Research Overview of Energy-Efficient Multimedia Systems Group

Vivienne Sze
Efficient Computing with Cross-Layer Design

**Algorithms**

- Convolutions
- Pooling
- Convs
- Linear Classifier
- C1 feature maps
- C2 feature maps
- C3 feature maps
- F4 maps

**Systems**

**Architectures**

- Input data
- DCNN Accelerator
- 14x12 PE Array
- Off-Chip DRAM
- Filter
- Input Image Decompression
- Output Image Computation
- Buffer SRAM 108KB
- 64 bits

**Circuits**

- On-Chip Buffer
- Spatial PE Array
Energy-Efficient Deep Neural Networks

Efficient and Flexible Hardware

Eyeriss Accelerator
minimize data movement

[ISSCC 2016, ISCA 2016]

http://eyeriss.mit.edu

Co-Design of Algorithms and Hardware

Energy Modeling for Design Exploration and Optimization

NetAdapt

Network Proposals

Platform

Pretrained Network

Empirical Measurements

Budget

Metric

Budget

Energy

Metric

Energy

NetAdapt

Network Proposals

[AISCC 2016, ISCA 2016]

[JETCAS 2019]

http://eyeriss.mit.edu

[Vivienne Sze @eems_mit

http://sze.mit.edu/

[CVPR 2017, ECCV 2018, CVPR 2021]

[Asilomar 2017, ICCAD 2019, ISPASS 2021]
Sparse Tensor Algebra Accelerators

Deep Learning
Graph Analytics
Circuit Simulations

Sparse Tensor Algebra

Exploit zero-based savings

0 x anything = 0
anything + 0 = anything

Large and diverse sparse tensor accelerator design space

In collaboration with Joel Emer
Enable Early Design Space Exploration

Workload

Mapping

Architecture

Sparseloop (Timeloop v2)

runtime behavior modeling [ISPASS2021]

Energy

Cycles

Fast analytical modeling frameworks for sparse tensor accelerators

Accelergy

energy estimation [ICCAD2019]

Vivienne Sze http://sze.mit.edu/ @eems_mit

Code and Tutorial material:
http://accelergy.mit.edu/sparse_tutorial.html

In collaboration with Joel Emer
Larger and More Diverse Design Space

Digital-Compute Accelerator Designs

- **Dense**
  - Eyeriss [JSSC2017]
  - Simba [MICRO2019]

- **Sparse**
  - ExTensor [MICRO2019]
  - Eyeriss V2 [JETCAS2019]
  - SCNN [ISCA2017]

Analog-Compute Accelerator Designs

- **Dense**
  - CASCADE [MICRO2019]
  - ISAAC [ISCA2016]
  - DrAcc [DAC2018]

- **Sparse**
  - SNrram [DAC2018]
  - SpaceA [HPCA2021]

Works we contributed to

- DianNao [ASPLOS2014]
- Simba [MICRO2019]

In collaboration with Joel Emer
Efficient Computing for Autonomous Navigation

Monocular Depth Estimation with FastDepth

RGB Prediction

http://fastdepth.mit.edu

~40fps on an iPhone

Visual Inertial Localization with Navion

http://navion.mit.edu

Robot Exploration with Mutual Information

In collaboration with Sertac Karaman
Orders of Magnitude Speed up Via Co-Design

For a 200x200 Map
(Note: Speed up increases for larger maps)

- **Shannon MI**
- **FSMI (CPU)**
- **FSMI (hardware)**
- **FCMI (CPU)**
- **FCMI (hardware)**

- **Optimize memory subsystem** (banking) for multi-beam parallel processing
- **Reformulate** using a continuous occupancy map framework and exploit recursive structure
- **Evaluate MI for all cells in entire beam** altogether removes numerical integration
- **Optimize memory subsystem**, time-interleave cores and approximate computing

Mutual Information for the **entire map** can be computed in real-time for the first time!

- **Shannon MI**
- **FSMI (CPU)**
- **FSMI (hardware)**
- **FCMI (CPU)**
- **FCMI (hardware)**

- [Julian, IJRR 2014]
- [Zhang, ICRA 2019]
- [Li, RSS 2019]
- [Henderson, ICRA 2020]
- [Gupta, IROS 2021]
Low-Energy Autonomy and Navigation (LEAN) Group

Group Website: [http://lean.mit.edu](http://lean.mit.edu)

A broad range of next-generation applications will be enabled by low-energy, miniature mobile robots including insect-size flapping wing robots that can help with search and rescue, chip-size satellites that can explore nearby stars, and balloons 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., nanosailor 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.
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
References

• Co-Design of Algorithms and Hardware for Deep Neural Networks
References

• Efficient Computing for Autonomous Navigation