Paper Review Assignment 3
AI Processors

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

Paper Readings and Review

- Paper related to AI Accelerators
  - To learn the basic architecture and dataflow of AI Accelerator
  - To understand state-of-the-art AI accelerator designs
- **Due**
  - **5/13 23:59**
- Requirement
  - Choose **at least one or more** papers
    - From recommended paper list
    - **Or any other paper as long as it related to the topics**
  - Summarize and write paper review in word/latex format
    - **LaTeX format is highly recommended**
  - Hand in **compiled pdf files** on moodle

Paper Readings and Review

- Reading reviews are free of format
- But the following review questions guide you through the paper reading process.
  - What are the **motivations** for this work?
  - What is the **proposed solution**?
  - What is the work's **evaluation** of the proposed solution?
  - What is your **analysis** of the identified problem, idea, and evaluation?
  - What are **future directions** for this research?
  - What **questions** are you left with?

Recommended Paper List

- Eyeriss
  - Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE journal of solid-state circuits*, *52*(1), 127-138.
- Eyeriss v2
  - Chen, Y. H., Yang, T. J., Emer, J., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *9*(2), 292-308.
- TPU
  - Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... & Yoon, D. H. (2017, June). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of the 44th annual international symposium on computer architecture* (pp. 1-12).
- SCNN
  - Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., ... & Dally, W. J. (2017). SCNN: An accelerator for compressed-sparse convolutional neural networks. *ACM SIGARCH computer architecture news*, *45*(2), 27-40.
- NVDLA
  - Nvidia, NVDLA Open Source Project, 2017. http://nvdla.org/

Recommended Paper List

- Gemmini
  - Genc, H., Kim, S., Amid, A., Haj-Ali, A., Iyer, V., Prakash, P., ... & Shao, Y. S. (2021, December). Gemmini: Enabling systematic deep-learning architecture evaluation via full-stack integration. In *2021 58th ACM/IEEE Design Automation Conference (DAC)* (pp. 769-774). IEEE.
- DaDianNao
  - Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., ... & Temam, O. (2014, December). Dadiannao: A machine-learning supercomputer. In *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 609-622). IEEE.
- neuFlow
  - Farabet, C., Martini, B., Corda, B., Akselrod, P., Culurciello, E., & LeCun, Y. (2011, June). Neuflow: A runtime reconfigurable dataflow processor for vision. In *CVPR 2011 workshops* (pp. 109-116). IEEE.
- GANPU: Multi-DNN Training Processor for GANs
  - Kang, S., Han, D., Lee, J., Im, D., Kim, S., Kim, S., & Yoo, H. J. (2020, February). 7.4 GANPU: A 135TFLOPS/W multi-DNN training processor for GANs with speculative dual-sparsity exploitation. In *2020 IEEE International Solid-State Circuits Conference-(ISSCC)* (pp. 140-142). IEEE.
- LNPU: Sparse DNN Learning Processor
  - Lee, J., Lee, J., Han, D., Lee, J., Park, G., & Yoo, H. J. (2019, February). 7.7 LNPU: A 25.3 TFLOPS/W sparse deep-neural-network learning processor with fine-grained mixed precision of FP8-FP16. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)* (pp. 142-144). IEEE.

Recommended Paper List

- CNPU: Mobile Deep RL Accelerator
  - Kim, C., Kang, S., Shin, D., Choi, S., Kim, Y., & Yoo, H. J. (2019, February). A 2.1 TFLOPS/W mobile deep RL accelerator with transposable PE array and experience compression. In 2019 IEEE International Solid-State Circuits Conference-(ISSCC) (pp. 136-138). IEEE.
- DNPU: Deep Neural Network SoC
  - Shin, D., Lee, J., Lee, J., & Yoo, H. J. (2017, February). 14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks. In 2017 IEEE International Solid-State Circuits Conference (ISSCC) (pp. 240-241). IEEE.
- Optimizing the Convolution Operation to Accelerate Deep Neural Networks on FPGA
  - Ma, Y., Cao, Y., Vrudhula, S., & Seo, J. S. (2018). Optimizing the convolution operation to accelerate deep neural networks on FPGA. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, *26*(7), 1354-1367.
- IBM 4-Core AI Chip
  - A. Agrawal *et al*., "9.1 A 7nm 4-Core AI Chip with 25.6TFLOPS Hybrid FP8 Training, 102.4TOPS INT4 Inference and Workload-Aware Throttling," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, 2021, pp. 144-146, doi: 10.1109/ISSCC42613.2021.9365791.
- QNAP: Quantized Network-Acceleration Processor
  - H. Mo *et al*., "A 12.1 TOPS/W Quantized Network Acceleration Processor With Effective-Weight-Based Convolution and Error-Compensation-Based Prediction," in *IEEE Journal of Solid-State Circuits*, doi: 10.1109/JSSC.2021.3113569.=

Recommended Paper List

- 8-Bit Shared Exponent Bias Floating Point and Multiple-Way Fused Multiply-Add Trees
  - J. Park, S. Lee and D. Jeon, "A Neural Network Training Processor With 8-Bit Shared Exponent Bias Floating Point and Multiple-Way Fused Multiply-Add Trees," in *IEEE Journal of Solid-State Circuits*, vol. 57, no. 3, pp. 965-977, March 2022, doi: 10.1109/JSSC.2021.3103603.
- A Multi-Mode 8K-MAC HW-Utilization-Aware Neural Processing Unit with a Unified Multi-Precision Datapath in 4nm Flagship Mobile SoC
  - J. -S. Park *et al*., "A Multi-Mode 8K-MAC HW-Utilization-Aware Neural Processing Unit with a Unified Multi-Precision Datapath in 4nm Flagship Mobile SoC," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 246-248, doi: 10.1109/ISSCC42614.2022.9731639.
- Systolic Neural CPU Processor for Combined Deep Learning and General-Purpose Computing with 95% PE Utilization, High Data Locality and Enhanced End-to-End Performance
  - Y. Ju and J. Gu, "A 65nm Systolic Neural CPU Processor for Combined Deep Learning and General-Purpose Computing with 95% PE Utilization, High Data Locality and Enhanced End-to-End Performance," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 1-3, doi: 10.1109/ISSCC42614.2022.9731757.
- Approximate-Computing-Based Transformer Processor
  - Y. Wang *et al*., "A 28nm 27.5TOPS/W Approximate-Computing-Based Transformer Processor with Asymptotic Sparsity Speculating and Out-of-Order Computing," *2022 IEEE International Solid- State Circuits Conference (ISSCC)*, 2022, pp. 1-3, doi: 10.1109/ISSCC42614.2022.9731686.
- Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC
  - C. -H. Lin *et al*., "7.1 A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 134-136, doi: 10.1109/ISSCC19947.2020.9063111.