Theme: Machine Intelligence

Subtheme: On-device Artificial Intelligence

Artificial Intelligence (AI) is playing an increasingly prominent role in our daily lives. However, cloud-based AI always suffer from a number of disadvantages: (1) since data is sent off-device, the long latency between data sent and response is serious in real time applications, and also largely deteriorates the user experience in many applications like voice assistant. (2) Due to the dependency on an outside network, AI applications are only as reliable as the Internet connection. (3) Transmission data to cloud could also lead to the leakage of user private and sensitive data such as voice fingerprints and face scans. These disadvantages limit the AI ability to provide compelling user experiences.

On-device AI is an essential component in enabling always-available, responsive and privacy-preserving intelligence: (1) Faster Response: on-device AI can process data on device provides faster response, which is particularly important for real-time applications; (2) Enhanced Reliability: on-device processing reduces reliance on network conditions for mission critical applications; (3) Bandwidth Efficiency: as more data is generated and utilized at the edge, on-device processing reduces unnecessary traffic; (4) User Privacy Protection: on-device processing allows us to keep our most personal data on device such that it’s not vulnerable to being compromised in the cloud. Recent complex algorithms and processes, such as machine learning and deep learning, are driving the transformation, boosting the need for more powerful on-device AI solutions.

As part of this program, various topics related to efficient on-device AI are of interest. These include, but are not limited to:

1. Novel on-device capabilities for natural language processing, vision, audio, and speech, including:
   a. Language understanding and conversational assistants on mobile and wearable devices
   b. Real-time computer vision applications on mobile devices
   c. Speech recognition on mobile and smart home devices
   d. Real-time and continuous on-device data analytics and predictions
2. Novel efficient machine learning model/architecture design for edge devices (e.g., MobileNet, ShuffleNet, etc.)
3. Machine learning techniques for model optimization (e.g., pruning, quantization, etc.)
4. Knowledge distilled neural networks to utilize knowledge to accelerate the neural network training and inference
5. Distributing AI capabilities across the whole system instead of performing all the compute in the cloud
6. On-device privacy-preserving learning
7. Neuroscience-inspired architectures
8. Machine learning techniques for system optimization (e.g., scheduling, caching, etc.)
9. Platform-aware model optimization (e.g., low-precision training/inference)
10. Efficient hardware accelerator design for neural computing on mobile devices
11. Tools for architecture modeling, design space exploration, and algorithm mapping
   a. Mobile-optimized neural network libraries (e.g., CoreML, TensorFlow Lite)
   b. Open-source software packages to build and deploy on-device models

※ The topics are not limited to the above examples and the participants are encouraged to propose original idea.

※ Funding: Up to USD $150,000 per year