Speaker Information:

SHAN, Haijun

Ph.D., is currently a researcher at Zhejiang Lab. He received his bachelor's degree from Zhejiang University in 2009, and then received his doctor's degree in Automatic Control Theory from Zhejiang University in 2015. From 2011 to 2012, he was a visiting scholar in University of Minnesota. Before joining in Zhejiang Lab, he was a senior engineer in Central Software Research Institute of Huawei and participated in the research and development of deep learning platforms, from 2015 to 2018.

His main research interests are machine learning, computer vision, and cloud computing. Dr. Shan has combined background in academic research and enterprise project development. He participated and led in 4 scientific research projects, including National Natural Science Foundation. As a subsystem architect, he has been in charge of two major technological innovation projects in Huawei. Published 5 SCI / EI papers and applied for 7 invention patents.

Title and Abstract

Deep network model compression, a way to enable terminal intelligence

Deep neural networks (DNN) have the characteristics of huge structure and numerous parameters. The model training and inference consume lots of computing resources, and thus most of DNN models need to be run in the cloud side. And it is unable to deploy and run models on terminal devices, including but not limited to mobile phones, robots, and wearable devices, etc. With the development of the Internet of Things and 5G, more and more scenarios require terminals to own local AI capabilities. That means DNN models should be light-weighted and run on terminal devices. One way to achieve this is model compression.

Currently there are four categories of model compression methods: parameter pruning and sharing, quantification and binary network, low rank approximation, and knowledge distillation. This presentation will give a deep inspection of the problem background, research status and future trends in model compression. At last, the latest progress about multi-level joint pruning, self-adaptive compression in my team will be shared.