Deep Model Compression and Acceleration toward On-Device AI
Changkyu Choi, Lab Director, Vice President, Computer Vision Lab
Samsung Advanced Institute of Technology (SAIT) / Samsung Electronics Co., Ltd.

After experiencing significant performance improvements in the major challenges in images and speeches, global IT companies have provided deep learning frameworks to gather a large pool of researchers and developers and tried to expand AI business using Cloud services. As services are getting popular, efforts have been made to reduce enormous Cloud costs, and as performance improves, applications where security and recognition speed matter are becoming more and more important. In line with these efforts the on-device AI technology, which is executed on the terminal or edge, has been attracting attention. This is a great opportunity for hardware companies producing processors, memories and sensors.

This talk will focus on the challenges and opportunities in on-device AI. Algorithms devoted to deep model compression and schemes elaborated to accelerate the massive computation will be discussed. Low energy consumption by balancing the load of computation between the sensors and processors is another crucial aspect to be shared.