Contribution ID: **40**

Type: **talk**

Efficient Multi-modal LLM on the Edge

Monday 14 October 2024 09:10 (25 minutes)

This talk presents efficient multi-modal LLM innovations with algorithm and system co-design. I'll first present VILA, a visual language model deployable on the edge. It is capable of visual in-context learning, multi-image reasoning, video captioning and video QA. Followed by SmoothQuant and AWQ for LLM quantization, which enables VILA deployable on edge devices, bringing new capabilities for mobile vision applications. Second, I'll present StreamingLLM, a KV cache optimization technique for long conversation and QUEST, leveraging sparsity for KV cache compression.

Author: Dr HAN, Song (MIT)

Presenter: Dr HAN, Song (MIT)

Session Classification: Morning Plenary