tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

"Physics-Aware Auto Tiny Machine Learning"

Swapnil Sayan Saha – Algorithm Development Engineer, STMicroelectronics Inc.

March 5, 2024

TINY ML

www.tinyML.org

Thank you, **tinyML Strategic Partners**,
for committing to take tinyML to the next Level, together

Executive Strategic Partners

Qualcomm AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency

Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization

Continuous learning, contextual, always-on, privacy-preserved, distributed learning

Efficient learning

Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception

Object detection, speech recognition, contextual fusion

Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

Edge cloud

Cloud

IoT/IIoT

Automotive

Mobile

Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

Platinum Strategic Partners

embedUR

DEPLOY VISION AI AT THE EDGE AT SCALE

Gold Strategic Partners

Build the
Future of tinyML

on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

BIG IDEAS FOR EVERY SPACE

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

synaptics®

ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com

Silver Strategic Partners

Join Growing tinyML Communities:

19.6k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

4.2k members
&
14.5k followers

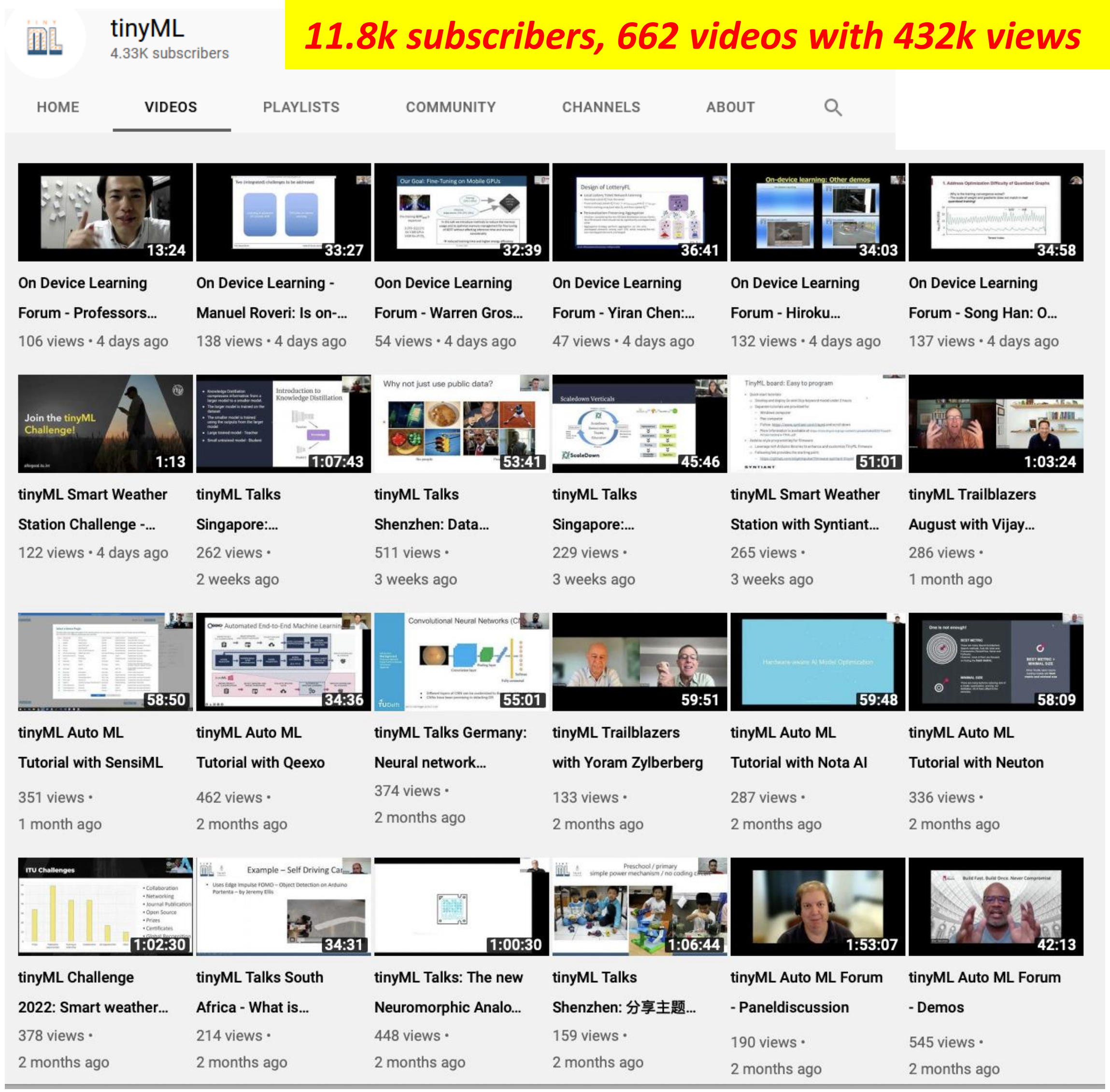
The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

YouTube

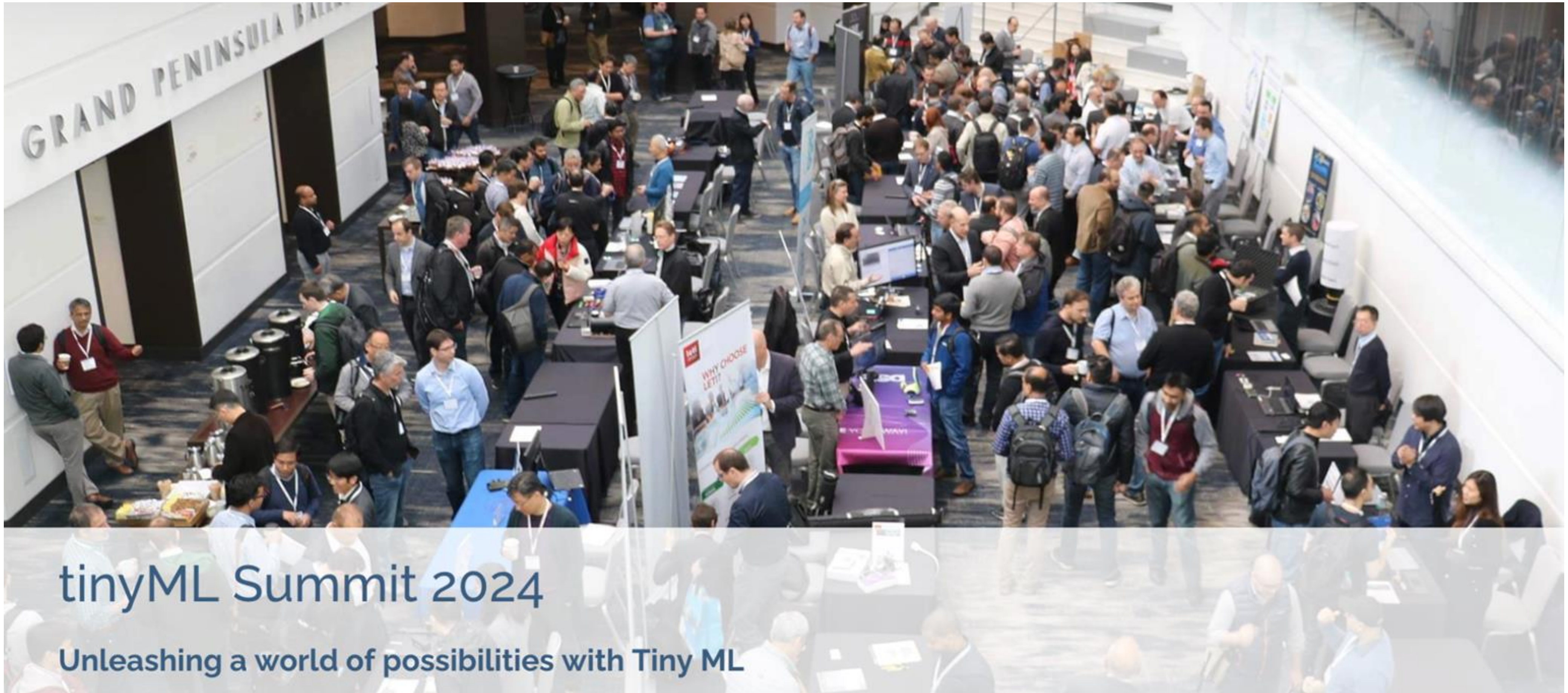
11.8k subscribers, 662 videos with 432k views



Subscribe to
tinyML YouTube Channel
for updates and notifications
(including this video)

www.youtube.com/tinyML

tinyML Summit
April 22 - 24, 2024 - Register now!



tinyML Awards 2024

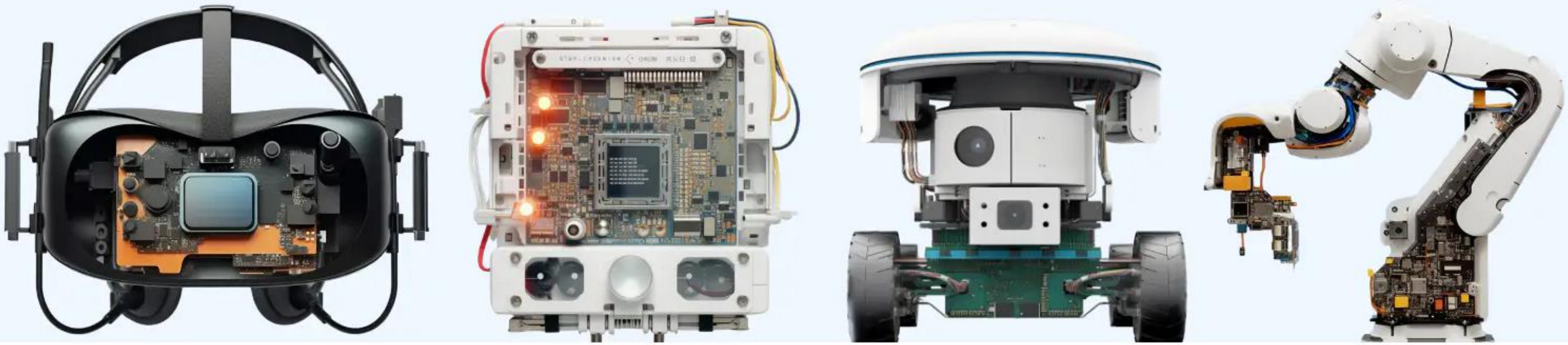
WEVOLVER

https://www.wevolver.com/article/2023-edge-ai-technology-report

TINY ML FOUNDATION

2023 Edge AI Technology Report

The guide to understanding the state of the art in hardware & software in Edge AI.



Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions

Swapnil Sayan Saha

Swapnil Sayan Saha is an algorithm development engineer at STMicroelectronics Inc. He received his Ph.D. and M.S. in Electrical and Computer Engineering from the University of California, Los Angeles in 2023 and 2021 respectively, and B.Sc. in Electrical and Electronics Engineering from the University of Dhaka in 2019. His research explores how rich, robust, and complex inferences can be made from sensors onboard low-end embedded systems within tight resource budgets in a platform-aware fashion. To date, he has published more than 25 peer-reviewed articles/patents and received more than 30 awards in robotics, technical, and business-case forums worldwide.

life.augmented

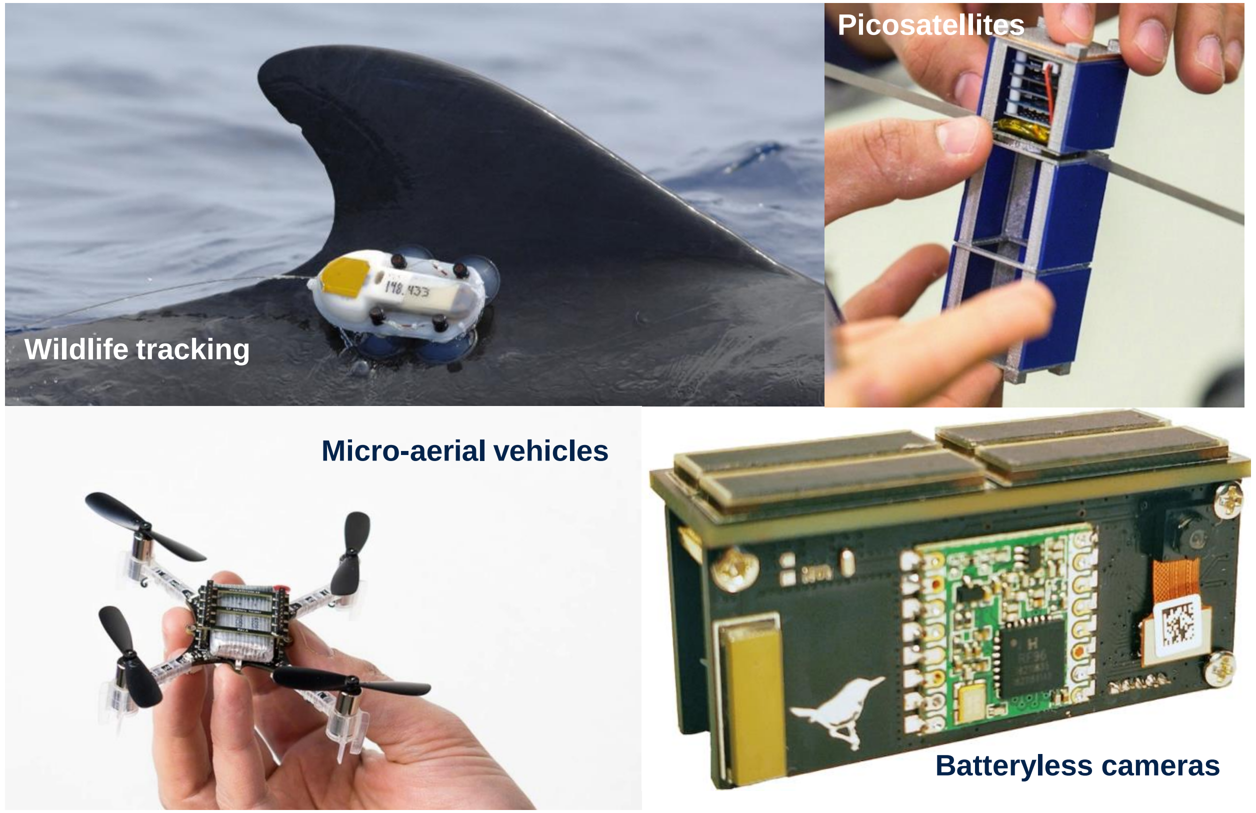
Physics-aware auto tiny machine learning

Swapnil Sayan Saha, Ph.D.
Algorithm Development Engineer
STMicroelectronics

TINY ML TALKS webcast

Tiny machine learning

Hardware and **software** suites that enable **always-on**, **ultralow power**, and **on-device** data analytics



Enables applications that need to make "**complex inferences**" for "**time-critical**" and "**remote**" applications from "**unstructured data**" independent of large systems.

life.augmented

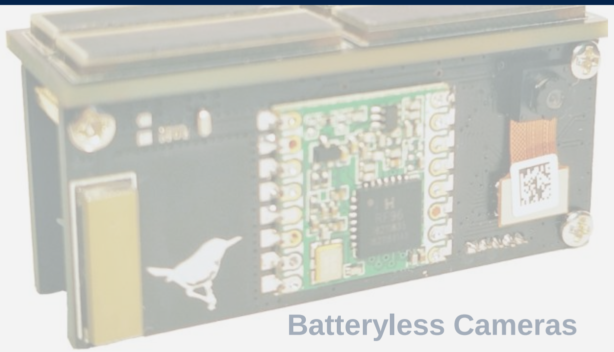
Saha, Swapnil Sayan, Sandeep Singh Sandha, and Mani Srivastava. "Machine Learning for Microcontroller-Class Hardware: A Review." in *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21362-21390, 15 Nov.15, 2022.

Tiny machine learning

Hardware and **software** suites that enable **always-on**, **ultra-low power**, and **on-device** data analytics.

First generation efforts focused on squeezing standalone neural networks within the resource bounds of tinyML platforms

Micro-Aerial Vehicles



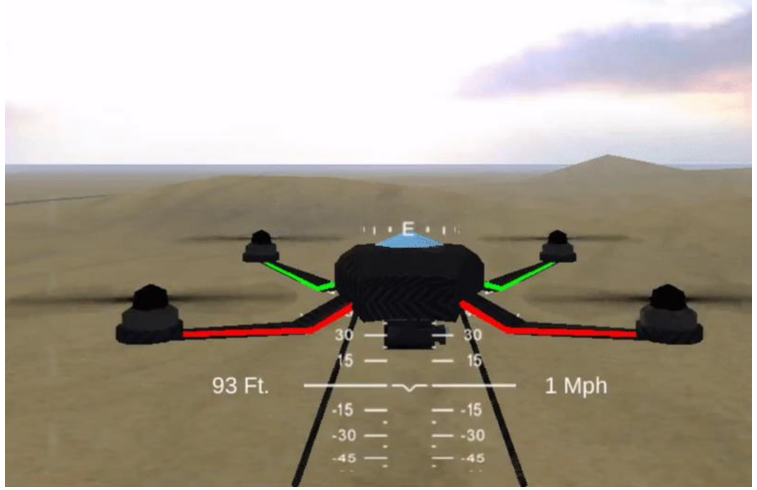
Batteryless Cameras

large systems.

Saha, Swapnil Sayan, Sandeep Singh Sandha, and Mani Srivastava. "Machine Learning for Microcontroller-Class Hardware: A Review." in *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21362-21390, 15 Nov.15, 2022.

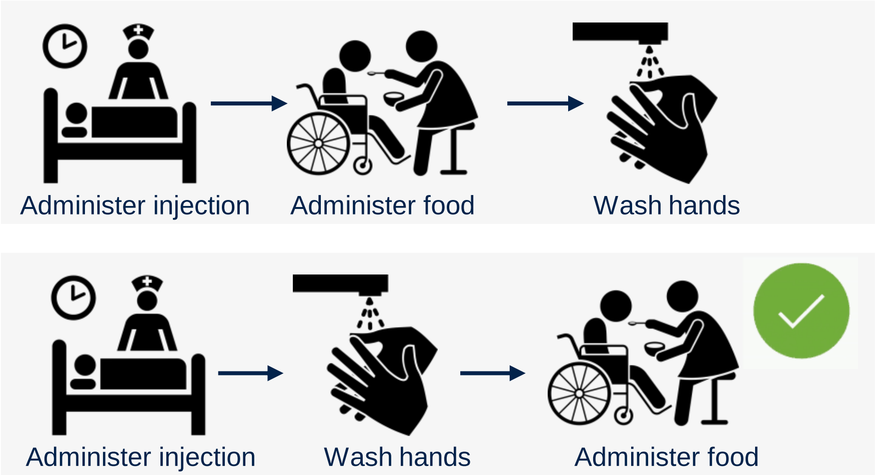
Challenge 1: obeying physics, rules, and constraints

Goal: land the quadrotor as fast as possible



Quadrotor crash due to unsafe bank angle caused by a neural flight controller.

Goal: guide a nurse perform correct action sequences



Invalid and valid sanitary protocol sequence in a nurse care setting for complex event processing.

Physics matter more at the edge; standalone neural networks cannot assure that the learned distributions obey the rules and physics of the underlying system

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Challenge 1: obeying physics, rules, and constraints

Goal: land the quadrotor as fast as possible

Goal: guide a nurse perform correct action sequences

Additional routines (called symbolic programs) must be jointly deployed with the neural network

Quadrotor crash due to unsafe bank angle caused by a neural flight controller.

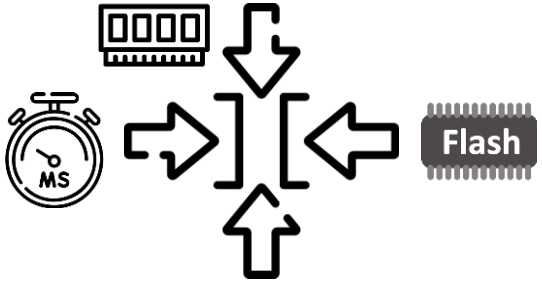
Invalid and valid sanitary protocol sequence in a nurse care setting for complex event processing.

Physics matter more at the edge; standalone neural networks cannot assure that the learned distributions obey the rules and physics of the underlying system

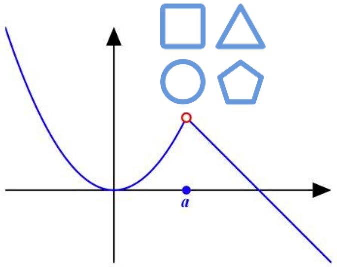
life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Challenge 2: synthesizing platform-aware neurosymbolic programs

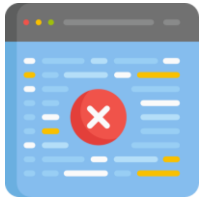


Fitting neural and symbolic routines within resource limits



Dealing with mixed and discontinuous parameter spaces

Arbitrary cost function formulation

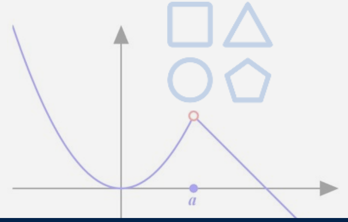


Dealing with runtime faults due to more moving parts

Finding proxies for so many program combinations

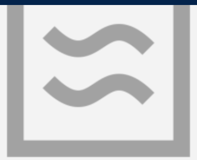
Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Challenge 2: synthesizing platform-aware neurosymbolic programs



Finding the optimal synergy between neural and symbolic components within the tight resource constraints is challenging

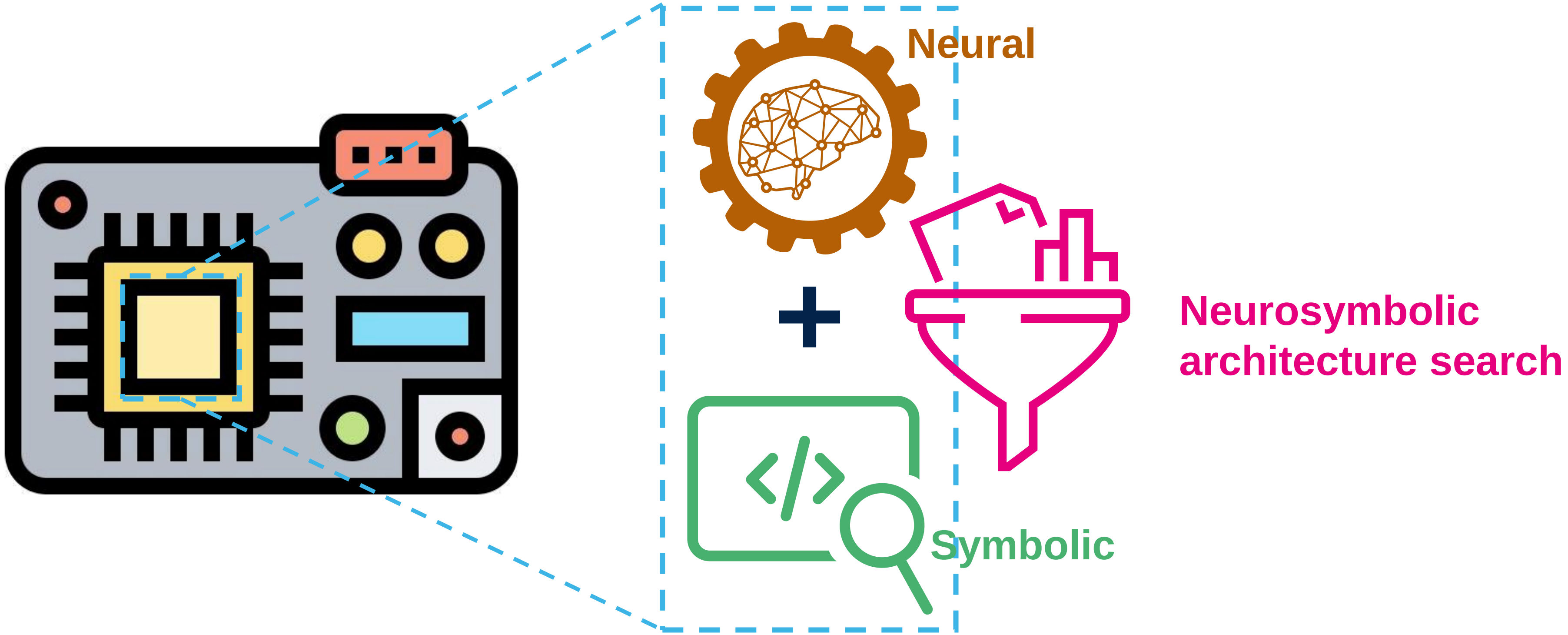
Dealing with runtime faults due to more moving parts



Finding proxies for so many program combinations

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Neurosymbolic auto tiny machine learning



Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

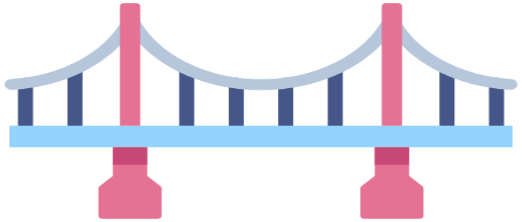
What is neurosymbolic artificial intelligence?

A program containing neural and human-readable (symbolic) code

Performant

Non-parametric (flexible and scalable)

Robust

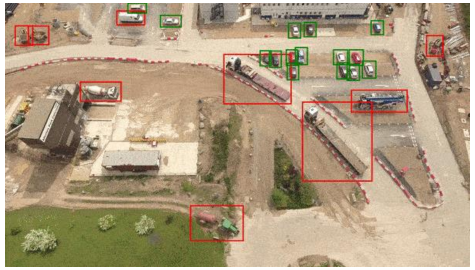


Neurosymbolic AI

Interpretable

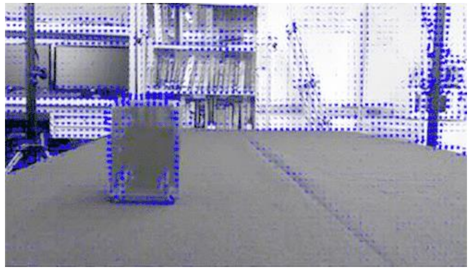
Data efficient

Physics and context-aware



Aerial object detection

Self-navigating surveillance quadrotor



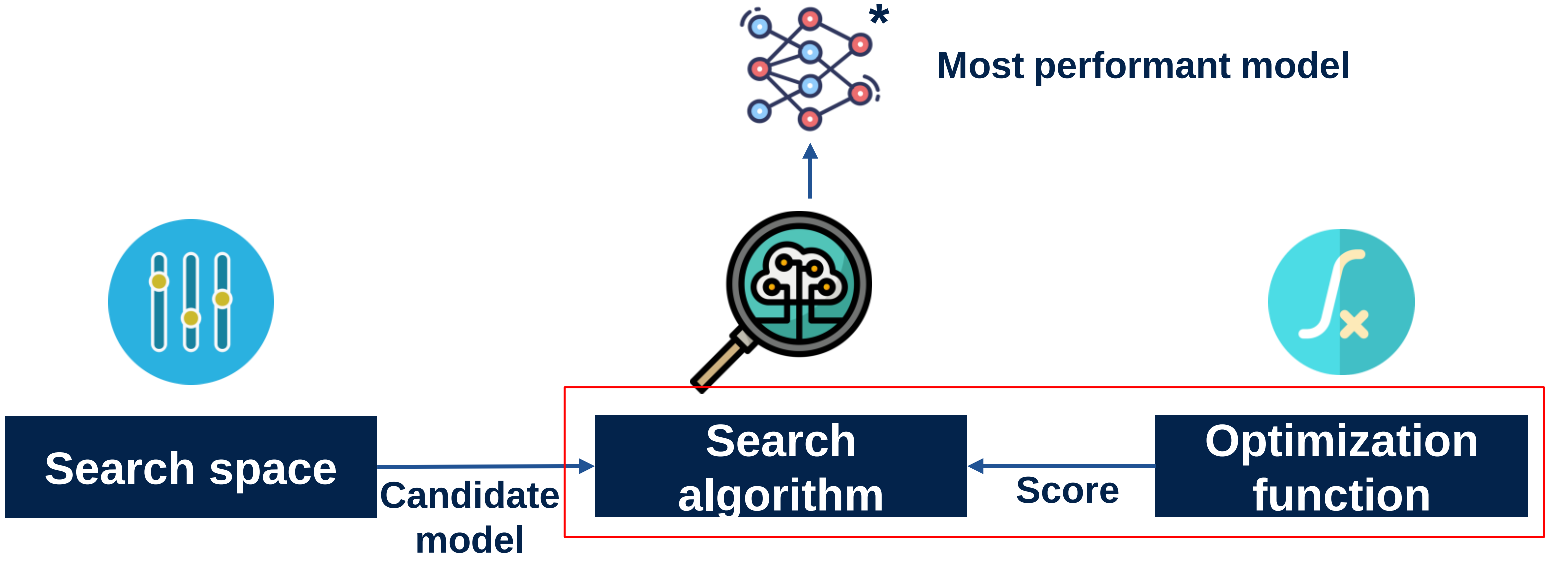
Optical flow

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

life.augmented

What is neural architecture search?

Automatically find the most performant neural network architecture from a hyperparameter space within some constraints



Saha, Swapnil Sayan, Sandeep Singh Sandha, and Mani Srivastava. "Machine Learning for Microcontroller-Class Hardware: A Review." in *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21362-21390, 15 Nov.15, 2022.

What is neural architecture search?

Automatically find the most performant neural network architecture from a hyperparameter space within some constraints

We adopt Mango, a black-box Bayesian optimizer, which can efficiently handle mixed and discontinuous search spaces

Search space

Candidate model

Search algorithm

Score

Optimization function

Saha, Swapnil Sayan, Sandeep Singh Sandha, and Mani Srivastava. "Machine Learning for Microcontroller-Class Hardware: A Review." in *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21362-21390, 15 Nov.15, 2022.

Gradient-free Bayesian optimization: surrogate function

Two components: surrogate function and acquisition function

$$\hat{f}(\Omega) \sim \mathcal{GP}(\mu(\Omega), k(\Omega, \Omega'))$$

A surrogate function approximates an optimization function, e.g., gaussian process

Gaussian process provides tractable assessment of uncertainty under data scarcity

Non-parametric model using mean μ and Matern kernel function *k* over the search space Ω

Sandha, Sandeep Singh, Mohit Aggarwal, Igor Fedorov, and Mani Srivastava. "Mango: A python library for parallel hyperparameter tuning." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3987-3991. IEEE, 2020.

Sandha, Sandeep Singh, Mohit Aggarwal, Swapnil Sayan Saha, and Mani Srivastava. "Enabling hyperparameter tuning of machine learning classifiers in production." In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 262-271. IEEE, 2021.

Gradient-free Bayesian optimization: acquisition function

Two components: surrogate function and acquisition function

$$\Omega_t = \arg\max_{\Omega}\left(\underline{\mu_{t-1}(\Omega)} + \underline{\beta^{0.5}\sigma_{t-1}(\Omega)}\right)$$

An acquisition function selects the next promising set of points to sample

Mango adopts Monte Carlo sampling with upper confidence bound, using adaptive exploration factor β

First term: goodness of sampled point (exploitation); second term: uncertainty of sampled point (exploration); does not get stuck in local optima

Sandha, Sandeep Singh, Mohit Aggarwal, Igor Fedorov, and Mani Srivastava. "Mango: A python library for parallel hyperparameter tuning." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3987-3991. IEEE, 2020.
Sandha, Sandeep Singh, Mohit Aggarwal, Swapnil Sayan Saha, and Mani Srivastava. "Enabling hyperparameter tuning of machine learning classifiers in production." In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 262-271. IEEE, 2021.

Gradient-free Bayesian optimization: acquisition function

Two components: surrogate function and acquisition function

$$\Omega_t = \arg\max_{\Omega}(\underline{\mu_{t-1}(\Omega)} + \underline{\beta^{0.5}\sigma_{t-1}(\Omega)})$$

Adaptive exploration factor finds near-optimal values at the boundary of violating deployability constraints with 90% theoretical guarantees

bound, using adaptive exploration factor β

First term: goodness of sampled point (exploitation); second term: uncertainty of sampled point (exploration); does not get stuck in local optima

Sandha, Sandeep Singh, Mohit Aggarwal, Igor Fedorov, and Mani Srivastava. "Mango: A python library for parallel hyperparameter tuning." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3987-3991. IEEE, 2020.
Sandha, Sandeep Singh, Mohit Aggarwal, Swapnil Sayan Saha, and Mani Srivastava. "Enabling hyperparameter tuning of machine learning classifiers in production." In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 262-271. IEEE, 2021.

life.augmented

Formulating the neurosymbolic optimization function

$$\min \mathbf{f}(\Omega), \quad \text{s.t.} \quad \mathbf{f}(\Omega) \leq \mathbf{b}$$

$$\min\ f_{\text{opt}}, \quad f_{\text{opt}} = \lambda_1 f_{\text{error}}(\Omega) + \lambda_2 f_{\text{flash}}(\Omega) + \lambda_3 f_{\text{SRAM}}(\Omega) + \lambda_4 f_{\text{latency}}(\Omega)$$

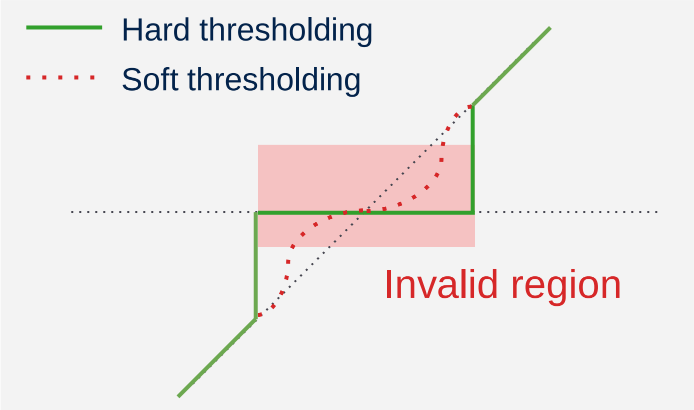
GP-UCB solves a non-linear program with constraints

Goal: construct a fault-free neurosymbolic program such that latency and error are minimized, while the memory usage is maximized within device memory limits

Search space Ω contains both neural and symbolic hyperparameters, trainable weights, neural operators, and symbolic program atoms

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Fast and guaranteed deployment: hard thresholding



To guarantee deployability and optimize at the execution level, we perform platform-in-the-loop search

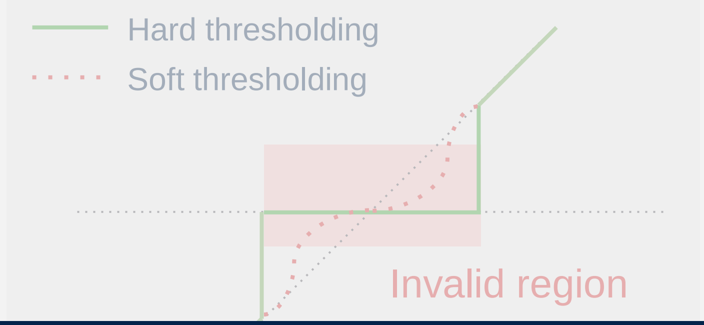
If a model induces faults, we do not train the model; the search algorithm is penalized by a constant large number (hard thresholding)

Thanks to GP-UCB, the search algorithm is able to observe the small valid linear region where memory usage and accuracy are proportional

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Fast and guaranteed deployment: hard thresholding



Platform-in-the-loop + hard thresholding =
50% faster than proxy + soft thresholding

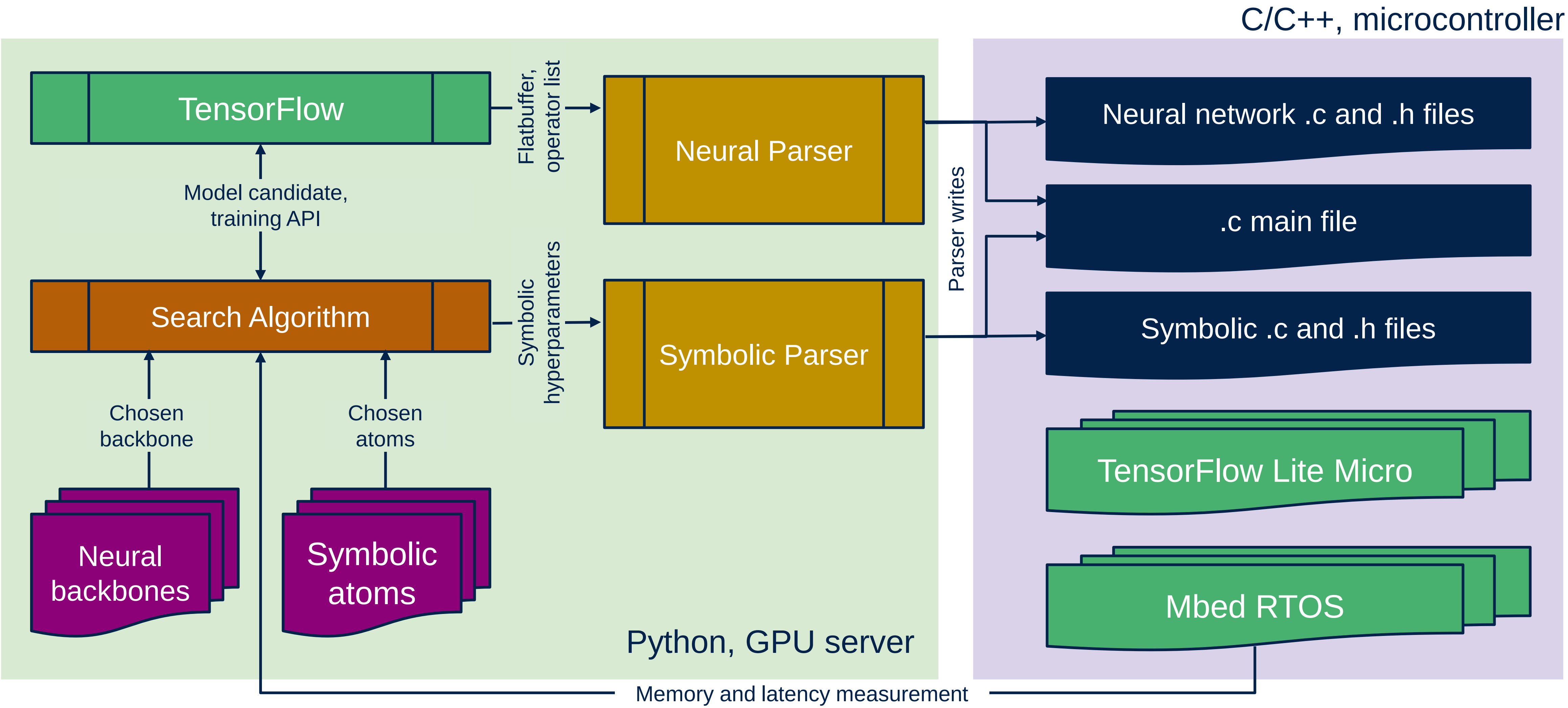
If a model induces faults, we do not train the model; the search algorithm is penalized by a constant large number (hard thresholding)

Thanks to GP-UCB, the search algorithm is able to observe the small valid linear region where memory usage and accuracy are proportional

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Implementing automatic platform-in-the-loop



Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

life.augmented

Comparing tinyML NAS strategies

Method	Search strategy	Profiler	Search space	Optimization terms
SpArSe	Gradient-driven Bayesian	Analytical	Conv2D	Error, SRAM, flash
MCUNet	Evolutionary (weight sharing)	Lookup tables	Conv2D	Error, SRAM, flash, latency
MicroNets	One-shot DNAS	Analytical	Conv2D	Error, SRAM, flash, latency
µNAS	Evolutionary (no weight sharing)	Analytical	Conv2D	Error, SRAM, flash, latency
iNAS	Reinforcement learning	Lookup tables	Conv2D (execution level)	Error, flash, latency, volatile buffer, power cycle energy
UDC	DNAS with exploration/exploitation	Analytical	Conv2D, pruning, quantization	Error, flash
TinyNS	**Gradient-free Bayesian with exploration/exploitation**	**Platform-in-the-loop**	**Any supported ML operator and symbolic atoms**	**Any scalar term**

Operates on mixed and arbitrary parameter spaces and optimization function terms

Platform-in-the-loop guarantees deployability by taking execution level dynamics into account

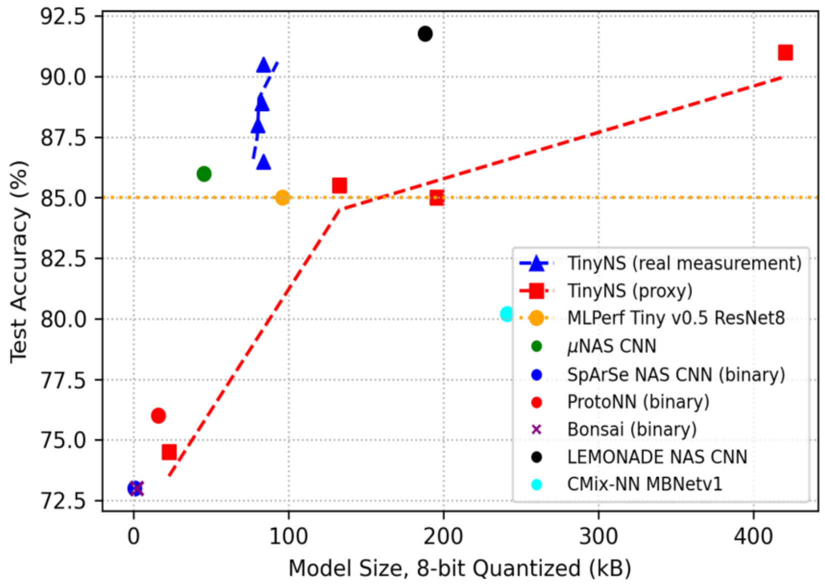
Efficient over RL, DNAS, and evolutionary search algorithms; combines the best features of other search algorithms in one package

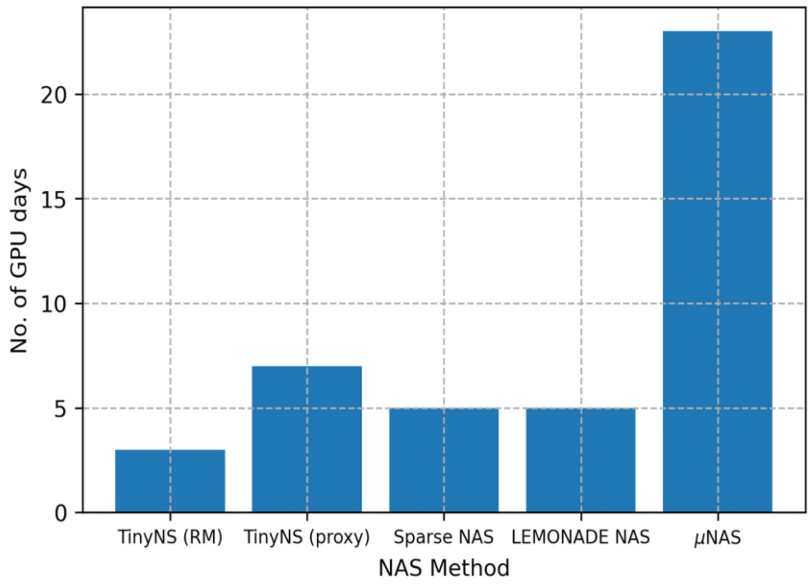
life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

MLPerf Tiny v0.5 inference benchmark (CIFAR10)

Backbone: ResNet8





Exceeds benchmark accuracy by 4.3%, outperforms baselines by 4.5%-17.5%

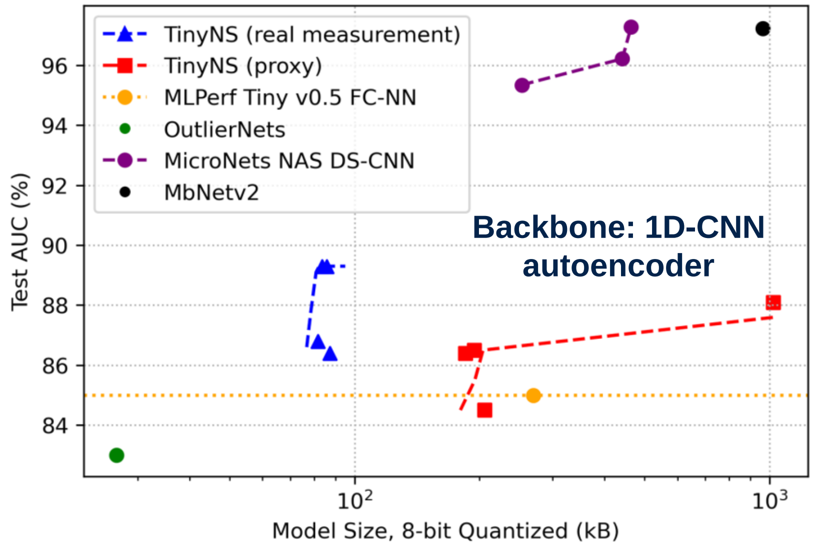
1.7x-7.7x lower convergence time than baselines

Platform-in-the-loop provides models that have 1.6%-5.5% higher accuracy and consumes 4.2x lower flash, while converging 2.3x faster

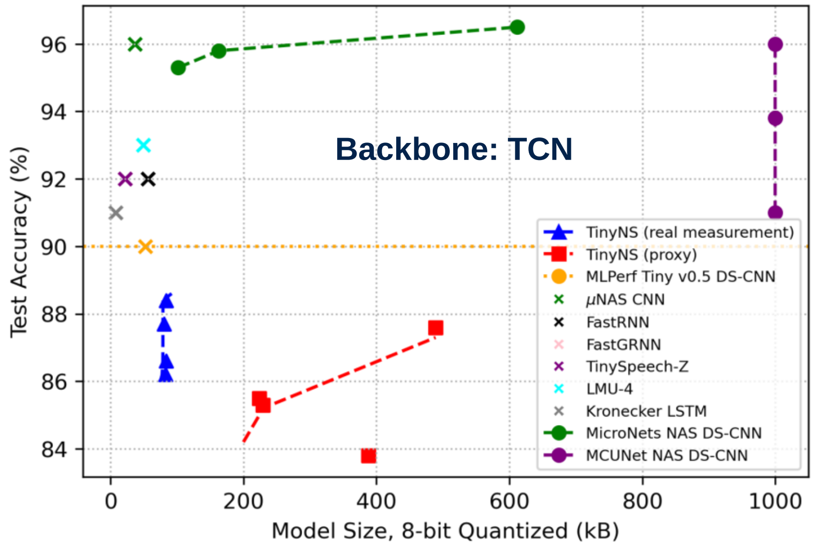
life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

MLPerf Tiny v0.5 inference benchmark – search space matters



Anomaly Detection



Keyword Spotting

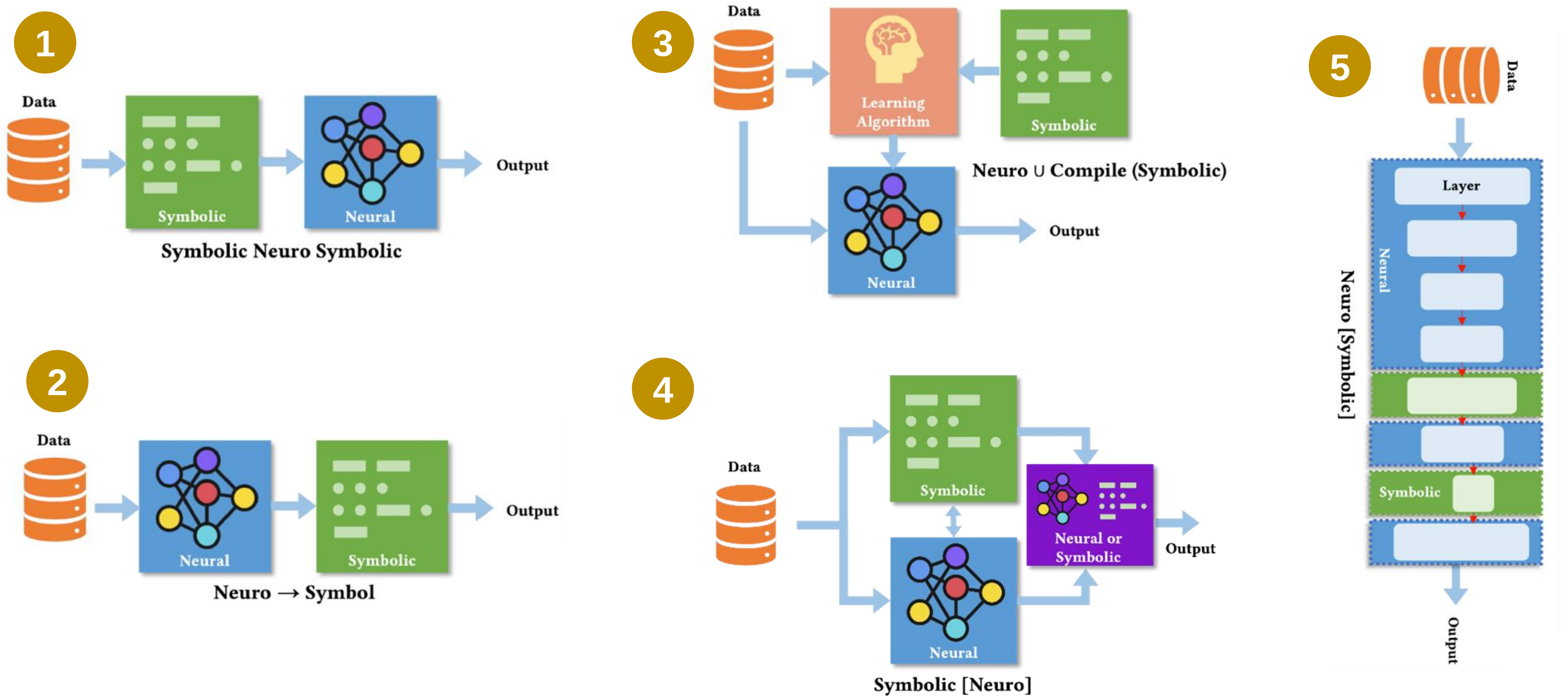
Anomaly Detection (ToyADMOS): exceeds benchmark accuracy by 4%, outperforms OutlierNets by 6.3%, guarantees deployability over MbNetv2

Keyword spotting (Speech): incorrect backbone leads to suboptimal pareto-frontier, stressing importance of a search space containing several models

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

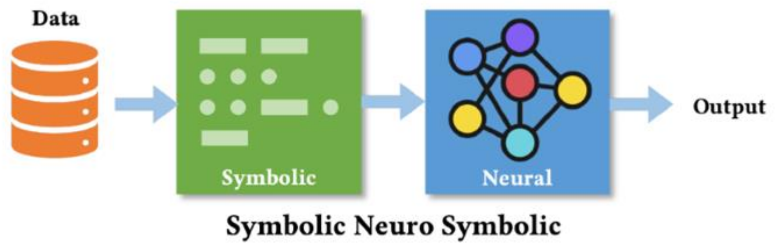
Neurosymbolic AI taxonomy



Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

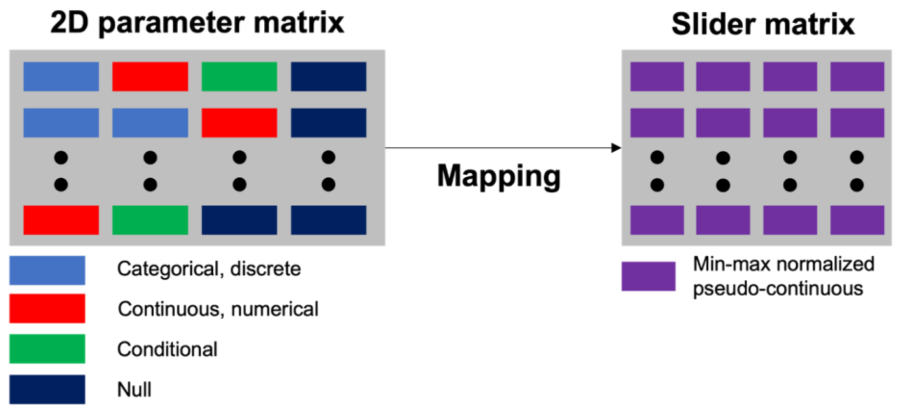
Symbolic neurosymbolic – problem formulation

A series of independent domain-engineered functions is applied on the input dataset X, followed by a single ML model



Symbolic search space: 2D hyperparameter matrix, with each row corresponding to the arguments of each function (binary mask)

Neural search space: multiple model backbones, each considered for use at each step using an ordinal mask



Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Example: co-optimizing features and a single backbone

Application: gesture recognition using a temporal CNN; operates on handcrafted features

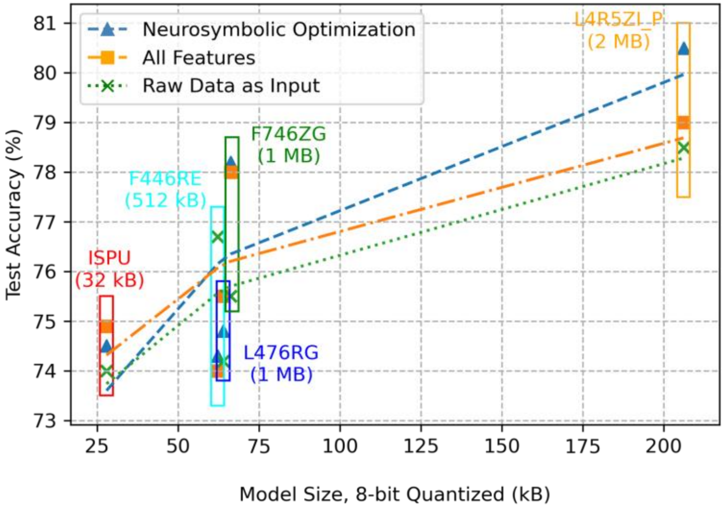
Features											
Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Example: co-optimizing features and a single backbone

Application: gesture recognition using a temporal CNN; operates on handcrafted features

Microcontroller (SRAM, Flash)	Features											
ISPU (8,32)	Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy
F446RE (128, 512)	Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy
L476RG (128, 1024)	Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy
F746ZG (320, 1024)	Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy
L4R5ZI_P (640, 2048)	Mean	IQR	Maximum	Median	Variance	MAD	Abs. Energy	Entropy	Peak-to-Peak	FFT mean coeff.	Fundamental Frequency	Abs. Energy

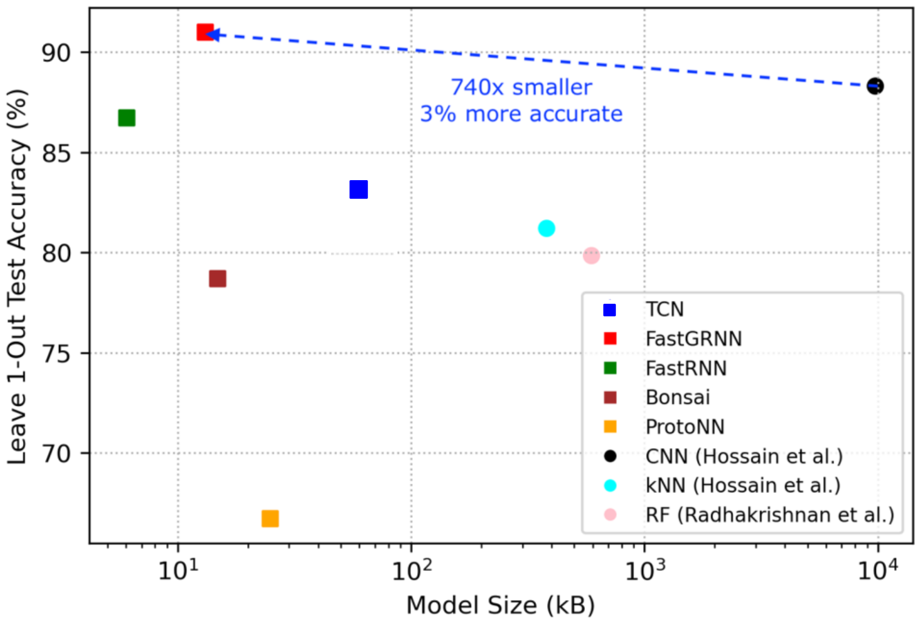


Extracting all features is computationally intensive. TinyNS picks the most important features when resources are scarce to maximize accuracy

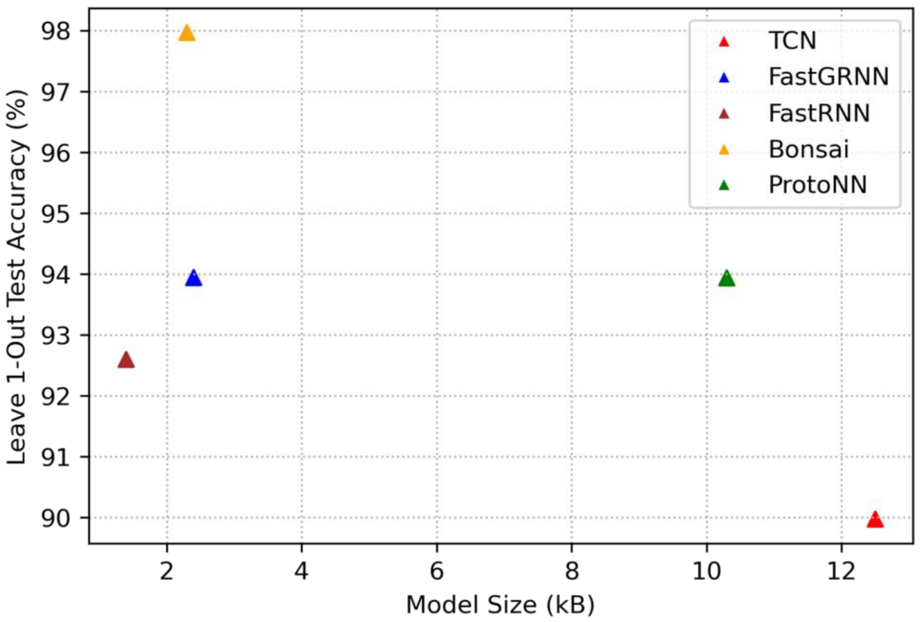
life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Example: optimizing over multiple backbones



Activity detection using earables



Fall detection using earables

98-740x smaller, 3-6% more accurate models for human activity detection using earables over baselines

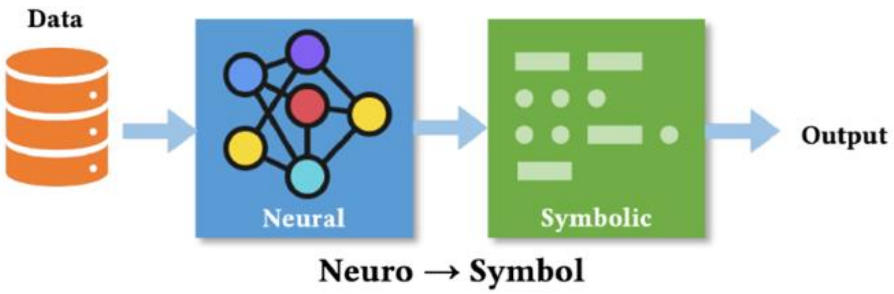
Activity detection under 6-12 KB of memory, fall detection under 2 KB of memory

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Neuro → symbol – problem formulation

A single ML model operates on the input data, followed by either a single domain-engineered function or a program graph



Given a collection of logical, relational, arithmetic, and conditional operators, program decision trees can be synthesized conditioned upon a finite tree count and depth

Tree enumeration algorithm to generate all possible paths to Decision A and B; or optimize parameters of a pre-defined tree

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Example: co-optimizing neural detector and symbolic tracker

Constraint	Flash Usage (MB)	Performance		Neural hyperparameters				Symbolic hyperparameters	
		MOTA	IDF	Kernel Size	Stacks	Head Convolution	Activations	Rendering	Confidence
None (handcrafted)	238	36.5	55.0	1	1	128	True	0.4	0.5
250 MB limit	238	36.1	54.6	1	1	150	True	0.3	0.4
500 MB limit	270	38.0	57.2	9	1	100	False	0.7	0.5

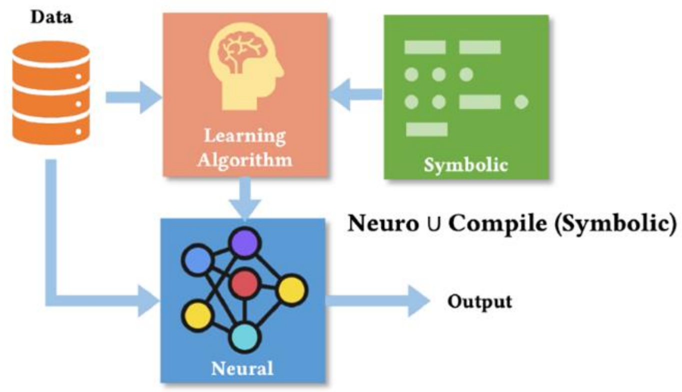
Achieves human-level performance (±1%) of program hand-tuned using hundreds of human hours in 3 GPU days

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Neuro ∪ compile (symbolic) – problem formulation

<u>Single ML model</u> operates on the data, while the symbolic rules are expressed in two ways



1. Add more regularizer terms in the NAS optimization function

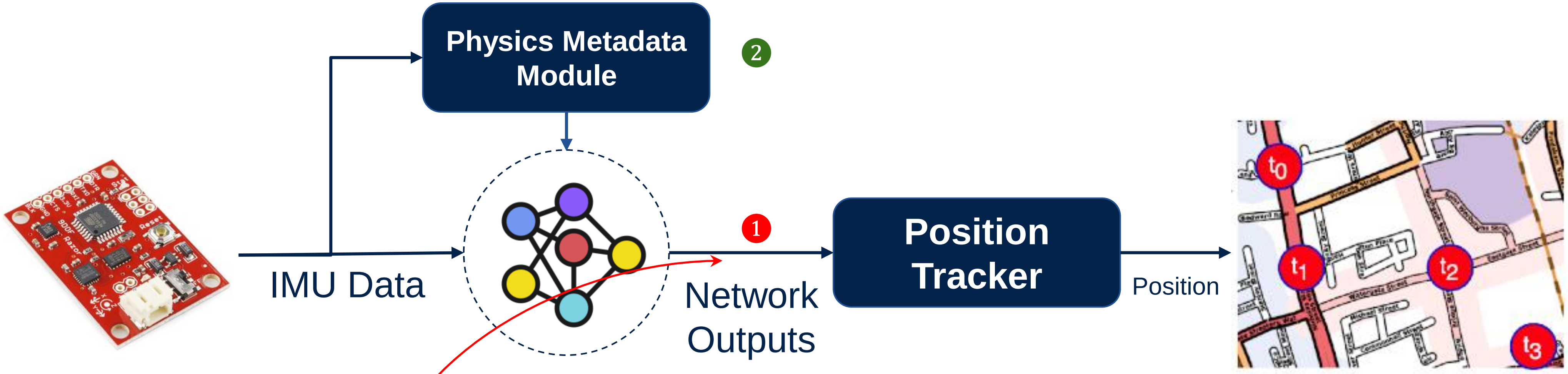
$$\min f_{\text{opt}}, \quad f_{\text{opt}} = \lambda_1 f_{\text{error}}(\Omega') + \lambda_2 f_{\text{flash}}(\Omega') + \lambda_3 f_{\text{SRAM}}(\Omega') + \lambda_4 f_{\text{latency}}(\Omega') + \boxed{\lambda_5 f_{\text{rule 1}}(\Omega')}$$

2. Add physics metadata channel as additional inputs to the model

life.augmented

Saha, Swapnil Sayan, Sandeep Singh Sandha, Mohit Aggarwal, Brian Wang, Liying Han, Julian de Gortari Briseno, and Mani Srivastava. "TinyNS: Platform-Aware Neurosymbolic Auto Tiny Machine Learning." *ACM Transactions on Embedded Computing Systems* (2023).

Example: physics-aware neural inertial navigation



$$\underbrace{\boxed{(v_{x,k}, v_{y,k})}}_{1} = y_{\theta^*}(\hat{\boldsymbol{a}}^I_{q:q+n}, \hat{\boldsymbol{w}}^I_{q:q+n}, \boxed{\hat{\boldsymbol{m}}^I_{q:q+n}}, \underbrace{\boxed{c_k(^I\hat{\boldsymbol{a}})}}_{2})$$

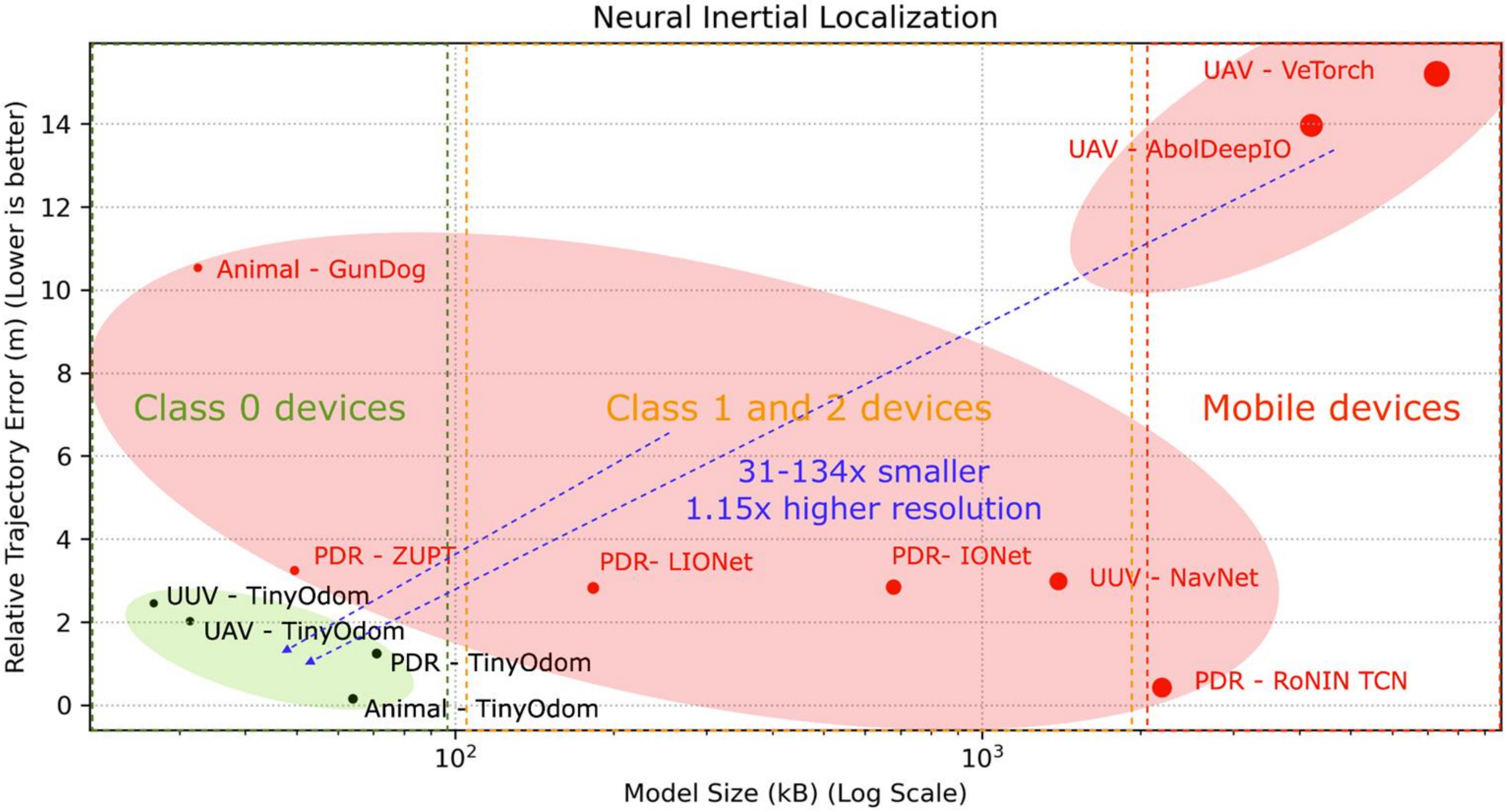
1 Velocity and magneto-centric DNN regresses velocities and uses magnetic North as an additional anchor point.

2 A physics metadata module supplies latent information about whether valid translational movements have occurred.

Saha, Swapnil Sayan, Sandeep Singh Sandha, Luis Antonio Garcia, and Mani Srivastava. "Tinyodom: Hardware-aware efficient neural inertial navigation." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, no. 2 (2022): 1-32.

Saha, Swapnil Sayan, Yayun Du, Sandeep Singh Sandha, Luis Antonio Garcia, Mohammad Khalid Jawed, and Mani Srivastava. "Inertial Navigation on Extremely Resource-Constrained Platforms: Methods, Opportunities and Challenges." In *2023 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 708-723. IEEE, 2023.

Example: physics-aware neural inertial navigation

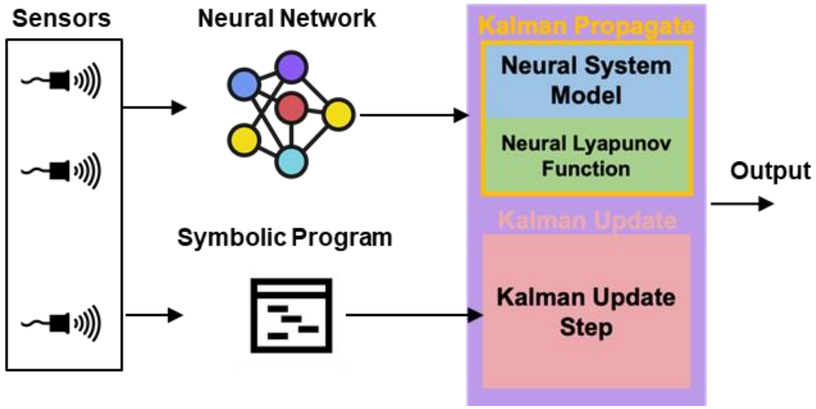


Saha, Swapnil Sayan, Sandeep Singh Sandha, Luis Antonio Garcia, and Mani Srivastava. "Tinyodom: Hardware-aware efficient neural inertial navigation." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, no. 2 (2022): 1-32.
Saha, Swapnil Sayan, Yayun Du, Sandeep Singh Sandha, Luis Antonio Garcia, Mohammad Khalid Jawed, and Mani Srivastava. "Inertial Navigation on Extremely Resource-Constrained Platforms: Methods, Opportunities and Challenges." In *2023 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 708-723. IEEE, 2023.

life.augmented

Symbolic [neuro] – problem formulation

Use Kalman filter theory to combine a noisy neural system model with noisy symbolic measurement updates



Separate neural and non-neural parts in Kalman propagate. Neural network provides a black box mapping

$$\hat{\mathbf{x}}_{k+1|k} = \boxed{\mathbf{A}}\hat{\mathbf{x}}_k + g(\mathbf{u}_{k+1}), \quad g(\cdot) = \boxed{f}(\boxed{y_\theta(\cdot)})$$

Linear state evolution (known)

Non-linear state evolution (known)

Use the linearized Jacobian of the neural network w.r.t the past state and inputs in the Lyapunov function

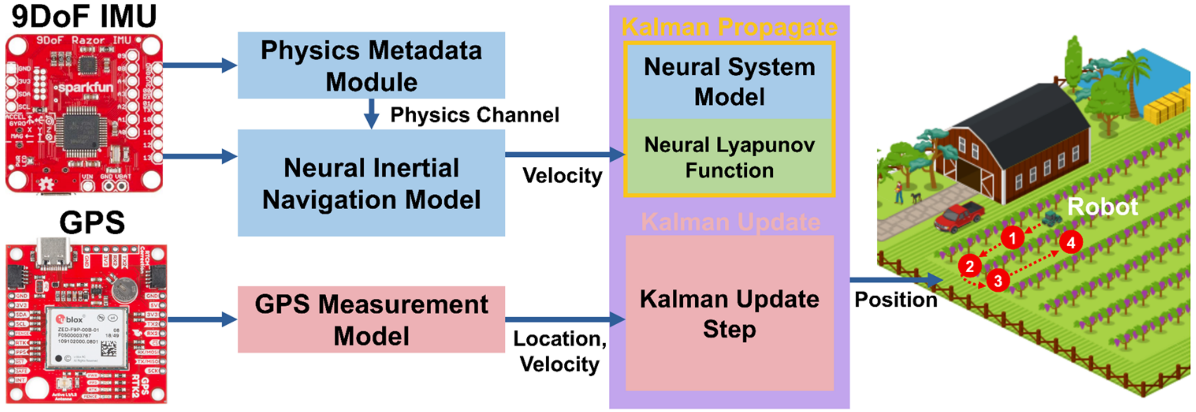
$$\mathbf{P}_{k+1|k} = \mathbf{A}\mathbf{P}_k\mathbf{A}^T + \mathbf{B}_{k+1}\boxed{\mathbf{U}_k}\mathbf{B}_{k+1}^T, \quad \boxed{\mathbf{B}_{k+1} = \frac{\partial g}{\partial u}\bigg|_{\hat{\mathbf{x}}_k,\mathbf{u}_{k+1}}}$$

Sensor Allan parameters

Jacobian term

Saha, Swapnil Sayan, Sandeep Singh Sandha, Luis Antonio Garcia, and Mani Srivastava. "Tinyodom: Hardware-aware efficient neural inertial navigation." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, no. 2 (2022): 1-32.
Saha, Swapnil Sayan, Yayun Du, Sandeep Singh Sandha, Luis Antonio Garcia, Mohammad Khalid Jawed, and Mani Srivastava. "Inertial Navigation on Extremely Resource-Constrained Platforms: Methods, Opportunities and Challenges." In *2023 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 708-723. IEEE, 2023.

Example: neural-Kalman filtering



Method (1 Hz GPS)	Median Absolute Trajectory Error (m)	Median Relative Trajectory Error (m)
UKF-M GPS/INS	4.35	0.21
EKF GPS/INS	2.24	0.35
GPS only	1.89	0.40
Neural-Kalman (ours)	**1.36**	**0.35**

Application: Tracking agricultural robots using neural inertial navigation and GPS; neural network provides a model-free evolution of the robot dynamics

Neural-Kalman filter combines smoothness and short-term accuracy of neural networks with long-term precision of noisy GPS/GNSS updates under 1 MB of memory

Saha, Swapnil Sayan, Sandeep Singh Sandha, Luis Antonio Garcia, and Mani Srivastava. "Tinyodom: Hardware-aware efficient neural inertial navigation." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, no. 2 (2022): 1-32.

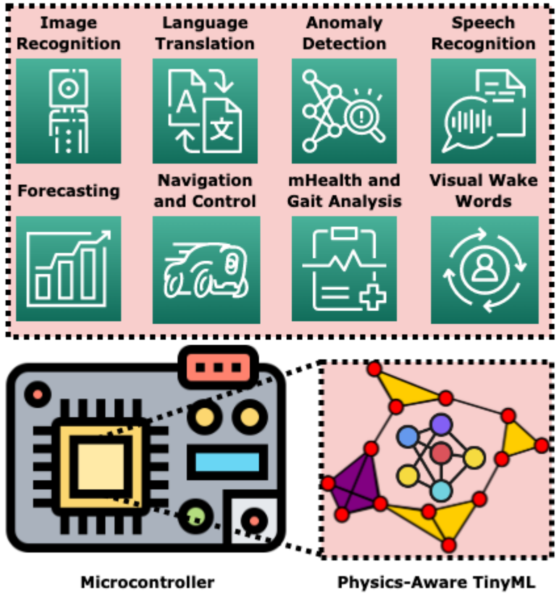
Saha, Swapnil Sayan, Yayun Du, Sandeep Singh Sandha, Luis Antonio Garcia, Mohammad Khalid Jawed, and Mani Srivastava. "Inertial Navigation on Extremely Resource-Constrained Platforms: Methods, Opportunities and Challenges." In *2023 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pp. 708-723. IEEE, 2023.

Conclusion

Neurosymbolic tiny machine learning enables context-aware, physics-aware, robust, interpretable, and performant edgeAI systems

TinyNS automates the process of generating neurosymbolic programs for TinyML platforms

Enables a broad spectrum of new applications for wearables, robots, automotives, and environmental sensors



life.augmented

Try TinyNS:
https://github.com/nesl/neurosymbolic-tinyml

Our technology starts with You

life.augmented

Copyright Notice

www.tinyml.org

Copyright Notice