@article{Vasilache:2019:NAL:3366460.3355606, author = {Vasilache, Nicolas and Zinenko, Oleksandr and Theodoridis, Theodoros and Goyal, Priya and Devito, Zachary and Moses, William S. and Verdoolaege, Sven and Adams, Andrew and Cohen, Albert}, title = {The Next 700 Accelerated Layers: From Mathematical Expressions of Network Computation Graphs to Accelerated GPU Kernels, Automatically}, journal = {ACM Trans. Archit. Code Optim.}, issue_date = {November 2019}, volume = {16}, number = {4}, month = oct, year = {2019}, issn = {1544-3566}, pages = {38:1–38:26}, articleno = {38}, numpages = {26}, url = {http://doi.acm.org/10.1145/3355606}, doi = {10.1145/3355606}, acmid = {3355606}, publisher = {ACM}, address = {New York, NY, USA}, keywords = {Deep learning layers, GPU acceleration, polyhedral compilation}, }
@inproceedings{Augustine:2019:GPC:3314221.3314615, author = {Augustine, Travis and Sarma, Janarthanan and Pouchet, Louis-No\"{e}l and Rodr\'{\i}guez, Gabriel}, title = {Generating Piecewise-regular Code from Irregular Structures}, booktitle = {Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation}, series = {PLDI 2019}, year = {2019}, isbn = {978-1-4503-6712-7}, location = {Phoenix, AZ, USA}, pages = {625–639}, numpages = {15}, url = {http://doi.acm.org/10.1145/3314221.3314615}, doi = {10.1145/3314221.3314615}, acmid = {3314615}, publisher = {ACM}, address = {New York, NY, USA}, keywords = {Polyhedral compilation, SpMV, sparse data structure, trace compression}, }
@inproceedings{Rawat:2016:ERM:2884045.2884047, author = {Rawat, Prashant Singh and Hong, Changwan and Ravishankar, Mahesh and Grover, Vinod and Pouchet, Louis-Noël and Sadayappan, P.}, title = {Effective Resource Management for Enhancing Performance of 2D and 3D Stencils on GPUs}, booktitle = {Proceedings of the 9th Annual Workshop on General Purpose Processing Using Graphics Processing Unit}, series = {GPGPU '16}, year = {2016}, isbn = {978-1-4503-4195-0}, location = {Barcelona, Spain}, pages = {92–102}, numpages = {11}, url = {http://doi.acm.org/10.1145/2884045.2884047}, doi = {10.1145/2884045.2884047}, acmid = {2884047}, publisher = {ACM}, address = {New York, NY, USA}, keywords = {GPGPU, resource management, stencil computations, tiling}, }

@article{DBLP:journals/corr/abs-1805-02566,

```
author    = {Hyoukjun Kwon and
             Michael Pellauer and
             Tushar Krishna},
title     = {Understanding Reuse, Performance, and Hardware Cost of DNN
Dataflows: A Data-Centric Approach},
journal   = {CoRR},
volume    = {abs/1805.02566},
year      = {2018},
url       = {http://arxiv.org/abs/1805.02566},
archivePrefix = {arXiv},
eprint    = {1805.02566},
timestamp = {Mon, 13 Aug 2018 16:46:45 +0200},
biburl    = {https://dblp.org/rec/bib/journals/corr/abs-1805-02566},
bibsource = {dblp computer science bibliography, https://dblp.org}
```

}

@inproceedings{Stock:2014:FED:2594291.2594342, author = {Stock, Kevin and Kong, Martin and Grosser, Tobias and Pouchet, Louis-Noël and Rastello, Fabrice and Ramanujam, J. and Sadayappan, P.}, title = {A Framework for Enhancing Data Reuse via Associative Reordering}, booktitle = {Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation}, series = {PLDI '14}, year = {2014}, isbn = {978-1-4503-2784-8}, location =

{Edinburgh, United Kingdom}, pages = {65–76}, numpages = {12}, url = {http://doi.acm.org/10.1145/2594291.2594342}, doi = {10.1145/2594291.2594342}, acmid = {2594342}, publisher = {ACM}, address = {New York, NY, USA}, }

@ARTICLE{7738524, author={Y. H. Chen and T. Krishna and J. S. Emer and V. Sze}, journal={IEEE Journal of Solid-State Circuits}, title={Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks}, year={2017}, volume={52}, number={1}, pages={127-138}, url = {http://ieeexplore.ieee.org/document/7738524/}, doi={10.1109/JSSC.2016.2616357}, ISSN={0018-9200}, month={Jan},}

@article{Vasilache:2019:NAL:3366460.3355606, author = {Vasilache, Nicolas and Zinenko, Oleksandr and Theodoridis, Theodoros and Goyal, Priya and Devito, Zachary and Moses, William S. and Verdoolaege, Sven and Adams, Andrew and Cohen, Albert}, title = {The Next 700 Accelerated Layers: From Mathematical Expressions of Network Computation Graphs to Accelerated GPU Kernels, Automatically}, journal = {ACM Trans. Archit. Code Optim.}, issue_date = {October 2019}, volume = {16}, number = {4}, month = oct, year = {2019}, issn = {1544-3566}, pages = {38:1–38:26}, articleno = {38}, numpages = {26}, url = {http://doi.acm.org/10.1145/3355606}, doi = {10.1145/3355606}, acmid = {3355606}, publisher = {ACM}, address = {New York, NY, USA}, keywords = {Deep learning layers, GPU acceleration, polyhedral compilation},