Acorns: A Framework for Accelerating Deep Neural Networks with Input Sparsity

Xiao Dong, Lei Liu, Peng Zhao, Guangli Li, et al

28th International Conference on Parallel Architectures and Compilation Techniques (PACT), 2019

Presenter: Won-Hyuk Lee

http://esoc.hanyang.ac.kr/people/wonhyuk_lee/index.html

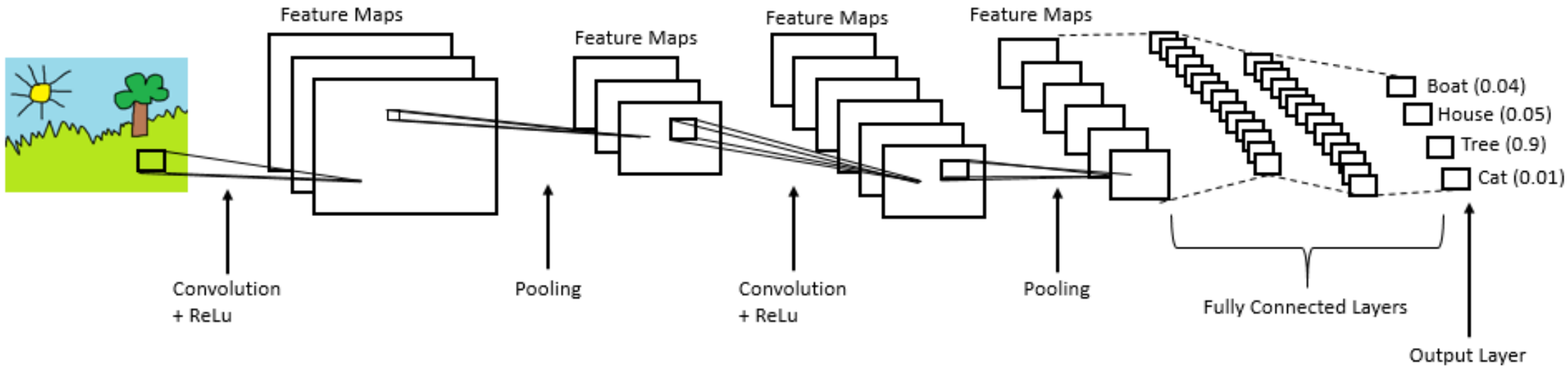
May 12, 2020

Neural Network Acceleration Study Season #2

Background

- **Operator**
  - Conv, ReLU, Pooling, Batch Norm, Scale

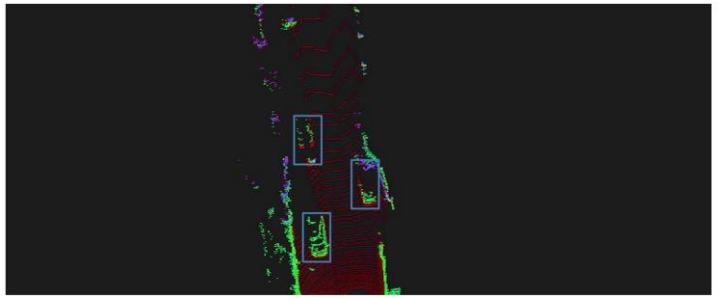
- **Input tensor's layout**
  - Original tensor layout (3D tensor)
  - Sparse tensor layout (***Proposed layout***)



Application with Sparse Input

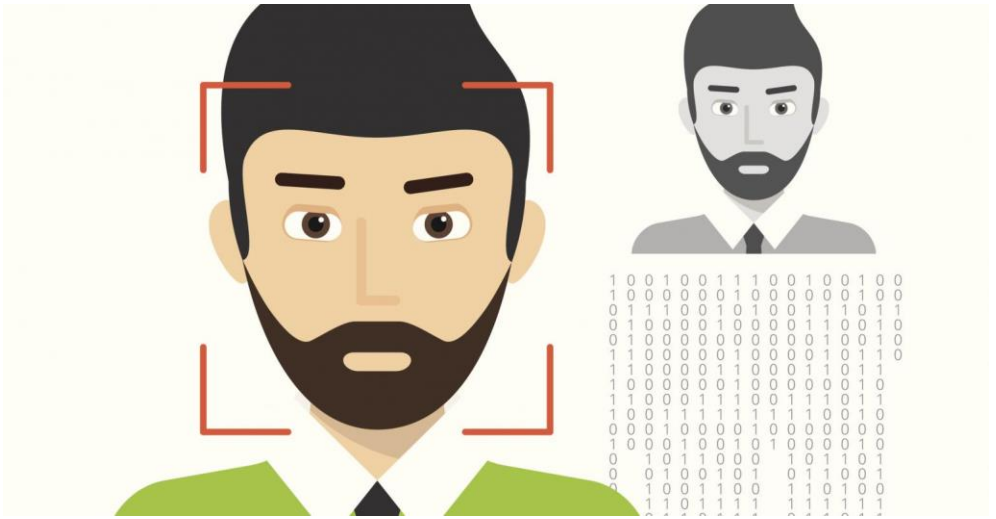
- **LiDAR-based Detection**
  - 75~95% Sparsity

(a) RGB image

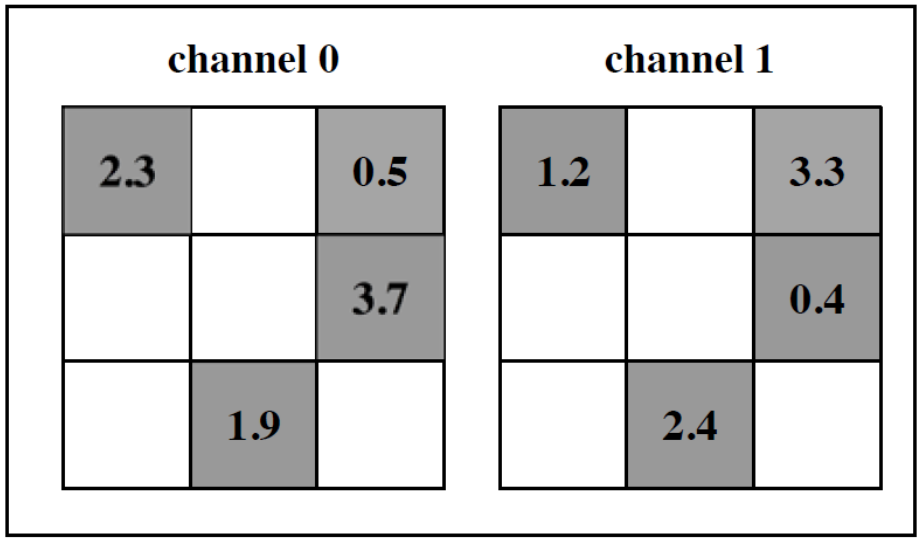


(b) bird-eye-view LiDAR image

- **Face Detection**
  - Only specific region is valid

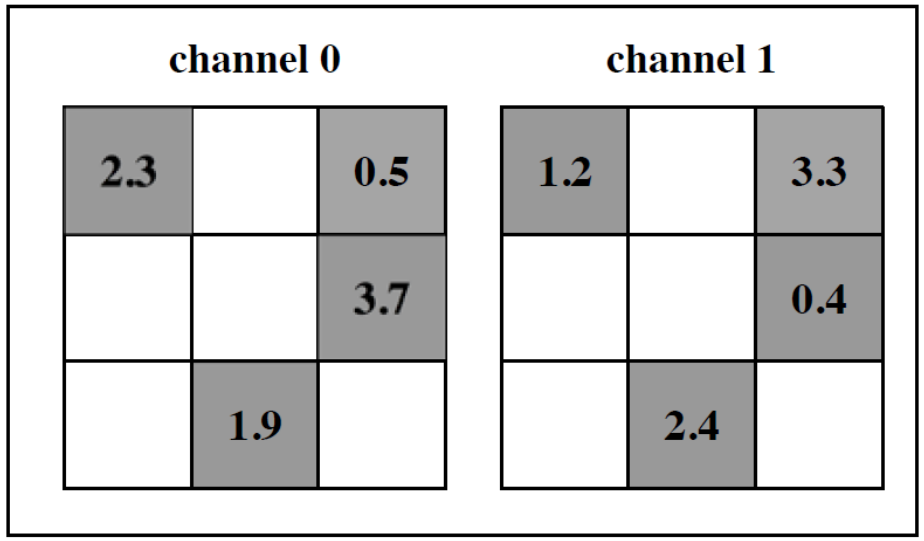


Channel consistency



(a) sparse tensor

Conventional Sparse formats



(a) sparse tensor

2.3	1.2
0.5	3.3
3.7	0.4
1.9	2.4

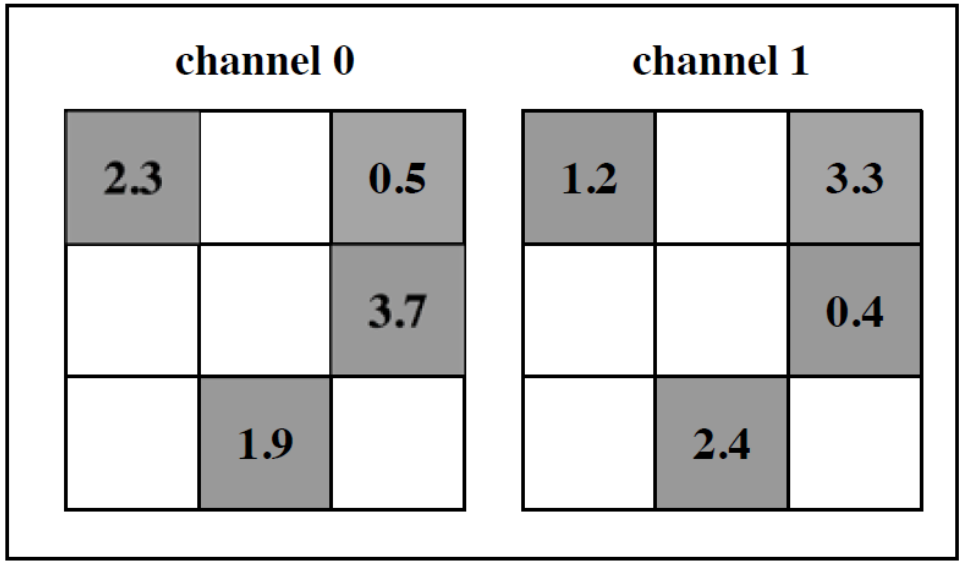
(c) unrolled sparse tensor

	2.3	1.2	0.5	3.3	3.7	0.4	1.9	2.4		nnzs
CSR	0	1	0	1	0	1	0	1		colidx
	0	2	2	4	4	4	6	6	8	8 rowptr

	2.3	1.2	0.5	3.3	3.7	0.4	1.9	2.4	nnzs
COO	0	0	2	2	5	5	7	7	rowidx
	0	1	0	1	0	1	0	1	colidx

(d) sparse matrix in different sparse formats

Proposed Sparse Data Layout

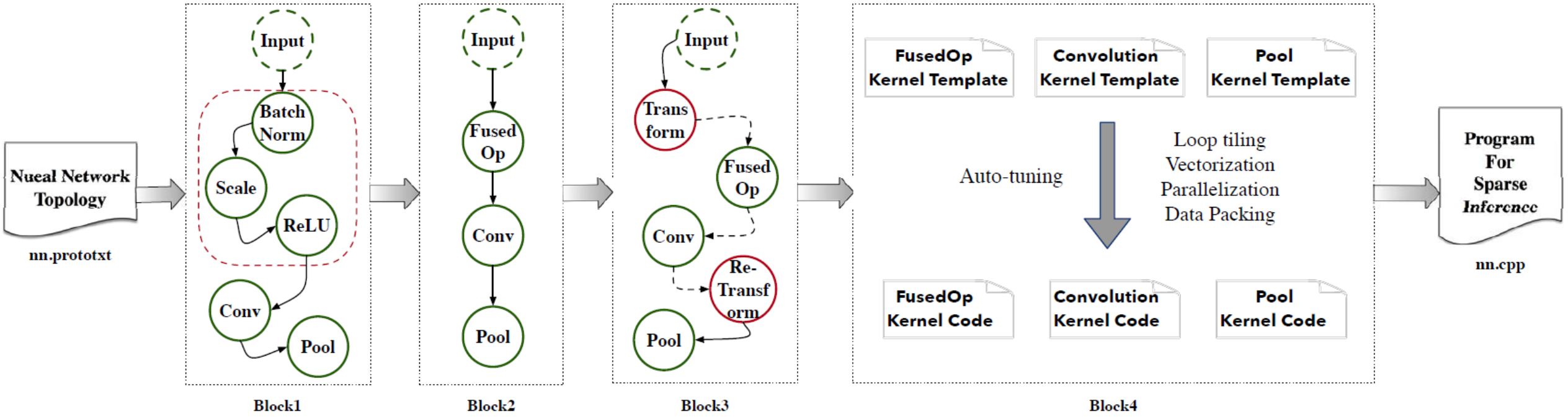


(a) sparse tensor

values		locations
2.3	1.2	(0, 0)
0.5	3.3	(0, 2)
3.7	0.4	(1, 2)
1.9	2.4	(2, 1)

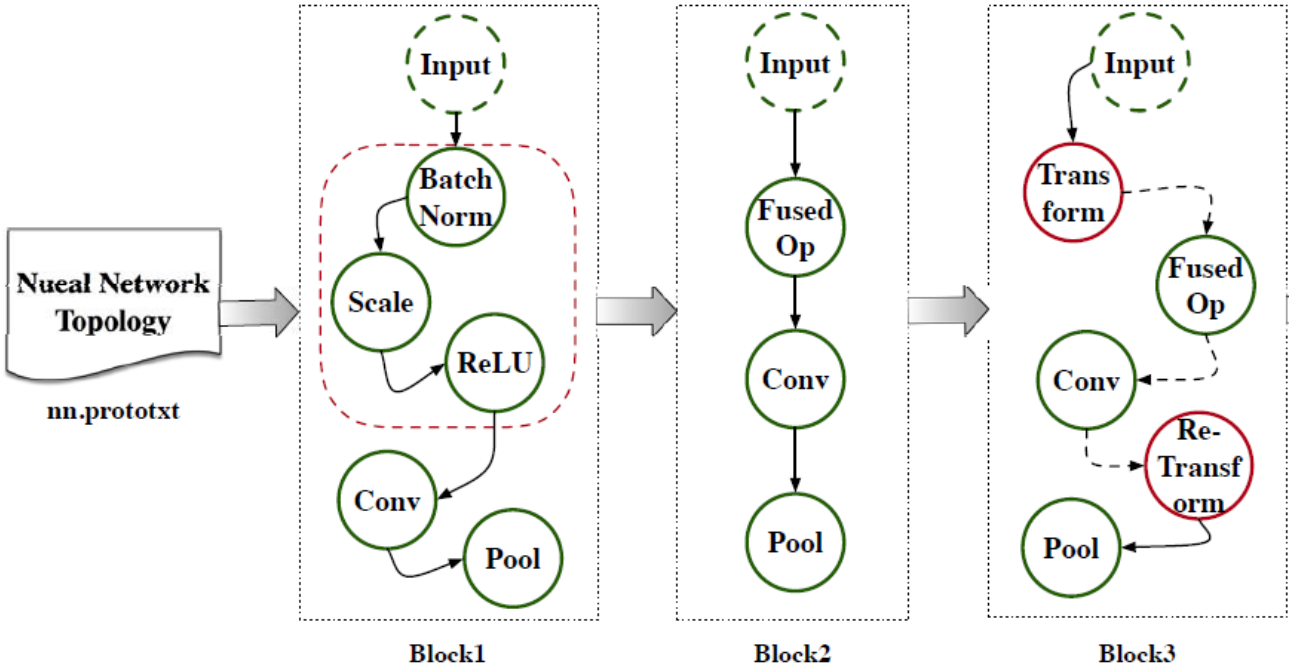
(b) sparse tensor in our format

Deep Learning Framework with Sparse Input



Workflow of Acorns

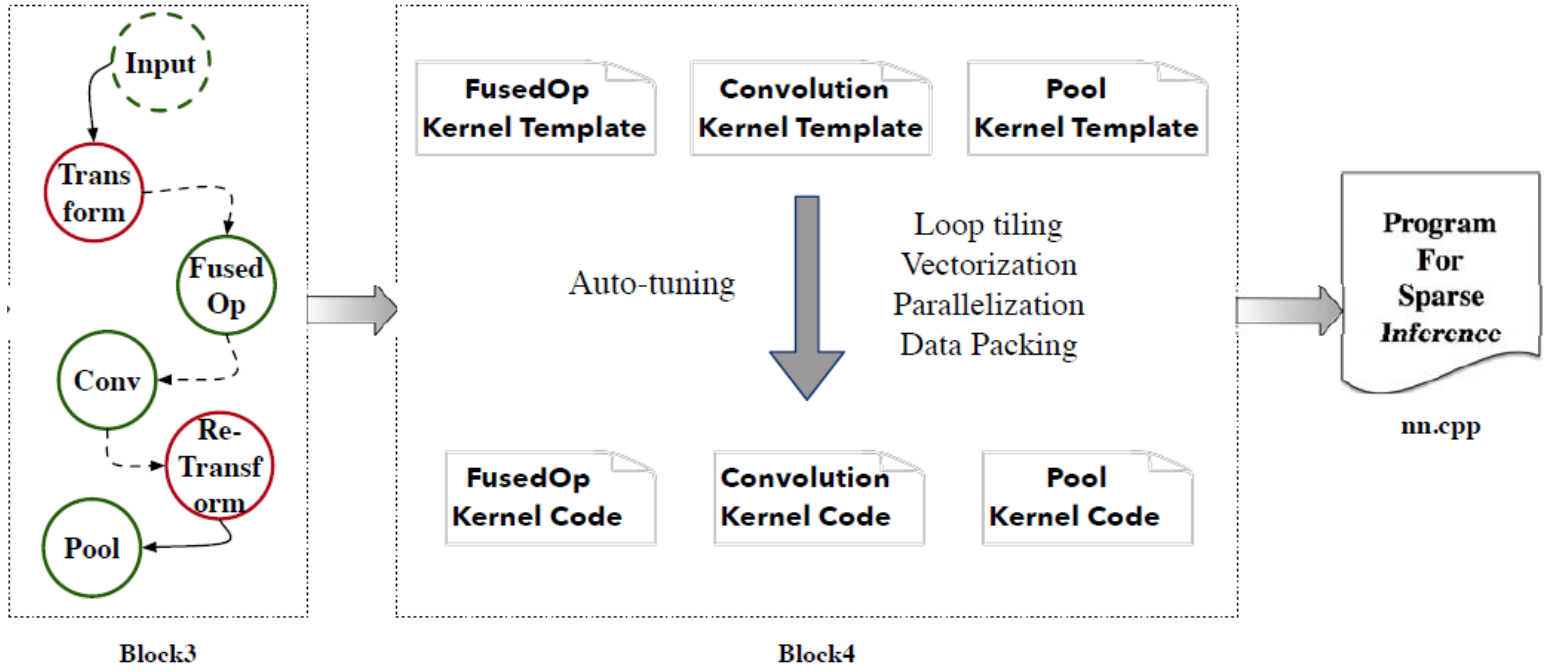
Inter-Operator Optimization



- **Operator fusion**
  - Consecutive operator can be merged and replacing
- **Sparse tensor layout conversion**
  - Proposed sparse tensor layout is preferred by most of operators

Kernel Code Optimization

- **Kernel Template**
  - Straightforward computation code annotated with optimizing directions
- **Optimizing Direction**
  - Loop tiling
  - Vectorization
  - Auto-tuning
  - Weight Packing
  - Multithreading



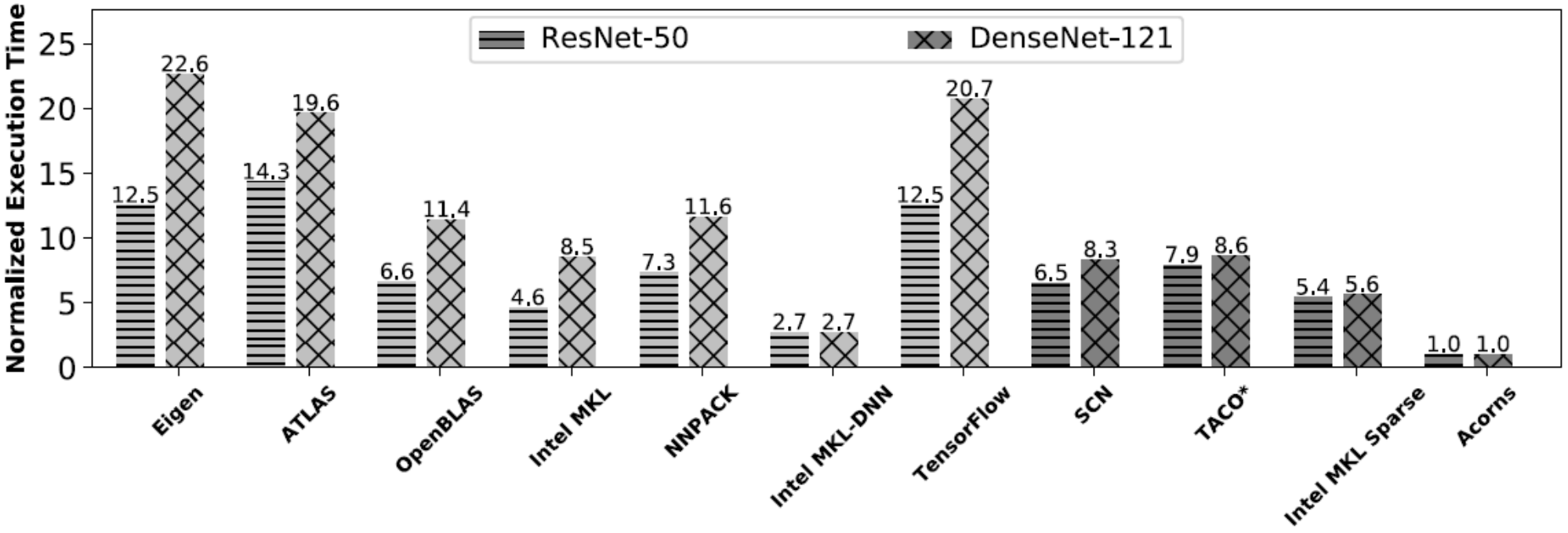
Methodology

- **Networks**
  - ResNet-50
  - DenseNet-121
- **Dataset**
  - KITTI dataset (average sparsity 79%)
- **CPU**
  - 32-core Intel Xeon E7-4809 v3
  - Supports AVX2

- **Sparsity-aware methods**
  - SparseConvNet (SCN)
  - Intel MKL-Sparse
  - TACO
- **Sparsity-unaware methods**
  - Intel MKL-DNN
  - NNPACK
  - Eigen
  - Intel MKL
  - OpenBLAS
  - ATLAS
  - Caffe
  - TensorFlow

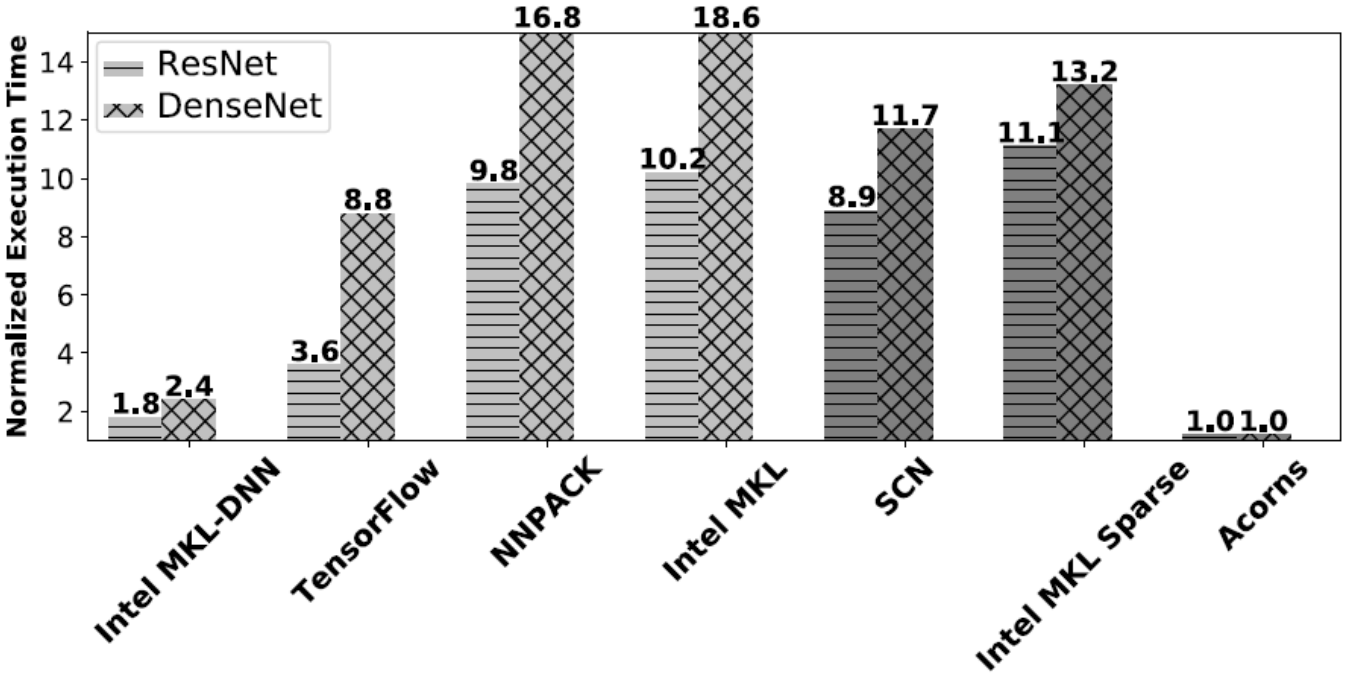
Evaluation: Single-Thread

- **Sparsity-unaware methods**
  - 2.7 to 22.7× speedups
- **Sparsity-aware methods**
  - 5.4 to 8.6× speedups



Evaluation: Multithreading

- **Multithreading performance**
  - 1.8 to 18.6× speedups



- **Speedups over single-thread**
  - Best multithreading speedup among the sparsity-aware methods

Speedup	ResNet	DenseNet
Intel MKL-DNN	5.4	3.3
TensorFlow	12.3	7.1
NNPACK	2.6	2.1
Intel MKL	1.6	1.4
SCN	2.6	2.2
Intel MKL-Sparse	1.7	1.3
Acorns	3.6	3.0

Thank you