UC San Diego

E-QUARTIC: Energy Efficient Edge Ensemble of Convolutional Neural Networks for Resource-Optimized Learning

Le Zhang, Onat Gungor, Flavio Ponzina, Tajana Rosing

Contact:
Flavio Ponzina
Postdoctoral scholar
fponzina@ucsd.edu

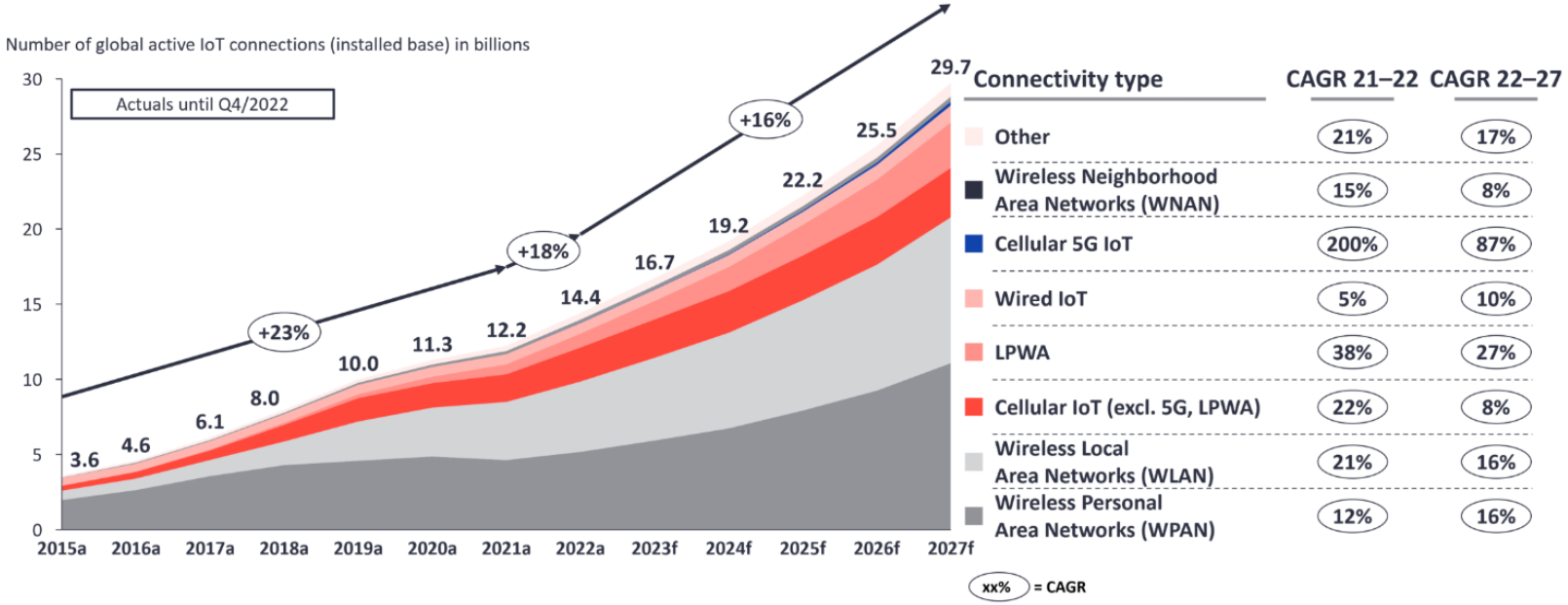
System Energy Efficiency Lab

seelab.ucsd.edu

see

The unstoppable IoT market

Global IoT market forecast (in billions of connected IoT devices)

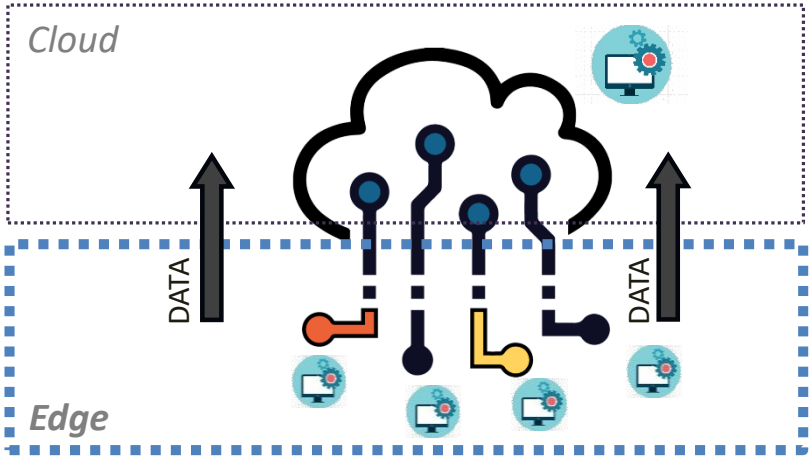


Note: IoT connections do not include any computers, laptops, fixed phones, cellphones, or consumers tablets. Counted are active nodes/devices or gateways that concentrate the end-sensors, not every sensor/actuator. Simple one-directional communications technology not considered (e.g., RFID, NFC). Wired includes ethernet and fieldbuses (e.g., connected industrial PLCs or I/O modules); Cellular includes 2G, 3G, 4G, 5G; LPWA includes unlicensed and licensed low-power networks; WPAN includes Bluetooth, Zigbee, Z-Wave or similar; WLAN includes Wi-Fi and related protocols; WNAN includes non-short-range mesh, such as Wi-SUN; Other includes satellite and unclassified proprietary networks with any range.
Source: IoT Analytics Research 2023. We welcome republishing of images but ask for source citation with a link to the original post and company website.

Source: IoT Analytics link

Empowering the edge

Computing at the edge



Applications


Healthcare


Virtual Reality


Space & Defense


Industrial IoT

Why edge computing?

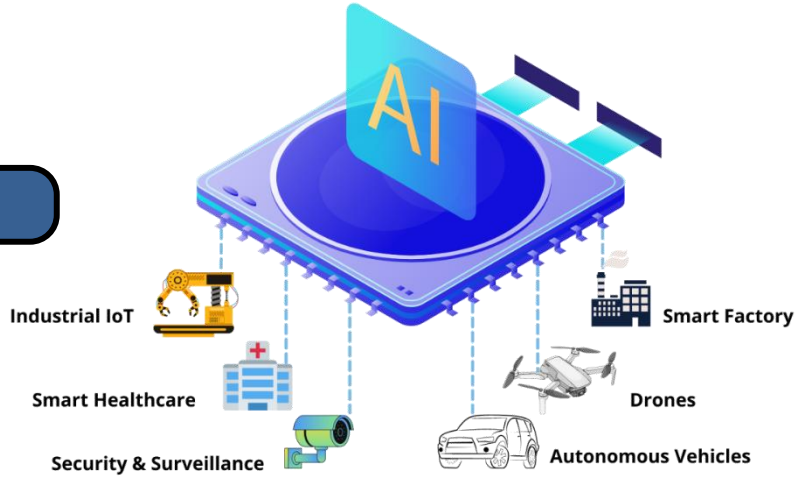

Autonomous Systems


Security and Privacy


Performance and Efficiency

Edge Artificial Intelligence (AI)

Deploy Artificial Intelligence in mobile devices



Edge AI challenges

Software

ChatGPT Gemini ResNet

- Memory-intensity
- Compute-intensity
- Energy-intensity

Hardware

- Small memories
- Limited computing resources
- Small batteries

High accuracy with limited resources!

Research efforts

- Software Optimization
- Hardware Optimization
- **SW/HW co-design**

Energy Harvesting (EH) IoT

Very limited energy budgets

- Harvesters extract solar, piezoelectric, or thermal energy
- They only get limited energy from the environment

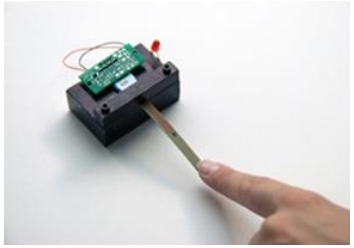
Constrained HW resources

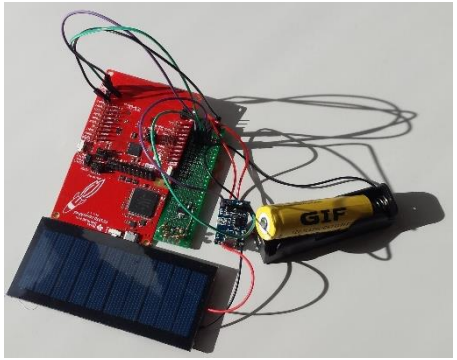
- Limited memory, storage (KB-MB), and compute resources

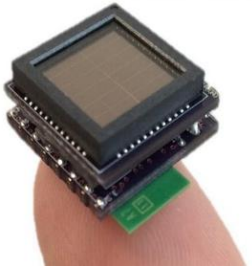
Dynamic energy availability

- Small batteries can provide more stable energy sources
- Extracted energy profiles have a dynamic shape

Can we empower EH IoT with AI capabilities?

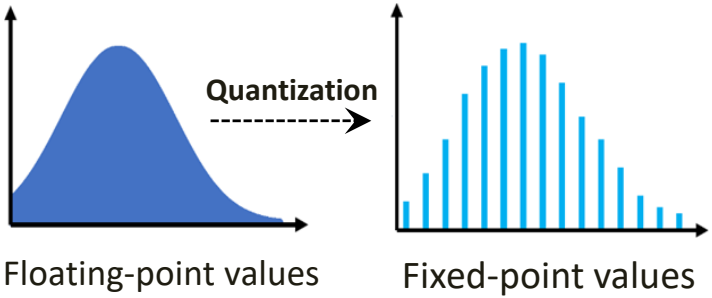






Optimizing convolutional neural networks for ultra-low power embedded systems

Quantization and Pruning

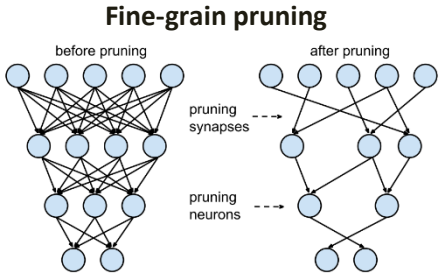


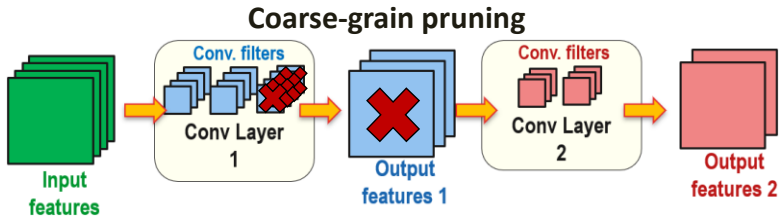
Quantization [ACM TIST'23]

- Reducing operands' bitwidth (e.g., from fp32 to int8)
- Integer arithmetic improves energy efficiency
- If too aggressive, may result in accuracy degradation

Pruning [IEEE TPAMI'24]

- Removes model parameters (e.g., *weights*)
- Fine-grain → removes individual weights
- Coarse-grain → removes entire filters





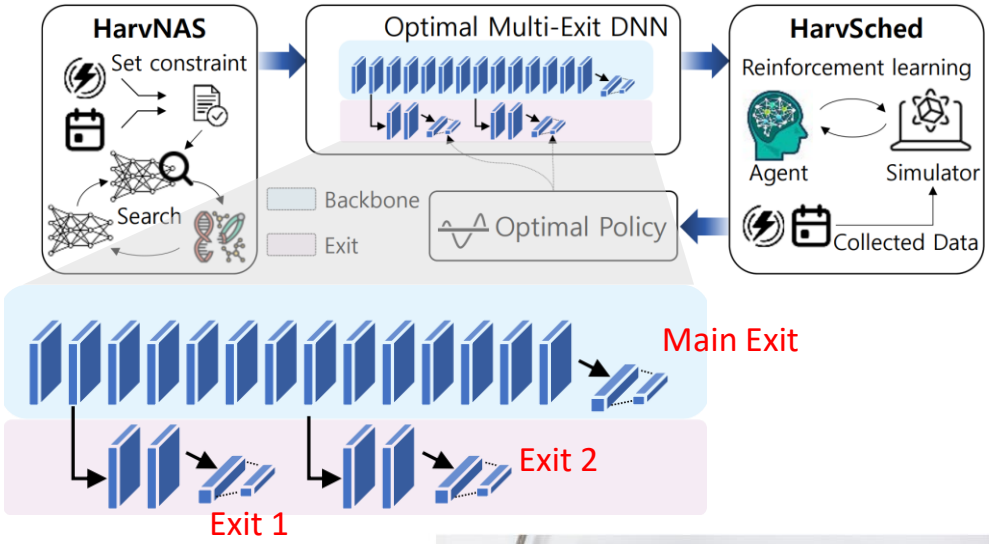
Rokh et al., "A Comprehensive Survey on Model Quantization for Deep Neural Networks in Image Classification", ACM TIST, 2023

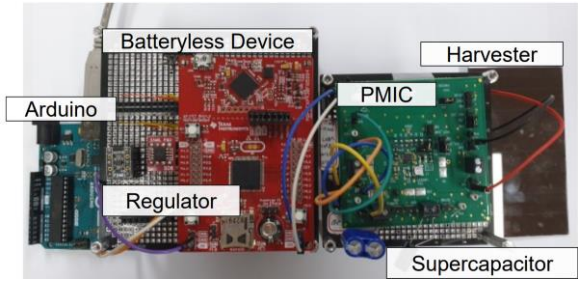
Cheng et al., "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations", IEEE TPAMI 2024

Early-exit CNNs

HarvNet [MobiSys'23]

- Neural Architecture Search with Reinforcement Learning for EH IoT optimization
- A lightweight DNN is enriched with multi-exit layers
- Early exits can be used when energy levels are low to reduce compute requirements!
- When to exit is determined via Reinforcement Learning
- Main limitation is **memory overhead**
  - Additional layers must be stored for each exit



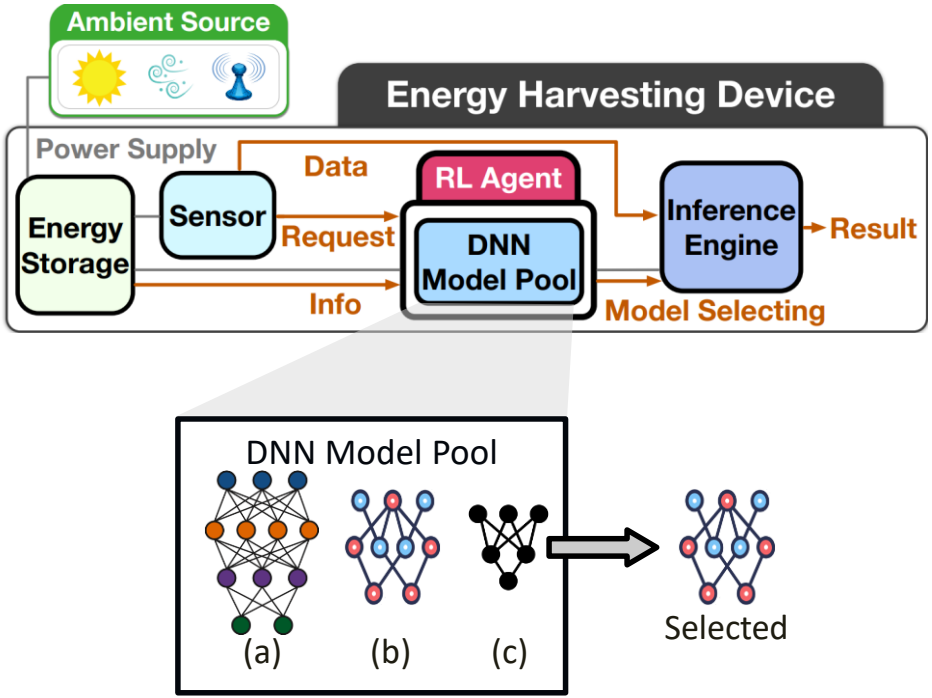


Jeon et al., "Harvnet: resource-optimized operation of multi-exit deep neural networks on energy harvesting devices." MobiSys 2023.

Using pools of Neural Network (NN) models

Adaptive inference for EH-based IoT [ISLPED'23]

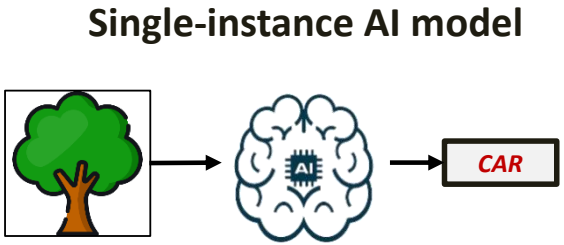
- Store in memory NNs of different complexity
- Get energy level info for each inference task
- Use RL to predict which NN should be used for inference given current energy levels
- Main limitation is **memory overhead**
  - Need to store multiple NNs

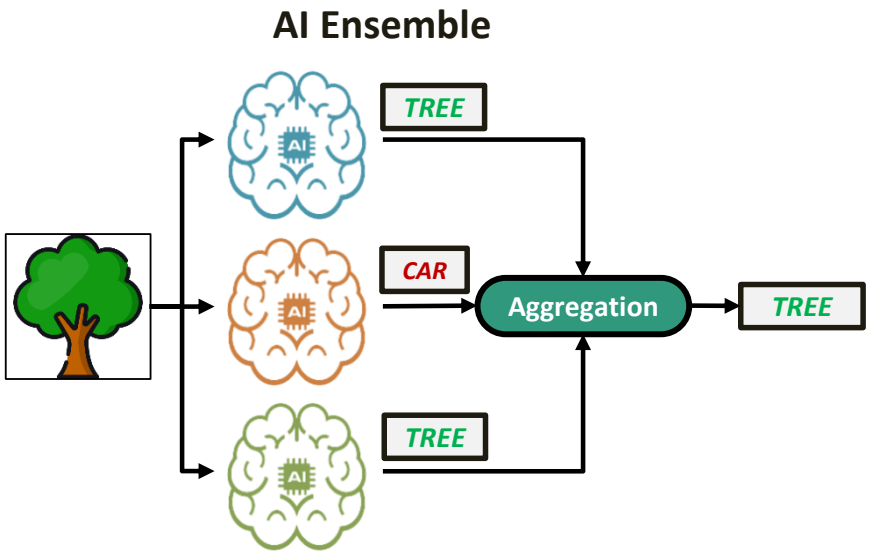


Park et al., "Energy-Harvesting-Aware Adaptive Inference of Deep Neural Networks in Embedded Systems." ISLPED 2023.

Hardware-aware ensembles of convolutional neural networks

AI Ensembling methods





Combine diverse AI models

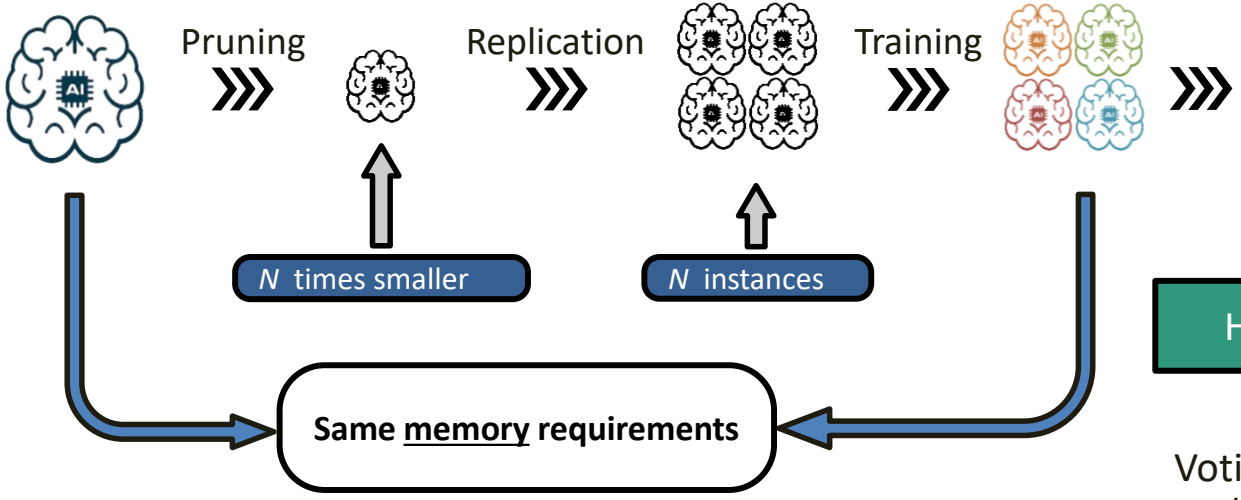
- Same input processed by multiple models
- **Diversity** enables higher accuracy
- Ensembles are **more robust**

High overhead

- **High memory** requirements
- **Increased computation**

E²CNN Methodology

Construct CNN Ensembles with no memory overhead



E²CNNs [IEEE TC'21]

How to introduce diversity?

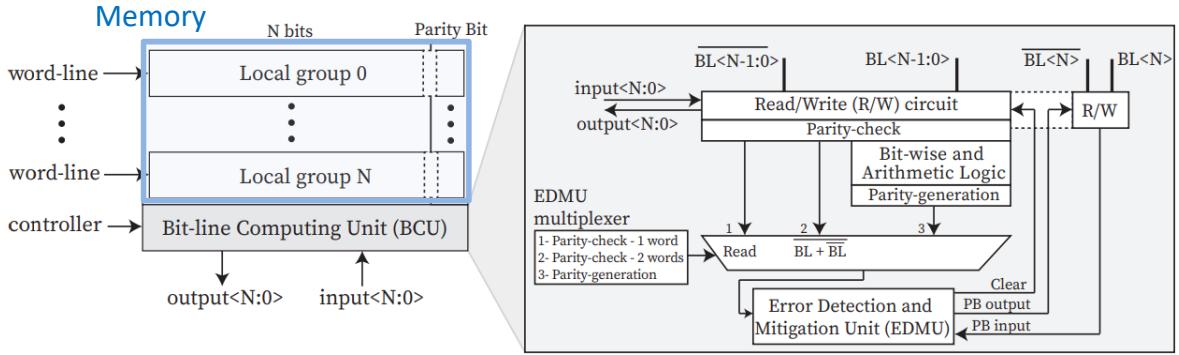
Voting-based ensembling method
(*Independent CNN training*)

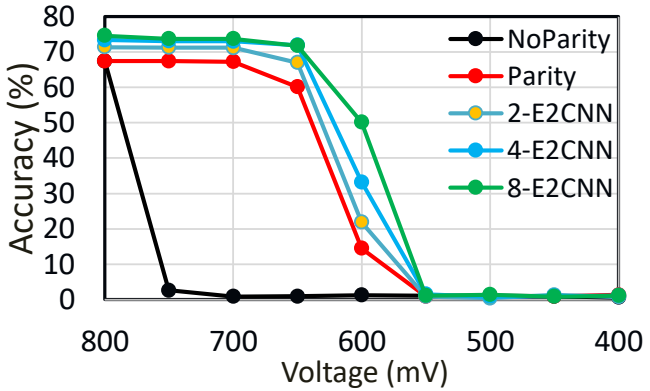
Ponzina et al., "E²CNNs: Ensembles of convolutional neural networks to improve robustness against memory errors in edge-computing devices", IEEE TC 2021

E²CNN to support memory voltage scaling

E²CNN for Error-Resilient PIM [GLSVLSI'22]

- Use of bitline computing for CNN acceleration
- Goal → Reduce voltage for higher energy efficiency
- Challenge → Operating at sub-nominal voltages introduces stuck-at-faults in memory cells → (accuracy degradation)
- Solution → Leverage the high robustness of E²CNN to compensate for memory errors
- Key results → **50% energy savings** (down to 650mV) with up to **5% accuracy improvement**





Rios *et al.*, "*Error resilient in-memory computing architecture for cnn inference on the edge*." Proceedings of the Great Lakes Symposium on VLSI, 2022.

From E²CNN to E-QUARTIC

Novel HW-aware pruning algorithm

- Memory *and* compute constraints
- Reduced memory needs

Boosting to generate the ensemble

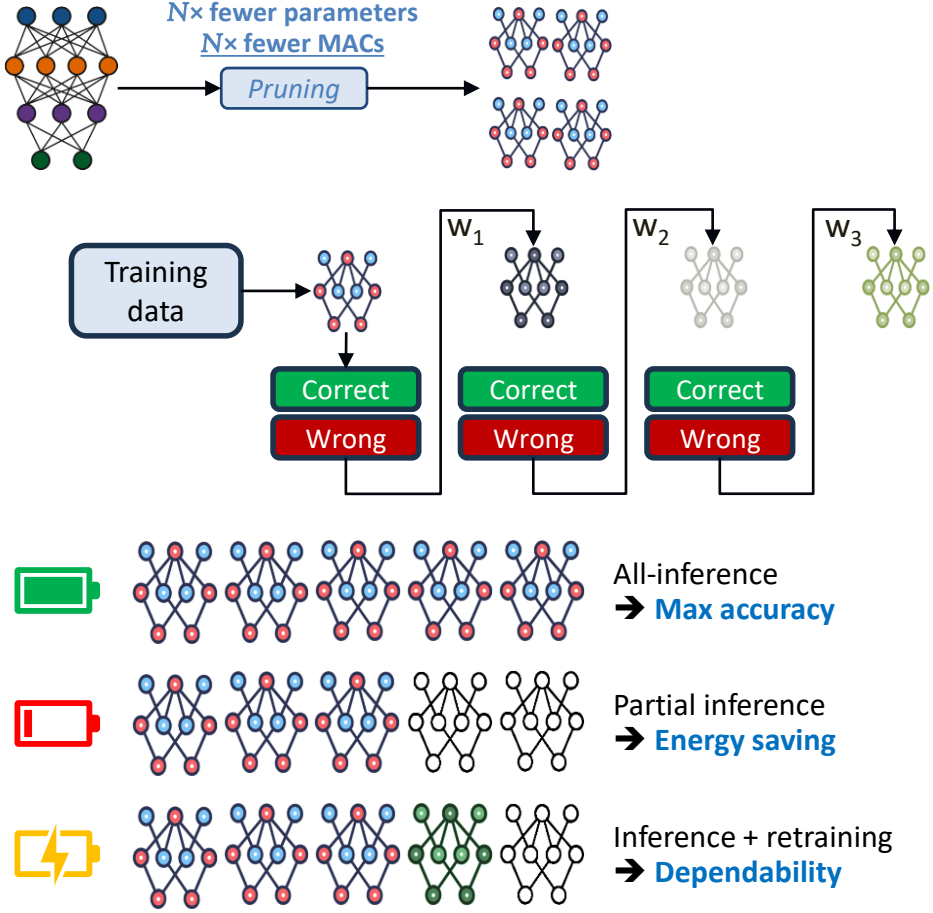
- Training focuses on previous misclassifications
- Higher accuracy

Exploit the multi-instance model

- Use a subset of instances only, if needed
- Energy/Accuracy trade-off
- Concurrent inference and on-device training

Energy-aware HW scheduler

- Determine E-QUARTIC execution
- Use energy info to decide what learners do



E-QUARTIC overview

HW-scheduler to leverage the ensemble-based model

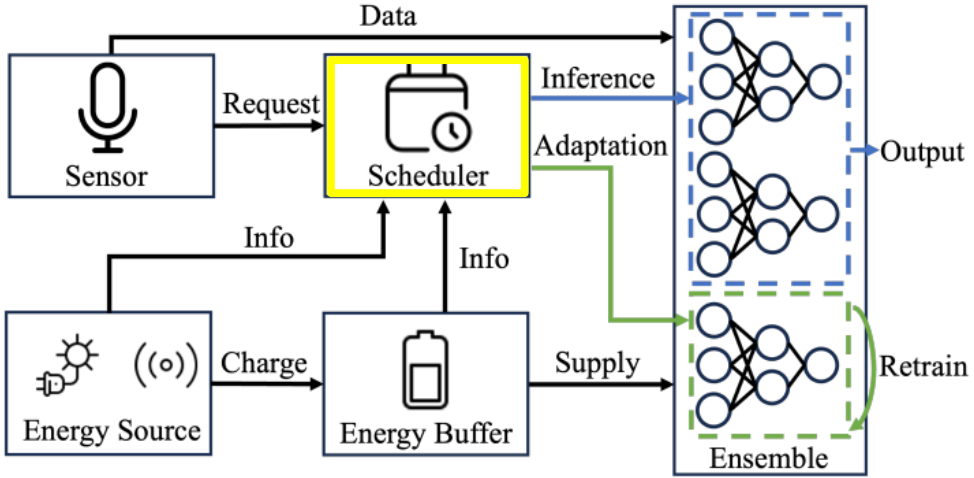
- Input from sensors, energy-harvesting system, and energy buffer
- Energy cost of inference depends on the number of learners to execute

Effective in energy-harvesting systems

- More accurate estimation of battery levels

Goal is to extend battery lifetime while achieving high accuracy

- Accuracy of a subset of learners is estimated from the validation set



E-QUARTIC: additional details

How to choose the $k < N$ learners to use?

- Use of a **Q-learning algorithm**
- Question: given *L* learners executed, do we want to use another one?
- Answer is based on an **expected reward**
- Reward evaluates the action taken (*execute vs. stop*)

State space definitions of our Q-learning algorithm

Symbols	Descriptions	Values
E_{now}	Current battery level	0: depleted, 1: low, 2: high, 3: full
E_{last}	Last 10 epochs mean energy	
P_{harv}	Harvesting power level	0: low, 1: mid, 2: high
L	Num of learners executed	$[0, 1, ..., N]$
R	Inference request	0: No, 1: Yes

How much accuracy do I expect to gain?

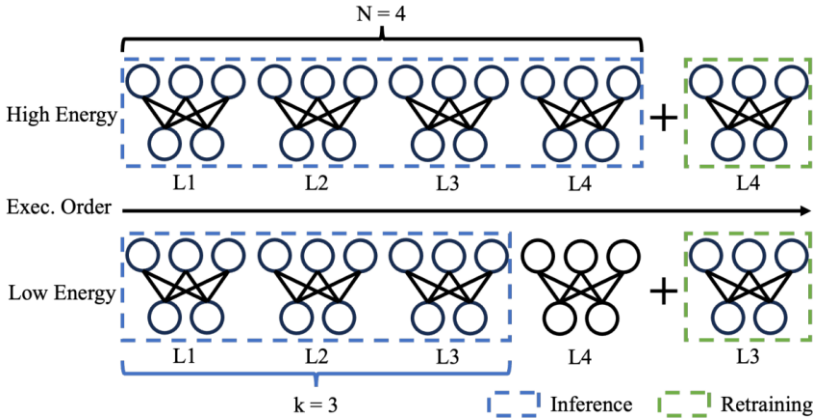
Try not to deplete stored energy

$$Reward(s,a) = \begin{cases} \Delta_{acc} - \beta(E_{max} - E_{now}) & a = 1 \\ -p_{miss} & R = 1 \text{ and } a = 0 \end{cases}$$

You should not avoid inference task

Concurrent inference-training

- Two options, based on energy levels
- Low-energy case
  - Runtime is still reduced by an average 25%
  - Negligible accuracy drop (0.4%)
  - Gradient storage is the main overhead



Experiments and Results

Experimental Setup

Real-hardware simulation

- Experiments run on the STM32L552ZE board - ARM Cortex-M33 (256KB SRAM, 512KB Flash)
- Indoor solar harvesting dataset [WDAA'19]

Baselines

- Single-instance CNN, Harvnet [MobiSys'23], Adaptive [ISLPED'23], E^2CNN [IEEE TC'21]

Metrics

- Accuracy, Memory, Performance, Energy

Benchmarks

- CNNs: ResNet-8, MobileNetV1, DSCNN
- Datasets: CIFAR-10, Visual Wake Word (VWW), Google Speech Commands (GSC-10)

Jeon et al., "Harvnet: resource-optimized operation of multi-exit deep neural networks on energy harvesting devices." MobiSys 2023.
Park et al., "Energy-Harvesting-Aware Adaptive Inference of Deep Neural Networks in Embedded Systems." ISLPED 2023.
Ponzina et al., "E^2CNNs: Ensembles of convolutional neural networks to improve robustness against memory errors in edge-computing devices", IEEE TC 2021
Sigrist et al., "Dataset: Tracing indoor solar harvesting." Proceedings of the 2nd Workshop on Data Acquisition to Analysis, 2019.

Results: Accuracy *vs*. Memory *vs*. Performance

Static evaluation

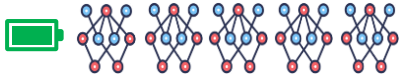
- No use of HW scheduler
- All learners are executed for inference
- Goal is to evaluate memory/performance and accuracy trade-off

Key results

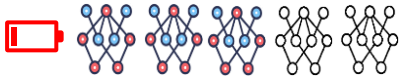
- Same compute as baseline models
- **Up to 54% memory reduction**
  thanks to the newly proposed pruning method
- Consistently achieving higher accuracy

Benchmark	SOTA	Memory (KB)	MAC (M)	Accuracy (%)
ResNet-8 (CIFAR10)	Single-instance	312	47	84.0
	E²CNN [IEEE TC'21]	311	52	78.2
	Harvnet [Mobisys'23]	339	47	83.5
	Adaptive [ISLPED'23]	430	47	84.0
	E-QUARTIC	**213**	**47**	**84.2**
MobileNetV1 (VWW)	Single-instance	279	34	82.0
	E²CNN [IEEE TC'21]	277	35	81.4
	Harvnet [Mobisys'23]	287	34	82.0
	Adaptive [ISLPED'23]	502	34	82.0
	E-QUARTIC	**125**	**34**	**82.8**
DSCNN (GSC-10)	Single-instance	44	56	94.0
	E²CNN [IEEE TC'21]	44	71	93.3
	Harvnet [Mobisys'23]	52	56	94.0
	Adaptive [ISLPED'23]	87	56	94.0
	E-QUARTIC	**31**	**55**	**94.9**

Results: Accuracy *vs*. Battery life



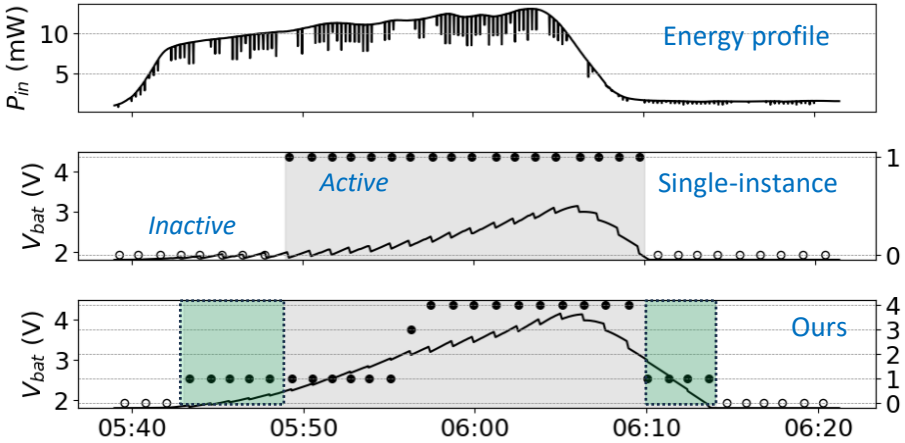
All-inference
➔ **Max accuracy**



Partial inference
➔ **Energy saving**

Dynamic evaluation

- Use the HW scheduler to save energy
- Definition: Failure rate is the number of times we have insufficient energy to perform an inference task



Comparison with baseline CNNs

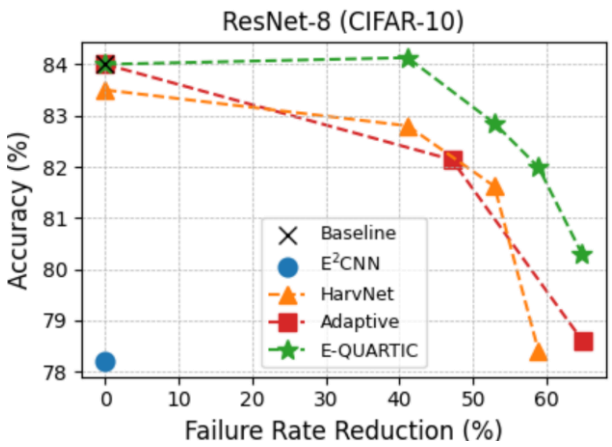
- E-QUARTIC with 4 instances
- Inference tasks start as soon as one instance can be run
- Extended battery life (green area)

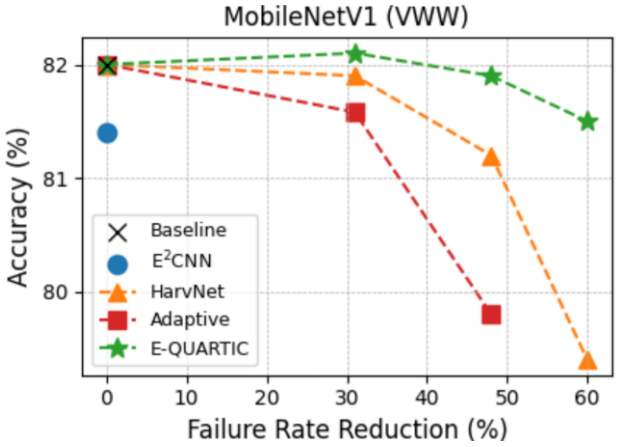
Results: Accuracy *vs.* Battery life

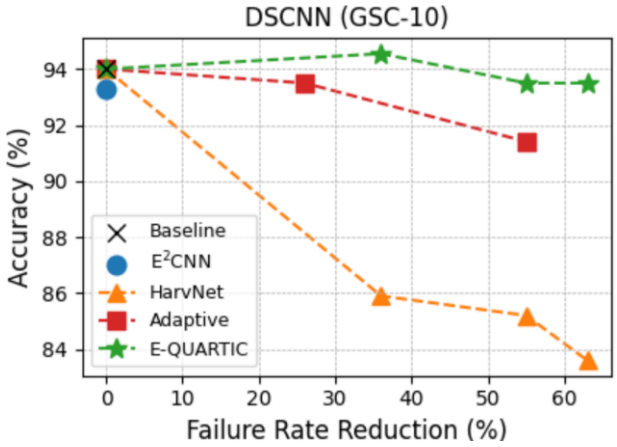
Key Results

- Average **+2.5% accuracy** than the best baseline for any failure rate reduction
- Average **+40% battery life** extension at iso-accuracy

Baselines.
Single-instance CNN
E^{2}CNN [IEEE TC'21]
Harvnet [MobiSys'23]
Adaptive [ISLPED'23]







Conclusions

Need for extremely high energy efficient in edge AI

- Empower edge devices with AI capabilities requires efficient computation
- HW resources are limited and EH IoT relies on energy extracted from the environment
- Current methods either incur **memory overhead** or **do not adapt to dynamic energy profiles**

E-QUARTIC

- **E-QUARTIC** uses HW-aware ensembles to improve efficiency/accuracy trade-off
- It leverages a multi-instance architecture **to dynamically adjust inference tasks to energy budgets**
- It also supports **concurrent inference and training stages**, key for high reliability/availability

Key insights

- New pruning algorithm enables up to **54% memory savings** compared to E^2CNN
- **Higher accuracy (+2.5%)** than state-of-the-art approaches for any failure rate levels
- E-QUARTIC achieves state-of-the-art accuracy while **extending battery life by 40%**