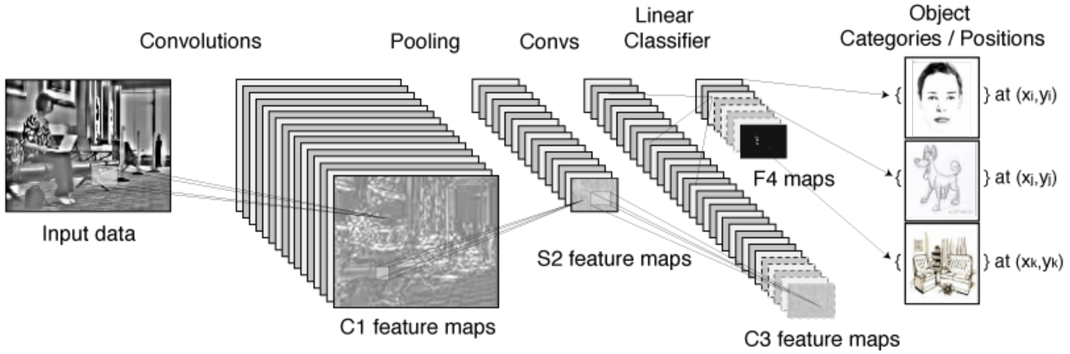
Research Overview of Energy-Efficient Multimedia Systems Group

Vivienne Sze

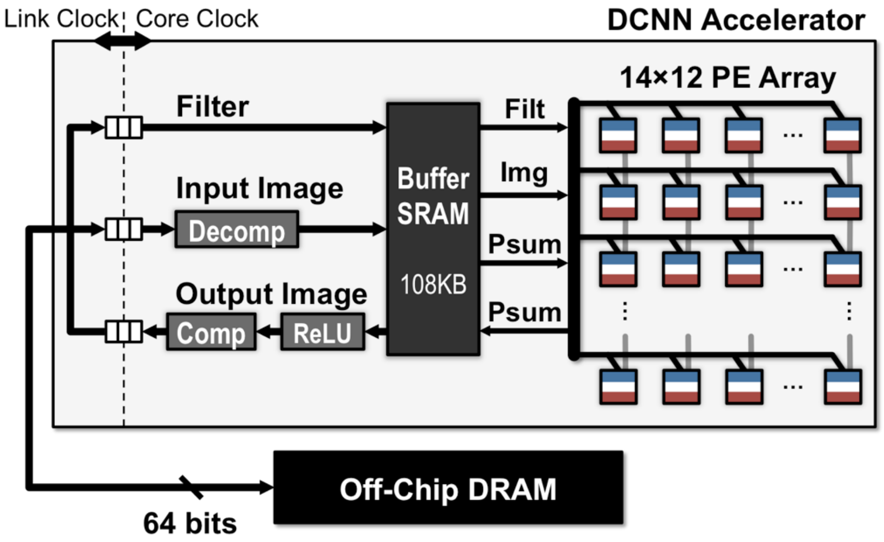
Efficient Computing with Cross-Layer Design

Algorithms

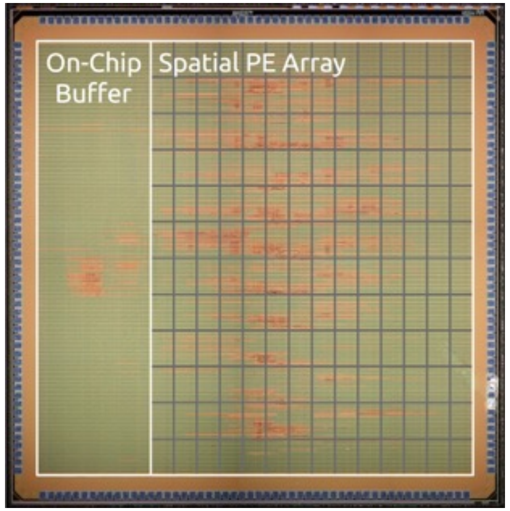


Systems

Architectures



Circuits

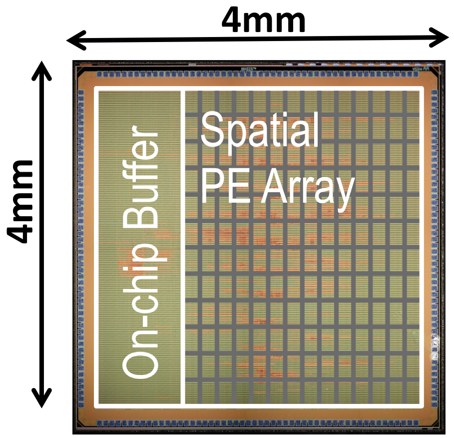


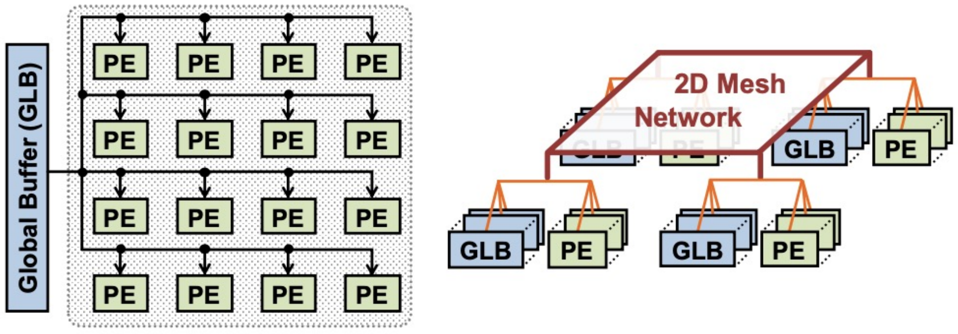
Energy-Efficient Deep Neural Networks

Efficient and Flexible Hardware

Eyeriss Accelerator
minimize data movement

[ISSCC 2016, ISCA 2016]



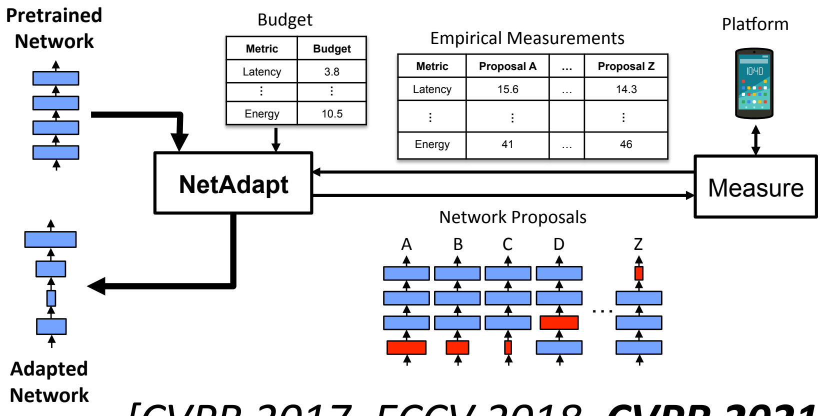


(a) Original Eyeriss (b) Eyeriss v2

[JETCAS 2019]

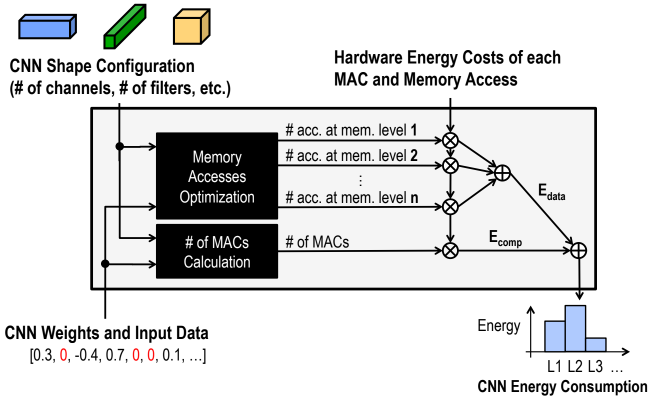
http://eyeriss.mit.edu

Co-Design of Algorithms and Hardware



*[CVPR 2017, ECCV 2018, **CVPR 2021**]*

***Energy Modeling** for Design Exploration and Optimization*

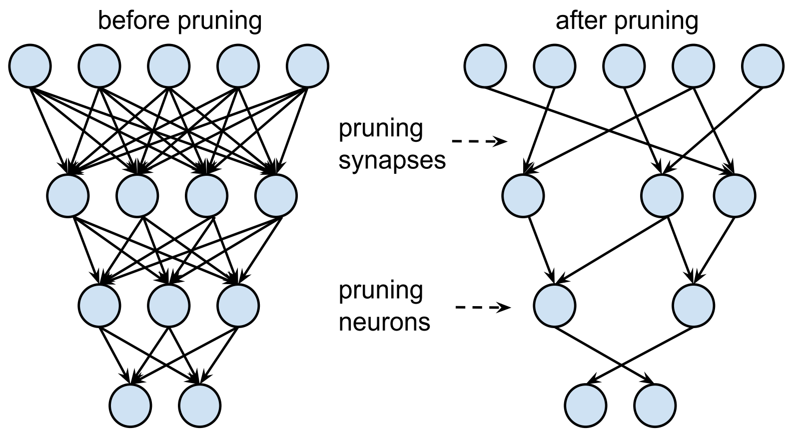


*[Asilomar 2017, ICCAD 2019, **ISPASS 2021**]*

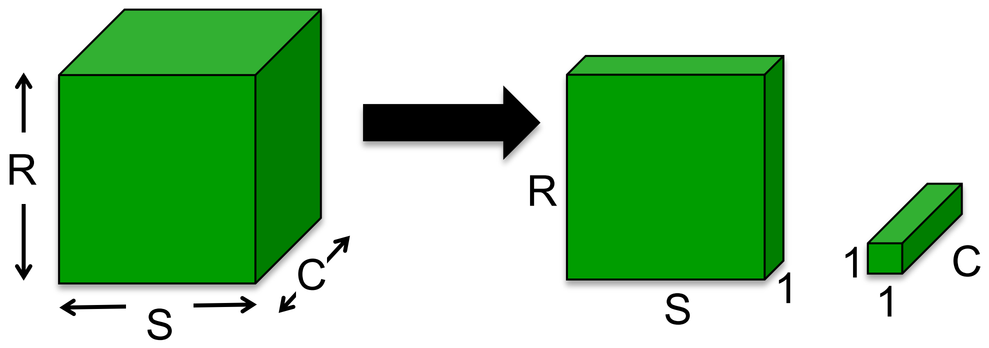
Design of Efficient DNN Algorithms

Popular efficient DNN algorithm approaches

Network Pruning



Efficient Network Architectures



Examples: SqueezeNet, MobileNet

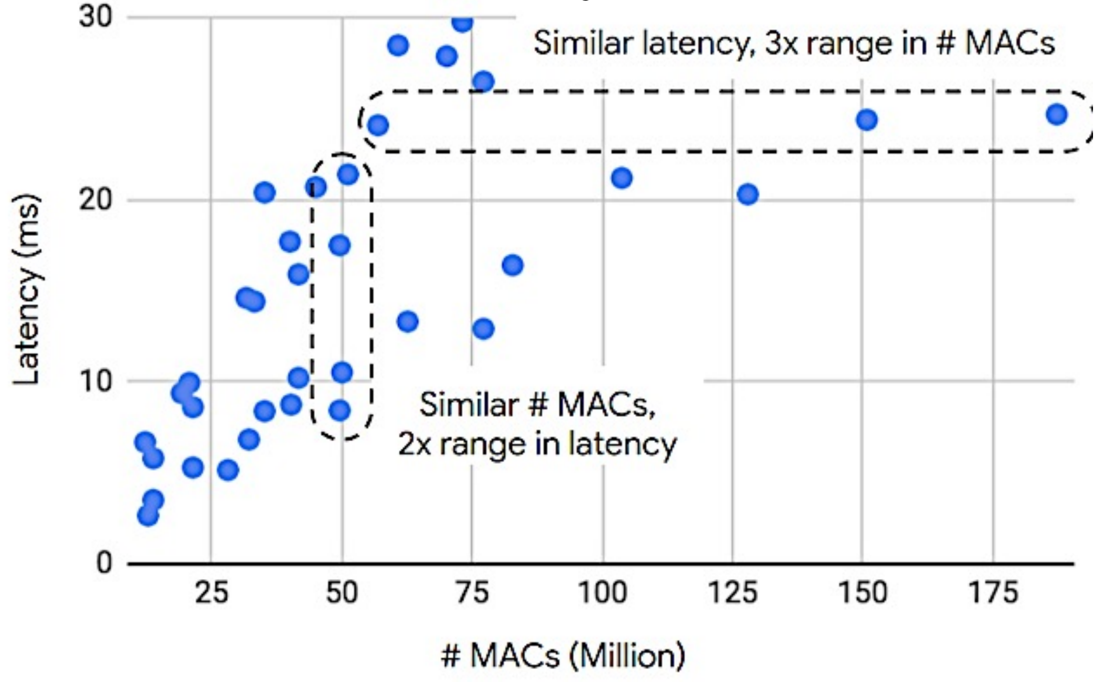
... also reduced precision

- Focus on reducing **number of MACs and weights**
- **Does it translate to energy savings and reduced latency?**

 [**Chen*, Yang***, *SysML* 2018]

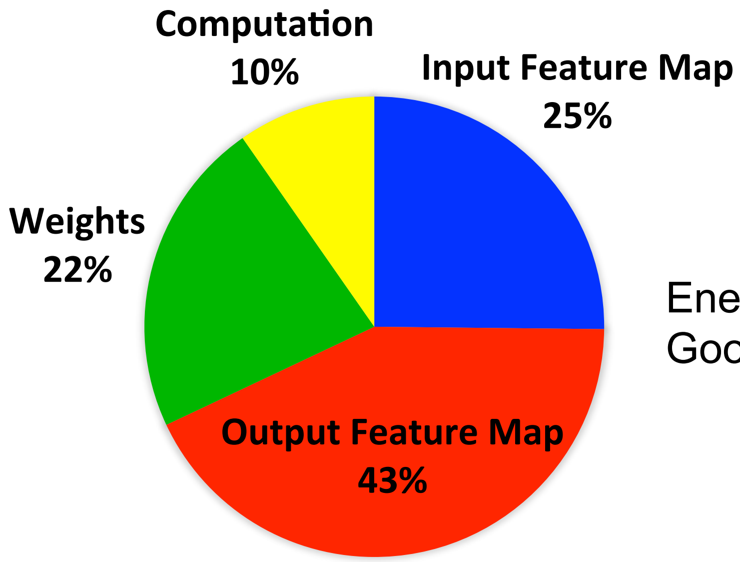
Number of MACs and Weights are Not Good Proxies

of operations (MACs) does not approximate latency well



Source: Google (https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html)

# of weights ***alone*** is not a good metric for energy (**All data types** should be considered)



Energy breakdown of GoogLeNet

https://energyestimation.mit.edu/

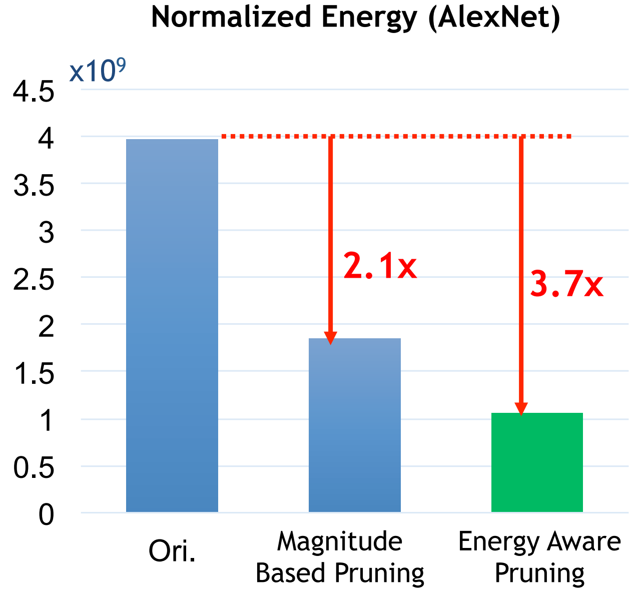
[**Yang**, *CVPR* 2017]

Design Hardware-Aware DNN Algorithms

Directly target energy and latency and incorporate it into the optimization of DNNs to provide better performance tradeoffs

Energy-Aware Pruning

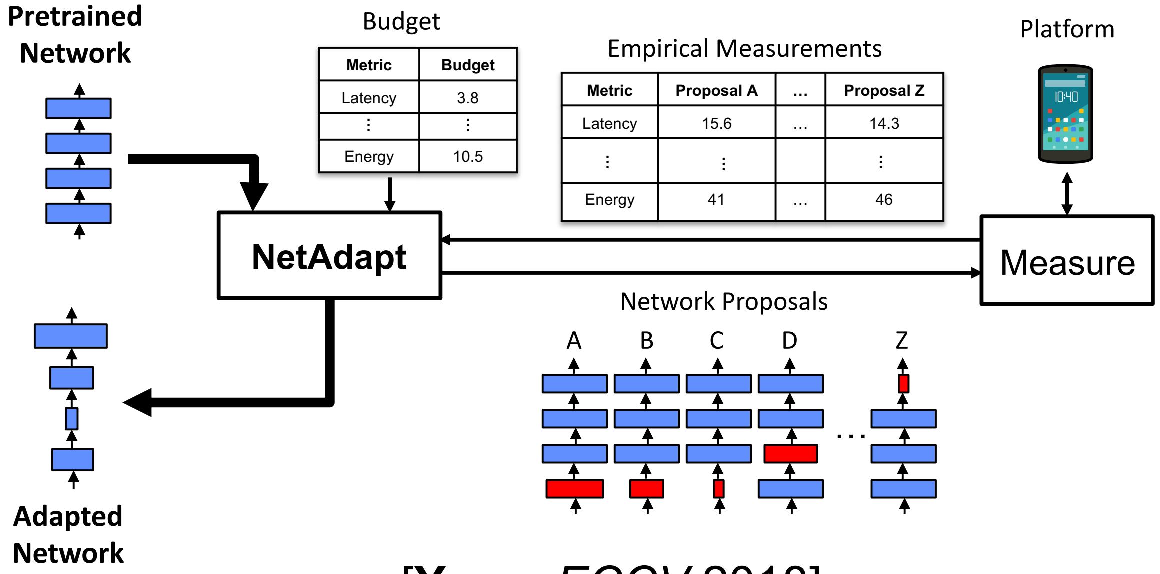
Remove weights based on energy consumption



[**Yang**, *CVPR* 2017]

NetAdapt: Platform-Aware DNN Adaptation

Automatically adapt DNN to reach target latency/energy



[**Yang**, *ECCV* 2018]

Code available at http://netadapt.mit.edu

NetAdapt v2: Reduce Adaption Time

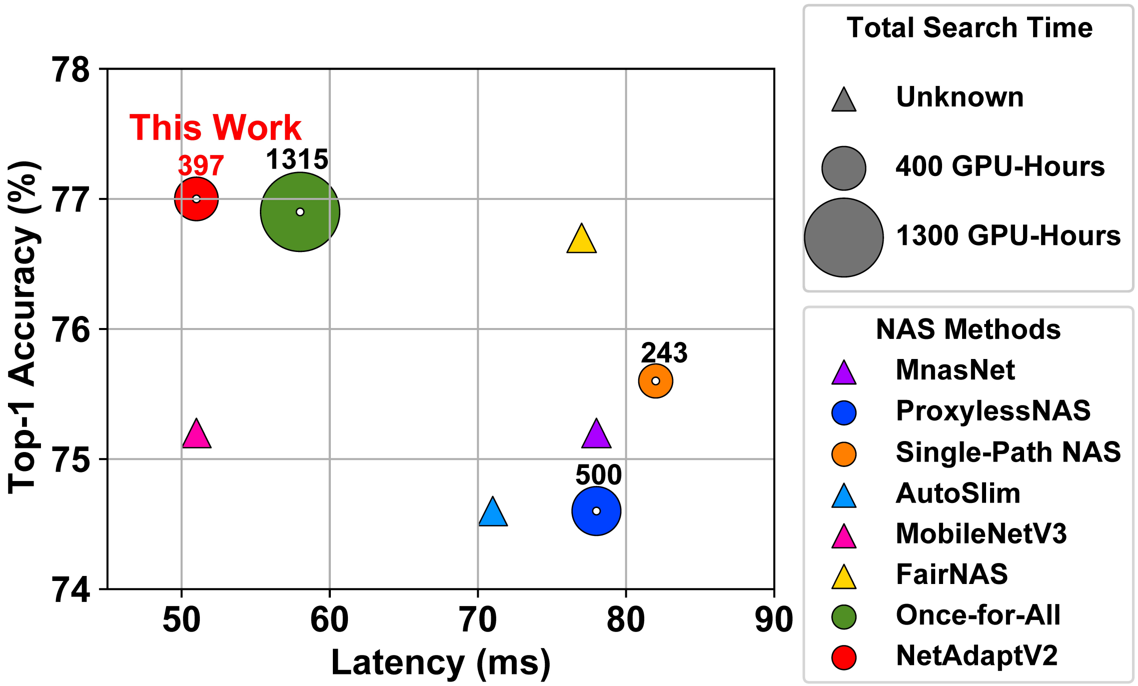
Reduce time to find efficient DNN that adapts to hardware by up to 5.8x

Typical Steps in Neural Architecture Search (NAS):

1) Train super-network (search space of DNNs)
2) Sample and evaluate different DNNs
3) Fine tune the final DNN

Contributions

- ***Ordered dropout:*** train multiple DNNs in *single* forward pass (reduce step 1)
- ***Channel-level bypass:*** merge layer depth and channel width into a *single* search dimension (reduce step 2)
- ***Multi-layer coordinate descent optimizer:*** consider joint effect of multiple layers (reduce step 2 & support non-differentiable metrics, e.g., latency)

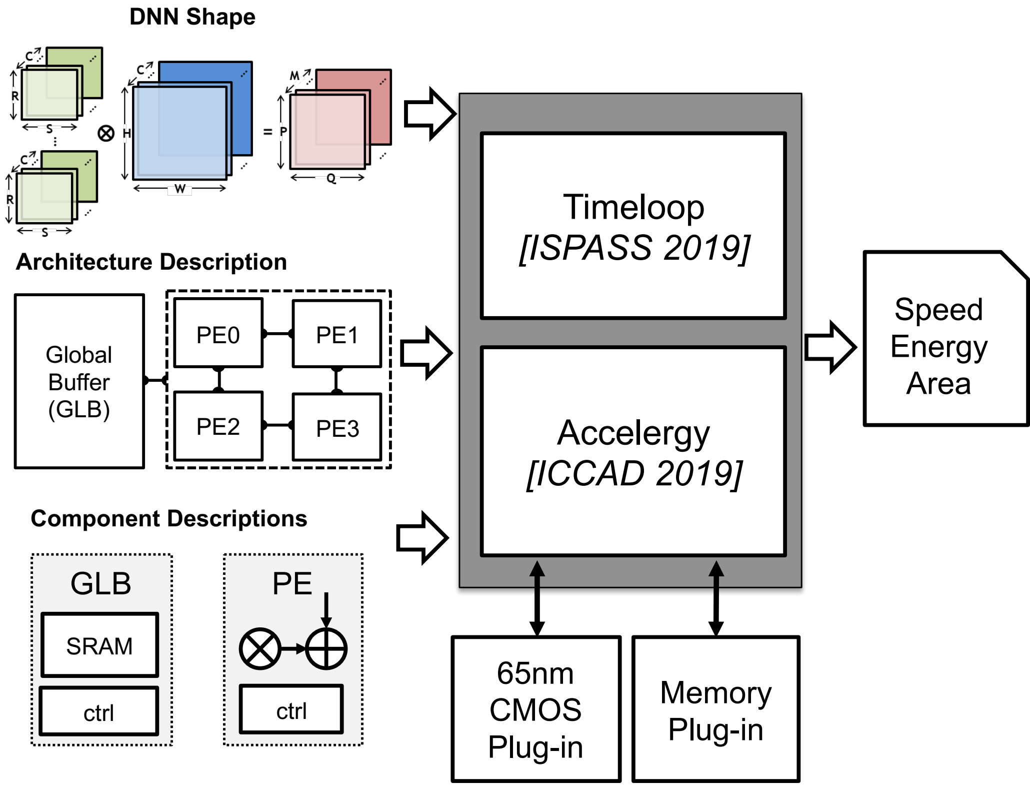


More info at http://netadapt.mit.edu

[**Yang**, *CVPR* 2021]

Energy Estimation for Accelerator Designs

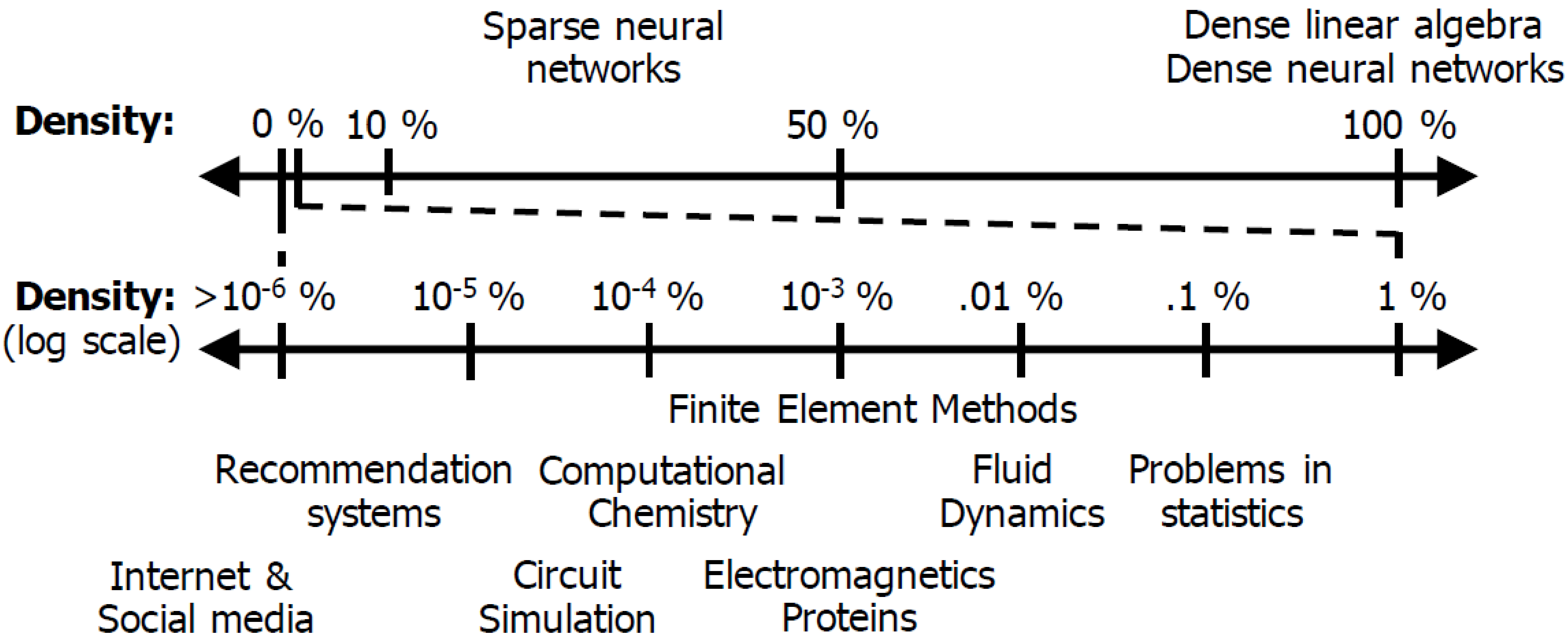
- **Accelergy** is an architecture-level energy estimator framework
  - Early-stage energy estimation
  - Rapid design space exploration
    - e.g., evaluate and compare different deep learning accelerator designs → performance modeling with Timeloop
- Provides flexibility to
  - Describe a diverse range of accelerator designs
  - Support different technologies
    - e.g., CMOS, RRAM, optical
- Validated on both digital and PIM based accelerators (95% accuracy)
- Bridge architecture, circuit and devices research



Open-source code available at:
http://accelergy.mit.edu

[**Wu**, *ICCAD* 2019],
[**Wu**, *ISPASS* 2020]

Applications that use Sparse Tensor

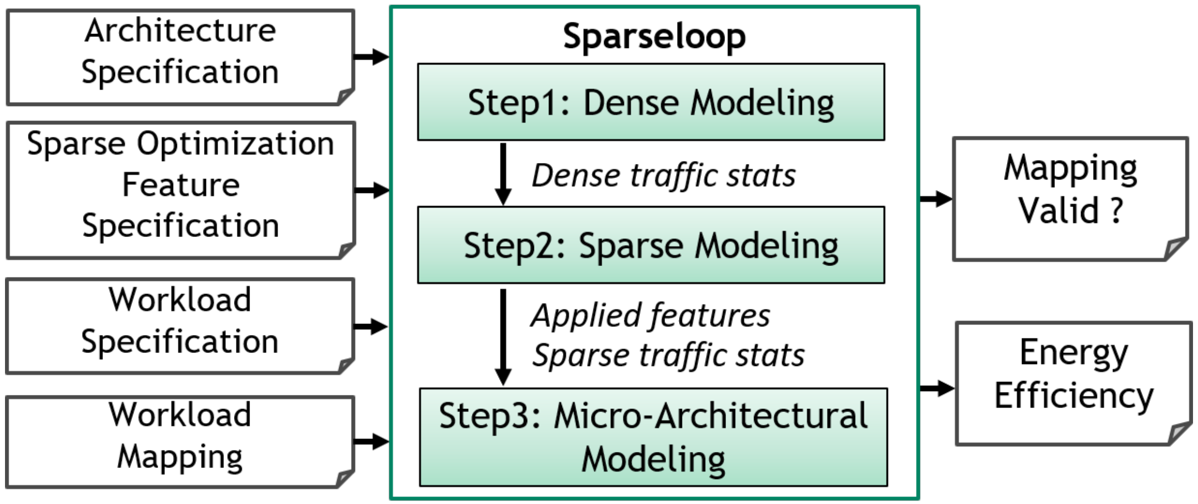


[**Hedge**, *MICRO* 2019]

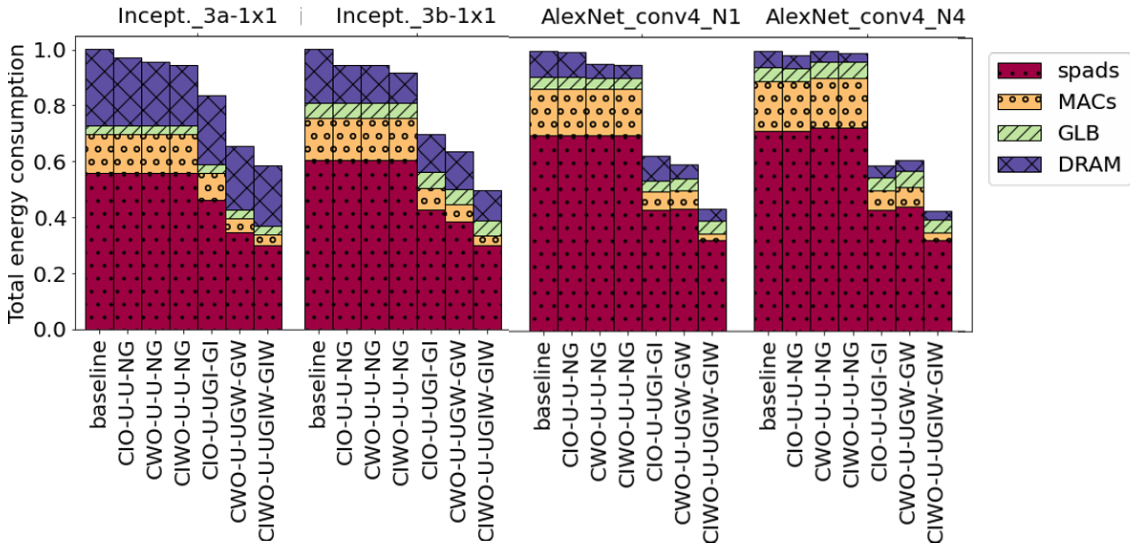
Sparseloop: Design Space Exploration for Sparse Tensor Accelerators

- An analytical design exploration framework that comprehends a wide range of
  - Sparse optimizations (e.g., zero-gating, zero-skipping, zero-compression)
  - Data representations (e.g., uncompressed, run length coding, bitmask)

Propose modularized three-step evaluation process



Energy impact of sparse optimizations at different levels of the memory hierarchy in Eyeriss-based topology

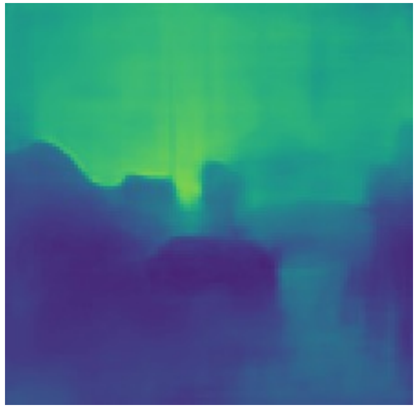


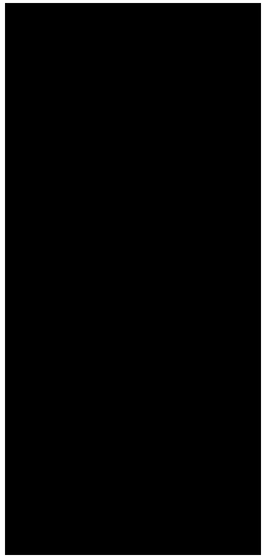
Tutorial at ISCA 2021 (June 19): http://accelergy.mit.edu/sparse_tutorial.html

Efficient Computing for Autonomous Navigation

Monocular Depth Estimation with FastDepth

RGB **Prediction**

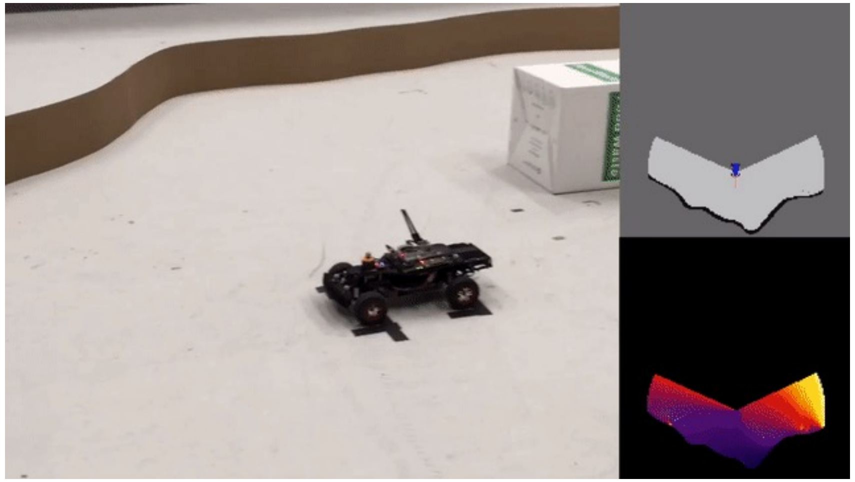


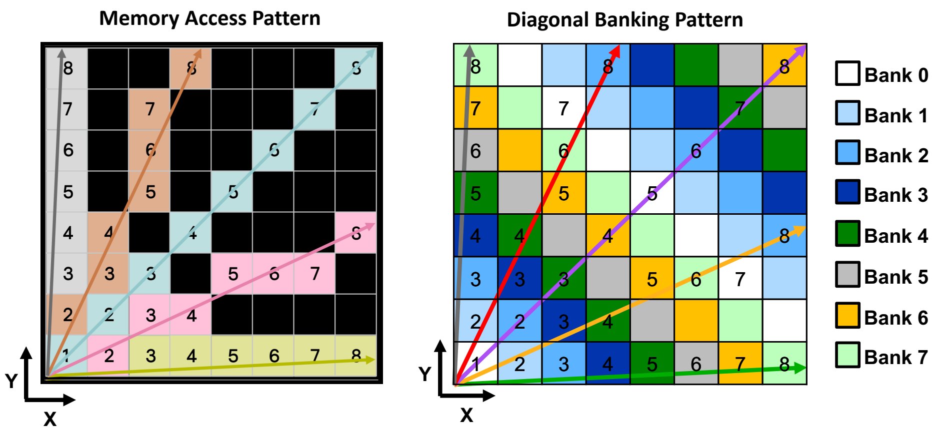


http://fastdepth.mit.edu

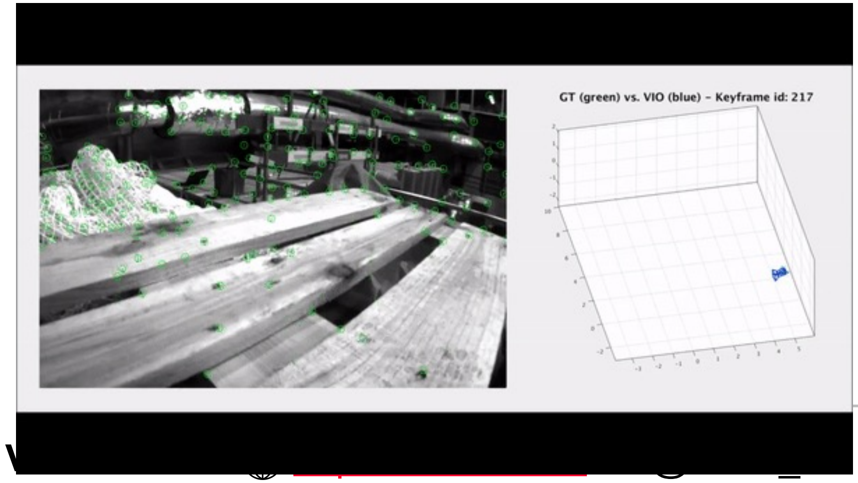
~40fps on an iPhone

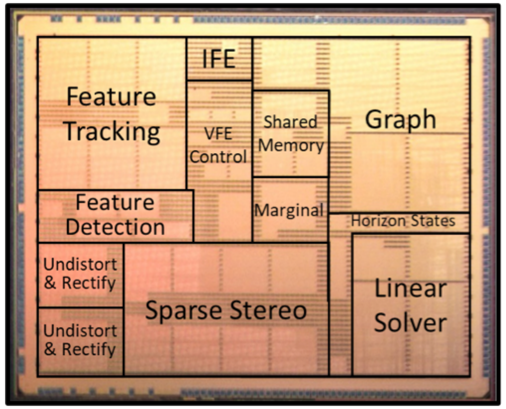
Robot Exploration with Mutual Information





Visual Inertial Localization with Navion





http://navion.mit.edu

In collaboration with Sertac Karaman

Low-Energy Autonomy and Navigation (LEAN) Group

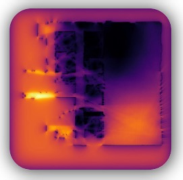
LEAN   HOME   TEAM   RESEARCH   PUBLICATIONS   PRESS   RECOGNITION

A broad range of next-generation applications will be enabled by low-energy, miniature mobile robotics including insect-size flapping wing robots that can help with search and rescue, chip-size satellites that can explore nearby stars, and blimps that can stay in the air for years to provide communication services in remote locations. While the low-energy, miniature actuation, and sensing systems have already been developed in many of these cases, the processors currently used to run the algorithms for autonomous navigation are still energy-hungry. Our research addresses this challenge as well as brings together the robotics and hardware design communities.

We enable efficient computing on various key modules of other autonomous navigation systems including perception, localization, exploration and planning. We also consider the overall system by considering the energy cost of computing in conjunction with actuation and sensing.

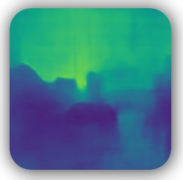
Motion Planning

Many motion planning and control algorithms aim to design trajectories and controllers that minimize actuation energy. However, in low-energy robotics, computing such trajectories and controls themselves may consume a large amount of energy. We develop algorithms that optimize this trade-off.



Mutual Information for Exploration

Computing mutual information between the map and future measurements is critical to efficient exploration. Unfortunately, mutual information computation is computationally very challenging. We develop new algorithms and hardware for efficient computation of mutual information, and demonstrate real-time computation for the whole map in a reasonably-sized map.



Depth Sensing and Perception

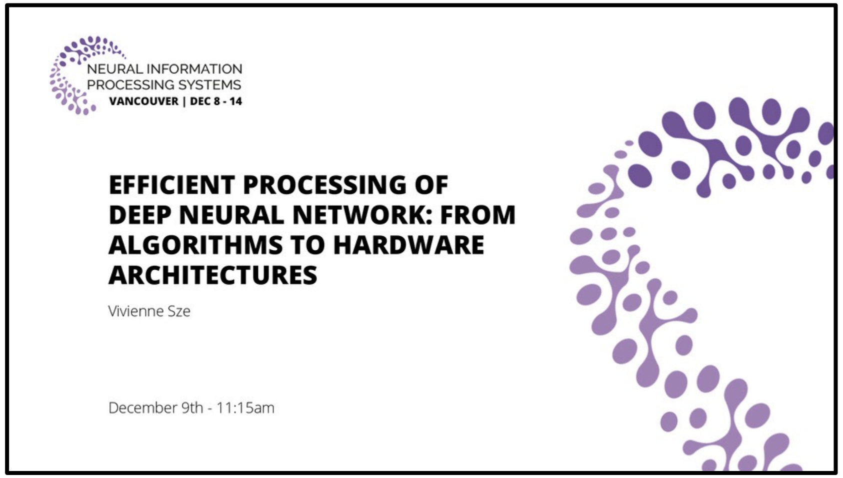
Depth sensing is a critical function for robotic tasks such as localization, mapping and obstacle detection. State-of-the-art single-view depth estimation algorithms are based on fairly complex deep neural networks that are too slow for real-time inference on an embedded platform, for instance, mounted on a micro aerial vehicle. We address the problem of fast depth estimation on embedded systems.

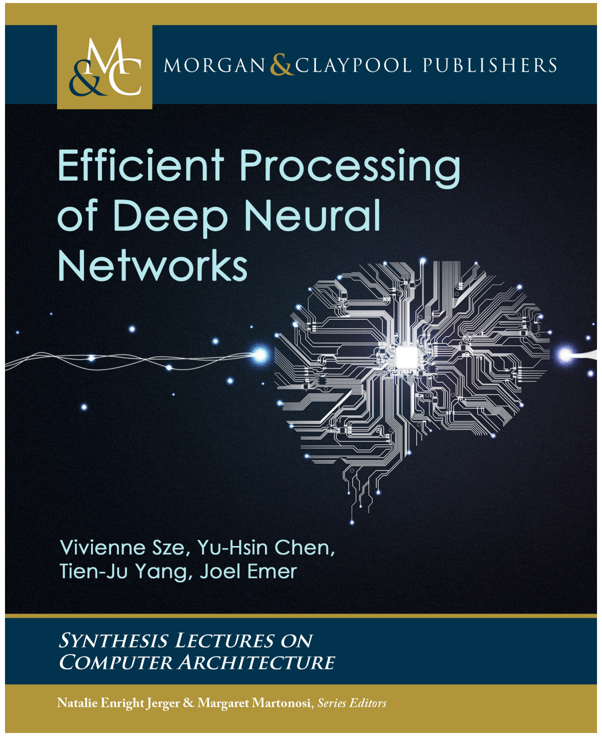
Localization and Mapping

Autonomous navigation of miniaturized robots (e.g., nano/pico aerial vehicles) is currently a grand challenge for robotics research, due to the need for processing a large amount of sensor data (e.g., camera frames) with limited on-board computational resources. We focus on the design of a visual-inertial odometry (VIO) system in which the robot estimates its ego-motion (and a landmark-based map) from on-board camera and IMU data.

Group Website: http://lean.mit.edu

Resources on Efficient Processing of DNNs





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

References

- **Efficient Hardware for Deep Neural Networks and Sparse Tensor Accelerators**

  - **Project website:** http://eyeriss.mit.edu
  - *Y.-H. Chen, T. Krishna, J. Emer, V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," IEEE Journal of Solid State Circuits (JSSC), ISSCC Special Issue, Vol. 52, No. 1, pp. 127-138, January 2017.*
  - *Y.-H. Chen, J. Emer, V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," International Symposium on Computer Architecture (ISCA), pp. 367-379, June 2016.*
  - *Y.-H. Chen, T.-J. Yang, J. Emer, V. Sze, "Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices," IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), June 2019.*
  - *Eyexam:* https://arxiv.org/abs/1807.07928
  - *Y. N. Wu, J. S. Emer, V. Sze, "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," International Conference on Computer Aided Design (ICCAD), November 2019.* http://accelergy.mit.edu
  - *Y. N. Wu, V. Sze, J. S. Emer, "An Architecture-Level Energy and Area Estimator for Processing-In-Memory Accelerator Designs," to appear in IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), April 2020*
  - *Y. N. Wu, P.-A. Tsai, A. Parashar, V. Sze, J. S. Emer, "Sparseloop: An Analytical, Energy-Focused Design Space Exploration Methodology for Sparse Tensor Accelerators," IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), March 2021*
  - *F. Wang, Y. N. Wu, M. Woicik, J. S. Emer, V. Sze, "Architecture-Level Energy Estimation for Heterogeneous Computing Systems," IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), March 2021*

References

- **Co-Design of Algorithms and Hardware for Deep Neural Networks**

  - *T.-J. Yang, Y.-H. Chen, V. Sze, "Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.*
  - *Energy estimation tool:* http://eyeriss.mit.edu/energy.html
  - *T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, V. Sze, H. Adam, "NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications," European Conference on Computer Vision (ECCV), 2018.* http://netadapt.mit.edu
  - *D. Wofk*, F. Ma*, T.-J. Yang, S. Karaman, V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," IEEE International Conference on Robotics and Automation (ICRA), May 2019.* http://fastdepth.mit.edu/
  - *T.-J. Yang, V. Sze, "Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators," IEEE International Electron Devices Meeting (IEDM), Invited Paper, December 2019.*
  - *T.-J. Yang, Y.-L. Liao, V. Sze. "NetAdaptV2: Efficient Neural Architecture Search with Fast Super-Network Training and Architecture Optimization," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.*

References

- **Efficient Computing for Autonomous Navigation**

  - *D. Wofk*, F. Ma*, T.-J. Yang, S. Karaman, V. Sze, "FastDepth: Fast Monocular Depth Estimation on Embedded Systems," IEEE International Conference on Robotics and Automation (ICRA), May 2019.*
  - *A. Suleiman, Z. Zhang, L. Carlone, S. Karaman, V. Sze, "Navion: A 2mW Fully Integrated Real-Time Visual-Inertial Odometry Accelerator for Autonomous Navigation of Nano Drones," IEEE Journal of Solid-State Circuits (JSSC), VLSI Symposia Special Issue, Vol. 54, No. 4, pp. 1106-1119, April 2019.*
  - *Z. Zhang, T. Henderson, V. Sze, S. Karaman, "FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping," IEEE International Conference on Robotics and Automation (ICRA), May 2019.*
  - *P. Li*, Z. Zhang*, S. Karaman, V. Sze, "High-throughput Computation of Shannon Mutual Information on Chip," Robotics: Science and Systems (RSS), June 2019*
  - *T. Henderson, V. Sze, S. Karaman, "An Efficient and Continuous Approach to Information-Theoretic Exploration," to appear in IEEE International Conference on Robotics and Automation (ICRA), May 2020.*
  - *Z. Zhang, T. Henderson, S. Karaman, V. Sze, "FSMI: Fast computation of Shannon Mutual Information for information-theoretic mapping," International Journal of Robotics Research (IJRR), Vol. 39, No. 9, pp. 1155-1177, August 2020.*