System-level design for efficient execution of CNNs at the edge

Minakova, S.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, 2015.

[2] M. Abadi, M. Isard, and D. G. Murray. A computational model for tensorflow: An introduction. In *1st ACM SIGPLAN International Workshop on Machine Learning and Programming (MAPL)*, MAPL 2017, page 1–7, New York, NY, USA, 2017. Association for Computing Machinery.

[3] M. S. Abdelfattah, L. Dudziak, T. Chau, R. Lee, H. Kim, and N. D. Lane. Best of both worlds: Automl codesign of a cnn and its hardware accelerator. In *Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

[4] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. V. Essen, A. A. S. Awwal, and V. K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *CoRR*, abs/1803.01164, 2018.

[5] M. Alwani, H. Chen, M. Ferdman, and P. Milder. Fused-layer cnn accelerators. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1–12, 2016.

[6] Y. Ando, S. Seiya, S. Honda, H. Tomiyama, and H. Takada. Automated identification of performance bottleneck on embedded systems

for design space exploration. *Embedded Systems: Design, Analysis and Verification*, pages 171–180, 2013.

[7] J. Bai, F. Lu, and K. Zhang. Open neural network exchange format (onnx) models zoo. https://github.com/onnx/models.

[8] A. Barbier and others at https://github.com/ARM software/ComputeLibrary/graphs/contributors. Arm compute library. https://github.com/ARM-software/ComputeLibrary.

[9] H. Benmeziane, K. E. Maghraoui, H. Ouarnoughi, S. Niar, M. Wistuba, and N. Wang. A comprehensive survey on hardware-aware neural architecture search. *CoRR*, abs/2101.09336, 2021.

[10] G. Bilsen, M. Engels, and R. Lauwereins. Cyclo-static dataflow. *IEEE Transactions on Signal Processing*, 44(2):397–408, 1996.

[11] D. Blalock, J. Ortiz, J. Frankle, and J. Guttag. What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems (MLSys)*, 2020.

[12] T. Bolukbasi, J. Wang, O. Dekel, and V. Saligrama. Adaptive neural networks for efficient inference. In *International Conference on Machine Learning (ICML)*, page 527–536, 2017.

[13] R. Bonna, D. S. Loubach, G. Ungureanu, and I. Sander. Modeling and simulation of dynamic applications using scenario-aware dataflow. *ACM Transactions on Design Automation of Electronic Systems*, 24(5), 2019.

[14] S. Branco, A. G. Ferreira, and J. Cabral. Machine learning in resource-scarce embedded systems, fpgas, and end-devices: A survey. *Electronics*, 8(11), 2019.

[15] J. M. Carroll. *Scenario-based design: envisioning work and technology in system development*. John Wiley and Sons Inc, 1995.

[16] A. Cheng, J. Dong, C. Hsu, S. Chang, M. Sun, S. Chang, J. Pan, Y. Chen, W. Wei, and D. Juan. Searching toward pareto-optimal device-aware neural architectures. In *International Conference On Computer-Aided Design (ICCAD)*. Association for Computing Machinery, 2018.

[17] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. *IEEE Signal Processing Magazine*, 2018.

[18] A. Chinchuluun et al. *Pareto Optimality, Game Theory And Equilibria*, volume 17. Springer, 01 2008.

[19] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[20] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas, and R. Ghosh. Dall·e mini. https://github.com/borisdayma/dalle-mini, 2021.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[22] M. Dhouibi, A. K. B. Salem, A. Saidi, and S. B. Saoud. Accelerating deep neural networks implementation: A survey. *IET Computers and Digital Techniques*, 2021.

[23] A. Diamant, A. Chatterjee, M. Vallieres, G. Shenouda, and J. Seuntjens. Deep learning in head and neck cancer outcome prediction. *Scientific Reports*, 9:27–64, 2019.

[24] T.-D. Do, M.-T. Duong, Q.-V. Dang, and M.-H. Le. Real-time self-driving car navigation using deep neural network. In *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, pages 7–12, 2018.

[25] T. Elsken, J. H. Metzen, and F. Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20:1–21, 2019.

[26] S. Even. *Graph Algorithms*. Cambridge University Press, 2 edition, 2011.

[27] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

[28] J. Farley and A. Gerstlauer. Memory-aware fusing and tiling of neural networks for accelerated edge inference. *CoRR*, abs/2107.06960, 2021.

[29] M. Garza-Fabre, G. T. Pulido, and C. A. Coello. Ranking methods for many-objective optimization. In *MICAI 2009: Advances in Artificial Intelligence*, pages 633–645, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[30] W. Gay. *Raspberry Pi Hardware Reference*. Apress, USA, 1st edition, 2014.

[31] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer. A survey of quantization methods for efficient neural network inference. *ArXiv*, abs/2103.13630, 2021.

[32] D. Gizopoulos, G. Papadimitriou, A. Chatzidimitriou, V. J. Reddi, B. Salami, O. S. Unsal, A. C. Kestelman, and J. Leng. Modern hardware margins: Cpus, gpus, fpgas recent system-level studies. In *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, pages 129–134, 2019.

[33] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge MA USA, 2016.

[34] A. Gordon, E. Eban, O. Nachum, B. Chen, T. Yang, and E. Choi. Morphnet: Fast & simple resource-constrained structure learning of deep networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1586–1595. IEEE Computer Society, 2018.

[35] J. Hanhirova, T. Kämäräinen, S. Seppälä, M. Siekkinen, V. Hirvisalo, and A. Ylä-Jääski. Latency and throughput characterization of convolutional neural networks for mobile computer vision. MMSys '18, page 204–215, New York, NY, USA, 2018. Association for Computing Machinery.

[36] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] S. Heath. *Embedded Systems Design 2nd Edition*. Newnes, 2002.

[38] C. Hsu, S. Chang, D. Juan, J. Pan, Y. Chen, W. Wei, and S. Chang. MONAS: multi-objective neural architecture search using reinforcement learning. *CoRR*, abs/1806.10332, 2018.

[39] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger. Multi-scale dense networks for resource efficient image classification. International Conference on Learning Representations (ICLR), 2018.

[40] G. Huang, Z. Liu, and K. Q. Weinberger. Densely connected convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[41] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18(1):6869–6898, 2017.

[42] L. N. Huynh, R. Balan, and Y. Lee. Deepsense: A gpu-based deep convolutional neural network framework on commodity mobile devices. In *Workshop on Wearable Systems and ApplicationsJune (WearSys'16)*, pages 25–30, 2016.

[43] L. N. Huynh, Y. Lee, and R. K. Balan. Deepmon: Mobile gpu-based deep learning framework for continuous vision applications. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2017.

[44] E. Jeong, J. Kim, and S. Ha. Tensorrt-based framework and optimization methodology for deep learning inference on jetson boards. *ACM Trans. Embed. Comput. Syst.*, dec 2022. Just Accepted.

[45] E. Jeong, J. Kim, S. Tan, J. Lee, and S. Ha. Deep learning inference parallelization on heterogeneous processors with tensorrt. *IEEE Embedded Systems Letters*, 14(1):15–18, 2022.

[46] W. Jiang, X. Zhang, E. H. Sha, L. Yang, Q. Zhuge, Y. Shi, and J. Hu. Accuracy vs. efficiency: Achieving both through fpga-implementation aware neural architecture search. In *Design Automation Conference (DAC)*, pages 1–6, 2019.

[47] H. Jin, B. Liu, W. Jiang, Y. Ma, X. Shi, B. He, and S. Zhao. Layer-centric memory reuse and data migration for extreme-scale deep learning on many-core architectures. *ACM Transactions on Architecture and Code Optimization*, 15(3), 2018.

[48] D. Kang, D. Kang, J. Kang, S. Yoo, and S. Ha. Joint optimization of speed, accuracy, and energy for embedded image recognition systems. In *2018 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, pages 715–720, 2018.

[49] D. Kang, E. Kim, I. Bae, B. Egger, and S. Ha. C-good: C-code generation framework for optimized on-device deep learning. In *International Conference On Computer-Aided Design (ICCAD)*, 2018.

[50] D. Kang, J. Oh, J. Choi, Y. Yi, and S. Ha. Scheduling of deep learning applications onto heterogeneous processors in an embedded device. *IEEE Access*, 8:43980–43991, 2020.

[51] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). http://www.cs.toronto.edu/ kriz/cifar.html, 2013.

[52] S. Kukkonen and J. Lampinen. Ranking-dominance and many-objective optimization. In *2007 IEEE Congress on Evolutionary Computation*, pages 3983–3990, 2007.

[53] C. Kyrkou, G. Plastiras, T. Theocharides, S. I. Venieris, and C.-S. Bouganis. Dronet: Efficient convolutional neural network detector for real-time uav applications. In *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, pages 967–972, 2018.

[54] L. Lai, N. Suda, and V. Chandra. Not all ops are created equal! In *SysML*, 2018.

[55] M. N. U. Laskar, L. G. S. Giraldo, and O. Schwartz. Correspondence of deep neural networks and the brain for visual textures. *ArXiv*, abs/1806.02888, 2018.

[56] Y. Lecun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[57] E. A. Lee and D. G. Messerschmitt. Synchronous data flow. *Proceedings of the IEEE*, 75(9):1235–1245, 1987.

[58] T. Lee, S. Mckeever, and J. Courtney. Flying free: A research overview of deep learning in drone navigation autonomy. *Drones*, 5(2), 2021.

[59] S. Lipschutz and M. Lipson. *Linear Algebra (Schaum's Outlines)*. McGraw Hill, 4 edition, 2009.

[60] D. Liu, H. Kong, X. Luo, W. Liu, and R. Subramaniam. Bringing AI To Edge: From deep learning's perspective. *Neurocomputing*, 485:297–320, 2022.

[61] L. Liu and J. Deng. Dynamic deep neural networks: Optimizing accuracy-efficiency trade-offs by selective execution. In *AAAI*, pages 3675–3682. AAAI Press, 2018.

[62] L. Lu, Y. Zheng, G. Carneiro, and L. Yang. *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2017.

[63] G. Martin. Overview of the mpsoc design challenge. In *Design Automation Conference (DAC)*, DAC '06, page 274–279, New York, NY, USA, 2006. Association for Computing Machinery.

[64] S. Minakova, D. Sapra, T. Stefanov, and A. D. Pimentel. Scenario based run-time switching for adaptive cnn-based applications at the edge. *ACM Transactions on Embedded Computing Systems (TECS)*, 21(2), 2022.

[65] S. Minakova and T. Stefanov. Buffer sizes reduction for memory-efficient cnn inference on mobile and embedded devices. In *Euromicro Conference on Digital System Design (DSD)*, pages 133–140. IEEE Xplore, 2020.

[66] S. Minakova and T. Stefanov. Memory-throughput trade-off for cnn-based applications at the edge. *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, 1(1), 2022.

[67] S. Minakova, E. Tang, and T. Stefanov. Combining task- and data-level parallelism for high-throughput cnn inference on embedded cpus-gpus mpsocs. In *International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*, pages 18–35, Cham, 2020. Springer International Publishing.

[68] O. Moreira. *Temporal analysis and scheduling of hard real-time radios running on a multi-processor*. PhD thesis, Technical University Eindhoven, 2012.

[69] F. Moya Rueda, R. Grzeszick, G. A. Fink, S. Feldhorst, and M. Ten Hompel. Convolutional neural networks for human activity recognition using body-worn sensors. *Informatics*, 5(2), 2018.

[70] L. Nanni, S. Ghidoni, and S. Brahnam. Ensemble of convolutional neural networks for bioimage classification. *Applied Computing and Informatics*, 17, 2021.

[71] NVIDIA. Jetson embedded platform. //https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-tx2.

[72] NVIDIA. Tensorrt framework. https://developer.nvidia.com/tensorrt, 2016.

[73] M. Olyaiy, C. Ng, and M. Lis. Accelerating dnns inference with predictive layer fusion. In *ICS*, page 291–303. Association for Computing Machinery, 2021.

[74] A. Parvat, J. Chavan, S. Kadam, S. Dev, and V. Pathak. A survey of deep-learning frameworks. In *International Conference on Inventive Systems and Control (ICISC)*, pages 1–7, 2017.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[76] Y. Pisarchyk and J. Lee. Efficient memory management for deep neural net inference. In *MLSys 2020 Workshop on Resource-Constrained Machine Learning (ReCoML 2020)*, 2020.

[77] B. Reagen, U. Gupta, R. Adolf, M. M. Mitzenmacher, A. M. Rush, G.-Y. Wei, and D. Brooks. Weightless: Lossy weight encoding for deep neural network compression. In *International Conference on Learning Representations (ICLR)*, 2018.

[78] P. Refaeilzadeh, L. Tang, and H. Liu. *Encyclopedia of Database Systems*, chapter Cross-Validation, pages 532–538. Springer US, 2009.

[79] A. Reiss. https://archive.ics.uci.edu/ml/datasets/PAMAP2 Physical Activity Monitoring, 2012.

[80] M. Richards and N. Ford. *Fundamentals of Software Architecture: An Engineering Approach*. O'Reilly Media, Incorporated, 2019.

[81] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[82] D. Sapra and A. D. Pimentel. Constrained evolutionary piecemeal training to design convolutional neural networks. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2020.

[83] K. Sastry, D. Goldberg, and G. Kendall. *Genetic Algorithms*, pages 97–125. Springer US, Boston, MA, 2005.

[84] B. Savelli, A. Bria, M. Molinara, C. Marrocco, and F. Tortorella. A multi-context cnn ensemble for small lesion detection. *Artificial Intelligence in Medicine*, 103:101749, 2020.

[85] L. M. Schmitt. Theory of genetic algorithms. *Theoretical Computer Science*, 259(1):1–61, 2001.

[86] M. Seeland and P. Mader. Multi-view classification with convolutional neural networks. *PLoS ONE*, 2021.

[87] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *IEEE International Conference on Machine Learning and Applications (IEEE ICMLA)*, pages 624–628, 2018.

[88] A. K. Singh, A. Prakash, K. R. Basireddy, G. V. Merrett, and B. M. Al-Hashimi. Energy-efficient run-time mapping and thread partitioning of concurrent opencl applications on cpu-gpu mpsocs. *ACM Trans. Embed. Comput. Syst.*, 16(5s), 2017.

[89] H. Sofaer, J. Hoeting, and C. Jarnevich. The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, 10, 12 2018.

[90] L. Song, J. Mao, Y. Zhuo, X. Qian, H. Li, and Y. Chen. Hypar: Towards hybrid parallelism for deep learning accelerator array. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 56–68, 2019.

[91] S. Stuijk, T. Basten, and M. Geilen. Sdf3: Sdf for free. In *Sixth International Conference on Application of Concurrency to System Design*, pages 276–278, Los Alamitos, CA, USA, jun 2006. IEEE Computer Society.

[92] M. Tan, B. Chen, R. Pang, V. Vasudevan, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2019.

[93] E. Tang, S. Minakova, and T. Stefanov. Energy-efficient and high-throughput cnn inference on embedded cpus-gpus mpsocs. In *International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*. Springer, 2021.

[94] L. Tang, Y. Wang, T. L. Willke, and K. Li. Scheduling computation graphs of deep learning models on manycore cpus. *CoRR*, abs/1807.09667, 2018.

[95] B. Taylor, V. S. Marco, W. Wolff, Y. Elkhatib, and Z. Wang. Adaptive selection of deep learning models on embedded systems. In *ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES)*, page 31–43. Association for Computing Machinery, 2018.

[96] I. Theodorakopoulos, V. K. Pothos, D. Kastaniotis, and N. Fragoulis. Parsimonious inference on convolutional neural networks: Learning and applying on-line kernel activation rules. *CoRR*, abs/1701.05221, 2017.

[97] J. Venugopalan, L. Tong, H. R. Hassanzadeh, and M. Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific Reports*, 11, 2021.

[98] M. P. Vestias. A survey of convolutional neural networks on edge with reconfigurable computing. *Algorithms*, 12(8), 2019.

[99] J. Vinu et al. A programmable approach to neural network compression. *IEEE Micro*, 40(5):17–25, 2020.

[100] C.-C. Wang, Y.-C. Liao, M.-C. Kao, W.-Y. Liang, and S.-H. Hung. Perfnet: Platform-aware performance modeling for deep neural networks. RACS '20, page 90–95, New York, NY, USA, 2020. Association for Computing Machinery.

[101] S. Wang, G. Ananthanarayanan, Y. Zeng, N. Goel, A. Pathania, and T. Mitra. High-throughput cnn inference on embedded arm big.little multicore processors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(10):2254–2267, 2020.

[102] Y. Wang, J. Shen, T.-K. Hu, P. Xu, T. Nguyen, R. Baraniuk, Z. Wang, and Y. Lin. Dual dynamic inference: Enabling more efficient, adaptive and controllable deep inference. *IEEE Journal of Selected Topics in Signal Processing*, 14:623–633, 2020.

[103] B. Wu, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, and K. Keutzer. FBNet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10734–10742. Computer Vision Foundation / IEEE, 2019.

[104] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. López. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, PP:1–11, 2020.

[105] Y. Xu, L. Xie, X. Zhang, X. Chen, B. Shi, Q. Tian, and H. Xiong. Latency-aware differentiable neural architecture search. *arXiv preprint arXiv:2001.06392*, 2020.

[106] T.-J. Yang, Y.-H. Chen, and V. Sze. Designing energy-efficient convolutional neural networks using energy-aware pruning. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6071–6079, 2017.

[107] J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang. Slimmable neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[108] J. T. Zhai, S. Niknam, and T. Stefanov. Modeling, analysis, and hard real-time scheduling of adaptive streaming applications. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2636–2648, 2018.

[109] Y. Zhao, W. Wang, Y. Li, C. Colman Meixner, M. Tornatore, and J. Zhang. Edge computing and networking: A survey on infrastructures and applications. *IEEE Access*, pages 1–1, 07 2019.

[110] Z. Zhao, K. M. Barijough, and A. Gerstlauer. Deepthings: Distributed adaptive deep learning inference on resource-constrained iot edge clusters. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(11):2348–2359, 2018.