IAF SYMPOSIUM ON INTEGRATED APPLICATIONS (B5)
Interactive Presentations - IAF SYMPOSIUM ON INTEGRATED APPLICATIONS (IP)

Author: Dr. Yuanhong Mao
Xi'an Microelectronics Technology Institute, CASC, China

Dr. Zhong Ma
Xi'an Microelectronics Technology Institute, China
Prof. Xi Liu
Xi'an Microelectronics Technology Institute, CASC, China
Dr. Liang Yang
Xi'an Microelectronics Technology Institute, China Aerospace Science and Technology Corporation (CASC), China
Mr. Pengchao He
Xi'an Microelectronics Technology Institute, CASC, China
Dr. Ning Wang
Xi'an Microelectronics Technology Institute, CASC, China

ENHANCING THE EFFICIENCY OF DEEP NEURAL NETWORKS FOR AEROSPACE APPLICATIONS

Abstract

Using deep neural networks (DNNs) in aerospace intelligence perception and decision-making has yielded remarkable advancements in recent years. DNNs have proven highly effective in handling complex and dynamic tasks such as space exploration, on-orbit calculations, visual navigation, and prognostic health management. However, aerospace computing devices' performance, power consumption, size, and weight pose significant challenges, as DNNs typically require large amounts of computation and memory, leading to increased power consumption and longer processing times. Consequently, achieving greater efficiency in DNN inference within aerospace systems has become a pressing concern. It is crucial to enhance the performance of the DNN algorithm by reducing its parameters and calculations. This paper proposes a platform that accelerates a trained DNN model without compromising performance. Firstly, the platform adopts more compact convolution structures than traditional convolution layers within the DNN architecture. These lightweight network structures effectively reduce the parameters and calculations required by DNNs. Secondly, network pruning is employed to simplify DNNs by selectively removing unimportant nodes, branches, weights, filters, channels, and layers based on their contribution to network performance. This process results in a smaller and faster network while maintaining network performance. Network fine-tuning and pruning can be performed alternately, ensuring the removal of redundant connections without compromising network performance. Thirdly, the platform utilizes the quantization method to convert the previous 32-bit float operations to more efficient fixed-point number operations such as 8-bit or 16-bit, and even binary or ternary operations. Fixed-point operations are significantly faster than float operations, and the reduced bit length translates to decreased memory access during DNN inference. Lastly, even if the accelerated model experiences a decrease in performance, knowledge distillation can be employed to restore its accuracy. Knowledge distillation involves learning from previous networks, akin to the relationship between teachers and students, to maintain accuracy. The acceleration above methods within our platform are not mutually exclusive and can be optimized collectively to enhance DNN architectures. By leveraging network acceleration, aerospace systems can efficiently carry out space missions

with limited memory and low power consumption. These approaches, when implemented in spacecraft, will facilitate the deployment of intelligence applications in the future.