it can recognize resting periods even with a highly approximated network [9]. The above-mentioned solutions, while using reconfigurability to optimize and adapt the hardware to the neural network that it executes, do not allow for dynamic, runtime adaptation of the network execution when the accuracy vs resource usage requirements change. The existence of software implementations of precision-scaling solutions, such as network quantization, has inspired research efforts targeting precision-reconfigurable hardware acceleration. One such example is [5] HyDRATE [10] – an accelerator based on a run-time configurable approximate multiplier. The employed multiplier

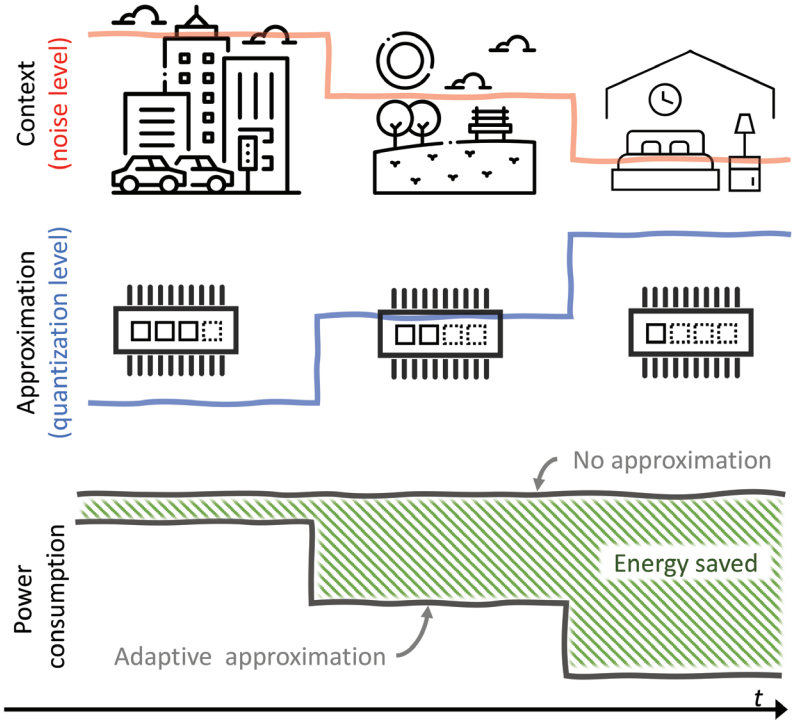


FIGURE 1. Mobile computing is characterized by context variability; tunable deep learning accelerators can harness approximation to match the amount of computation and the contextual requirements; the difference in resource usage between the fully precise and the approximated execution represents savings determined by the actual opportunities for approximation.

performs approximate blockwise multiplication, where blocks of input operands are multiplied to generate partial products. Approximate multipliers are also used in [11] where Xuan et al. introduce an accelerator-based on an approximate multiplier supporting two multiplication modes: the exact 4-bit and the approximate sub-8-bit multiplication. The low precision of input operands and small reconfigurability choices represent the biggest drawback of the presented accelerator.

Coarse-grained reconfigurable architectures (CGRAs) have been introduced to reach a balance between performance, power, and programmability in the quest for deploying energy-efficient deep learning on edge and IoT devices. An energy-efficient approximate CRA is proposed by Akbari et al. [12]: X-CGRA employs configurable arithmetic-logic units (ALUs), where tunability is achieved by choosing between exact and approximate addition and multiplication. The arithmetic circuit for multiplication and addition are composed of approximate and supplementary parts, where the supplementary part is power gated in approximate mode, and the structure and functionality of the other architectural components can be dynamically adapted using several quality-scalable operating modes. To ensure this dynamic run-time accuracy reconfigurability, X-CGRA's PEs exhibit the ability to switch between exact and different approximate configurations for the arithmetic units. This enables runtime adaptable accuracy, and a considerable energy consumption reduction at the price of modest quality degradation. The price, however, is paid in the placement overhead, as both the approximate and the supplementary part of arithmetic circuits have to be integrated in the design.

OVERVIEW OF THE ACurracy Tunable Accelerator (ACTA) Platform

In the mobile realm, the opportunities for adaptable computation tend to be gradual – in direct sunlight, a mobile app for object detection from camera images might perform well even with significant approximation; as the day passes and the lighting conditions change, the app might

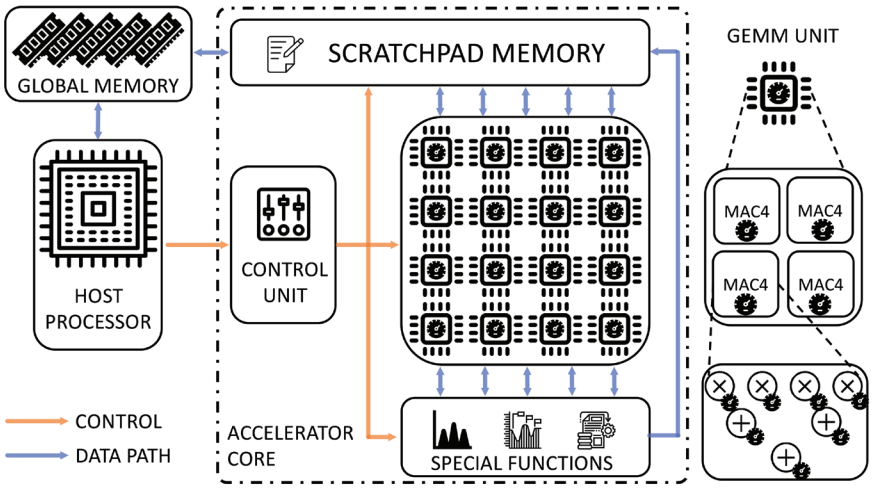


FIGURE 2. The ACTA's architecture. The orange lines depict the control paths, while the blue ones represent data paths. The right side illustrates envisioned accuracy tunability and energy consumption of the approximate GEMM unit.

require more accurate computation to preserve the inference accuracy; at dusk, it might need to harness all the resources possible, i.e., use the fully-accurate network, to perform object recognition. The existing reconfigurable accelerators, including those presented in the previous section, offer either two (i.e., accurate or approximate) or a small number of accuracy modes. Furthermore, another defining property of mobile computing is the need for portability. The existing accuracy-tunable accelerators usually host several arithmetical units with different accuracy, which renders them prohibitively large for small form factor of devices.

To overcome the limitations of the existing solutions, we devise *ACcuracy Tuneable Accelerator (ACTA)*, a platform that provides ultra-fast dynamic adaptation for mobile and IoT environments. The core element of the ACTA represents a dedicated Approximate General matrix multiply hardware Unit (AGU), whose accuracy can be changed on the fly. The envisioned AGU is based on the iterative logarithmic product approximation proposed by Babić et al. [13] and the design of an approximate GEMM unit presented by Pilipović et al. [14]. The AGU does not duplicate the functionality in multiple accuracy versions but incorporates a simple logic to approximate addition and multiplication, constituting the GEMM operation. The AGU harnesses specially crafted iterative computation, applied over the input operands for several clock cycles to achieve an arbitrary accuracy. Although AGU requires additional cycles to meet the desired accuracy, its compact design is preferable for mobile computing systems, where portability often trumps processing speed.

Figure 2 illustrates our design. During runtime, the host processor collects contextual data from sensors, e.g., audio data for noise assessment, motion, or battery level for overall power consumption, and determines the required computation accuracy level. The host processor then instructs the accelerator to meet the current accuracy demands. For the host processor, we envision the RISC-V processor due to open instruction set and the rising popularity of such processors in IoT applications. On the accelerator side, the control unit decodes the processor's instructions and, by controlling the number of iterations the computation is refined with, tunes the accuracy of the AGUs. More accurate processing brings increased energy consumption and vice versa. The intermediate results and input operands are stored in scratchpad memory, which acts as the accelerator's cache memory. Finally, the accelerator encompasses a unit that calculates special functions needed to process machine learning algorithms, such as activation, pooling, batch normalization, and others.

The ACTA is currently in early stages of development. However, we already have rough assessments of the improvements it brings. The synthesis results showed that the envisioned multiplier in the AGU design provides five times lower power-delay-product energy consumption with almost two times smaller design than the exact multiplier. The loss of accuracy stands at 8% for a single iteration and drops down to 1% after two multiplication iterations. With this, ACTA is well suited not only for NN inference, but for the training as well. We plan to employ the ACTA platform in a federated learning scenario, where the accuracy tuning can be used to both bring training to low-end devices, and to ensure that the training duration is balanced with the quality of the trained model. Moreover, while the model updates may represent numerically sensitive operations, so ACTA would tune AGUs to complete the training with a higher accuracy, the aggregate model could be interpreted with different accuracy at different devices to meet the device's capabilities and user needs.

IN THIS ARTICLE, WE ARGUE THAT MOBILE DEEP LEARNING NEEDS A FLEXIBLE HARDWARE PLATFORM

HORIZONS

The integration of accuracy-tuneable accelerators is bound to spur another wave of DL proliferation. Not only will DL become introduced to a wider set of devices, but will also become more efficient, potentially opening new application avenues. For instance, the adaptation might enable a low-power wearable worn by an elderly person to perform fall detection continuously, whereas such detection could have been done only periodically previously to ensure the battery would last throughout the day.

Nevertheless, the tuneable accelerator hardware is only one piece of the puzzle. With it in place, to start, we still need to quantify the relationship between the approximation levels and the resulting inference accuracy. Charting this relationship requires well designed validation datasets that closely mimic the inputs that will be observed in the wild. Second, we need to identify opportunities for approximation. What are the dimensions of the context that impact the accuracy of a (approximated) deep learning model? In ubiquitous computing these can vary widely – from a device's location, over noise levels, to outside brightness, input quality, and other factors. Moreover, the end user's requirements can vary with the context. For instance, a user may be willing to repeat a misclassified spoken keyword in one situation, but not in another. We believe that light models based on reinforcement learning can be used to develop the understanding of a user's context-dependent accuracy requirements. Finally, even if knowing whether a certain approximation fulfils a user's requirements at a particular point in time, we need a broader view of the resources and the expected context to ensure that the operation is driven towards the final goal, which could be, for example, "to ensure that the battery lasts until the next time the device is charging." For this, we envision novel model predictive control systems to be designed for approximate edge deep learning.

In this article, we argue that mobile deep learning needs a flexible hardware platform. We presented the accuracy-tuneable accelerator as a likely candidate and briefly surveyed the state of the art in this field, before detailing our recent research efforts founded in dynamically adaptable approximate multipliers. As we move beyond conceptualization, we look forward to solutions for the challenges towards the realization of accuracy-tuneable accelerator, as well as to innovative applications enabled by resource efficient adaptive learning on the edge. Thus, we use this opportunity to call the research community to join us in making deep learning truly ubiquitous. ■

Acknowledgements

This work was partly supported by the Slovenian Research Agency (research core funding No. P2-0098).

Ratko Pilipović received his BSc and MSc in electrical engineering from the Faculty of Electrical Engineering, University in Banjaluka, Bosnia, and Hercegovina in 2015 and 2017, respectively. In 2021, he obtained his PhD at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. He has published several conference articles and papers in prestigious journals. His research interests include approximate computing, arithmetic circuit design, FPGA design, embedded processing and machine vision. Currently, he is a research and teaching assistant at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

Veljko Pejović received his PhD in computer science from the University of California Santa Barbara. He is now an assistant professor at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, and leading research on mobile computing, focusing on resource-efficient mobile systems, human-computer interaction, and cybersecurity in ubiquitous systems. His awards include the best paper nomination at ACM UbiComp and the first prize at Orange D4D challenge for his work on epidemics modeling. For more information, visit http://lrss.fri.uni-lj.si/Veljko/

Octavian-Mihai Machidon obtained his PhD in electronic engineering and telecommunications from Transilvania University of Brasov, Romania. In 2020, he joined the Faculty of Computer and Information Science, University of Ljubljana, Slovenia, where he is currently an assistant professor and researcher, and also an adjunct research professor at the Faculty of Mathematics and Computer Science, Babes-Bolyai University of Cluj-Napoca, Romania. His current research focuses on context-adaptive energy-efficient mobile deep learning solutions. Other research interests include ubiquitous and pervasive computing, embedded systems and intelligent systems. He has published more than 50 peer-reviewed scientific papers in journals, conference proceedings, or book chapters and co-authored one patent.

REFERENCES

[1] M. Weiser, 1991. The computer for the 21st century. *Scientific American*, vol 265, no. 3, 94-105.

[2] B. Parilusyan, M. Teyssier, V. Martinez-Missir, C. Duhart and M. Serrano. 2022. Sensurfaces: A novel approach for embedded touch sensing on everyday surfaces. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, 1-19.

[3] C.-W. You, N.D. Lane, F. Chen, R. Wang, Z. Chen, T.J. Bao, M. Montes-de-Oca and et. al. 2013. Carsafe app: Alerting drowsy and distracted drivers using dual cameras on smartphones. 2013. *Proceedings of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, Taipei, Taiwan.

[4] T. Cejrowski, J. Szymański, H. Mora and D. Gil. 2018. Detection of the bee queen resence using sound analysis. *Asian Conference on Intelligent Information and Database Systems*., Dong Hoi City, Vietnam.

[5] J.L. Hennessy and D.A. Patterson. 2019. A new golden age for computer architecture. *Communications of the ACM*, vol. 62, no. 2, 48-60.

[6] O. Machidon and V. Pejovic. 2022. Energy-efficient adaptive keyword spotting using slimmable convolutional neural networks. *TinyML Summit*, San Francisco, CA.

[7] B. Zhao, J. Li, H. Pan and M. Wang. 2018. A high-performance reconfigurable accelerator for convolutional neural networks. *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*, Shenzhen, China.

[8] A. Firuzan, M. Modarressi and M. Reshadi. 2022. Reconfigurable network-on-chip-based convolutional neural network accelerator. *Journal of Systems Architecture*, 102567.

[9] T. Knez, O. Machidon and V. Pejović. 2021. Self-adaptive approximate mobile deep learning. *Electronics*, vol. 10, no. 23, 2958.

[10] R. Elangovan, S. Jain and A. Raghunathan. 2022. Ax-BxP: Approximate blocked computation for precision-reconfigurable deep neural network acceleration. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 27, no. 3, 1-20.

[11] L. Xuan, K.-F. Un, C.-S. Lam and R.P. Martins. 2022. An FPGA-based energy-efficient reconfigurable depthwise separable convolution accelerator for image recognition. *IEEE Transactions on Circuits and Systems II: Express Briefs*., vol. 69, no. 10, 4003-4007.

[12] O. Akbari, M. Kamal, A. Afzali-Kusha, M. Pedram and M. Shafique. 2019. X-CGRA: An energy-efficient approximate coarse-grained reconfigurable architecture. 2019. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 10, 2558-2571.

[13] Z. Babić, A. Avramović and P. Bulić. 2011. An iterative logarithmic multiplier. *Microprocessors and Microsystems*, vol. 35, no. 1, 23-33.

[14] R. Pilipović, V. Risojević, J. Božič, P. Bulić and U. Lotrič. 2021. An approximate GEMM unit for energy-efficient object detection. *Sensors*, vol. 21, no. 12, 4195.