Mobile NPUs for Intelligent Human Computer Interaction

Hoi-Jun Yoo, KAIST

Abstract: Recently, Deep Neural Networks are widely used for realization of the AI services in cyber and physical world. In this presentation, firstly, the status of AI and DNN SoCs will be reviewed from the viewpoint of the mobile and edge AIs, and the evolution of DNN Accelerators. Especially, mobile, embedded and IoT deep learning hardware, low power NPU and reconfigurable NPU will be introduced. In addition, “Dynamically Reconfigurable Processor” architecture will be explained in detail with the real chip measurement results, such as human emotion recognition for intelligent HCI. Secondly, the On-Chip Training, which will be the next wave over the current AI revolution, will be explained for personalized and privacy protected AI applications. The personalization and autonomous adaptation to the environmental changes are possible with the on-chip supervised and reinforcement learning capability. Low power training processors will be explained with algorithm and hardware co-optimization methods. Real-time training, GAN and DRL examples will be introduced with dedicated training chips which are implemented for the many exciting low-power and high-performance applications such as object recognition and the object tracking applications in Gaming, AR/VR, Intelligent Robotics, Drones, Autonomous Driving, Security, Camera, Health Monitoring and IoT.