From small to tiny: How to co-design ML models, computational precision and circuits in the energy-accuracy trade-off space

Prof. Marian Verhelst, MICAS laboratories (MICro-electronics And Sensors), Electrical Engineering Department, KU Leuven

Deep narrow networks, shallow wide networks, low precision networks, and even binary networks. NN model designers have so many degrees of freedom. Yet, also chip and circuit architects have many design options: from MAC-centric streaming architectures, to multi-level memory-hierarchies. From variable precision digital processing, over binary compute units, to analog or even in memory-processing. To achieve truly tiny ML, one has to make smart design decisions across this complete algorithm-architecture stack, while judging design decisions on their impact at the final system and application level metrics.

This talk will show design options at these different layers, and assess their impact. Subsequently, we will discuss how to achieve the required cross-layer trade-offs in a methodological way. This will allow to gain insights from such co-optimization, illustrated with resulting state-of-the-art implementations. Finally, it is very important to judge all optimizations at the system and application level, which is the only one that really matters to the user. The impact on these system level metrics will be assessed in light of 2 applications: always-on face recognition and keyword spotting.