Theme: Machine Intelligence

- Subtheme: Tiny Machine Learning

Artificial Intelligence (AI) is playing an increasingly prominent role in our daily lives. However, cloud-based AI suffers from some disadvantages, including high bandwidth, heavy dependence on Internet, latency, and very importantly also user data privacy. These disadvantages make it almost impossible to enable always-on AI services that need continuous data transmission and processing, such as wakeup detection, audio de-noising, video surveillance, etc.

TinyML is a field of study in machine learning and embedded systems that explores the types of models that can run on small, low-powered devices like microcontrollers. It enables low-latency, low power, and low bandwidth AI services at edge devices, especially for always-on services. (1) Low Latency: Since the model runs on the edge, the data doesn't have to be sent to a server to run inference. This reduces the latency of the output. (2) Low Power Consumption: As we discussed before, microcontrollers consume very little power. This enables them to run without being charged for a very long time. (3) Low Bandwidth: As the data does not have to send to the server constantly, less internet bandwidth is used. (4) Privacy: Since the model is running on the edge, user data is not stored on any server. These advantages of TinyML enable AI services and applications to run on edge devices unplugged and batteries for weeks, months, and in some cases, even years.

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

1. Novel on-device capabilities for vision, audio, and speech, including:
   a. Real-time and always-on computer vision applications on mobile and IoT devices
   b. Real-time and always-on audio processing on mobile and IoT devices
   c. 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. Novel model compression techniques (e.g., pruning, quantization, knowledge distillation, SVD, etc.)

4. Novel signal/image processing incorporated lightweight deep/machine learning models
5. Novel software-hardware co-design for efficient machine/deep learning
6. Optimization of deep/machine learning models on microcontrollers, IoT devices, etc.
7. Platform and device aware model optimization (e.g., low-precision training/inference)
8. Neuroscience-inspired architectures
9. Machine learning techniques for system optimization (e.g., scheduling, caching, etc.)
10. Efficient hardware accelerator design for neural computing on mobile devices
11. Tools for TinyML
    a. Microcontrollers/IOT-optimized neural network libraries (e.g., TensorFlow Lite for Microcontrollers)
    b. Open-source software packages to build and deploy TinyML models

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

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