# RETROSPECTIVE: ShiDianNao: Shifting Vision Processing Closer to the Sensor

Zidong Du, Tianshi Chen, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen
ICT, China

Robert Fasthuber, Paolo Ienne
EPFL, Switzerland

Olivier Temam
Google

## I. Context & History

This article was published at ISCA in 2015. In 2010, Olivier Temam gave a keynote at ISCA [4] on the occurrence of three simultaneous trends: growing pertinence of hardware customization due to Moore's Law plateauing, remarkable recent progress in Deep Neural Networks (DNNs) two decades after an initial spark of interest in Artificial Neural Networks, and growing needs in visual recognition applications. These three simultaneous trends led to the conclusion and proposition that NN accelerators would become a sensible development path forward for our domain and industry. Olivier Temam published a first paper at ISCA [1] on the topic in 2012. Because the concept of NN accelerators was initially met with a lot of skepticism at the time, he tried to find partners to implement a prototype chip in order to convincingly demonstrate the higher compute density and efficiency (w.r.t. GPUs at the time). Both academic institutions and companies in Europe and the US were not interested at the time. Eventually, he met people at an institute in Beijing, China (ICT), who were not working on NN accelerators, but were part of a research and engineering team developing a processor series for the Chinese market (Loongson). Some people showed interest in the NN accelerator concept after attending ISCA 2010, and an Inria/ICT collaboration was started, later on formalized with a joint lab on accelerators for AI, which was inaugurated in Beijing by the then presidents of Inria and ICT. That collaboration led to a series of joint papers. The first one [2] was a follow up to [1] where activations and parameters were stored in memory to enable a smaller area footprint. The second one [3] was an AI supercomputer based on the node defined in [2] with both inference and training capacity. The third one is the present paper, which is also derived from [2] but targets embedded systems and especially shows the benefit of locating an AI accelerator next to a visual sensor. The present paper was a collaboration with Paolo Ienne and his team at EPFL, who was a visitor at ICT at the time.

[2] and [3] were both best paper awards at, respectively, ASPLOS and MICRO in 2014. The natural next step to develop this academic work was to create industrial products which AI researchers and engineers could effectively use. Discussions for a potential joint startup occurred in 2013 and 2014, but the ICT partners eventually declined. At that time, Google was just starting its TPU effort and proposed Olivier Temam to join to co-lead the first TPU supercomputer for inference and training (a.k.a. Cloud TPU; it is the second TPU chip developed by Google). Later on, he created the Edge TPU project at Google for the embedded and low power markets. In 2016, the ICT partners finally decided to leverage the DianNao accelerators work to create a company for the Chinese market and founded Cambricon.

## II. NN Accelerator and Smart Vision Sensors

In the first years after NN accelerators emerged, their high performance and high energy efficiency led to a quickly growing popularity for a fast developing AI domain. Yet, both accelerator and energy performance were (and still are) dominated by long-latency/low-bandwidth memory accesses. At the same time, for image recognition applications, a special type of DNNs was typically used: Convolutional Neural Networks (CNNs). A nice property of CNNs is that they reuse the same parameters across an image through the notion of feature maps, and the number/size of feature maps can be tailored. These properties mean fewer parameters than other forms of DNNs, and high locality properties. Both were conducive to SRAM implementation, and to forgo expensive DRAM accesses. For smart visual sensors, these notions were particularly attractive. Grafting SRAM and CNN logic next to the sensor allowed to overcome most of the performance and energy limitations of smart visual sensors.

## III. What is the heritage in 2023

The popularity of AI accelerators at large is now well known, whether it is TensorCores in NVIDIA chips since Volta, the many large AI chip startups, or the AI accelerator programs in Google, Amazon, Microsoft and Meta. In parallel with what were largely data center accelerators, AI accelerators for the embedded market are also numerous, with the embedded AI accelerator Movidius being the first Intel acquisition of an AI accelerator company. Still, most of these embedded accelerators remained separated from the visual sensor itself, even though many of the applications were related to smart cameras. It is only in 2020 that Sony (which has about half of the CMOS image sensor market) introduced the IMX500 smart sensor which combines an AI accelerator with a vision sensor. Going forward, it is likely that such integration will become the norm due to the aforementioned performance and energy advantages, but also privacy, latency and bandwidth benefits.

## References

[1] O. Temam,, “A defect-tolerant accelerator for emerging high performance applications”, 39th Annual International Symposium on Computer Architecture (ISCA), vol. 00, no. c, pp. 356–367, 2012.

[2] T. Chen, Z. Du, N. Sun, J. Wang, and C. Wu, “DianNao: a small-footprint high-throughput accelerator for ubiquitous machinelearning”, in Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Salt Lake City, UT, USA, 2014, pp. 269–284.

[3] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, “DaDianNao: A Machine-Learning Supercomputer,” in Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2015, pp. 609–622.

[4] Olivier Temam: The Rebirth of Neural Networks, International Symposium on Computer Architecture (ISCA), June 2010 (Keynote).