Edge computing platforms for AI-driven services: from enhanced security to the semiconductor industry

M. Perea,[1] J.A. Álvarez-García,[1] O. Deniz,[2] P. Szanto,[3] L. Liss,[4] J.J. Vegas Olmos[4], O. Bouchmal[5], I. Tafur Monroy[5]

[1] Universidad de Sevilla, Dept. Lenguaje y Sistemas Informáticos, Sevilla, Spain
[2] University of Castilla-La Mancha, VISILAB, E.T.S.I.I, Ciudad Real, Spain
[3] Faculty of Electrical Engineering and Informatics, Budapest University, Budapest, Hungary
[4] Software Architecture NVIDIA Corporation, Yokneam, Israel
[5] Eindhoven University of Technology, Department of Electrical Engineering, Eindhoven, The Netherlands

This paper presents user scenarios that benefit from edge computing platforms and the ability to offload artificial intelligence intense processes from the end user subsystems. Then, we briefly describe one of the main challenges in edge computing, the parametrization of the resources in terms of processing power and network capacity, to effectively execute services and applications.

I. Introduction

In the first decades of the 21st century, through developments in photonic technologies being integrated into transceivers and co-processors, linerate processing capacity and transmission capacity have been finally balanced. This is leading to a massive introduction of artificial intelligence (AI), which will transform most aspects of the end-users ecosystems in a variety of verticals, from smart-cities, automobile, factory 4.0, healthcare to mobility. These developments are forcing networks to complete a full circle: as the initial internet was local-based with small-servers in-house, we migrated to cloud services massively over the last 20 years. However, requirements of AI processes in terms of processing and latency needs and ever-growing bandwidth increase in the communication systems are pushing back to local servers feeding the users. This is now called edge computing.

As a result, networks can be now split in three segments: cloud computing (for intensive processing or storage service), edge computing (light AI and services) and artificial-internet-of-things (AIoT) (decentralized communications). In particular, technology paradigms that require real-time processing and low reaction times (i.e. autonomous vehicles, electric grid control) are particularly interested in having processing power closer in terms of networking cycles. Hence, the effective deployment of AI services and applications should now be measured not only in terms of processing power available but also in terms of efficient networked access to such resources.

This paper provides an overview of some of the key use cases being developed in the contexts of European and national projects developing edge computing enabling AI. Then, it will elaborate on one of the main challenges facing the community at the crossroads between AI researchers and network researchers, namely the parametrization of needed resources to effectively execute the services and applications offered to the users.

II. Architectural reference for Deep Learning Edge Computing

To meet the computational requirements of Deep Learning (DL), a common approach is to leverage cloud computing, but moving the data from the source to the cloud introduces several challenges that normally cannot satisfy real-time requirements: latency (queuing and propagation delays), scalability (network access can become a bottleneck) and privacy (sensitive data to the cloud) [1].

Edge computing [2] is becoming the standard computing paradigm for latency-sensitive real-time IoT workloads since it addresses the aforementioned limitations related to centralized cloud-computing models. Such a paradigm brings the service and utilities of cloud computing closer to the end user and is characterized by quick processing and fast application response time, solving cloud computing issues: latency is decreased given the proximity to data sources (enabling real-time services), and scalability bottlenecks are eliminated using a hierarchical architecture avoiding process thousands of devices data per node and privacy concerns are reduced passing the information to the local trusted edge server, thus avoiding traversal of the public Internet.

On the other hand, a new challenge appears: how to accommodate the high resource requirements of DL on less powerful edge compute resources, how to allocate and dimension the hierarchical architecture to serve an adequate number of edge devices to satisfy the client requirements (i.e. frames per second, real-time response, etc.), and what level of Edge intelligence [3] is needed, or how training and inference are shared between device, edge and cloud maintaining privacy.

Therefore, edge computing emerges as an attractive alternative to cloud computing, without hindering the possibility of edge and cloud co-existing: some works have combined edge computing with cloud computing, resulting in hybrid edge-cloud architectures. The main advantages of this are the backbone network relief, distributed edge computing nodes that can handle many computation tasks, the agile service response,

reducing the delay of data transmissions and improving the speed of response by edge-hosted services, and strong cloud backup, which provides powerful processing and massive storage capabilities when the edge cannot afford it.

III. Use cases

This section describes use cases that are currently being evaluated in BRAINE [4] and ID2PPAC European Projects and VICTORY national project [5].

A. Hyperconnected smart cities

The interactions between Internet of Things (IoT) devices and new smart sensors are growing exponentially [6], creating massive and continuous flows of events from the most diverse origins. Edge computing allows implementing a Smart city Edge infrastructure where various sensor data is processed, and services based on this data are offered by multiple tenants. By creating multi-tenant platforms for smart service providers costs can be saved on private single-tenant infrastructure, data can be shared among service providers to develop new services and most importantly privacy and data sovereignty can be safeguarded on the platform compliant. AI-assisted edge computing platforms are able to anticipate changes in the city (for example weather conditions, public events, emergencies, etc.) and in line with those changes redistribute instances or place new instances on the camera's infrastructure. This adds pressure on the edge

computing platforms, that need to be able to process large volumes of audio/video from HW resources with different constraints (QoS, Latency, resolution), from different services (logistics planning, city space planning, pollution, security and emergency response, crowd management, retail, infrastructure maintenance, etc.).

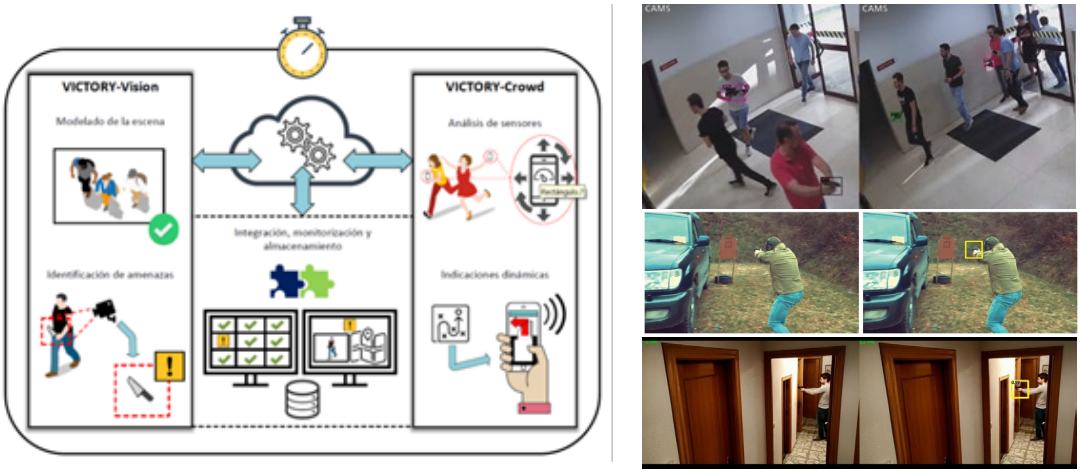

Figure 1. (Left) An edge computing platform conducts live processing for scene modelling, weapon and threats identification, crowd analysis in emergency situations and dynamic emergency signals and instructions to users. (Right) VICTORY results. Top: study and dataset related to CCTV handgun detection [7]. Middle [8] and bottom [9]: Combination of pose and object detection (left – only object detection, right – proposed in VICTORY).

A representative example of this paradigm is the surveillance system depicted in Figure 4, developed as part of a national project in Spain. The VICTORY system (VIsion and Crowdsensing Technology for an Optimal Response in physical-securitY) combines actionable data obtained from surveillance cameras and wearable sensors (smartphones, smartwatches, etc.) to provide rapid detection of dangerous situations inside public buildings. When smartphones are used, instructions can be sent so that people can evacuate through less risky routes. Such a system has developed advanced AI-based techniques for detecting the presence of handguns in the image. The technique is based not only on the appearance of the handgun but also on the pose of the individual, which may be more visible than the gun itself, see Figure 5.

B. Semiconductor sector digitalization

The recent introduction of AIoT integrated circuit devices has led to an additional need to produce small series of specialized circuitry devices with a very short life cycle compared to past technologies. Therefore, the beyond state-of-the-art industry 4.0 digitization innovations for semiconductor manufacturing are associated to the electronic components and systems end markets which require cost reductions and in consequence: 1) Re-define Moore's law from spatial shrinking to novel 3D technology and materials, 2) The requirement for cycle time reduction and, 3) Improvement of the process and metrology equipment throughput, reproducibility, accuracy and efficiency.
Hence, the semiconductor sector is now transitioning from in-line metrology (fabricate, AI-process, discard) to real-life metrology (AI-process while fabricating, with zero production waste). This migration requires a move from cloud computing to almost

embedded edge computing, where data from integration equipment is feed life and processes in-line. ID2PPAC, led by ASML in Eindhoven is examining together with NVIDIA and TUE how to integrate AI-enabling technologies into the production lines for semiconductors.

IV. Future work

As it can be perceived, the calculus to dimension the number of processing units to deal with scalability is not a trivial problem and the literature shows the implementation of specific situations, frameworks or even use case systems, but normally this calculus is overestimated in order to avoid bottlenecks in computation process but, these bottlenecks can appear due to other reasons such as the DL model selected. Furthermore, there exist some techniques such as operators fusion, quantization or optimization stacks that can be included to improve the computation power for DL models. Is here where we would like to advance the state-of-the-art, generating an Edge DL computing formula to include all the possible variables and obtain the best fit for each problem and configuration.

Acknowledgment

This work was partly funded under the BRAINE project ('Edge computing and AI technology for big data processing'- Horizon 2020 framework Grant Agreement 876967), ID2PPAC ('Integration of processes and moDules for the 2 nm node meeting Power Performance Area and Cost requirements' - Horizon 2020 framework ECSEL Grant Agreement 101007254), MADEin4 ('Metrology Advances for Digitized ECS industry 4.0' – Horizon 2020 framework ECSEL Grant Agreement 826589) and DISARM project ('Automatic Detection of Armed Individuals' - Grant n. PDC2021-121197, funded by MCIN/AEI/ 10.13039/501100011033, IoTalentum project (EU Horizon 2020 Marie Skłodowska-Curie European Training Network on Internet of Things with grant number 953442) and by the "European Union NextGenerationEU/PRTR").

References

[1] J. Chen et al., "Deep Learning With Edge Computing: A Review." Proceedings of the IEEE 1655-1674, 2019.
[2] M. Satyanarayanan, "The emergence of edge computing." Computer 50.1 (2017): 30-39.
[3] Z. Zhou et al.,"Edge intelligence: Paving the last mile of artificial intelligence with edge computing". Proceedings of the IEE, 1738-1762, 2019.
[4] F. Paolucci, et al. "Industrial Cyber Security at de Network Edge: The BRAINE Project Approach." Cybersecurity workshop by European Steel Technology Platform (2020): 119-131.
[5] F. Enríquez, et al. "Vision and crowdsensing technology for an optimal response in physical-security." International Conference on Computational Science (2019): 15-26.
[6] Y. Perwej, et al. "The internet of things (IoT) and its application domains." International Journal of Computer Applications 975.8887 (2019): 182.
[7] J.L. Salazar, et al. "Real-time gun detection in CCTV: An open problem." Neural networks 132 (2020): 297-308.
[8] J. Ruiz-Santaquiteria, et al. "Handgun detection using combined human pose and weapon appearance." IEEE Access 9 (2021): 123815-123826.
[9] A. Velasco-Mata, et al. "Using human pose information for handgun detection." Neural Computing and Applications 33.24 (2021): 17273-17286.