My research interests focus on the intersection of machine learning (ML) and edge computing that could make ML services ubiquitous in the future. To realise this vision, ML systems, both for inference and training, should not be limited to only running at the cloud but in a synergistic cooperation with end devices and edge nodes. However, it currently suffers from issues such as intensive computation/communication overhead, limited resources and heterogeneous environment. To bridge the gap, my PhD project aims to build efficient, robust and flexible ML systems by exploring advanced techniques, e.g. federated learning, edge computing and model compression.