Universal CNN Accelerator Intended for Edge-Based AI Inference

Prof. Dr. Rastislav Struharik

University of Novi Sad und IDS Imaging Development Systems d.o.o.

Deep learning, and particularly Convolutional Neural Networks (CNNs), are currently one of the most intensively and widely used machine learning predictive models. CNNs are not a new concept, but after recent breakthrough applications in the fields of image processing, and speech recognition, they have returned to the academic and industrial focus. Today, different types of CNNs are being employed in a wide range of applications, ranging from autonomous driving, medical, and even to playing complex games. In many of these application domains, CNNs are now able to exceed human levels of performance.

However, the superior accuracy of CNNs comes at a high cost because of their computational and storage complexity. State-of-the-art CNNs are described by hundreds of millions of parameters and require billions of computations in order to classify single input instance. It is highly likely that future CNNs will be even larger, deeper, will process larger input instances, requiring even more computations per input instance, and will be used to perform more intricate classification tasks at faster speeds, ever-increasingly in real-time, within low-power operating conditions. Because of this, careful selection of appropriate computing platform for the implementation of CNN-based applications is of great importance. This becomes even more important if we are to deploy CNNs in edge-based applications.

In this talk we will present hardware options available for implementing CNN acceleration on an edge device and discuss what are their weak and strong points. We will also present IDS deep ocean core, FPGA-based CNN hardware accelerator intended for edge-based CNN inference. We will discuss basic operating principles of it, present some use cases, and compare its performance to several competing solutions.