Published in the proceedings of 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), 29-31 October 2019, London, United Kingdom, which should be cited to refer to this work.
DOI: https://doi.org/10.1109/IPCCC47392.2019.8958742

IoT meets distributed AI – Deployment scenarios of Bonseyes AI applications on FIWARE

1st Lucien Moor
Haute Ecole Arc
HES-SO
St-Imier, Switzerland
lucien.moor@he-arc.ch

2nd Lukas Bitter
Haute Ecole Arc
HES-SO
St-Imier, Switzerland
lukas.bitter@he-arc.ch

3rd Miguel De Prado
Haute Ecole Arc
HES-SO
St-Imier, Switzerland
miguel.deprado@he-arc.ch

4th Nuria Pazos
Haute Ecole Arc
HES-SO
St-Imier, Switzerland
nuria.pazos@he-arc.ch

4th Nabil Ouerhani
Haute Ecole Arc
HES-SO
St-Imier, Switzerland
nabil.ouerhani@he-arc.ch

***Abstract*—Bonseyes is an Artificial Intelligence (AI) platform composed of a Data Marketplace, a Deep Learning Toolbox, and Developer Reference Platforms with the aim of facilitating tech and non-tech companies a rapid adoption of AI as an enabler for their business. Bonseyes provides methods and tools to speed up the development and deployment of AI solutions on low power Internet of Things (IoT) devices, embedded computing systems, and data centre servers. In this work, we address the deployment and the integration of Bonseyes AI applications in a wider enterprise application landscape involving different applications and services. We leverage the well-established IoT platform FIWARE to integrate the Bonseyes AI applications into an enterprise ecosystem. This paper presents two AI application deployment and integration scenarios using FIWARE. The first scenario addresses use cases where edge devices have enough compute power to run the AI applications and there is only need to transmit the results to the enterprise ecosystem. The second scenario copes with use cases where an edge device may delegate most of the computation to an external/cloud server. Further, we employ FIWARE IoT Agent generic enabler to manage all edge devices related to Bonseyes AI applications. Both scenarios have been validated on concrete use cases and demonstrators.**

***Index Terms*—Artificial Intelligence, Machine Learning, Internet of Things, Edge Computing**

I. Bonseyes - Distributed AI Platform

Artificial Intelligence (AI) is literally pervading all industrial sectors, from finance to retails and even to manufacturing. Deep Learning, a branch of AI, has been an important enabler in this spectacular evolution of AI systems. However, most of the innovations and value creation in the field of Deep Learning systems are dominated by big players like Google, Facebook, Amazon, Apple and IBM. Two reasons among many others could explain this monopoly. First, Deep Learning based systems need huge amount of training data in order to achieve acceptable accuracy and performance. Secondly, designing and tuning Deep Learning based systems, which can be deployed for real world applications, requires considerable effort.

This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 732204 (Bonseyes). This work is supported by the Swiss State Secretariat for Education Research and Innovation (SERI) under contract number 16.0159. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies.

The Bonseyes project [1], funded by the EU commission through its Horizon 2020 Research and Innovation Program, has the objective to address these challenges in order to allow also small players to build Deep Learning based systems by providing a data market place for data and models (Fig. 1). It also by decreasing the needed development and deployment effort on a variety of computing platforms such as low power Internet of Things (IoT) devices (”edge computing”), embedded computing systems, and data centre servers (”cloud computing”) [2].

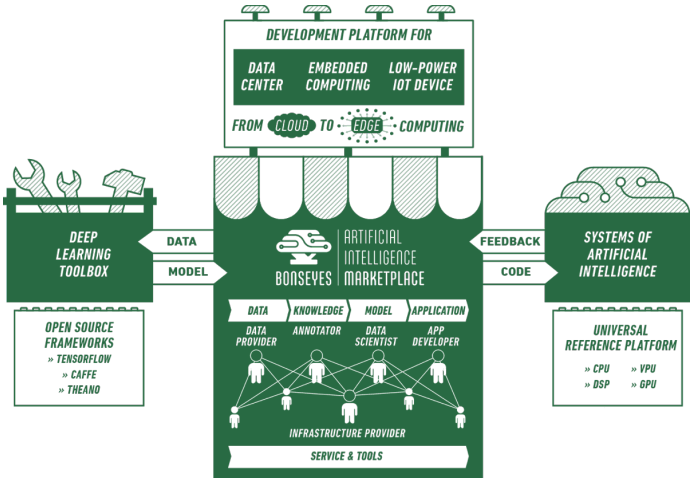

Fig. 1. Bonseyes general concept (from project web site: www.bonseyes.eu/).

Concretely, Bonseyes leverages an end-to-end AI pipeline concept [3] to address the three major embedded AI challenges, namely 1) data collection and annotation, 2) model training, and 3) deployment on resource constrained platforms. The third challenge, which is the most relevant for this work, is addressed by Bonseyes through LPDNN (Low Power Deep Neural Network). LPDNN is an inference engine optimizer aiming at accelerating the deployment of neural networks on embedded devices [4]. The main component of LPDNN is an inference module whose role is to generate efficient and portable code for DNN models across the wide range of compute platforms.

The rest of the paper describes how the generated LPDNN

inference engine can be integrated into a broader application ecosystem, in particular using IoT architectures.

II. FIWARE - The EU reference platform for IoT

FIWARE [5] is an independent open community with the mission of building an open and sustainable IoT ecosystem with public and royalty free access to facilitate the development of new smart applications in multiple fields such as industry, smart cities, manufacturing, media, and agrifood. The following FIWARE Generic Enablers (GE) have been used in our project to deploy LPDNN-based AI applications:

- Orion context broker: is the backbone of any FIWARE project. It allows to record descriptors of the entities of the system and holds their context. Each Bonseyes edge device is considered as an Orion entity providing data.
- Short Term Historic (STH): manages and stores the history of context changes related to entities e.g. the Bonseyes edge devices.
- IoT Agent: Bonseyes edge devices need to collect data from various sources with different acquisition mechanisms and with different communication protocols. The FIWARE IoT Agent facilitates data collection from heterogeneous sources. The related FIWARE backend IoT Device Management GE ensures the remote management of the edge devices.
- Kurento Media Server: helps the development of web-based media applications for browsers or smartphones by providing ready to use bricks of media processing algorithms such as computer vision, video indexing, augmented reality and speech analysis. In the Bonseyes project, we created a dedicated media processing module capable of calling a Deep Learning inference engine.
- Wirecloud: Wirecloud is an end-user centred web application mashup platform that allows to create web applications and dashboards/cockpits to visualize data of interest or to control IoT environments. It is used by Bonseyes to develop the user interfaces of the applications.

III. AI applications deployment scenarios

Distributed AI, as promoted by Bonseyes, is expected to unlock the innovation potential in many industries. However, it brings also a set of challenges like dealing with a very heterogeneous set of edge devices. The difference between edge devices in terms of computation power, battery autonomy, network bandwidth, etc makes it almost impossible to develop AI applications fitting to all edge devices. To cope with this situation, we propose a flexible AI application deployment concept that can deal with various usage scenarios (Fig 2).

A. Edge processing scenario

In this scenario, we are working with devices providing a high computational power (i.e. nvidia Xavier) allowing an embedded edge processing. The benefits of this situation is a light payload exchange between the device and the context broker because only the valuable results are sent. The downside is the requirement of a high capability device, that may consume more energy. We developed two mechanisms in order to control the AI application running on the edge device:.

- The edge device acts as an autonomous system controlling the start/stop and processing of the deployed AI application.
- The edge device running the AI application is controlled by a remote service managed by the FIWARE IoT Device Management GE.

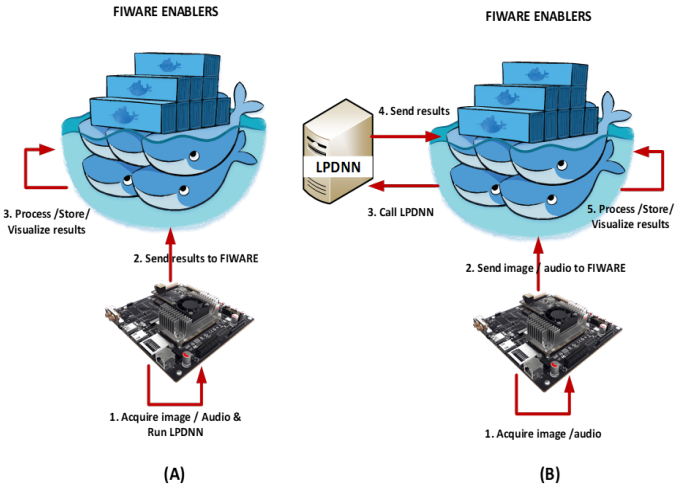

Fig. 2. LPDNN deployment scenarios. (A) Edge processing scenario and (B) Server/Cloud processing scenario

B. Server/cloud processing scenario

In this scenario we are using a low capability edge device that is not able to run complex algorithms efficiently. In this case, the data has to be sent to a server where LPDNN is deployed. The data can be transmitted in different manners depending on the use case. If there is a close to real-time requirement, it is possible to set up a media stream between the processing server and the edge device using Kurento media server. If there is no specific requirement, the image could be sent over FTP or other data transmission protocol.

References

[1] T. Llewellynn, M. M. Fernández-Carrobles, O. Deniz, S. Fricker, A. Storkey, N. Pazos, G. Velikic, K. Leufgen, R. Dahyot, S. Koller, G. Goumas, P. Leitner, G. Dasika, L. Wang, and K. Tutschku, "Bonseyes: Platform for open development of systems of artificial intelligence: Invited paper," in *Proceedings of the Computing Frontiers Conference*, ser. CF'17. New York, NY, USA: ACM, 2017, pp. 299–304. [Online]. Available: http://doi.acm.org/10.1145/3075564.3076259

[2] N. Ouerhani, S. Carola, M. D. Pardo, L. Bitter, L. Moor, F. Tiche, and N. Pazos, "Hybrid and flexible computing architectures for deep learning systems," in *Proceeding Zoom Innovation on Consumer Electronics (ZINC)*, 2017.

[3] M. de Prado, J. Su, R. Dahyot, R. Saeed, L. Keller, and N. Vállez, "AI pipeline - bringing AI to you. end-to-end integration of data, algorithms and deployment tools," in *CoRR*, vol. abs/1901.05049, 2019. [Online]. Available: http://arxiv.org/abs/1901.05049

[4] M. de Prado, M. Denna, L. Benini, and N. Pazos, "Quenn: Quantization engine for low-power neural networks," in *Proceedings of the 15th ACM International Conference on Computing Frontiers*. ACM, 2018, pp. 36–44.

[5] "Fi-ware project," 2018. [Online]. Available: https://www.fiware.org/