Maurice Peemen received the MSc degree in electrical engineering from the Eindhoven University of Technology, The Netherlands in 2011. Currently, he is working toward the PhD degree at the Eindhoven University of Technology. His main research topics are efficiency improvements for Deep Convolutional Networks by algorithmic changes, data access optimizations, custom accelerators, and optimizing compilers.

Title: Accelerating Deep Learning Applications

Abstract: Throughout the past decade, Deep Learning and Convolutional Networks (ConvNets) have dramatically improved the state-of-the-art in object detection, speech recognition, and many other pattern recognition domains. The recent success of these deep learning models motivates researchers to further improve their accuracy by increasing model size and depth. Consequently the computational and data transfer workloads have grown tremendously. For beating accuracy records using huge compute clusters this is not yet a big issue; e.g. the introduction of GP-GPU computing improved the raw compute power of these server systems tremendously. However, for consumer applications in the mobile or wearable domain these impressive ConvNets are not used. Their execution requires far too much compute power and energy.
This lecture will introduce the key aspects that drive the success of deep learning for different application domains. The efficiency challenges of deep learning are addressed and we will go over several methodologies that substantially improve the energy-efficiency of deep convolutional networks. These techniques focus on the embedded system domain where platform based design is very important. We look at data movement optimization for custom memory hierarchies, by advanced tiling techniques. Further we investigate different dedicated accelerator architectures. And explore algorithmic modifications. Finally we will also discuss automatic code generation for custom accelerators. The above optimization parts significantly improve the efficiency and programmability of deep Convolutional Networks; it thereby enables their applicability to the mobile and wearable use cases.