Abstract

The research activity described in this thesis aims to demonstrate the possibility to embed Artificial Intelligence (AI) capabilities in wearable and portable devices by deploying and executing Neural Network (NN) models close to the sensing element. Among AI models, Deep Learning (DL) and Deep Neural Networks can achieve high performance in many tasks, e.g. image classification, activity recognition, and so on. However, DL models usually require a huge amount of memory resources and high-performance digital architecture to be executed. These specifications are hardly met by wearable and portable devices, which have to be as small as possible and guarantee a satisfactory battery lifetime. For this reason, the cloud computing strategy is often used. However, higher latencies occur in this case, which can be unacceptable in many latency-sensitive applications, such as autonomous vehicles or assisted microsurgery. Moreover, the data transfer consumes network bandwidth and energy. In this context, moving the computation close to the device is highly demanded, and it is named edge-computing. However, deploying DL models on edge devices is still a challenge. General-purpose platforms (i.e. CPUs, GPUs) are not the best solution in terms of energy efficiency, especially for wearable and battery-powered devices, where the device lifetime is a major concern. Thus, a lot of research is being made about the design of custom HW accelerators for DL and to move the circuitry needed to implement the computation closer to the sensing element, thus obtaining a smart sensor. In this thesis, a novel Hybrid Binary Neural Network (HBN) model is proposed, which exploits the advantages of Binarized Neural Networks (BNNs). Human Activity Recognition (HAR) based on inertial sensors has been selected as a case study. Also, a pre-processing algorithm has been developed to solve the device-orientation problem for 3-axis accelerometers. The pre-processing operations can improve the accuracy of the proposed system in some conditions when it is used in conjunction with the HBN model. The results show an accuracy of up to 99% in recognizing 5 human activities. After having developed the model, a custom ultra-low power HW accelerator has been designed and implemented with both FPGA and CMOS standard cells. Due to the very low operating frequency required by HAR applications, power consumption has been reduced by reducing the number of resources. The design can implement both the pre-processing operation and the HBN model. The results show that the HW accelerator has a power consumption of 6.3 μW and an area occupation of 0.20 mm2 when synthesized with CMOS 65 nm Low-Power (LP) High Voltage Threshold (HVT) standard cells. The proposed design has at least 7.3 times lower power consumption than the state-of-the-art solution. Also, a FPGA-based demo board has been developed to demonstrate the real-time operation of the system.