Title: Deep Neural Networks for Resource Constraint IoT Devices

Speaker: Dr Khalid Pandit, Ph.D. from the National Institute of Technology, Srinagar.

Date and Time: July 15, 2021 (Thu). 11:00 AM

Venue: Online

Abstract:

Deep Neural Networks (DNNs) is the computing paradigm that has achieved remarkable success in various fields of engineering in recent years, primarily visual recognition. DNNs owe their success to many weight parameters (and increased depth), which led to huge computation and memory costs for implementation. These limiting factors hinder the scalability of such algorithms on resource-constrained IoT devices. In general, DNNs are believed to be over parametrized i.e., the parameters are highly redundant, thus can be structurally removed without significant loss of performance.

In this talk, I will focus on the methods for reducing the computational complexity of the deep learning models and the ways of filling these models onto resource constraint devices. I will present the method of embedding the intelligence in resource-constrained devices by splitting the layers of the neural network models and placing them on different compute resources.

I will also present the method for developing computationally sparse deep neural network models to fit on the devices themselves. I will talk about two new regularization schemes based on Non-convex transformed l1 (Tl1) and variance-guided approaches. Experimental results demonstrate the efficacy of the proposed regularizers trained on benchmark datasets like MNIST, CIFAR-10, CIFAR-100, ImageNet, etc. It is worth mentioning that the effectiveness of proposed regularizers is evaluated against state-of-art neural network pruning methods. Various evaluation criteria, like the number of parameters, floating-point operations (FLOPs), and accuracy, are used to evaluate the approach's efficacy. The development of neural network models based on the proposed schemes is a substantial step towards embedding intelligence into resource-constrained devices. These techniques can be instrumental in creating truly smart IoT devices with decision-making capabilities in the fast-growing IoT domain.

Finally the talk will conclude with the idea of investigating the use of third-generation neural networks (spiking neural networks (SNNs)) for use in resource constraint environments. These neural networks are biologically plausible and closely mimic the working of the biological brain. SNNs on neuromorphic hardware exhibit favorable properties such as low power consumption, fast inference, and event-driven information processing. This makes them interesting candidates for the efficient implementation of deep neural networks on resource constraint devices.

Bio:

Khalid pursued Ph.D. from the National Institute of Technology, Srinagar, under the guidance of Prof. Roohie Naaz and Prof. Mohammad Ahsan Chishti. He worked on investigating the methods for embedding intelligence into resource constraint devices by reducing the computational complexity of deep neural network models. His research interests include artificial intelligence, machine learning, IoT, and edge computing. He has published papers in journals like IEEE Transactions, IET Electronics letters, and Sparsity in neural networks workshop among others. He has also have been awarded 3 funded research projects through TEQIP, NPIU (Ministry of Human Resource Development).