Task-oriented Communications for Edge AI

Abstract:

Deep learning has achieved remarkable successes in many application domains, such as computer vision, image processing, and natural language processing. However, deploying powerful deep learning models on resource-constrained mobile devices (e.g., wearable or IoT devices) faces great challenges. Recently, edge AI techniques that rely on the emerging mobile edge computing platforms have been proposed, which forward intermediate features to be processed by a powerful edge server. To achieve high-accuracy and low-latency inference, effective feature encoders with low complexity and high compression capability will be needed. This calls for a paradigm shift in wireless communications, from “data-oriented communications”, which maximize data rates, to “task-oriented communications”, where the data transmission is an intermediate step to be optimized for the downstream inference task. This talk will introduce recent progresses on task-oriented communication for edge inference. An effective design principle based on information bottleneck will be firstly introduced, which will then be extended to multi-device cooperative perception based on a distributed information bottleneck framework. Use cases on edge video analytics and edge-assisted localization for mobile robots will be presented, followed by introduction of EdgeGPT, an autonomous edge AI system empowered by large language models.

Bio:

Jun Zhang received his Ph.D. degree in Electrical and Computer Engineering from the University of Texas at Austin. He is an IEEE Fellow and an IEEE ComSoc Distinguished Lecturer. He is an Associate Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology. His research interests include wireless communications and networking, mobile edge computing and edge AI, and cooperative AI. Dr. Zhang co-authored the book Fundamentals of LTE (Prentice-Hall, 2010). He is a co-recipient of several best paper awards, including the 2021 Best Survey Paper Award of IEEE Communications Society, the 2019 IEEE Communications Society & Information Theory Society Joint Paper Award, and the 2016 Marconi Prize Paper Award in Wireless Communications. He also received the 2016 IEEE ComSoc Asia-Pacific Best Young Researcher Award. He is an Editor of IEEE Transactions on Communications and IEEE Transactions on Machine Learning in Communications and Networking, and was an editor of IEEE Transactions on Wireless Communications (2015-2020).