References

ISCA Tutorial (2017)

Website: http://eyeriss.mit.edu/tutorial.html

Joel Emer, Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang

References (Alphabetical by Author)

- Albericio, Jorge, et al. "Cnvlutin: ineffectual-neuron-free deep neural network computing," ISCA, 2016.
- Alwani, Manoj, et al., "Fused Layer CNN Accelerators," MICRO, 2016
- Chakradhar, Srimat, et al., "A dynamically configurable coprocessor for convolutional neural networks," ISCA, 2010
- Chen, Tianshi, et al., "DianNao: a small-footprint high-throughput accelerator for ubiquitous machine-learning," ASPLOS, 2014
- Chen, Yu-Hsin, et al. "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," ISSCC, 2016.
- Chen, Yu-Hsin, Joel Emer, and Vivienne Sze. "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," ISCA, 2016.
- Chen, Yunji, et al. "Dadiannao: A machine-learning supercomputer," MICRO, 2014.
- Chi, Ping, et al. "PRIME: A Novel Processing-In-Memory Architecture for Neural Network Computation in ReRAM-based Main Memory," ISCA 2016.
- Cong, Jason, and Bingjun Xiao. "Minimizing computation in convolutional neural networks." International Conference on Artificial Neural Networks. Springer International Publishing, 2014.
- Courbariaux, Matthieu, and Yoshua Bengio. "Binarynet: Training deep neural networks with weights and activations constrained to+ 1 or-1." arXiv preprint arXiv:1602.02830 (2016).
- Courbariaux, Matthieu, Yoshua Bengio, and Jean-Pierre David. "Binaryconnect: Training deep neural networks with binary weights during propagations," NIPS, 2015.

References (Alphabetical by Author)

- Dean, Jeffrey, et al., "Large Scale Distributed Deep Networks," NIPS, 2012
- Denton, Emily L., et al. "Exploiting linear structure within convolutional networks for efficient evaluation," NIPS, 2014.
- Dorrance, Richard, Fengbo Ren, and Dejan Marković. "A scalable sparse matrix-vector multiplication kernel for energy-efficient sparse-blas on FPGAs." Proceedings of the 2014 ACM/SIGDA international symposium on Field-programmable gate arrays. ACM, 2014.
- Du, Zidong, et al., "ShiDianNao: shifting vision processing closer to the sensor," ISCA, 2015
- Eryilmaz, Sukru Burc, et al. "Neuromorphic architectures with electronic synapses." ISQED, 2016.
- Esser, Steven K., et al., "Convolutional networks for fast, energy-efficient neuromorphic computing," PNAS 2016
- Farabet, Clement, et al., "An FPGA-Based Stream Processor for Embedded Real-Time Vision with Convolutional Networks," ICCV 2009
- Gokhale, Vinatak, et al., "A 240 G-ops/s Mobile Coprocessor for Deep Neural Networks," CVPR Workshop, 2014
- Govoreanu, B., et al. "10× 10nm 2 Hf/HfO x crossbar resistive RAM with excellent performance, reliability and low-energy operation," IEDM, 2011.
- Gupta, Suyog, et al., "Deep Learning with Limited Numerical Precision," ICML, 2015
- Gysel, Philipp, Mohammad Motamedi, and Soheil Ghiasi. "Hardware-oriented Approximation of Convolutional Neural Networks." arXiv preprint arXiv:1604.03168 (2016).
- Han, Song, et al. "EIE: efficient inference engine on compressed deep neural network," ISCA, 2016.

References (Alphabetical by Author)

- Han, Song, et al. "Learning both weights and connections for efficient neural network," NIPS, 2015.
- Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," ICLR, 2016.
- He, Kaiming, et al. "Deep residual learning for image recognition," CVPR, 2016.
- Horowitz, Mark. "Computing's energy problem (and what we can do about it)," ISSCC, 2014.
- Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size," ICLR, 2017.
- Ioffe, Sergey, and Szegedy, Christian, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," ICML, 2015
- Jermyn, Michael, et al., "Neural networks improve brain cancer detection with Raman spectroscopy in the presence of operating room light artifacts," Journal of Biomedical Optics, 2016
- Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit," ISCA, 2017.
- Judd, Patrick, et al. "Reduced-precision strategies for bounded memory in deep neural nets." arXiv preprint arXiv:1511.05236 (2015).
- Judd, Patrick, Jorge Albericio, and Andreas Moshovos. "Stripes: Bit-serial deep neural network computing." IEEE Computer Architecture Letters (2016).
- Kim, Duckhwan, et al. "Neurocube: a programmable digital neuromorphic architecture with high-density 3D memory." ISCA 2016
- Kim, Yong-Deok, et al. "Compression of deep convolutional neural networks for fast and low power mobile applications." ICLR 2016

References (Alphabetical by Author)

- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks," NIPS, 2012.
- Lavin, Andrew, and Gray, Scott, "Fast Algorithms for Convolutional Neural Networks," arXiv preprint arXiv: 1509.09308 (2015)
- LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.
- LeCun, Yann, et al. "Optimal brain damage," NIPS, 1989.
- Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." arXiv preprint arXiv:1312.4400 (2013).
- Mathieu, Michael, Mikael Henaff, and Yann LeCun. "Fast training of convolutional networks through FFTs." arXiv preprint arXiv:1312.5851 (2013).
- Merola, Paul A., et al. "Artificial brains. A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, 2014
- Moons, Bert, and Marian Verhelst. "A 0.3–2.6 TOPS/W precision-scalable processor for real-time large-scale ConvNets." Symposium on VLSI Circuits (VLSI-Circuits), 2016.
- Moons, Bert, et al. "Energy-efficient ConvNets through approximate computing," WACV, 2016.
- Parashar, Angshuman, et al. "SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks." *ISCA,* 2017.
- Park, Seongwook, et al., "A 1.93TOPS/W Scalable Deep Learning/Inference Processor with Tetra-Parallel MIMD Architecture for Big-Data Applications," ISSCC, 2015
- Peemen, Maurice, et al., "Memory-centric accelerator design for convolutional neural networks," ICCD, 2013
- Prezioso, Mirko, et al. "Training and operation of an integrated neuromorphic network based on metal-oxide memristors." Nature 521.7550 (2015): 61-64.

References (Alphabetical by Author)

- Rastegari, Mohammad, et al. "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," ECCV, 2016
- Reagen, Brandon, et al. "Minerva: Enabling low-power, highly-accurate deep neural network accelerators," ISCA, 2016.
- Rhu, Minsoo, et al., "vDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design," MICRO, 2016
- Russakovsky, Olga, et al. "Imagenet large scale visual recognition challenge." International Journal of Computer Vision 115.3 (2015): 211-252.
- Sermanet, Pierre, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks," CVPR, 2014.
- Shafiee, Ali, et al. "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars." Proc. ISCA. 2016.
- Shen, Yichen, et al. "Deep learning with coherent nanophotonic circuits." *Nature Photonics*, 2017.
- Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
- Szegedy, Christian, et al. "Going deeper with convolutions," CVPR, 2015.
- Vasilache, Nicolas, et al. "Fast convolutional nets with fbfft: A GPU performance evaluation." arXiv preprint arXiv:1412.7580 (2014).
- Yang, Tien-Ju, et al. "Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning," CVPR, 2017
- Yu, Jiecao, et al. "Scalpel: Customizing DNN Pruning to the Underlying Hardware Parallelism." *ISCA*, 2017.
- Zhang, Chen, et al., "Optimizing FPGA-based Accelerator Design for Deep Convolutional Neural Networks," FPGA, 2015