Edge and Fog Computing Enabled AI for IoT -An Overview

Zhuo Zou[1], Yi Jin[1], Paavo Nevalainen[2], Yuxiang Huan[1], Jukka Heikkonen[2], Tomi Westerlund[2]
[1]Fudan University, Shanghai, China
[2]Dept. of Future Technologies, University of Turku, Turku, Finland
Email: {zhuo, jin_y16, yhuan13}@fudan.edu.cn; {paavo.nevalainen, jukhei, tovewe}@utu.fi

***Abstract*—In recent years, Artificial Intelligence (AI) has been widely deployed in a variety of business sectors and industries, yielding numbers of revolutionary applications and services that are primarily driven by high-performance computation and storage facilities in the cloud. On the other hand, embedding intelligence into edge devices is highly demanded by emerging applications such as autonomous systems, human-machine interactions, and the Internet of Things (IoT). In these applications, it is advantageous to process data near or at the source of data to improve energy & spectrum efficiency and security, and decrease latency. Although the computation capability of edge devices has increased tremendously during the past decade, it is still challenging to perform sophisticated AI algorithms in these resource-constrained edge devices, which calls for not only low-power chips for energy efficient processing at the edge but also a system-level framework to distribute resources and tasks along the edge-cloud continuum. In this overview, we summarize dedicated edge hardware for machine learning from embedded applications to sub-mW "always-on" IoT nodes. Recent advances of circuits and systems incorporating joint design of architectures and algorithms will be reviewed. Fog computing paradigm that enables processing at the edge while still offering the possibility to interact with the cloud will be covered, with focus on opportunities and challenges of exploiting fog computing in AI as a bridge between the edge device and the cloud.**

***Index Terms*—Internet of Things, Artificial Intelligence, Edge AI, Machine Learning, Fog computing, Edge computing, Embedded Processor**

I. Introduction

It is estimated that there will be over 30 billion connected devices on the Internet by 2020 [1]. These smart devices are attached to or embedded in physical objects with sensors, actuators, processors, and wireless connectivity, such as wearables, environmental sensors, home appliances, smart cameras, and vehicles [2]. This new era of the Internet of Things (IoT) enables a smart society by interconnecting cyberspace to the physical world. At the same time, artificial intelligence (AI), especially Machine Learning (ML) [3] are widely spread in a variety of business sectors and industries. Numbers of revolutionary applications in computer vision, game, speech recognition, medical diagnostics, and many others are reshaping our everyday life, which are primarily driven by the big data, advances in ML algorithms, and high-performance computation and storage facilities in the cloud.

Combing IoT and AI, namely the AI-enabled IoT, is the key pillar of realizing the vision of ubiquitous intelligence

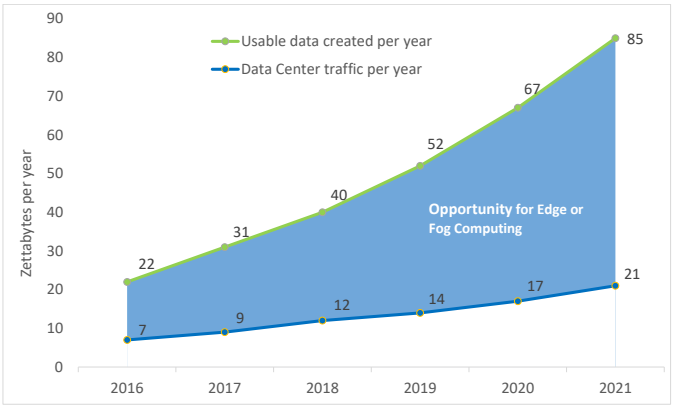

Fig. 1. The gap between data created by users/things and data processed by Cloud. (source: Cisco Global Cloud Index) [4]

[5]. Billions of IoT devices produce big data for ML, yielding intelligent decisions, analytics, and other added values of data as the return. On the other hand, massive amounts of data may overwhelm storage systems and crowd the wireless bandwidth. As shown in Fig. 1 from the Cisco Global Cloud Index [4], the gap between user- and device-created data and processed data is growing which brings new challenges and opportunities to the service architecture of the network as well as the hardware design. Processing near or at the source of data is highly demanded in many IoT applications. These IoT application can utilise local processing to i) improve spectrum efficiency for bandwidth-demanded applications such as surveillance cameras, ii) decrease the latency in time-sensitive applications such as human-robotic interactions, and iii) secure privacy and trustworthy such as in healthcare and medical applications. To this end, the new fog computing paradigm can be incorporated to provide a system-level framework to distribute resources and tasks along the edge-cloud continuum [6].

Edge devices and fog nodes are characterized by heterogeneity and specialty in terms of computational resource and tasks to be performed for given applications. Therefore, deploying data-intensive AI processing on edge devices and at the fog layer is relying on dedicated hardware. Recent advances in low-power ML processors exhibit orders of magnitude improvement in energy efficiency, which can be embedded in radio base station, local servers, routers, mobile platforms, gateways, and "always on" IoT sensors, as shown

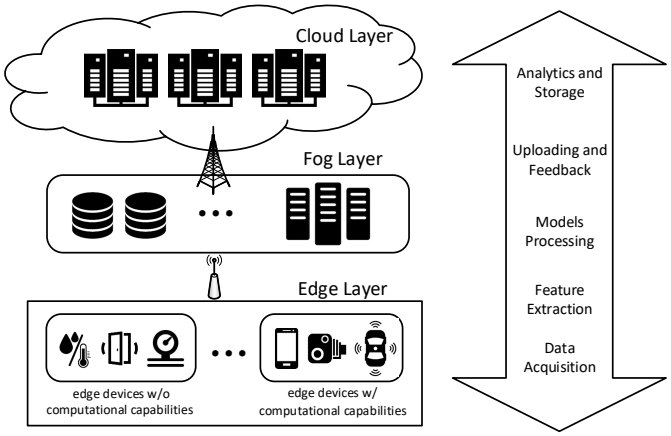

Fig. 2. A simplified architecture of edge-to-cloud network with fog computing

in Fig. 2. Such designs provide possibilities to map a variety of computing tasks along the edge-cloud network hierarchies to achieve different levels of intelligence at different costs and energy budgets.

In this overview, an introduction of edge and fog computing is provided in Section II, followed by related works of fog-enabled AI for IoT in Section III. Section IV briefly reviews state-of-the-art AI hardware. Then we conclude the paper in discussing the opportunities and future aspects of edge and fog enabled AI for IoT in Section V.

II. Computing at the Edge and Fog of the Network

Computing at the edge of the network has several names. The most common ones are Edge Computing (EC) and Fog Computing (FC). The latter is an umbrella for technologies to bring cloud computing capabilities to the edge of the network, whereas the former is one of the computing elements in FC. Although EC is a subset of the other, they are sometimes used interchangeably [7], [8] because they both are general terms defining computing at the network edge. Further ambiguity is caused by the lack of a universally accepted definition for EC [7], [9]. Where EC does not have a clear definition, FC has more solid one. F. Bonomi *et al*. defined FC as "an extension of the cloud platform from the core to the edge of the network" [10]. The definition has transformed overtime [11], [12], but its core has remained the same: geographically distributed resources located at the edge of a network for collaborative computation and communication. In this paper, we distinguish between edge and fog when appropriate. For simplicity, we define the "edge" as the end-device of the network, while "fog" as the intermediate nodes between edge and cloud.

Although FC has raised much attention from industry and academia, cloud computing is a viable solution, for example, when one needs scalable storage and processing services. In addition to scalability, cloud provides easier maintainability than distributed IoT based solutions can offer. However, in many latency sensitive applications, for example, in autonomous vehicles, the delay waiting for a result from a cloud-based AI is unacceptable. In addition, sending huge amount of data to the cloud for storage and processing might consume all network bandwidth making it a non-scalable and energy-hungry solution especially in image recognition. Given the continuously increasing amount of data, fog computing can increase the processing efficiency by distributed intelligence [13]. Fog computing reduces the data sizes and provides high quality data for further data analytics. In [14], the authors reported that fog computing system-based architecture has a better response time and generated less Internet traffic when compared to the cloud environment. A more comprehensive study reported that when the number of applications demanding real-time services, the fog computing paradigm outperforms traditional cloud computing providing 50 % decrease in the overall service latency [15]. In [16], the authors reported that with utilizing fog computing, wearable cognitive-assistance systems improves response times by between 80 and 200 ms and reduces energy consumption by 30 to 40 percent. The benefits of using fog computing are evident based on these reported results.

Specifications, reference models and architectures for FC are developed to gain benefits in larger scale. For example, the Open Connectivity Foundation announced in 2017 as a reference architecture for FC [17]. The reference architecture was targeted to industry to create and maintain fog computing elements in their applications. It provides different views to FC by providing deployment models, system architecture view as well as how containerization can be used for application support. Containerization can also be used in increasing flexibility of the FC layer by allowing live migration of containers horizontally at the edge level or vertically between the edge and cloud levels [18], [19]. A natural restriction in live migration is that the receiving fog node must have similar or better hardware support than the sending fog node.

Another active party is Open Connectivity Foundation that published specification for architecture, interfaces, protocols and services for IoT devices [20]. An open source reference implementation of OCF specification is provided by IoTivity. The reference implementation operates as middleware acting as a bridge across all operating systems and connectivity platforms. Combination of fog and cloud computing with a distributed data modeling at the sensor device for wireless sensor networks can be beneficial as reported in [21].

III. Fog Computing enabling ML in IoT

The ML models and their corresponding functions such as clustering, feature extraction, and classification for IoT data analytics have extensively investigated in [22], including k-means, k-nearest neighbors (k-NN), support vector machine (SVM), liner regression, and DNN. The resource constrained environment poses new problems to ML algorithm design. There are some generic distributable algorithms, *e.g.*, k-NN and some special neural network methods, which yield to distribution and resource constraints. The latter requires often laborious tuning of the granularity of the environment sensor data and the environment model, though. Early sensor fusion [23] using a 3D voxel environment model is one solution.

The fog-level IoT system results in faster and more efficient knowledge transfer to the cloud, and they also use edge intelligence to reduce the amount data transfer and storage need on the cloud.

In [24], the authors state that the future IoT platforms should include intelligence at the fog layer to enable local IoT networks to perform edge analytics. Also the importance of data management can be improved by fog-layer analytics and decision making as well as local storage. To make this happen, we need to have resource-efficient algorithms that consumes little energy and memory. In a recent work [25], Microsoft presented ML models that can run on tiny IoT devices. The reported the developed algorithm can be deployed on a Arduino Uno board. In [26], k-means is deployed in fog nodes using Intel Edison and Raspberry Pi for pathological speech data clustering. In [27], the authors presented a fog-level IoT system utilizing neural networks in analyzing sensor data. Disturbed DNNs (DDNNs) is proposed by Teerapittayanon *et al*. [28], that can scale up in neural network size and scale out in geographical span, and allows early exit points in a DNN. Data can be classified and exited locally when the system is confident and offloaded to the fog and the cloud when additional processing is required [29].

IV. Enabling Hardware for Edge and Fog AI

The hardware that supports a variety of AI algorithms is usually categorized as i) general-purpose computing platforms such as CPU, GPU, DSP, and FPGA; ii) customized AI processors or accelerators by application-specific circuits (ASICs) design; and iii) other novel architectures that often associate with emerging devices like neuromorphic architectures using memristors. Machine learning, especially Deep Learning (DL), is still the driving force among AI fields nowadays. Therefore, the efforts on hardware designs are mainly focused on improving the performance and efficiency for the computationally and memory-intensive Deep Neural Network and its derivatives such as the Convolutional Neural Network (CNN), the Fully-Connected Network (FCN), and the Recurrent Neural Network (RNN). Different Figure of Merits (FoMs) and optimization goals have been proposed, *e.g.*, GOPS/s for performance, GOPs/W for efficiency, operations/weights for the balance between ALU and memory bandwidth, etc. It is also widely accepted that training is performed on the cloud and inference can be executed locally alone the edge-cloud network.

It remains a challenge to perform sophisticated ML algorithms in resource and power constrained edge devices and fog nodes. On one hand, the intelligent algorithms become increasingly complex for improved accuracy. For example, the model size has increased by 16X from AlexNet to ResNet, with computational tasks increase from 1.4 GOPS to 22.6 GOPS for interface [52]. Commercially-available edge AI hardware [53], [54] still consumes multi-watt power even for inference. On the other hand, performance/watt is no longer natively scaled with the process technology due to the ending of Moore's Law and Dennard Scaling [55].

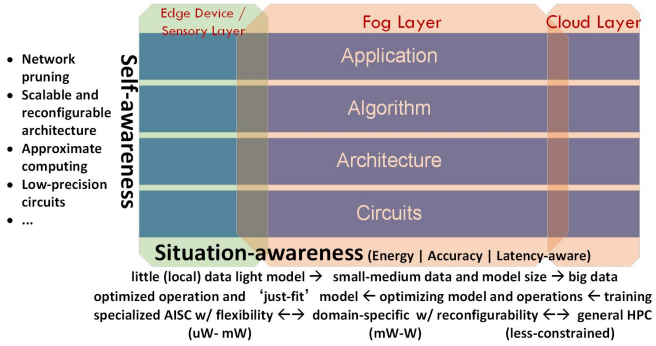

Fig. 3. A 2-D awareness and adaptivity design space.

Specialized hardware, especially the domain-specific architecture is therefore widely considered in embedded ML processors/accelerator designs. Several representative works exhibit orders of magnitude power reduction through model-architecture-circuits co-design, pushing the energy efficiency from 100s of GOPS/W up to 10s of TOPS/W in the sub-W power consumption region. They are basically exploiting redundancy and resiliency natures of neural networks to offload computational operations and data movements. For examples, EIE uses network pruning, weight sharing, and model compression to significantly simplify the network [52]; Eyersis introduces a row-stationary data flow with the NoC architecture to maximize data reuse [56]; ENVISION proposes the dynamic-precision SIMD architecture providing energy-precision scalabilities [57]. More recently, several low-power techniques are further explored aiming at sub-mW region power, which can potentially be deployed in "always-ON" IoT nodes while trading performance and generality to power. To list a few, low-precision quantization [38] and non-linear quantization [58], analog/mixed-signal computing with binary CNN [59] and CIM architecture [60], as well as time-domain neuromorphic chip for reinforced learning [61].They reduce the power consumption to the sub-mW region which can potentially be deployed in "always-ON" IoT nodes, while trading performance and generality to power. The current state of the arts of low-power ML processors are summarized in Table. I.

V. Discussion and Outlook

The convergence of IoT and AI relies on deploying data-intensive intelligent processing tasks along the cloud-to-thing continuum. Fog computing brings cloud features close to the edge of the network enabling local storage, data processing and analysis, which provides better support for time-critical and bandwidth-limited applications. On the other hand, there still exists a gap between hardware designers and network architects, *i.e.*, state-of-the-art hardware designs usually focus on efficiency as well as the related trade-offs and scalability between performance/accuracy/generality and power/cost, whereas fog computing platforms and applications normally consider the conventional general-purpose embedded platform

TABLE I
SUMMARY OF STATE OF THE ART ML PROCESSORS

Architecture	Algorithms for Intelligence	Tasks/Applications	Power/Energy Efficiency
Boltzmann machine (RBM) processor [30]	NN	Energy-efficient restricted RBM processor(MNIST)	310mW,1.45TOPS/W
Neuromorphic [31]	TDNN(BNN/CNN)	Apply TDNN technique to BNN(MNIST)	48.2 TSOP/S/W
DNN accelerator [32]	DNN	DNN classifier(MNIST)	33.7mW, 0.36 J/inference
DNN accelerator [33]	DNN	Fully-variable weight bit-precision(Alexnet/VGG-16)	3.2mW@0.63V/297mW@1.1V, 345.6GOPS@16b weights
DCNN processor [34]	CNN	CNN for intelligent embedded system (AlexNet)	39mW, 676GOPS(CAs) /76GOPS(DSP)
FC-DNN accelerator [35]	FC-DNN	Accelerator (MNIST) with fault tolerance	0.56μW/Decision
CNN processor (Eyeriss) [36]	DNN/CNN	Accelerator (Imagenet with Row Stationary)	278mW, 7.94mJ/Decision
CNN FR processor [37]	CNN	Face detection and recognition	620μW
CNN-RNN processor [38]	CNN, RNN, general purpose DNN	General purpose DNN (quantization-table-based multiplier)	63mW
SoC with a CNN accelerator [39]	CNN	IoT Edge Mote, image and enviromental data processing	sub-mW
Mixed-signal binary CNN processor [40]	Binary CNN	Image classification (CIFAR-10), near memory computing	899μW
Hybrid-NN processor (Thinker) [41]	CNN/FCN	Hybrid-NN (AlexNet/LRCN)	4-447mW, 409.6GOPS
Neuromorphic [42]	Embedded Reinforcement Learning: CNN/DNN/SVM	Mobile self-driving micro-robot at the edge of the cloud	690μW
Computing-in-memory [43]	CNN	Real-time 3D hand-gesture recognition processor	6.57mW, 11.8GOPS
Computing-in-memory [44]	CNN-based Machine Learning	LeNet-5 CNN	28.1 TOPS/W
Computing-in-memory [45]	SVM	In-memory machine learning classifier (SVM)	0.042nJ/Decision
Computing-in-memory [46]	binary DNN	Binary-based CIM-SRAM macro(MNIST)	55.8TOPS/W
Computing-in-memory [47]	Versatile DNN	Binary/Ternary reconfigurable in-memory DNN Accelerator	0.6W, 1.4TOPS
VAD system [48]	DNN	Voice activity detector (AFE and a digital BNN classifier)	1μW(voice detector)/ 0.38μW(feature extraction)
AI SoC [49]	DNN	3D-Stacked log-quantization for DNN inference (CNN/MLP/RNN)	3.3W, 1.96TOPS@4b weights
ConvNet processor [50]	CNN	Face recognition (VGG-16/AlexNet)	7.5-300mW, 10TOPS/W
DLA(deep learning accelerator) [51]	DNN	Keyword spotting and face detection	288μW

for hardware. Some possible directions that may further enhance fog/edge enabled AI are envisioned in Fig. 3 for discussion. The design space can be extended to 2 dimensions that are crossed coupled each other, *i.e.*, not only horizontally through edge-cloud optimization driven by situation-awareness to archive the overall quality-of-service (QoS), but also across different abstraction layers from a hardware perspective driven by the self-awareness method [62].

The fog architecture is expected to coordinate and support distributed intelligence over the edge devices and fog nodes collectively and cooperatively. Thus, algorithms with distributed natures will be of great interests, for example, from simple the k-Means [26] to more advanced DDNNs [28]. The new models are ideal to be i) computations already performed on lower-layer devices can be useful for further processing at higher layers, ii) supporting multiple models on multiple nodes that can be aggregated together for coordinated learnings, and iii) minimizing communication overhead in transferring intermediate data. Meanwhile, reinforced learning, transferred learning and incremental learning can be considered in fog nodes to further refine the operations of edge device for better accuracy and performance.

From a hardware perspective, more aggressively utilization of approximate computing and probabilistic computing [63] can be utilized when such a 2-D awareness and adaptivity platform is evolved. This will bring more synergic effects between the non-deterministic hardware and the resilience and fault-tolerance nature of bio-inspired approaches such as the

DNNs. For example, uncertainties generated by probabilistic computing platforms at low levels are recovered by either (both) higher abstraction levels and higher fog and cloud layers explicitly to meet the application goal. Meanwhile, the scalable and reconfigurable hardware will continue to be of interests. Specifically, the flexibility at low-level edge device is expected to support the run-time adaptation with the higher-layer fog nodes, and scalability will be a merit to accommodate the various scale of edge devices and fog nodes while providing agile application programming interfaces.

In response to these opportunities, this special session invites contributions from circuits and systems to applications, aiming to bring chip designers and system architects for edge-fog-cloud framework together, to discuss different approaches to edge and fog computing to enable AI in IoT applications.

ACKNOWLEDGMENT

This work was supported by NFSC under grant 61876039, Shanghai Pujiang Program (17PJ1400800), Shanghai Institute of Intelligent Electronics and Systems, and Shanghai Science and Technology Innovation Program (No. 17JC1401400).

REFERENCES

[1] Gartner, "Gartner says 8.4 billion connected "things" will be in use in 2017, up 31 percent from 2016," https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016, accessed: 2018-09-17.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[4] Cisco. Cisco Global Cloud Index: Forecast and Methodology, 2016–2021 White Paper. (2018, Sept. 10). [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.html

[5] L.-R. Zheng, H. Tenhunen, and Z. Zou, *Smart Electronic Systems: Heterogeneous Integration of Silicon and Printed Electronics*. John Wiley & Sons, 2018.

[6] B. Negash, T. Westerlund, and H. Tenhunen, "Towards an interoperable internet of things through a web of virtual things at the fog layer," *Future Generation Computer Systems*, vol. 91, pp. 96 – 107, 2019.

[7] B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D. S. Nikolopoulos, "Challenges and opportunities in edge computing," in *2016 IEEE International Conference on Smart Cloud (SmartCloud)*, 2016.

[8] W. Hu, Y. Gao, K. Ha, J. Wang, B. Amos, Z. Chen, P. Pillai, and M. Satyanarayanan, "Quantifying the impact of edge computing on mobile applications," in *Proceedings of the 7th ACM SIGOPS Asia-Pacific Workshop on Systems*, ser. APSys '16, 2016.

[9] L. Geng, M. Zhang, M. McBride, and B. Liu, "Problem statement of edge computing beyond access network for industrial iot," https://tools.ietf.org/pdf/draft-geng-iiot-edge-computing-problem-statement-00.pdf, accessed: 2018-09-29.

[10] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC '12. ACM, 2012.

[11] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, 2014.

[12] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proceedings of the 2015 Workshop on Mobile Big Data*, ser. Mobidata '15, 2015.

[13] W. Wang, S. De, Y. Zhou, X. Huang, and K. Moessner, "Distributed sensor data computing in smart city applications," in *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2017.

[14] Y. N. Krishnan, C. N. Bhagwat, and A. P. Utpat, "Fog computing — network based cloud computing," in *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, Feb 2015.

[15] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the suitability of fog computing in the context of internet of things," *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, Jan 2018.

[16] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, May 2016.

[17] OpenFog Consortium, "Openfog reference architecture for fog computing," https://www.openfogconsortium.org/wp-content/uploads/OpenFog_Reference_Architecture_2_09_17-FINAL.pdf, accessed: 2018-09-29.

[18] S. R. U. Kakakhel, L. Mukkala, T. Westerlund, and J. Plosila, "Virtualization at the network edge: A technology perspective," in *2018 Third International Conference on Fog and Mobile Edge Computing (FMEC)*, April 2018, pp. 87–92.

[19] Z. Tang, X. Zhou, F. Zhang, W. Jia, and W. Zhao, "Migration modeling and learning algorithms for containers in fog computing," *IEEE Transactions on Services Computing*, 2018.

[20] Open Connectivity Foundation, "Ocf specification 2.0," https://openconnectivity.org/developer/specifications, accessed: 2018-09-29.

[21] M. Lavassani, S. Forsström, U. Jennehag, and T. Zhang, "Combining fog computing with sensor mote machine learning for industrial iot," *Sensors*, vol. 18, no. 5, 2018.

[22] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, "Machine learning for internet of things data analysis: A survey," *Digital Communications and Networks*, 2017.

[23] X. Qin and Y. Gu, "Data fusion in the internet of things," vol. 15, pp. 3023–3026, 12 2011.

[24] J. Mineraud, O. Mazhelis, X. Su, and S. Tarkoma, "A gap analysis of internet-of-things platforms," *Computer Communications*, vol. 89-90, 2016.

[25] A. Kumar, S. Goyal, and M. Varma, "Resource-efficient machine learning in 2 kb ram for the internet of things," 2017. [Online]. Available: https://www.microsoft.com/en-us/research/publication/resource-efficient-machine-learning-2-kb-ram-internet-things/

[26] D. Borthakur, H. Dubey, N. Constant, L. Mahler, and K. Mankodiya, "Smart fog: Fog computing framework for unsupervised clustering analytics in wearable internet of things," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 472–476.

[27] H. M. Raafat, M. S. Hossain, E. Essa, S. Elmougy, A. S. Tolba, G. Muhammad, and A. Ghoneim, "Fog intelligence for real-time iot sensor data analytics," *IEEE Access*, vol. 5, 2017.

[28] S. Teerapittayanon, B. McDanel, and H. T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, June 2017, pp. 328–339.

[29] ——, "Branchynet: Fast inference via early exiting from deep neural networks," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 2464–2469.

[30] C. Tsai, W. Yu, W. H. Wong, and C. Lee, "A 41.3/26.7 pj per neuron weight rbm processor supporting on-chip learning/inference for iot applications," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 10, pp. 2601–2612, Oct 2017.

[31] D. Miyashita, S. Kousai, T. Suzuki, and J. Deguchi, "A neuromorphic chip optimized for deep learning and cmos technology with time-domain analog and digital mixed-signal processing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 10, pp. 2679–2689, Oct 2017.

[32] P. N. Whatmough, S. K. Lee, D. Brooks, and G. Wei, "Dnn engine: A 28-nm timing-error tolerant sparse deep neural network processor for iot applications," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 9, pp. 2722–2731, Sept 2018.

[33] J. Lee, C. Kim, S. Kang, D. Shin, S. Kim, and H. Yoo, "Unpu: A 50.6tops/w unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 218–220.

[34] G. Desoli, N. Chawla, T. Boesch, S. Singh, E. Guidetti, F. D. Ambroggi, T. Majo, P. Zambotti, M. Ayodhyawasi, H. Singh, and N. Aggarwal, "14.1 a 2.9tops/w deep convolutional neural network soc in fd-soi 28nm for intelligent embedded systems," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 238–239.

[35] P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, and G. Wei, "14.3 a 28nm soc with a 1.2ghz 568nj/prediction sparse deep-neural-network engine with gt;0.1 timing error rate tolerance for iot applications," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 242–243.

[36] Y. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan 2017.

[37] K. Bong, S. Choi, C. Kim, S. Kang, Y. Kim, and H. Yoo, "14.6 a 0.62mw ultra-low-power convolutional-neural-network face-recognition processor and a cis integrated with always-on haar-like face detector," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 248–249.

[38] D. Shin, J. Lee, J. Lee, and H. Yoo, "14.2 dnpu: An 8.1tops/w reconfigurable cnn-rnn processor for general-purpose deep neural networks," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 240–241.

[39] T. Karnik, D. Kurian, P. Aseron, R. Dorrance, E. Alpman, A. Nicoara, R. Popov, L. Azarenkov, M. Moiseev, L. Zhao, S. Ghosh, R. Misoczki, A. Gupta, M. Akhila, S. Muthukumar, S. Bhandari, Y. Satish, K. Jain, R. Flory, C. Kanthapanit, E. Quijano, B. Jackson, H. Luo, S. Kim, V. Vaidya, A. Elsherbini, R. Liu, F. Sheikh, O. Tickoo, I. Klotchkov, M. Sastry, S. Sun, M. Bhartiya, A. Srinivasan, Y. Hoskote, H. Wang, and V. De, "A cm-scale self-powered intelligent and secure iot edge mote featuring an ultra-low-power soc in 14nm tri-gate cmos," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 46–48.

[40] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8j/86in 28nm cmos," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 222–224.

[41] S. Yin, P. Ouyang, S. Tang, F. Tu, X. Li, S. Zheng, T. Lu, J. Gu, L. Liu, and S. Wei, "A high energy efficient reconfigurable hybrid neural network processor for deep learning applications," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 968–982, April 2018.

[42] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, "A 55nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 124–126.

[43] S. Choi, J. Lee, K. Lee, and H. Yoo, "A 9.02mw cnn-stereo-based real-time 3d hand-gesture recognition processor for smart mobile devices," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 220–222.

[44] A. Biswas and A. P. Chandrakasan, "Conv-ram: An energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 488–490.

[45] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42pj/decision 3.12tops/w robust in-memory machine learning classifier with on-chip training," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 490–492.

[46] W. Khwa, J. Chen, J. Li, X. Si, E. Yang, X. Sun, R. Liu, P. Chen, Q. Li, S. Yu, and M. Chang, "A 65nm 4kb algorithm-dependent computing-in-memory sram unit-macro with 2.3ns and 55.8tops/w fully parallel product-sum operation for binary dnn edge processors," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 496–498.

[47] K. Ando, K. Ueyoshi, K. Orimo, H. Yonekawa, S. Sato, H. Nakahara, S. Takamaeda-Yamazaki, M. Ikebe, T. Asai, T. Kuroda, and M. Motomura, "Brein memory: A single-chip binary/ternary reconfigurable in-

[49] K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, J. Kadomoto, T. Miyata, M. Hamada, T. Kuroda, and M. Motomura, "Quest: A 7.49tops multi-purpose log-quantized dnn inference engine stacked on 96mb 3d sram using inductive-coupling technology in 40nm cmos," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 216–218.

memory deep neural network accelerator achieving 1.4 tops at 0.6 w," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 983–994, April 2018.

[48] M. Yang, C. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "A 1w voice activity detector using analog feature extraction and digital deep neural network," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 346–348.

[50] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 246–247.

[51] S. Bang, J. Wang, Z. Li, C. Gao, Y. Kim, Q. Dong, Y. Chen, L. Fick, X. Sun, R. Dreslinski, T. Mudge, H. S. Kim, D. Blaauw, and D. Sylvester, "14.7 a 288µw programmable deep-learning processor with 270kb on-chip weight storage using non-uniform memory hierarchy for mobile intelligence," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2017, pp. 250–251.

[52] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "Eie: efficient inference engine on compressed deep neural network," in *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. IEEE, 2016, pp. 243–254.

[53] Nvidia. NVIDIA Jetson TX2 Module. (2018, Sept. 10). [Online]. Available: https://developer.nvidia.com/embedded/buy/jetson-tx2

[54] Google. Edge TPU. (2018, Sept. 10). [Online]. Available: https://cloud.google.com/edge-tpu/

[55] N. P. Jouppi, C. Young, N. Patil, and D. Patterson, "A domain-specific architecture for deep neural networks," *Communications of the ACM*, vol. 61, no. 9, pp. 50–59, 2018.

[56] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.

[57] B. Moons, R. Uytterhoeven, W. Dehaene, and M. Verhelst, "14.5 envision: A 0.26-to-10tops/w subword-parallel dynamic-voltage-accuracy-frequency-scalable convolutional neural network processor in 28nm fdsoi," in *Solid-State Circuits Conference (ISSCC), 2017 IEEE International*. IEEE, 2017, pp. 246–247.

[58] K. Ueyoshi, K. Ando, K. Hirose, S. Takamaeda-Yamazaki, J. Kadomoto, T. Miyata, M. Hamada, T. Kuroda, and M. Motomura, "Quest: A 7.49 tops multi-purpose log-quantized dnn inference engine stacked on 96mb 3d sram using inductive-coupling technology in 40nm cmos," in *Solid-State Circuits Conference-(ISSCC), 2018 IEEE International*. IEEE, 2018, pp. 216–218.

[59] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, "An always-on 3.8 μj/86% cifar-10 mixed-signal binary cnn processor with all memory on chip in 28nm cmos," in *Solid-State Circuits Conference-(ISSCC), 2018 IEEE International*. IEEE, 2018, pp. 222–224.

[60] A. Biswas and A. P. Chandrakasan, "Conv-ram: An energy-efficient sram with embedded convolution computation for low-power cnn-based machine learning applications," in *Solid-State Circuits Conference-(ISSCC), 2018 IEEE International*. IEEE, 2018, pp. 488–490.

[61] A. Amravati, S. B. Nasir, S. Thangadurai, I. Yoon, and A. Raychowdhury, "A 55nm time-domain mixed-signal neuromorphic accelerator with stochastic synapses and embedded reinforcement learning for autonomous micro-robots," in *Solid-State Circuits Conference-(ISSCC), 2018 IEEE International*. IEEE, 2018, pp. 124–126.

[62] A. Jantsch, N. Dutt, and A. M. Rahmani, "Self-awareness in systems on chip – a survey," *IEEE Design Test*, vol. 34, no. 6, pp. 1–19, December 2017. [Online]. Available: http://jantsch.se/AxelJantsch/papers/2017/AxelJantsch-DesignTest.pdf

[63] M. Alioto, V. De, and A. Marongiu, "Guest editorial energy-quality scalable circuits and systems for sensing and computing: From approximate to communication-inspired and learning-based," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 361–368, 2018.