In this presentation, I will present several novel quantization techniques for deep neural networks. Throughout the research period, we have devoted to providing stronger theoretical guarantees for the ASkewSGD algorithm, proposed by Leconte et al., by optimization approaches. Besides, we aim to devise new approaches to tackle the smoothed optimization problem.

Recent advancements in deep neural network theory have evoked the demand for energy-efficient, lighter neural networks with shorter inference time. Many methods are developed for these purposes, e.g. hardware-architecture co-design, knowledge distillation,