Methods of Deep Neural Network Compression —Zh-Zhg 2017-6-15

Notes:
cnn-benchmarks
convnet-benchmarks
Benchmarking DNN Processors
Deep Neural Network Energy Estimation Tool

[152] Efficient Processing of Deep Neural Networks: A Tutorial and Survey
[153] An analysis of deep neural network models for practical applications

1. JETC: ACM Journal on Emerging Technologies in Computing Systems
2. SiPS: Signal Processing Systems
3. NC: Neurocomputing-Elsevier
4. JMLR: Journal of Machine Learning Research
5. CoRR: Computing Research Repository
6. ISCA: International Symposium on Computer Architecture
7. CDICS: Computer-Aided Design of Integrated Circuits and Systems
8. PDPS: Parallel and Distributed Processing Symposium
9. FPGA: International Symposium on Field-Programmable Gate Arrays
10. IS: Interspeech
11. ICASSP: Acoustics, Speech and Signal Processing

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[2] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[3] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. *arXiv preprint arXiv:1603.06318*, 2016.

[4] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[5] Gregor Urban, Krzysztof J Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Rich Caruana, Abdelrahman Mohamed, Matthai Philipose, and Matt Richardson. Do deep convolutional nets really need to be deep and convolutional? *arXiv preprint arXiv:1603.05691*, 2016.

[6] William Chan, Nan Rosemary Ke, and Ian Lane. Transferring knowledge from a rnn to a dnn. *arXiv preprint arXiv:1504.01483*, 2015.

[7] Ping Luo, Zhenyao Zhu, Ziwei Liu, Xiaogang Wang, Xiaoou Tang, et al. Face model compression by distilling knowledge from neurons. In *AAAI*, pages 3560–3566, 2016.

[8] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[9] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv preprint arXiv:1707.01220*, 2017.

[10] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. 2016.

[11] Zhenyang Wang, Zhidong Deng, and Shiyao Wang. Accelerating convolutional neural networks with dominant convolutional kernel and knowledge pre-regression. In *European Conference on Computer Vision*, pages 533–548, 2016.

[12] Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.

[13] Babak Hassibi, David G Stork, et al. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, pages 164–164, 1993.

[14] Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pages 2148–2156, 2013.

[15] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, pages 1135–1143, 2015.

[16] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[17] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, et al. Dsd: Dense-sparse-dense training for deep neural networks. 2016.

[18] Suraj Srinivas and R Venkatesh Babu. Data-free parameter pruning for deep neural networks. *arXiv preprint arXiv:1507.06149*, 2015.

[19] Yiwen Guo, Anbang Yao, and Yurong Chen. Dynamic network surgery for efficient dnns. In *Advances In Neural Information Processing Systems*, pages 1379–1387, 2016.

[20] Jongsoo Park, Sheng Li, Wei Wen, Ping Tak Peter Tang, Hai Li, Yiran Chen, and Pradeep Dubey. Faster cnns with direct sparse convolutions and guided pruning. 2016.

[21] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient transfer learning. *arXiv preprint arXiv:1611.06440*, 2016.

[22] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[23] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. Structured pruning of deep convolutional neural networks. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 13(3):32, 2017.

[24] Mohammad Babaeizadeh, Paris Smaragdis, and Roy H Campbell. A simple yet effective method to prune dense layers of neural networks. 2016.

[25] Tien-Ju Yang, Yu-Hsin Chen, and Vivienne Sze. Designing energy-efficient convolutional neural networks using energy-aware pruning. *arXiv preprint arXiv:1611.05128*, 2016.

[26] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. *arXiv preprint arXiv:1705.08665*, 2017.

[27] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

[28] Zelda Mariet and Suvrit Sra. Diversity networks. *arXiv preprint arXiv:1511.05077*, 2015.

[29] Jian-Hao Luo and Jianxin Wu. An entropy-based pruning method for cnn compression. *arXiv preprint arXiv:1706.05791*, 2017.

[30] Huizi Mao, Song Han, Jeff Pool, Wenshuo Li, Xingyu Liu, Yu Wang, and William J Dally. Exploring the regularity of sparse structure in convolutional neural networks. *arXiv preprint arXiv:1705.08922*, 2017.

[31] Vadim Lebedev and Victor Lempitsky. Fast convnets using group-wise brain damage. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2554–2564, 2016.

[32] Jian-Hao Luo, Jianxin Wu, and Weiyao Lin. Thinet: A filter level pruning method for deep neural network compression. *arXiv preprint arXiv:1707.06342*, 2017.

[33] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

[34] Shuai Zheng, Abhinav Vishnu, and Chris Ding. Accelerating deep learning with shrinkage and recall. In *Parallel and Distributed Systems (ICPADS), 2016 IEEE 22nd International Conference on*, pages 963–970. IEEE, 2016.

[35] Xin Li and Changsong Liu. Prune the convolutional neural networks with sparse shrink. *Electronic Imaging*, 2017(6):97–101, 2017.

[36] Fernando Moya Rueda, Rene Grzeszick, and Gernot A Fink. Neuron pruning for compressing deep networks using maxout architectures. 2017.

[37] Frederick Tung, Srikanth Muralidharan, and Greg Mori. Fine-pruning: Joint fine-tuning and compression of a convolutional network with bayesian optimization. 2017.

[38] Kirill Neklyudov, Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Structured bayesian pruning via log-normal multiplicative noise. 2017.

[39] Yunhe Wang, Chang Xu, Jiayan Qiu, Chao Xu, and Dacheng Tao. Towards evolutional compression. 2017.

[40] Sam Leroux, Steven Bohez, Cedric De Boom, Elias De Coninck, Tim Verbelen, Bert Vankeirsbilck, Pieter Simoens, and Bart Dhoedt. Lazy evaluation of convolutional filters. 2016.

[41] Arash Ardakani, Carlo Condo, and Warren J Gross. Sparsely-connected neural networks: Towards efficient vlsi implementation of deep neural networks. 2016.

[42] Alireza Aghasi, Nam Nguyen, and Justin Romberg. Net-trim: A layer-wise convex pruning of deep neural networks. 2016.

[43] Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. 2017.

[44] Denis A. Gudovskiy and Luca Rigazio. Shiftcnn: Generalized low-precision architecture for inference of convolutional neural networks. 2017.

[45] Lei Deng, Peng Jiao, Jing Pei, Zhenzhi Wu, and Guoqi Li. Gated xnor networks: Deep neural networks with ternary weights and activations under a unified discretization framework. 2017.

[46] Alexander G Anderson and Cory P Berg. The high-dimensional geometry of binary neural networks. 2017.

[47] Kyuyeon Hwang and Wonyong Sung. Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1. In *Signal Processing Systems (SiPS), 2014 IEEE Workshop on*, pages 1–6. IEEE, 2014.

[48] Darryl Lin, Sachin Talathi, and Sreekanth Annapureddy. Fixed point quantization of deep convolutional networks. In *International Conference on Machine Learning*, pages 2849–2858, 2016.

[49] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems*, pages 3123–3131, 2015.

[50] Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

[51] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *arXiv preprint arXiv:1609.07061*, 2016.

[52] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[53] Minje Kim and Paris Smaragdis. Bitwise neural networks. *arXiv preprint arXiv:1601.06071*, 2016.

[54] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014.

[55] Guillaume Soulié, Vincent Gripon, and Maëlys Robert. Compression of deep neural networks on the fly. In *International Conference on Artificial Neural Networks*, pages 153–160. Springer, 2016.

[56] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

[57] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[58] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

[59] Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.

[60] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng. Quantized convolutional neural networks for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2016.

[61] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *International Conference on Machine Learning*, pages 2285–2294, 2015.

[62] Ryan Spring and Anshumali Shrivastava. Scalable and sustainable deep learning via randomized hashing. *arXiv preprint arXiv:1602.08194*, 2016.

[63] Lei Shi, Shikun Feng, et al. Functional hashing for compressing neural networks. *arXiv preprint arXiv:1605.06560*, 2016.

[64] Wenlin Chen, James T Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks. *arXiv preprint arXiv:1506.04449*, 2015.

[65] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1475–1484, 2016.

[66] Zhouhan Lin, Matthieu Courbariaux, Roland Memisevic, and Yoshua Bengio. Neural networks with few multiplications. *arXiv preprint arXiv:1510.03009*, 2015.

[67] Zhiyong Cheng, Daniel Soudry, Zexi Mao, and Zhenzhong Lan. Training binary multilayer neural networks for image classification using expectation backpropagation. *arXiv preprint arXiv:1503.03562*, 2015.

[68] Vincent Vanhoucke, Andrew Senior, and Mark Z Mao. Improving the speed of neural networks on cpus. In *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*, volume 1, page 4, 2011.

[69] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

[70] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Towards the limit of network quantization. *arXiv preprint arXiv:1612.01543*, 2016.

[71] Alexander Novikov, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. Tensorizing neural networks. In *Advances in Neural Information Processing Systems*, pages 442–450, 2015.

[72] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *ICML*, pages 1737–1746, 2015.

[73] Jordan L Holi and J-N Hwang. Finite precision error analysis of neural network hardware implementations. *IEEE Transactions on Computers*, 42(3):281–290, 1993.

[74] Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. *arXiv preprint arXiv:1702.00953*, 2017.

[75] Zhaofan Qiu, Ting Yao, and Tao Mei. Deep quantization: Encoding convolutional activations with deep generative model. *arXiv preprint arXiv:1611.09502*, 2016.

[76] Junwhan Ahn Eunhyeok Park and Sungjoo Yoo. Weighted-entropy-based quantization for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4820–4828, 2017.

[77] Cong Leng, Hao Li, Shenghuo Zhu, and Rong Jin. Extremely low bit neural network: Squeeze the last bit out with admm. *arXiv preprint arXiv:1707.09870*, 2017.

[78] Yinpeng Dong, Renkun Ni, Jianguo Li, Yurong Chen, Jun Zhu, and Hang Su. Learning accurate low-bit deep neural networks with stochastic quantization. *arXiv preprint arXiv:1708.01001*, 2017.

[79] Jong Hwan Ko, Duckhwan Kim, Taesik Na, Jaeha Kung, and Saibal Mukhopadhyay. Adaptive weight compression for memory-efficient neural networks. In *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 199–204. IEEE, 2017.

[80] Shu-Chang Zhou, Yu-Zhi Wang, He Wen, Qin-Yao He, and Yu-Heng Zou. Balanced quantization: An effective and efficient approach to quantized neural networks. *Journal of Computer Science and Technology*, 32(4):667–682, 2017.

[81] Bert Moons, Bert De Brabandere, Luc Van Gool, and Marian Verhelst. Energy-efficient convnets through approximate computing. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[82] Jeng Hau Lin, Tianwei Xing, Ritchie Zhao, Zhiru Zhang, Mani Srivastava, Zhuowen Tu, and Rajesh K. Gupta. Binarized convolutional neural networks with separable filters for efficient hardware acceleration. 2017.

[83] Wenjun Zhang XiaoKang Yang Wen Gao Zefan Li, Bingbing Ni. Performance guaranteed network acceleration via high-order residual quantization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 0–0, 2017.

[84] Lu Hou, Quanming Yao, and James T Kwok. Loss-aware binarization of deep networks. *arXiv preprint arXiv:1611.01600*, 2016.

[85] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[86] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. Flattened convolutional neural networks for feedforward acceleration. *arXiv preprint arXiv:1412.5474*, 2014.

[87] Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. Learning separable filters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2754–2761, 2013.

[88] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *arXiv preprint arXiv:1405.3866*, 2014.

[89] Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, pages 1269–1277, 2014.

[90] Vadim Lebedev, Yaroslav Ganin, Maksim Rakhuba, Ivan Oseledets, and Victor Lempitsky. Speeding-up convolutional neural networks using fine-tuned cp-decomposition. *arXiv preprint arXiv:1412.6553*, 2014.

[91] Xiangyu Zhang, Jianhua Zou, Xiang Ming, Kaiming He, and Jian Sun. Efficient and accurate approximations of nonlinear convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1984–1992, 2015.

[92] Shuchang Zhou and Jia-Nan Wu. Compression of fully-connected layer in neural network by kronecker product. *arXiv preprint arXiv:1507.05775*, 2015.

[93] Jian Xue, Jinyu Li, and Yifan Gong. Restructuring of deep neural network acoustic models with singular value decomposition. In *Interspeech*, pages 2365–2369, 2013.

[94] Yu Cheng, Felix X Yu, Rogerio S Feris, Sanjiv Kumar, Alok Choudhary, and Shi-Fu Chang. An exploration of parameter redundancy in deep networks with circulant projections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2857–2865, 2015.

[95] Cheng Tai, Tong Xiao, Yi Zhang, Xiaogang Wang, et al. Convolutional neural networks with low-rank regularization. *arXiv preprint arXiv:1511.06067*, 2015.

[96] Zichao Yang, Marcin Moczulski, Misha Denil, Nando de Freitas, Alex Smola, Le Song, and Ziyu Wang. Deep fried convnets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1476–1483, 2015.

[97] Yani Ioannou, Duncan Robertson, Jamie Shotton, Roberto Cipolla, and Antonio Criminisi. Training cnns with low-rank filters for efficient image classification. *arXiv preprint arXiv:1511.06744*, 2015.

[98] Min Wang, Baoyuan Liu, and Hassan Foroosh. Factorized convolutional neural networks. *arXiv preprint arXiv:1608.04337*, 2016.

[99] Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

[100] Peisong Wang and Jian Cheng. Accelerating convolutional neural networks for mobile applications. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 541–545. ACM, 2016.

[101] Jose Alvarez and Lars Petersson. Decomposeme: Simplifying convnets for end-to-end learning. 2016.

[102] Vikas Sindhwani, Tara Sainath, and Sanjiv Kumar. Structured transforms for small-footprint deep learning. In *Advances in Neural Information Processing Systems*, pages 3088–3096, 2015.

[103] Min Wang, Baoyuan Liu, and Hassan Foroosh. Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial bottleneck structure.

[104] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6655–6659. IEEE, 2013.

[105] Sek Chai, Aswin Raghavan, David Zhang, Mohamed Amer, and Tim Shields. Low precision neural networks using subband decomposition. *arXiv preprint arXiv:1703.08595*, 2017.

[106] Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Beyond filters: Compact feature map for portable deep model. In *International Conference on Machine Learning*, pages 3703–3711, 2017.

[107] Liang Zhao, Siyu Liao, Yanzhi Wang, Jian Tang, and Bo Yuan. Theoretical properties for neural networks with weight matrices of low displacement rank. *arXiv preprint arXiv:1703.00144*, 2017.

[108] Wei Wen, Cong Xu, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Coordinating filters for faster deep neural networks. *arXiv preprint arXiv:1703.09746*, 2017.

[109] Timur Garipov, Dmitry Podoprikhin, Alexander Novikov, and Dmitry Vetrov. Ultimate tensorization: compressing convolutional and fc layers alike. *arXiv preprint arXiv:1611.03214*, 2016.

[110] Jaeyong Chung and Taehwan Shin. Simplifying deep neural networks for neuromorphic architectures. In *Design Automation Conference (DAC), 2016 53nd ACM/EDAC/IEEE*, pages 1–6. IEEE, 2016.

[111] Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814, 2015.

[112] Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

[113] Soravit Changpinyo, Mark Sandler, and Andrey Zhmoginov. The power of sparsity in convolutional neural networks. *arXiv preprint arXiv:1702.06257*, 2017.

[114] Benjamin Graham. Spatially-sparse convolutional neural networks. *arXiv preprint arXiv:1409.6070*, 2014.

[115] Guoliang Kang, Jun Li, and Dacheng Tao. Shakeout: A new approach to regularized deep neural network training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[116] Markus Thom and Günther Palm. Sparse activity and sparse connectivity in supervised learning. *Journal of Machine Learning Research*, 14(Apr):1091–1143, 2013.

[117] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2074–2082, 2016.

[118] Mikhail Figurnov, Aizhan Ibraimova, Dmitry P Vetrov, and Pushmeet Kohli. Perforatedcnns: Acceleration through elimination of redundant convolutions. In *Advances in Neural Information Processing Systems*, pages 947–955, 2016.

[119] Shengjie Wang, Haoran Cai, Jeff Bilmes, and William Noble. Training compressed fully-connected networks with a density-diversity penalty. 2016.

[120] Mohammad Javad Shafiee, Parthipan Siva, and Alexander Wong. Stochasticnet: Forming deep neural networks via stochastic connectivity. *IEEE Access*, 4:1915–1924, 2016.

[121] Yani Ioannou, Duncan Robertson, Roberto Cipolla, and Antonio Criminisi. Deep roots: Improving cnn efficiency with hierarchical filter groups. *arXiv preprint arXiv:1605.06489*, 2016.

[122] Hao Zhou, Jose M Alvarez, and Fatih Porikli. Less is more: Towards compact cnns. In *European Conference on Computer Vision*, pages 662–677. Springer, 2016.

[123] Xuanyi Dong, Junshi Huang, Yi Yang, and Shuicheng Yan. More is less: A more complicated network with less inference complexity. *arXiv preprint arXiv:1703.08651*, 2017.

[124] Maxwell D Collins and Pushmeet Kohli. Memory bounded deep convolutional networks. *arXiv preprint arXiv:1412.1442*, 2014.

[125] Jaehong Yoon and Sung Ju Hwang. Combined group and exclusive sparsity for deep neural networks. In *International Conference on Machine Learning*, pages 3958–3966, 2017.

[126] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition.

[127] Shaohuai Shi and Xiaowen Chu. Speeding up convolutional neural networks by exploiting the sparsity of rectifier units. *arXiv preprint arXiv:1704.07724*, 2017.

[128] Hessam Bagherinezhad, Mohammad Rastegari, and Ali Farhadi. Lcnn: Lookup-based convolutional neural network. *arXiv preprint arXiv:1611.06473*, 2016.

[129] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. *arXiv preprint arXiv:1608.06049*, 2016.

[130] Kaiming He and Jian Sun. Convolutional neural networks at constrained time cost. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5353–5360, 2015.

[131] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[132] Chunpeng Wu, Wei Wen, Tariq Afzal, Yongmei Zhang, Yiran Chen, and Hai Li. A compact dnn: Approaching googlenet-level accuracy of classification and domain adaptation. *arXiv preprint arXiv:1703.04071*, 2017.

[133] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*, 2017.

[134] Nadav Cohen, Or Sharir, and Amnon Shashua. Deep simnets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4782–4791, 2016.

[135] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through ffts. *arXiv preprint arXiv:1312.5851*, 2013.

[136] Andrew Lavin and Scott Gray. Fast algorithms for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4021, 2016.

[137] Nicolas Vasilache, Jeff Johnson, Michael Mathieu, Soumith Chintala, Serkan Piantino, and Yann LeCun. Fast convolutional nets with fbfft: A gpu performance evaluation. *arXiv preprint arXiv:1412.7580*, 2014.

[138] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A Horowitz, and William J Dally. Eie: efficient inference engine on compressed deep neural network. In *Proceedings of the 43rd International Symposium on Computer Architecture*, pages 243–254. IEEE Press, 2016.

[139] Renzo Andri, Lukas Cavigelli, Davide Rossi, and Luca Benini. Yodann: An architecture for ultra-low power binary-weight cnn acceleration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2017.

[140] Aleksandar Zlateski, Kisuk Lee, and H Sebastian Seung. Znn–a fast and scalable algorithm for training 3d convolutional networks on multi-core and many-core shared memory machines. In *Parallel and Distributed Processing Symposium, 2016 IEEE International*, pages 801–811. IEEE, 2016.

[141] Yaman Umuroglu, Nicholas J Fraser, Giulio Gambardella, Michaela Blott, Philip Leong, Magnus Jahre, and Kees Vissers. Finn: A framework for fast, scalable binarized neural network inference. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 65–74. ACM, 2017.

[142] Juyong Kim, Yookoon Park, Gunhee Kim, and Sung Ju Hwang. Splitnet: Learning to semantically split deep networks for parameter reduction and model parallelization. In *International Conference on Machine Learning*, pages 1866–1874, 2017.

[143] Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *International Conference on Machine Learning*, pages 527–536, 2017.

[144] Minsik Cho and Daniel Brand. Mec: Memory-efficient convolution for deep neural network. *arXiv preprint arXiv:1706.06873*, 2017.

[145] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. Cnnpack: packing convolutional neural networks in the frequency domain. In *Advances in Neural Information Processing Systems*, pages 253–261, 2016.

[146] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[147] Manoj Alwani, Han Chen, Michael Ferdman, and Peter Milder. Fused-layer cnn accelerators. In *Microarchitecture (MICRO), 2016 49th Annual IEEE/ACM International Symposium on*, pages 1–12. IEEE, 2016.

[148] Lingxi Xie and Alan Yuille. Genetic cnn. 2017.

[149] Marcin Moczulski, Misha Denil, Jeremy Appleyard, and Nando De Freitas. Acdc: A structured efficient linear layer. *Computer Science*, 2015.

[150] Zhe Li, Xiaoyu Wang, Xutao Lv, and Tianbao Yang. Sep-nets: Small and effective pattern networks. 2017.

[151] Sanjay Ganapathy, Swagath Venkataramani, Balaraman Ravindran, and Anand Raghunathan. Dyvedeep: Dynamic variable effort deep neural networks. 2017.

[152] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel Emer. Efficient processing of deep neural networks: A tutorial and survey. *arXiv preprint arXiv:1703.09039*, 2017.

[153] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

Class		Conference	Cite	Detailed	Superiority	Weakness
Transfer	[1]	arXiv2015	337	teach small model	distill knowledge	-
	[2]	ICLR2015	172	deeper+thinner	quickly+accuracy	-
	[3]	arXiv2016	22	FOL rules	iterative distill	-
	[4]	NIPS2014	204	shallow nets	CIFAR+TIMIT	-
	[5]	ICLR2017	12	yes,they do	CIFAR10	-
	[6]	arXiv2015	23	RNN to DNN	soft alignments	-
	[7]	AAAI2016	5	face-model-com	51c+90s tea	-
	[8]	arXiv2017	0	NTS	-	-
	[9]	arXiv2017	0	DarkRank	-	-
	[10]	ICLR2017	8	pay more attention	-	-
	[11]	ECCV2016	1	pre-regression	-	-
Pruning	[12]	NIPS1990	2202	brain damage	removing weights	diagonal Hessian
	[13]	NIPS1993	952	brain surgeon	remove right	high computation
	[14]	NIPS2013	154	95% redundancy	learning weights	-
	[15]	NIPS2015	206	prune connect	AlexNet 9*	no-acceleration
	[16]	ICLR2016	239	deep compression	AlexNet 35*	-
	[17]	ICLR2017	2	DSD	better accuracy	-
	[18]	arXiv2015	27	similar neurons	data-free	fc-layer
	[19]	NIPS2016	17	DNS error	connec splicing	no-acceleration
	[20]	ICLR2017	3	sparse+prune	AlexNet 3-7*	-
	[21]	ICLR2017	2	prune conv-kern	Taylor expan	-
	[22]	ICLR2017	14	prune filters	VGG 32*	-
	[23]	JETC2017	20	structured sparse	-	no-imagenet
	[24]	arXiv2017	0	NoiseOut	correl-neuron	insufficient-exp
	[25]	arXiv2016	6	energy-aware	prune weight	-
	[26]	arXiv2017	0	bayesian-compre	prune nodes	-
	[27]	ICML2015	101	weight pruning	Bayes-by-BP	-
	[28]	ICLR2016	16	DivNet	DPP+fuse-neur	just-fully
	[29]	arXiv2017	0	Entropy-based	VGG 3.3s+16c	-
	[30]	NIPS2017	0	explore prune	coarse-grained	-
	[31]	CVPR2016	29	group bra-dam	-	-
	[32]	arXiv2017	1	ThiNet	prune filter	-
	[33]	arXiv2016	16	Data-Driven	prune neurons	-
	[34]	PADS2016	3	prune data	speed train	-
	[35]	EI2017	0	prune channel	Sparse Shrink	-
	[36]	arXiv2017	0	prune maxout	-	-
	[37]	arXiv2017	0	Fine-Pruning	-	-
	[38]	arXiv2017	0	StructuredBP	-	no imagenet
	[39]	arXiv2017	0	Evolutional	prune filters	-
	[40]	ICMLW16	1	prune filters	absolute sum	-
	[41]	ICLR2017	1	sparsely-fc	-	-
	[42]	arXiv2016	2	Net-Trim	-	-
	[43]	arXiv2017	0	theoratical-view	training quan	-
	[44]	arXiv2017	0	ShiftCNN	Generalized	-
	[45]	arXiv2017	0	Gated-XNOR	-	-
	[46]	arXiv2017	0	High-dimen	-	-

Class		Conference	Cite	Detailed	Superiority	Weakness
Quantization	[47]	SiPS2014	44	+1,0,-1	little loss	just-fully
	[48]	ICML2016	25	fix-point quantiza	bit-width alloc	-
	[49]	NIPS2015	141	binary-connc	-	no-imagenet
	[50]	CoRR2016	96	BinaryNet	MNIST 7*faster	no-save-param
	[51]	arXiv2016	28	QNN	AlexNet 51% acc	-
	[52]	ECCV2016	118	XNOR-Net	58* faster	-
	[53]	arXiv2016	34	all-binary	-	no-conv
	[54]	arXiv2014	104	vector-quantiz	compress 16*	-
	[55]	ICANN2016	5	on-the-fly	extra-regulariz	no-imagenet
	[56]	arXiv2016	25	DoReFa-net	diff-bitwidth	-
	[57]	arXiv2016	21	-w,0,w	compression-32*	-
	[58]	ICLR2017	10	train-ternary confuse	16* smaller	-
	[59]	ICLR2017	5	INQ	AlexNet 89*	-
	[60]	CVPR2016	28	QCNN–PQ	6speed-20comp	-
	[61]	ICML2015	112	HashNet	randomly-group	-
	[62]	arXiv2016	4	sustainable-LSH	5% multip	-
	[63]	arXiv2016	0	FunHashNN	-	-
	[64]	NIPS2015	18	DCT+Hash	-	-
	[65]	KDD2016	1	FreshNet	feature hashing	-
	[66]	ICLR2016	59	few-multip	quantized-BP	-
	[67]	arXiv2015	14	train-binary	expect-BP	-
	[68]	NIPS2011	180	fixed-point	CPU-speed	-
	[69]	ICLR2017	6	soft-weight-share	-	-
	[70]	ICLR2017	2	Hessian-weight kms	2.4% of AlexNet	-
	[71]	NIPS2015	55	Tensorizing-NN	-	just-fully
	[72]	ICML2015	155	Stocha-rounding	16 bits	-
	[73]	ToC1993	218	necessary precision	theoretical-analys	-
	[74]	CVPR2017	1	HWGQ-Net	train-low-precisi	-
	[75]	CVPR2017	5	Deep Quantiza	FV-VAE	-
	[76]	CVPR2017	0	weighted-entropy	multi-bit Q	-
	[77]	arXiv2017	0	low-bit ADMM	-	-
	[78]	arXiv2017	0	Stochastic Q error	-	-
	[79]	DATE2017	1	JPEG encoding	-	-
	[80]	JCST2017	0	balanced-Q	-	-
	[81]	WACV2016	0	energy efficient	-	-
	[82]	arXiv2017	0	BCNNw/SF	-	-
	[83]	ICCV2017	0	residual Q	xnor	-
	[84]	ICLR2017	4	loss-aware Bin	CNN,RNN	-

Class	Conference	Cite	Detailed	Superiority	Weakness
Decomposition and Low-Rank					
[85]	arXiv2017	5	MobileNets	depthwise	-
[86]	arXiv2015	13	flattened	3-1D-kernel	-
[87]	CVPR2013	59	separable filters	Separable-conv	-
[88]	BMVC2014	149	3D-to-2conv	speed 4.5*	-
[89]	NIPS2014	164	Biclustering	speed 2*	no whole-model
[90]	ICLR2015	55	CP-decomposition	CPU 8.5*speed	a single-layer
[91]	CVPR2015	33	approx-nonlinear	speed 4*	-
[92]	ICACI2016	1	Kronecker Product	Alex-10*-reduce	just-fully
[93]	IS2013	117	restruct-svd	reduc-80%	just-fully
[94]	ICCV2015	36	Circulant-Project	-	just-fully
[95]	ICLR2016	12	low-rank regula	speed 2*	-
[96]	ICCV2015	67	Fastfood transform	train-scratch	just-fully
[97]	ICLR2016	17	train low-rank	diffshapefilter	-
[98]	arXiv2016	7	factorized-CNN	single-in-channe	-
[99]	ICLR2016	45	Tucker-decompos	-	-
[100]	ACMMM16	4	BTD	6.6*VGG-speed	-
[101]	arXiv2016	2	DecomposeMe	-	-
[102]	NIPS2015	35	Structure-Transform	3.5% compres	-
[103]	arXiv2017	0	Topolo-Subdivision	-	-
[104]	ICASSP13	165	final-weight-layer	-	speech-recog
[105]	arXiv2017	2	subband-decom	fusion better	-
[106]	ICML2017	0	Beyond Filters	feature maps	-
[107]	arXiv2017	0	LDR	theoretical	-
[108]	ICCV2017	1	force-regular	training	-
[109]	arXiv2016	4	ultimate tensor	-	-
[110]	DAC2016	6	factor+prune	-	-

Class		Conference	Cite	Detailed	Superiority	Weakness
Sparse	[111]	CVPR2015	63	sparse-CNN	90% sparse	-
	[112]	NC2017	8	Group sparsity	-	-
	[113]	arXiv2017	2	power-of-sparsity	sparse-random	-
	[114]	arXiv2014	57	Spatially-sparse	-	-
	[115]	TPAMI2017	0	Shakeout	-	-
	[116]	JMLR2013	18	sparsen projection	-	-
	[117]	NIPS2016	26	SSL	structured-sparse	-
	[118]	NIPS2016	15	PerforatedCNNs	skip-spatial-pos	-
	[119]	ICLR2017	0	density-diversity	-	especial-fully
	[120]	Access16	8	Stochasticnet	stochastic-connec	no-imagenet
	[121]	arXiv2016	4	Deep Roots	Hier-Filt-Group	-
	[122]	ECCV2016	3	Less is More	neuron-reduct	-
	[123]	arXiv2017	1	More is Less	skip-0-position	-
	[124]	arXiv2014	41	memory-bounded	just-fully	store indexes
	[125]	ICML2017	0	Exclusive Sparsity	+Group Sparsity	-
	[126]	CVPR2017	0	low-rank+sparse	GreBdec	-
	[127]	arXiv2017	2	skip-0-neuron	no-whole-net	just CPU
Design	[128]	arXiv2016	3	LCNN	lookup-based	-
	[129]	arXiv2016	2	LBCNN	9-169 save-param	-
	[130]	CVPR2015	78	constrain-time	-	-
	[131]	arXiv2016	99	SqueezedNet	AlexNet 50*fewer	-
	[132]	arXiv2017	1	A Compact DNN	Domain Adaptat	-
	[133]	arXiv2017	0	ShuffleNet	group-conv	-
	[134]	CVPR2016	13	Deep SimNets	-	-
	[135]	CoRR2013	111	ffts	-	power of 2
	[136]	CVPR2016	68	Winograd	-	kernel 3×3
	[137]	ICLR2015	75	fbfft	-	-
	[138]	ISCA2016	93	EIE	-	-
	[139]	CDICS2017	1	YodaNN	-	-
	[140]	PDPS2016	11	ZNN	-	-
	[141]	FPGA2017	4	FINN	-	-
	[142]	ICML2017	0	SplitNet	-	-
	[143]	ICML2017	0	early-exit	2.8*speed	-
	[144]	ICML2017	0	MEC	memory-efficient	-
	[145]	NIPS2016	10	CNNpack	Frequency Domain	-
	[146]	CVPR2017	155	Densely	-	-
	[147]	MICRO16	9	Fused-Layer	-	-
	[148]	arXiv2017	3	GeneticCNN	-	-
	[149]	ICLR2016	23	ACDC	new layer	fc-layer
	[150]	arXiv2017	1	SEP-Nets	binary 3x3	-
	[151]	arXiv2017	0	DyVEDeep	-	-