Socionext Prototypes Low-Power AI Chip with Quantized Deep Neural Network Engine

Delivers Significant Expansion of Edge Computing Capabilities, Performance and Functionality

Yokohama, March 17, 2020 --- Socionext Inc. has developed a prototype chip that incorporates newly-developed quantized Deep Neural Network (DNN) technology, enabling highly-advanced AI processing for small and low-power edge computing devices.

The prototype is a part of a research project on “Updatable and Low Power AI-Edge LSI Technology Development” commissioned by the New Energy and Industrial Technology Development Organization (NEDO) of Japan. The chip features a "quantized DNN engine" optimized for deep learning inference processing at high speeds with low power consumption.

Today’s edge computing devices are based on conventional, general-purpose GPUs. These processors are not generally capable of supporting the growing demand for AI-based processing requirements, such as image recognition and analysis, which need larger devices at higher cost due to increases in power consumption and heat generation. Such devices and their limited performance are not desirable for state-of-the art AI processing.

Quantized DNN Engine

In their place, Socionext has developed a proprietary architecture based on "quantized DNN technology" for reducing the parameter and activation bits required for deep learning. The result is improved performance of AI processing along with lower power consumption. The architecture incorporates bit reduction including 1-bit (binary) and 2-bit (ternary) in addition to the conventional 8-bit, as well as the company’s original parameter compression technology, enabling a large amount of computation with fewer resources and significantly less amounts of data.

In addition, Socionext has developed a novel on-chip memory technology that provides highly efficient data delivery, reducing the need for extensive large capacity on-chip or external memory typically required for deep learning.

Integrating these new technologies, Socionext has prototyped an AI chip with its “DNN engine” and has confirmed its functionality and performance. The prototype chip achieved object detection by “YOLO v3” at 30fps, while consuming less than 5W of power. This is 10 times

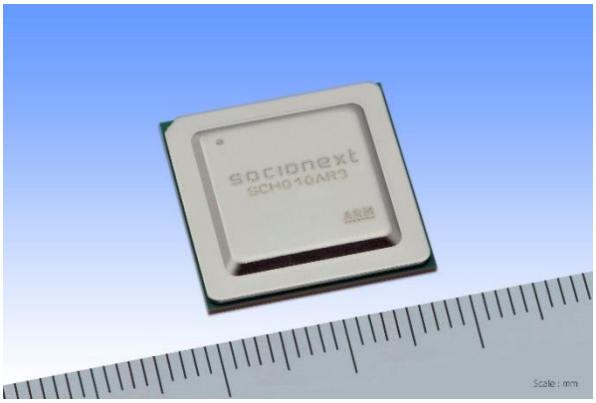
For Press Inquiry:
Socionext Inc.
Tel: +81-45-568-1006 https://www.socionext.com/en/contact/

more efficient than conventional, general-purpose GPUs. The chip is also equipped with a high-performance, low-power Arm Cortex-A53 quad-core CPU. Unlike other “accelerator” chips, it can perform the entire AI processing without external processors.

Deep Learning Software Development Environment

Socionext has also built a deep learning software development environment. Incorporating TensorFlow as the base framework, it allows developers to perform original, low-bit "quantization-aware training" or "post-training quantization". When used in combination with the new chip, users can choose and apply the optimal quantization technology to various neural networks and execute highly accurate processing. The new chip will add the most advanced computer vision functionality to small form factor, low-power edge devices. Target applications include advanced driver assistance system (ADAS), security camera, and factory automation among others.

Socionext is currently conducting circuitry fine-tuning and performance optimization through the evaluation of this prototype chip. The company will continue working on research and development with the partner companies towards the completion of the NEDO-commissioned project, to deliver the AI Edge LSI as the final product.



Prototype Chip with Quantized DNN Engine
(View Larger Image)

NEDO Project:
Project for Innovative AI Chips and Next-Generation Computing Technology Development
Development of Innovative AI Edge Computing Technologies
Updatable and Low Power AI-Edge LSI Technology Development

About Socionext

Socionext is a global, innovative enterprise that designs, develops and delivers System-on-Chip solutions to customers worldwide. The company is focused on technologies that drive today's leading-edge applications in consumer, automotive and industrial markets. Socionext combines world-class expertise, experience, and an extensive IP portfolio to provide exceptional solutions and ensure a better quality of experience for customers. Founded in 2015, Socionext Inc. is headquartered in Yokohama, and has offices in Japan, Asia, United States and Europe to lead its product development and sales activities. For more information, visit www.socionext.com.