PRESS RELEASE

Source: Tokyo Institute of Technology

For immediate release: August 23, 2021

Cutting "Edge": A Tunable Neural Network Framework Towards Compact and Efficient Models

(Tokyo, August 23) **Convolutional neural networks (CNNs) have enabled numerous AI-enhanced applications, such as image recognition. However, the implementation of state-of-the-art CNNs on low-power edge devices of Internet-of-Things (IoT) networks is challenging because of large resource requirements. Researchers from Tokyo Institute of Technology have now solved this problem with their efficient sparse CNN processor architecture and training algorithms that enable seamless integration of CNN models on edge devices.**

With the proliferation of computing and storage devices, we are now in an information-centric era in which computing is ubiquitous, with computation services migrating from the cloud to the "edge," allowing algorithms to be processed locally on the device. These architectures enable a number of smart internet-of-things (IoT) applications that perform complex tasks, such as image recognition.

Convolutional neural networks (CNNs) have firmly established themselves as the standard approach for image recognition problems. The most accurate CNNs often involve hundreds of layers and thousands of channels, resulting in increased computation time and memory use. However, "sparse" CNNs, obtained by "pruning" (removing weights that do not signify a model's performance), have significantly reduced computation costs while maintaining model accuracy. Such networks result in more compact versions that are compatible with edge devices. The advantages, however, come at a cost: sparse techniques limit weight reusability and result in irregular data structures, making them inefficient for real-world settings.

Addressing this issue, Prof. Masato Motomura and Prof. Kota Ando from Tokyo Institute of Technology (Tokyo Tech), Japan, along with their colleagues, have now proposed a novel 40 nm sparse CNN chip that achieves both high accuracy and efficiency, using a Cartesian-product MAC (multiply and accumulate) array (Figures 1 and 2), and "pipelined activation aligners" that spatially shift "activations" (the set of input/output values or, equivalently, the input/output vector of a layer) onto regular Cartesian MAC array.

"Regular and dense computations on a parallel computational array are more efficient than irregular or sparse ones. With our novel architecture employing MAC array and activation aligners, we were able to achieve dense computing of sparse convolution," says Prof. Ando, the principal researcher, explaining the significance of the study. He adds, "Moreover, zero weights could be eliminated from both storage and computation, resulting in better resource utilization." The findings will be presented at the *33rd Annual Hot Chips Symposium*.

One important aspect of the proposed mechanism is its "tunable sparsity." Although sparsity can reduce computing complexity and thus increase efficiency, the level of sparsity has an influence on prediction accuracy. Therefore, adjusting the sparsity to the desired accuracy and efficiency helps unravel the accuracy-sparsity relationship. In order to obtain highly efficient "sparse and quantized" models, researchers applied "gradual pruning" and "dynamic quantization" (DQ) approaches on CNN models trained on standard image datasets, such as CIFAR100 and ImageNet. Gradual pruning involved pruning in incremental steps by dropping the smallest weight in each channel (Figure 3), while DQ helped quantize the weights of neural networks to low bit-length numbers, with the activations being quantized during inference. On testing the pruned and quantized model on a prototype CNN chip, researchers measured 5.30 dense TOPS/W (tera operations per second per watt—a metric for assessing performance efficiency), which is equivalent to 26.5 sparse TOPS/W of the base model.

"The proposed architecture and its efficient sparse CNN training algorithm enable advanced CNN models to be integrated into low-power edge devices. With a range of applications, from smartphones to industrial IoTs, our study could pave the way for a paradigm shift in edge AI," comments an excited Prof. Motomura.

It certainly seems that the future of computing lies on the "edge"!

Reference

Conference:	*Hot Chips 33*
Poster Session:	Edge Inference Engine for Deep & Random Sparse Neural Networks with 4-bit Cartesian-Product MAC Array and Pipelined Activation Aligner
Authors:	Kota Ando*, Jaehoon Yu, Kazutoshi Hirose, Hiroki Nakahara, Kazushi Kawamura, Thiem Van Chu, and Masato Motomura*
Affiliations:	Tokyo Institute of Technology

*Corresponding author's email: motomura.m.ab@m.titech.ac.jp;
ando.kota@artic.iir.titech.ac.jp

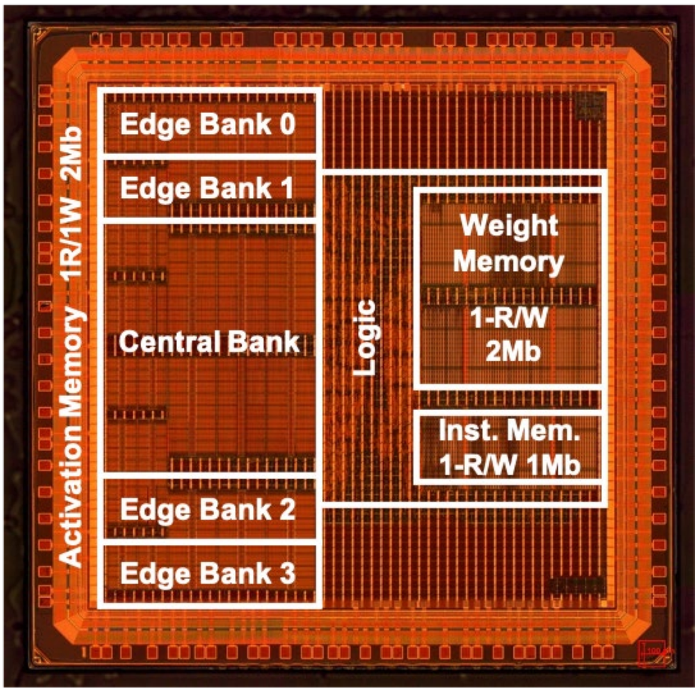

Figure 1. The prototype chip fabricated in 40 nm technology

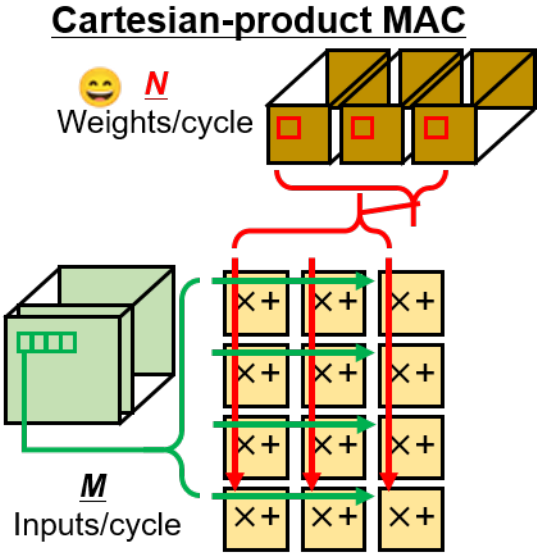

Figure 2. The Cartesian product MAC array for maximizing arithmetic intensity of pointwise convolution

Researchers from Tokyo Tech proposed a novel CNN architecture using Cartesian product MAC (multiply and accumulate) array in the convolutional layer.

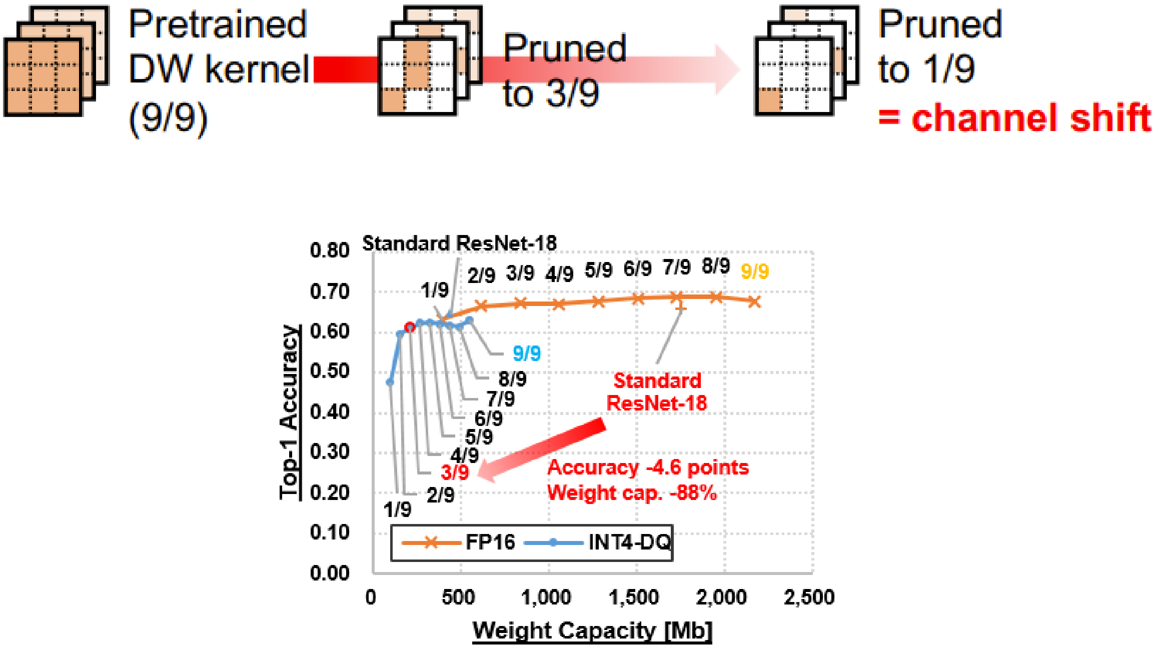

Figure 3. Using gradual pruning and dynamic quantization to control the accuracy-efficiency trade-off

The trained model was pruned by removing the lowest weight in each channel. Only one element remains after 8 rounds of pruning (pruned to 1/9). Each of the pruned models is then subjected to dynamic quantization.

Contact

Emiko Kawaguchi
Public Relations Division,
Tokyo Institute of Technology
media@jim.titech.ac.jp
+81-3-5734-2975

About Tokyo Institute of Technology

Tokyo Tech stands at the forefront of research and higher education as the leading university for science and technology in Japan. Tokyo Tech researchers excel in fields ranging from materials science to biology, computer science, and physics. Founded in 1881, Tokyo Tech hosts over 10,000 undergraduate and graduate students per year, who develop into scientific leaders and some of the most sought-after engineers in industry. Embodying the Japanese philosophy of “monotsukuri,” meaning “technical ingenuity and innovation,” the Tokyo Tech community strives to contribute to society through high-impact research.
https://www.titech.ac.jp/english/