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, yuanhongmao@hotmail.com

Mr. Pengchao He
Xi'an Microelectronics Technology Institute, CASC, China, hepengchao@163.com
Dr. Xi Liu
Xi'an Microelectronics Technology Institute, CASC, China, liuxi_771@qq.com
Prof. Haifeng He
Xi'an Microelectronics Technology Institute, CASC, China, 15349269176@163.com

MORE EFFICIENT DEEP NEURAL NETWORKS FOR AEROSPACE APPLICATIONS

Abstract

In recent years, artificial intelligence promoted by deep neural networks (DNNs) has achieved state-of-the-art performance in aerospace intelligence perception and decision-making. More aerospace applications need DNNs to handle complex and dynamic tasks such as space exploration, on-orbit calculations, visual navigation, prognostics health management and more. Intelligent systems aimed at aerospace usually need the faster operation and lower power consumption. Compared with the traditional algorithm, deep neural networks usually require more calculations and memory, which means more power consumption and longer delay. Therefore, making DNN's inference more efficient in aerospace systems has become challenging. In addition to improving the hardware performance, reducing the parameters and calculation in the DNN algorithm is essential to addressing this challenge. This paper proposes a platform to accelerate a trained DNN model without performance degradation. Firstly, more compact convolution structures are adopted instead of the traditional convolution layers in DNN architecture. These lightweight network structures can significantly reduce the parameters and calculations in DNNs. Secondly, network pruning can further simplify DNNs by different granularity, such as weights, filters, channels, and layers. By measuring their contribution to the network performance, we can remove the unimportant nodes or branches to get a smaller and faster network. Network fine-tuning and pruning can perform alternately. While removing the redundant connection, network performance is also guaranteed. Thirdly, we can use the quantization method to convert the previous 32-bit float operation to 8-bit or 16-bit fixed-point number operation, even binary operation or ternary operation. Fixed point operation is much faster than float operation. Moreover, the shorter bits mean less memory access in DNN inference. Finally, even if the performance of the accelerated model decreases, we can also use knowledge distillation to restore its accuracy. Knowledge distillation can maintain accuracy by learning from previous networks, just like students and teachers. The acceleration methods in our platform above are not exclusive. They can optimize DNN architectures together. Aerospace systems can efficiently implement space missions with limited memory and low power consumption by using network acceleration. These approaches in our platform will make intelligence applications easier to deploy in spacecraft in the future.