Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Future Artificial Intelligence Research

Spoke 7 - Edge and Exascale AI

Giuseppe Averta, Barbara Caputo (PI)
Politecnico di Torino

Rome, Oct 20, 2023

Politecnico di Torino

Mainstream AI is inefficient

Energy

"Training GPT-3 would cost over $4.6M. […] With the increase in parameters, there's an exponential increase in energy."[1] Lambda Labs

Hardware

"The computational demands of AI presents an emerging problem for its implementation on different hardware platform"[2] Nature

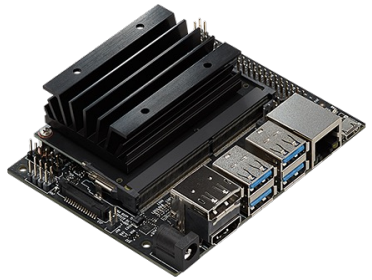
Data

"Data labeling takes anywhere from 35 to 80% of project budgets."[3] Forbes

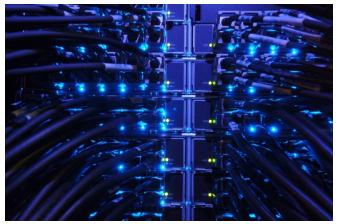
Environment

"Training a single AI model can emit as much carbon as five cars in their lifetimes"[4] MIT Technology Review

[1]Source: https://lambdalabs.com/blog/demystifying-gpt-3
[2]Source: https://www.nature.com/articles/s41928-018-0068-2
[3]Source: https://www.forbes.com/sites/cognitiveworld/2022/08/06/no-youre-not-alone-google-is-also-making-this-big-mistake-on-ai
[4]Source: https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/



Edge and Exascale AI



Tiny AI

Parallel AI

Foundation

WP 7.1 | WP 7.3

- ✓ Higher energy efficiency
- ✓ Environment friendly
- ✓ Lower hardware requirement
- ✓ Faster training and inference

WP 7.2 | WP 7.3

Downstream tasks

WP 7.4

WP 7.5

Implication

WP 7.6

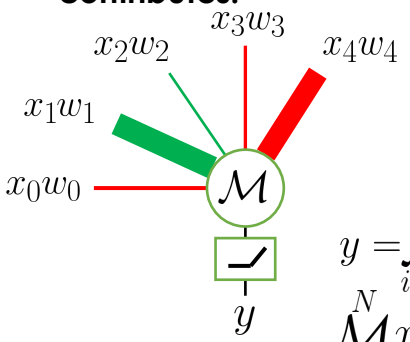
Spotlight works

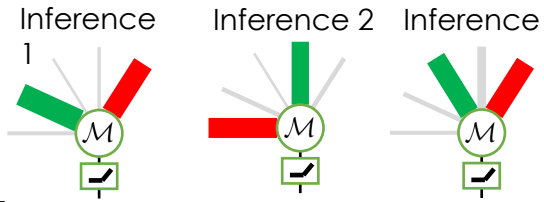
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

We study how to better build neural models

The Multiply and Max/Min (MAM) –based neuron

MAM-based neurons do not accumulate all the weighted inputs, but **sum together only the maximum and minimum contributes.**



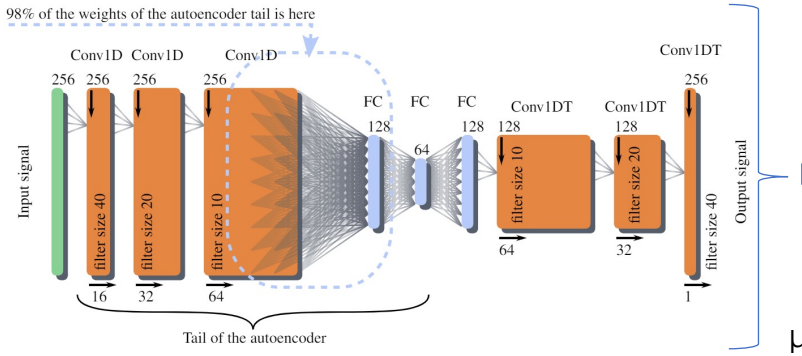


$$y = \overset{N}{\underset{i=1}{\mathcal{M}}} x_i w_i + b$$

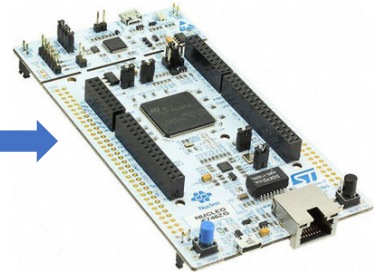
$$\overset{N}{\underset{i=1}{\mathcal{M}}} x_i w_i \triangleq \max_{i \in \{1,..,N\}} x_i w_i + \min_{i \in \{1,..,N\}} x_i w_i$$

Neural networks built with MAM learn to use **only a small subset of interconnections during inference.**

Case study: MAM for ECG autoencoder tail

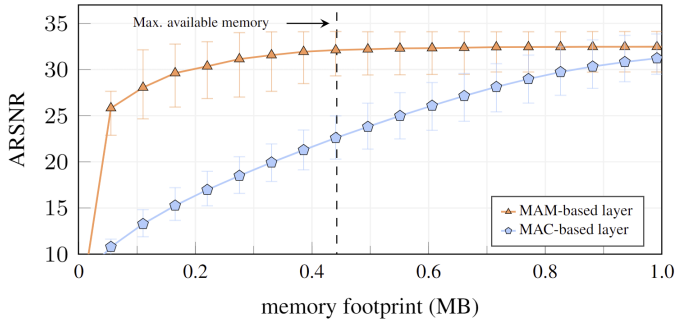


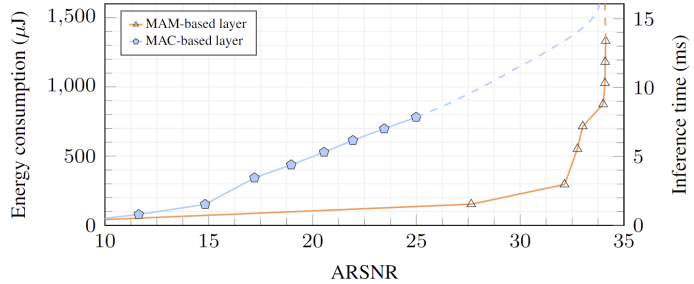
STM32F767ZI



µC unit (512 kB RAM, 216 MHz Clock)

The FC layer is pruned by removing the parameters with the lowest magnitude.



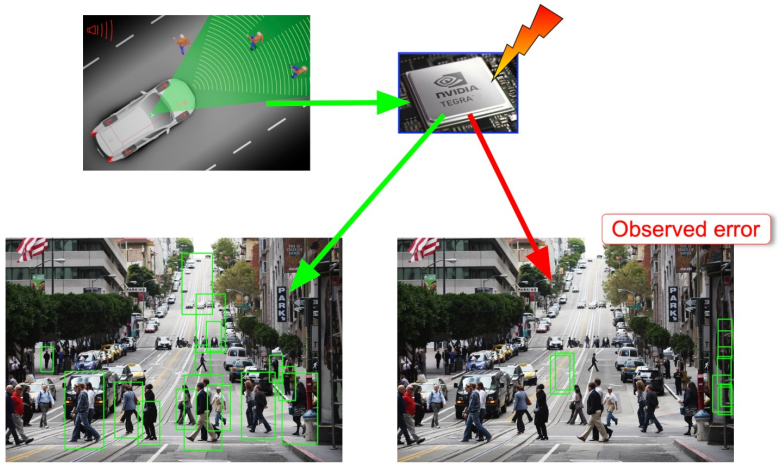


By using all the available memory on device, MAM achieves 33 dB reconstruction performance (where a standard DNN would achieve 18 dB)

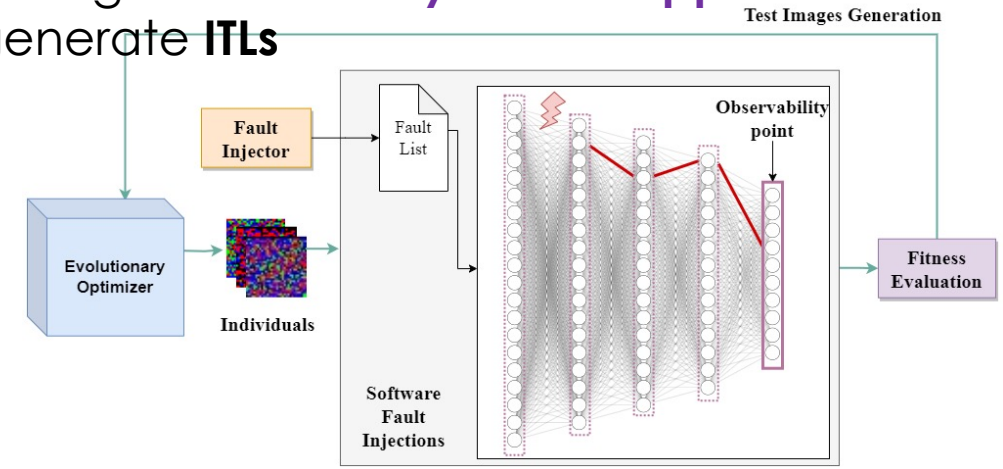
Prono, Luciano; Bich, Philippe; Mangia, Mauro; Pareschi, Fabio; Rovatti, Riccardo; Setti, Gianluca (2023). A Multiply-And-Max/min Neuron Paradigm for Aggressively Prunable Deep Neural Networks. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.22561567.v1

We assess the reliability of existing architectures

Robustness w.r.t. external disturbance (e.g. neutron strikes)



Robustness w.r.t. permanent hardware faults through **Evolutionary-based approach** to generate **ITLs**



Robustness w.r.t. information representation (e.g. POSIT)

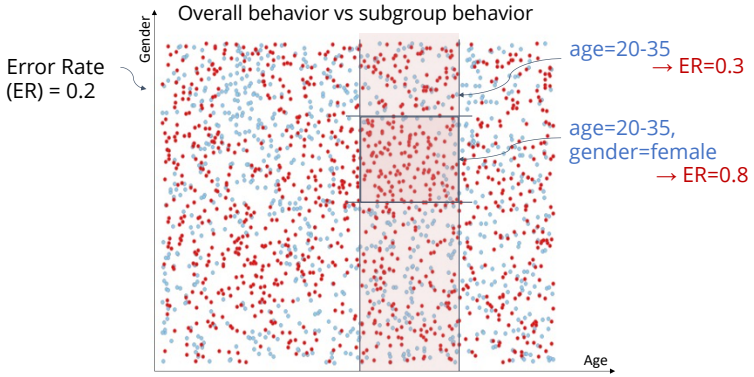

IEEE Spectrum Posits, a New Kind of Number, Improves the Math of AI

Posits, a New Kind of Number, Improves the Math of AI › The first posit-based processor core gave a ten-thousandfold accuracy boost

- V.Turco, A. Ruospo, G. Gavarini, E. Sanchez, M. Sonza Reorda, "*Uncovering hidden vulnerabilities in CNNs through evolutionary-based Image Test Libraries*" IEEE Int. Symp on DFT 2023
- Cavagnero, Niccolò, et al. "*Transient-Fault-Aware Design and Training to Enhance DNNs Reliability with Zero-Overhead.*" IEEE IOLTS 2022.

And study how and what models are learning

amazon alexa



Divergence as a measure of anomalous behavior of a data subgroup S w.r.t. overall dataset D for a function f

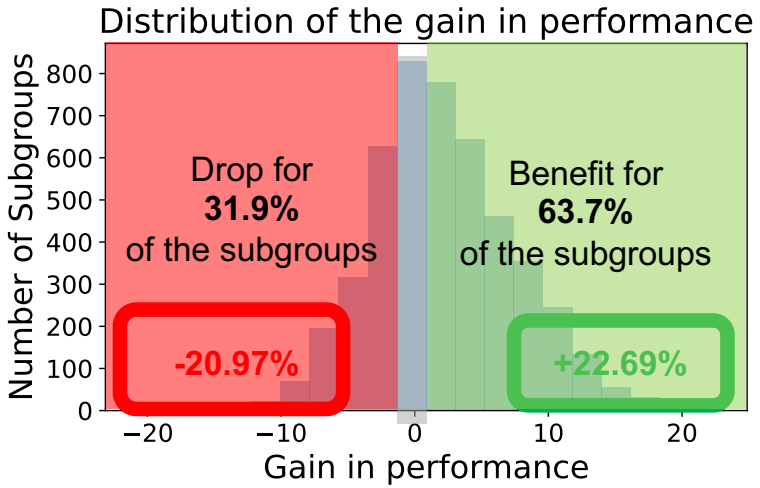
$$\Delta(S) = f(S) - f(D)$$

Subgroup	*Sup*	*acc*	Δ_{acc}	t
{*age=22-40, gender=male, loc=none, speakRate=high, tot_silence=high*}	0.03	74.79	-18.38	4.7
{*action=increase, gender=male, speakRate=high*}	0.03	74.81	-18.36	4.9

Miss-classification (in red) rate of **speech models** (intent classif.) also depend on subgroups – **how to detect low-accuracy subgroups without supervision?**

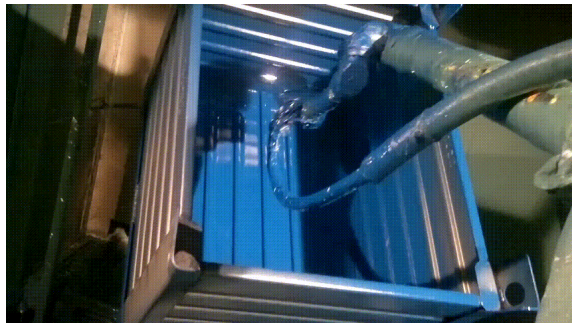
Subgroup gain as the difference in performance of two models M_1, M_2 on a specific subgroup S for f

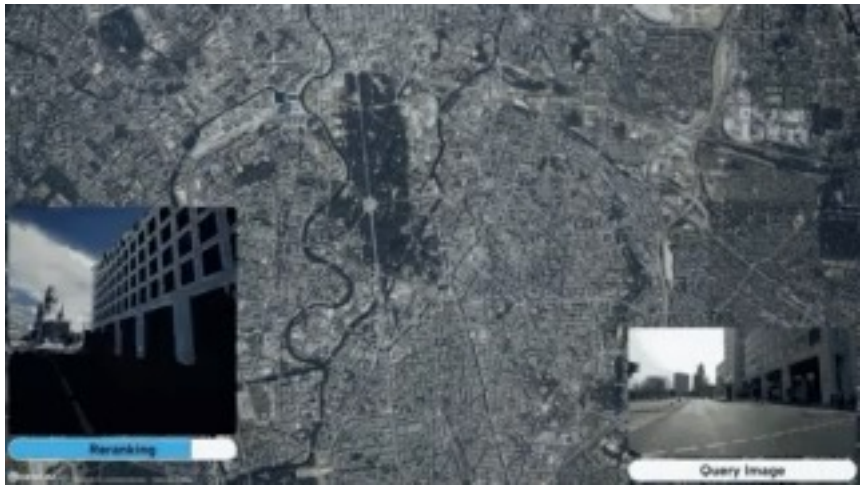
$$gain_f(S, M_1, M_2) = f(S, M_2) - f(S, M_1)$$

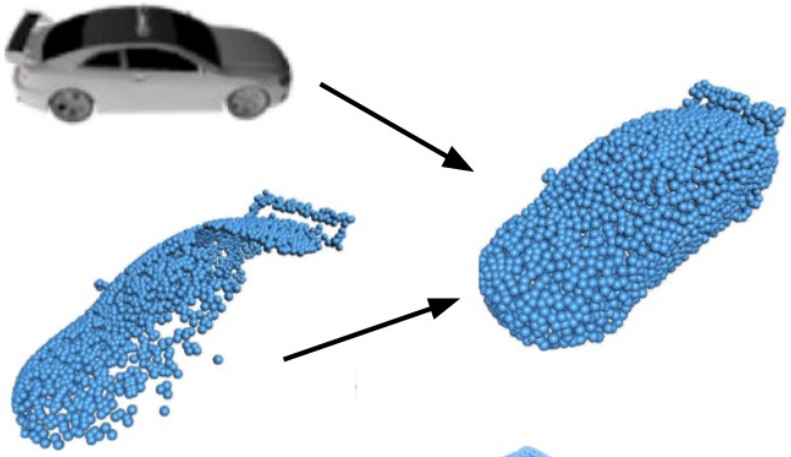


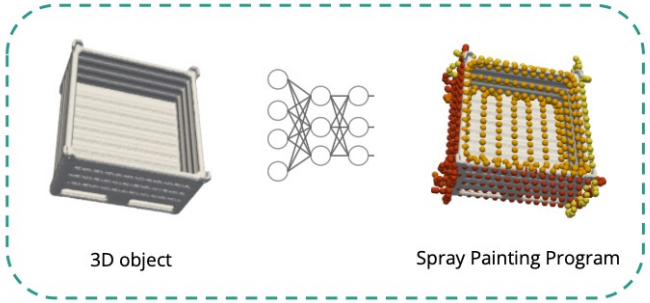
Koudounas, Alkis, et al. "*Exploring subgroup performance in end-to-end speech models.*" ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

We care about applications:
Efficient 2D and 3D computer vision on edge and very large scale





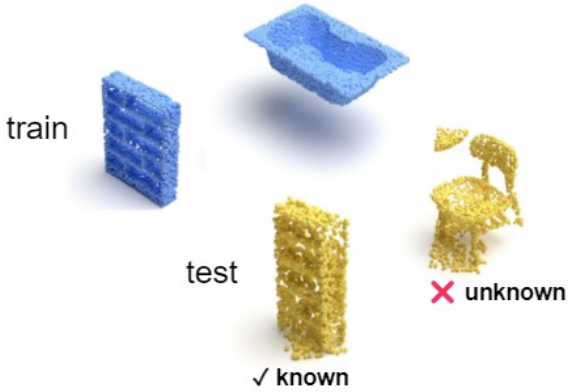




Execution

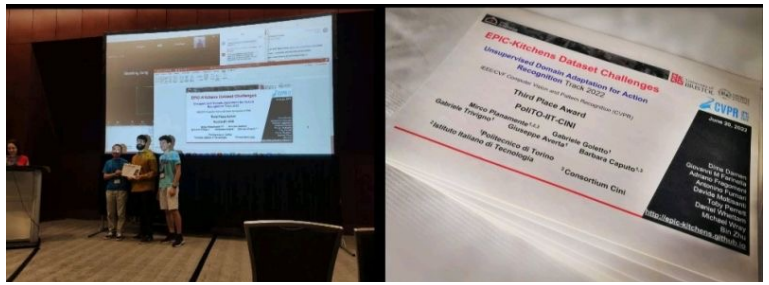
Painted Object

Our task

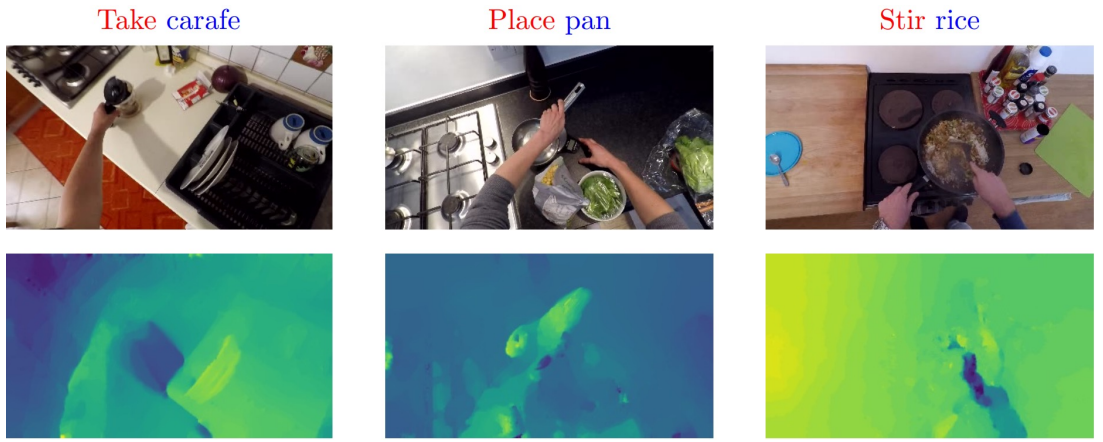


We care about applications: Understanding human behavior from egocentric data

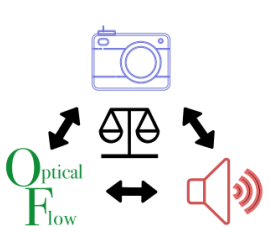
- **Mobile models for first person action recognition**
- Challenges:
  - Untrimmed videos
  - Robust to Multi-modal domain gap (time, space)
  - Model footprint for edge deployment



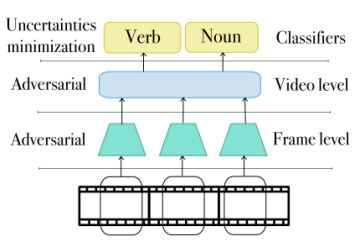
3rd place EPIC KITCHENS UDA challenge at CVPR2022 & CVPR2023



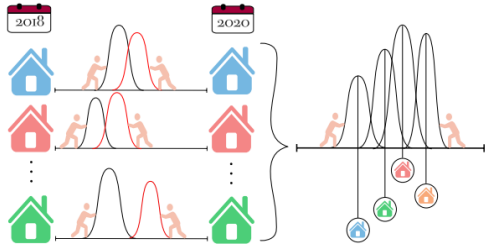
Domain Generalization → Unsupervised Domain Adaptation → Multi-Source Multi-Target Domain Adaptation → Ensemble Domain Adaptation



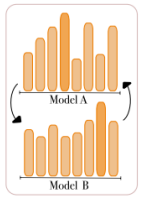
RNA-Net



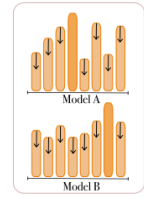
Multi Level Adversarial Alignment



Multiple Spatial-Temporal Adversarial Alignment (MSTAA)



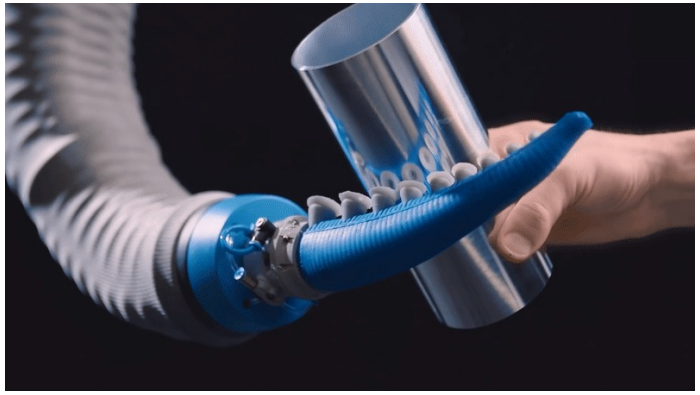
Min Entropy Consensus (MEC loss)

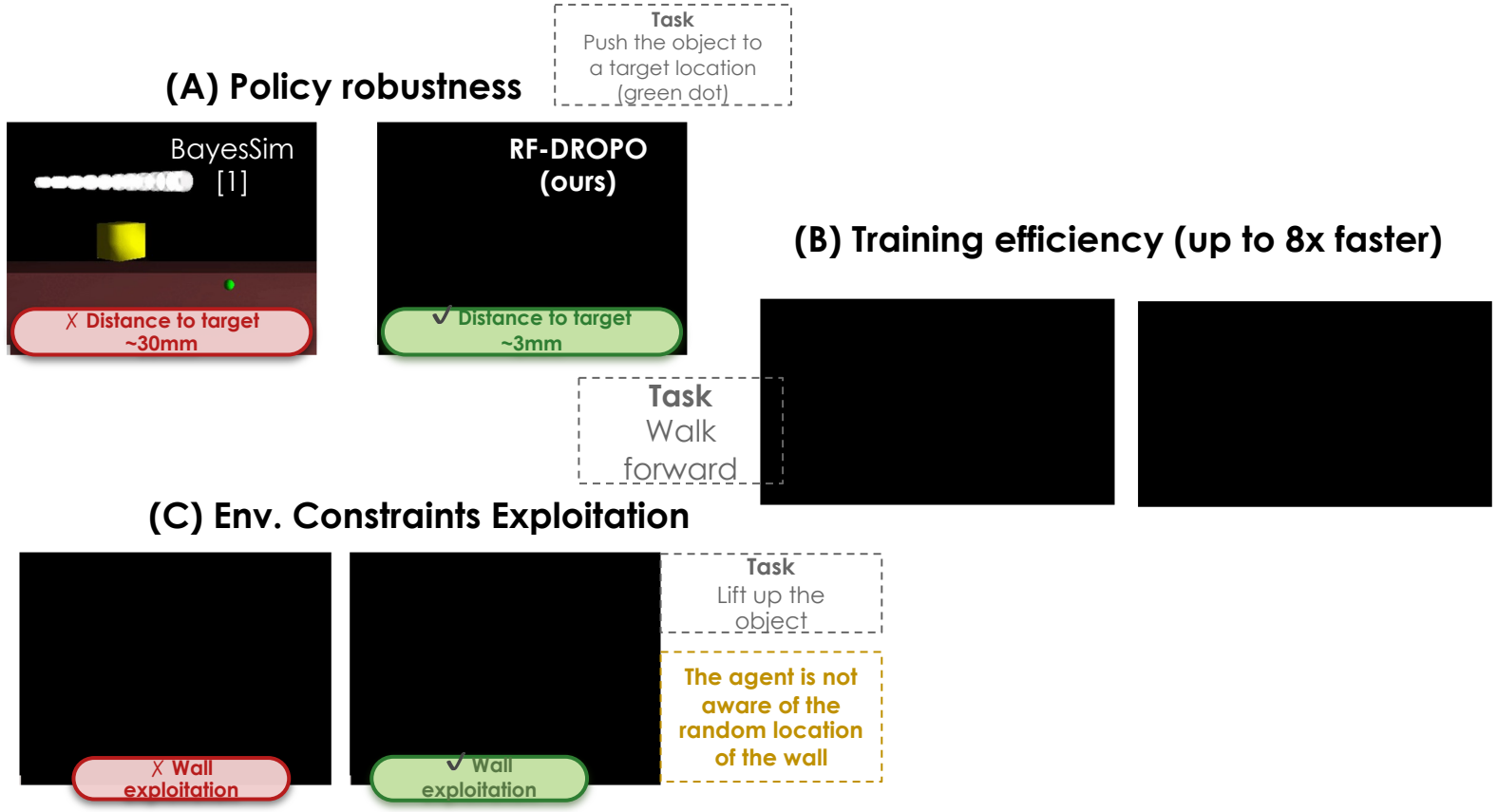


Complement Entropy (CENT loss)

Goletto, G., et al. 2023 *RA-L*,
Peirone et al. 2023, *under review*

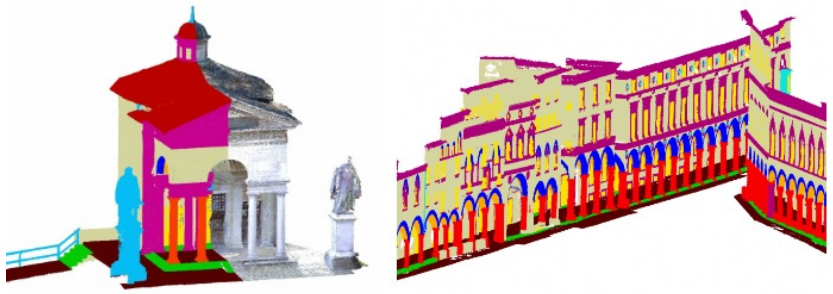
Not only “supervised”: RL for (soft) intelligent manipulators





Tiboni, G., Protopapa, A., Tommasi, T., & Averta, G. (2023). *Domain Randomization for Robust, Affordable and Effective Closed-loop Control of Soft Robots*. In 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

AI and its societal impact



Semantic segmentation of cultural heritage point clouds

Dependence	SMALL
	Range [0,1]
Contingency coefficient	0.1413
Effect size	0.1427

Diverseness	
	Probability
Target variable	Range [0,1]
0	0.545
1	0.455
Protected attribute	
Asian	0.005
Black	0.514
Caucasian	0.341

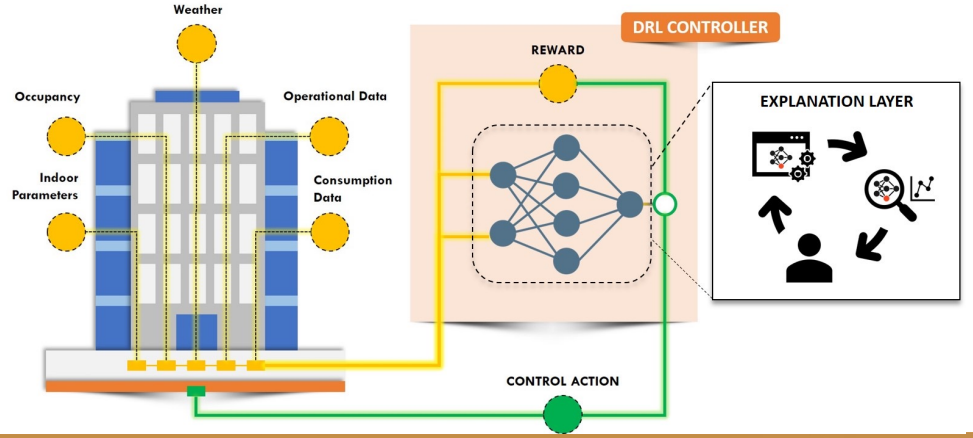
Inclusiveness	
	Probability
	Range [0,1]
P(Asian ∩ 0)	0.0023
P(Asian ∩ 1)	0.0008
P(Black ∩ 0)	0.1514
P(Black ∩ 1)	0.1661
P(Caucasian ∩ 0)	0.1281
P(Caucasian ∩ 1)	0.0822
P(Hispanic ∩ 0)	0.0320
P(Hispanic ∩ 1)	0.0189
P(Native american ∩ 0)	0.0006
P(Native american ∩ 0)	0.0005
P(other ∩ 0)	0.0219
P(other ∩ 1)	0.0124

Training likelihood	
	Probability
	Range [0,1]
P(Caucasian | 1)	0.293
P(Caucasian | 0)	0.381
P(0 | Caucasian)	0.609
P(1 | Caucasian)	0.391
P(Black | 1)	0.591
P(Black | 0)	0.450
P(0 | Black)	0.477
P(1 | Black)	0.523

Ethically-sensitive dataset labeling

Predictive maintenance on district heating networks

Balancing the complexity and interpretability of AI-based energy management strategies



Spoke 7 in Numbers

- 30+ professors
- 8 Assistant Professors enrolled (more coming soon)
- 4 PhD students enrolled (more coming soon)
- 6 research assistants (all on board)
- 9+ journal papers with peer review
- 40+ conference papers with peer review

Dissemination activity

IROS 2023, Detroit, USA

ICIAP 2023, Udine, Italy

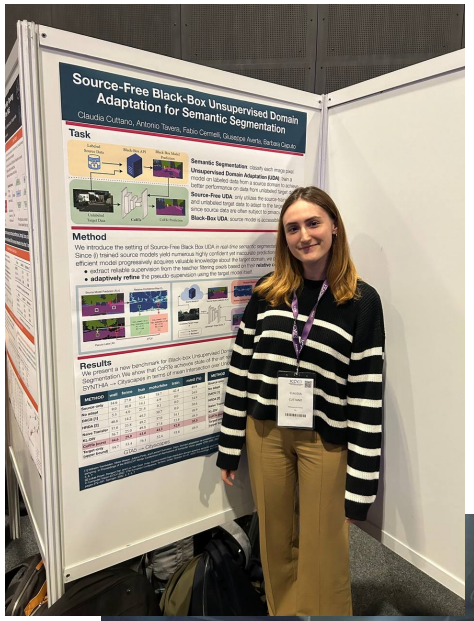
IEEE TPTR, Anaheim, USA

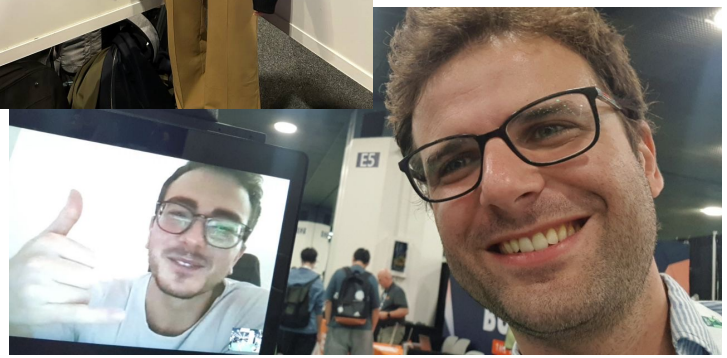
ECML-PKDD 2023, Turin, Italy

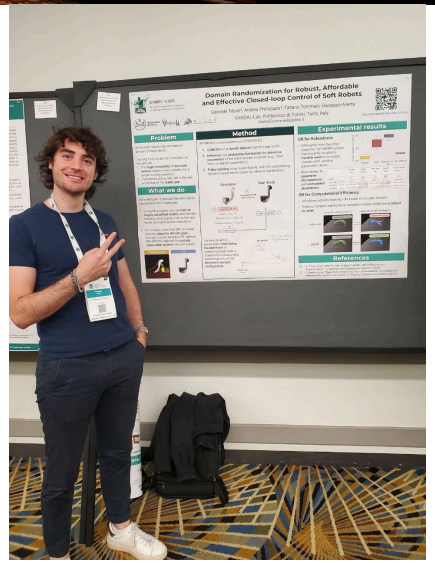
ECML-PKDD 2023, Turin, Italy

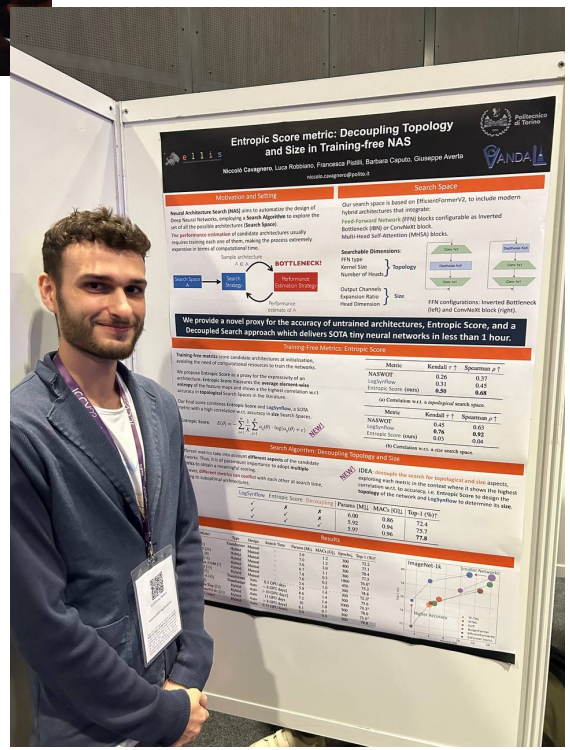
AUTOMATICA 2023, Catania, Italy

Spoke 7 in the world









Not only rocket science

- We do have extensive experience in academia-industry collaborations
- Many already existing collaborations
- Keen to use FAIR as flywheel to enhance current relationships and open new opportunities

life.augmented

INTESA SANPAOLO

LEONARDO

Thanks

- Happy to chat in the networking session
- Come and visit us in Turin!

- giuseppe.averta@polito.it
- barbara.caputo@polito.it

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.