Accelerating Video Analytic Processing on Edge Intelligence

Pedro Fernández, Jaime Jiménez and Armando Astarloa
Departamento de Tecnología Electrónica
University of the Basque Country (UPV/EHU)
Bilbao, Spain
armando.astarloa@ehu.eus

Mikel Idirin and Sergio Salas
System-on-Chip *engineering*
Ribera de Axpe 50 Erandio, Spain
mikel.idirin@soc-e.com

***Abstract*—The most demanding Artificial Video analytic applications require in-edge inference of the AI model to ensure low-latencies to obtain the result. In general, the training process of the model can be executed on the cloud taking benefit from the high-performance computing capabilities available on those premises.**

This work presents an AI Video analytic application implemented on an Edge-computing device. This device is capable of accelerating the inference of AI models and Video compression by dedicated hardware.

This paper presents the architecture designed to implement Image, Networking and Deep Learning Processing functionalities on a reconfigurable System-on-Chip. Additionally, the design tools and design flow followed to generate all software and hardware configuration is detailed.

This Edge Intelligence platform is currently in-service, providing the preliminary results for the targeted applications. The proposed solution can process 33 times more video data volume in real-time than the software GPU accelerated implementation for the testing conditions described in the paper.

***Index Terms*—AI, NN, DNN, CNN, , FCN, RNN, DPU, SOC**

I. Introduction

Edge Intelligence (EI) combines edge computing and Artificial Intelligence (AI). Real-time video analytics is a killer application for edge computing [1], [2], [3]. Critical systems like automotive or Aerospace&Defence demand low latency in the analysis of the videos. Additionally, the resolution of these video sources increases continuously. Thus, the traditional AI Cloud-based approach requires higher bandwidth and speed networking, making this approach inviable for many applications. One viable approach is edge computing with dedicated hardware acceleration for Video coding, decoding, and Deep Learning Processing units to accelerate AI computing.

The most demanding AI Video analytic applications require in-edge inference of the AI model to ensure low-latencies to obtain the result. In general, the training process of the model can be executed on the cloud taking benefit from the high-performance computing capabilities available in those premises.

This work was partially supported by ECSEL Joint Undertaking in H2020 project IMOCO4.E, grant agreement No.10100731 and by the Department of Education of the Basque Government within the fund for research groups of the Basque university system IT1440-22.

This work presents an AI Video analytic application implemented on an Edge-computing device that specifically targets real-time object detection, in this case face detection. The main point of interest is to demonstrate how this device is capable of accelerating the inference of AI models and Video compression by dedicated hardware.

In Section II an introduction to Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs) and Densely Connected Convolutional Networks (DenseNet) is presented. These models are the base for the AI application implemented. The hardware acceleration is achieved using FPGA technology embedded on a reconfigurable SoC. Section III-B introduces briefly the use of these technology for this purpose.

Section IV details the SoC architecture implemented to run the accelerated AI Video analytic application. In this section, the AI/ML Design Flow followed to implement the design, the High level Architecture and the Data Flow path inside the SoC are presented.

Section V summarizes the preliminary results obtained in the set-up and the paper ends with the conclusions and future work in Section VI.

II. State-of-the-art

A. Deep Neural Networks (DNNs)

Deep Neural Networks (DNNs) are currently the base for many modern Artificial Intelligence (AI) applications [4]–[6]. DNNs are employed in applications from selfdriving cars [7], tumor detection [8] to gaming [9]. In many of these domains, DNNs are now able to exceed human accuracy. DNNs are able to extract high-level features from raw sensory data after using statistical learning over a large amount of data to obtain an effective representation of an input space.

B. Convolutional Neural Networks (CNNs)

A weight-shared and windowed DNN layer implements the computation as a convolution. The weighted sum for each output activation is computed using a small neighborhood of input activations and the same set of weights are shared for every output. This is named as the 'receptive field'. These convolution-based layers are referred to as convolutional (CONV) layers.