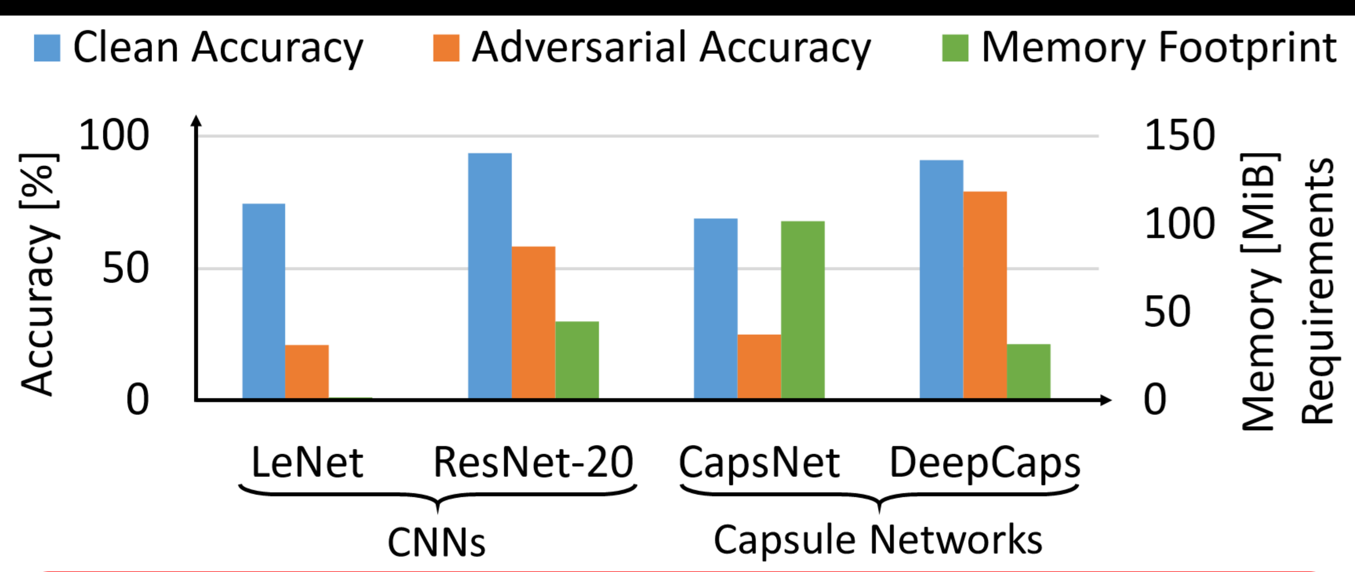
HARNAS: Neural Architecture Search Jointly Optimizing for Hardware Efficiency and Adversarial Robustness of Convolutional and Capsule Networks

Alberto Marchisio[1], Vojtech Mrazek[2], Andrea Massa[3], Beatrice Bussolino[3], Maurizio Martina[3], Muhammad Shafique[4]

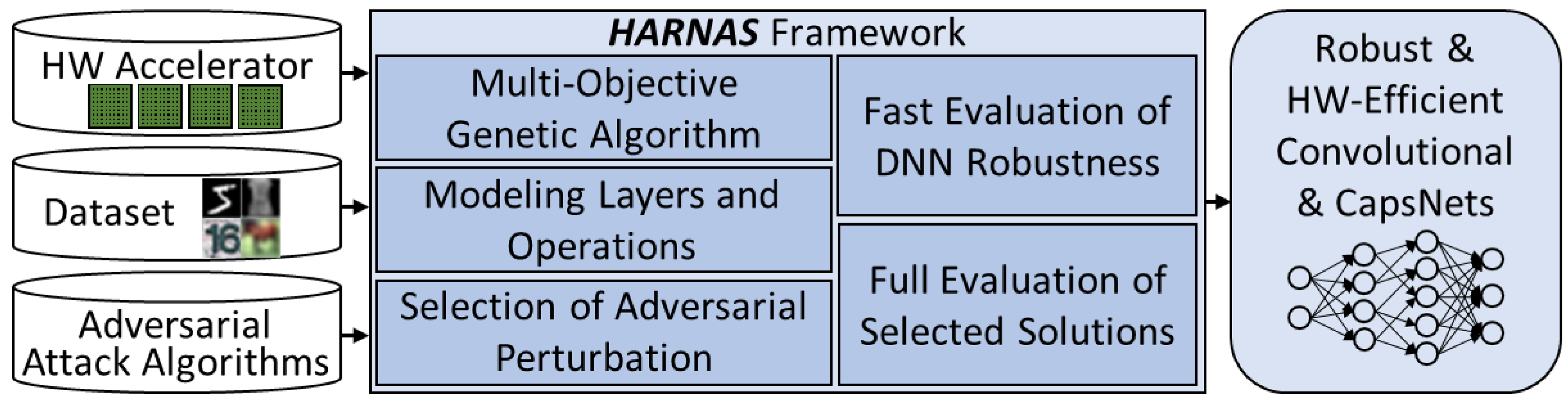
[1]Institute of Computer Engineering, Technische Universität Wien (TU Wien), Vienna, Austria
[2]Faculty of Information Technology, Brno University of Technology, Brno, Czechia
[3]Department of Electronics and Telecommunications, Politecnico di Torino, Turin, Italy
[4]eBrain Lab, Division of Engineering, New York University Abu Dhabi, United Arab Emirates

Motivations and Novel Contributions



Motivation Analysis

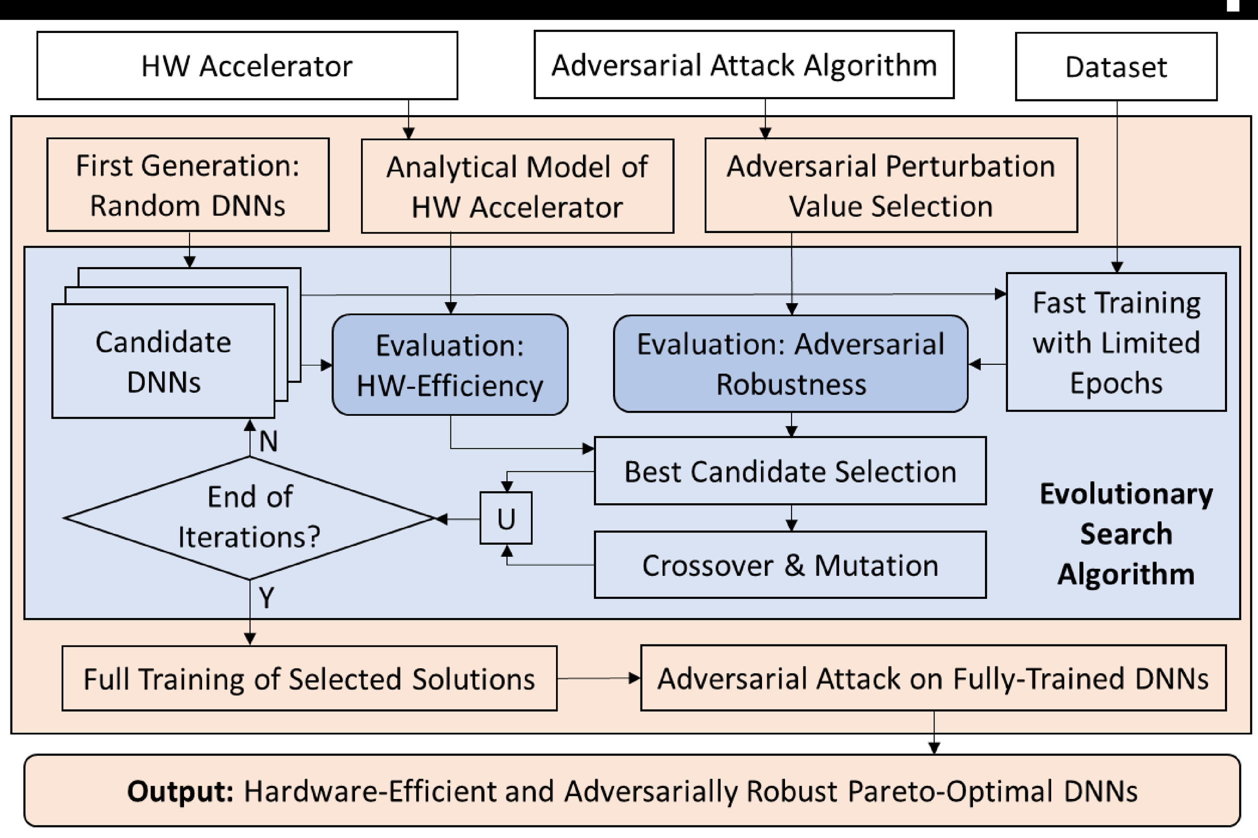
- ❑ The LeNet, which is relatively small and shallow, is hardware efficient due to its low memory footprint, but relatively more vulnerable to attacks.
- ❑ A more complex DNN such as the ResNet-20 has a higher memory footprint but it also exhibits higher adversarial accuracy than the LeNet.
- ❑ The DeepCaps, despite having a smaller memory footprint than the ResNet-20, is also relatively more robust against adversarial attacks.



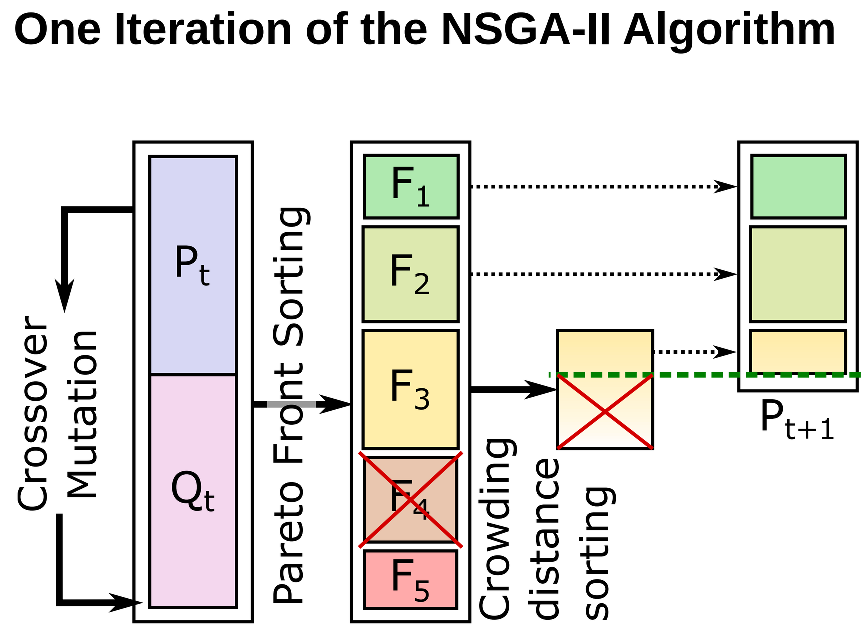
Our Novel Contributions

- ❑ Analytical models of DNN and CapsNet layers and operations for architectural flexibility and fast hardware estimation.
- ❑ Analysis and selection of the adversarial perturbations values to employ in the NAS for a fast robustness evaluation.
- ❑ Specialized evolutionary algorithm, based on the principles of the NSGA-II method, to perform a multi-objective Pareto-frontier selection, with conjoint optimization for adversarial robustness, energy, memory, and latency of DNNs.
- ❑ Fast evaluation methodology for DNNs trained for a limited number of epochs to reduce the training time.
- ❑ Full-training evaluation of the Pareto-optimal solutions to obtain the exact results.

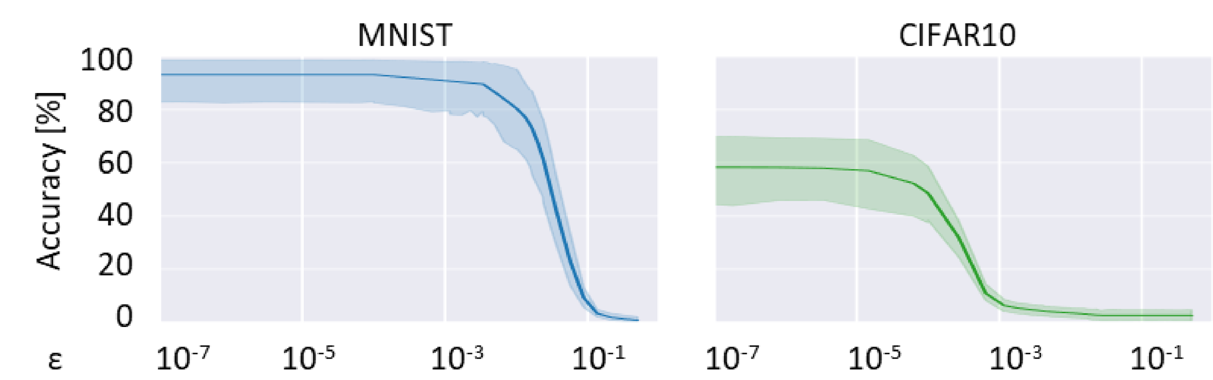
Proposed HARNAS Framework



One Iteration of the NSGA-II Algorithm



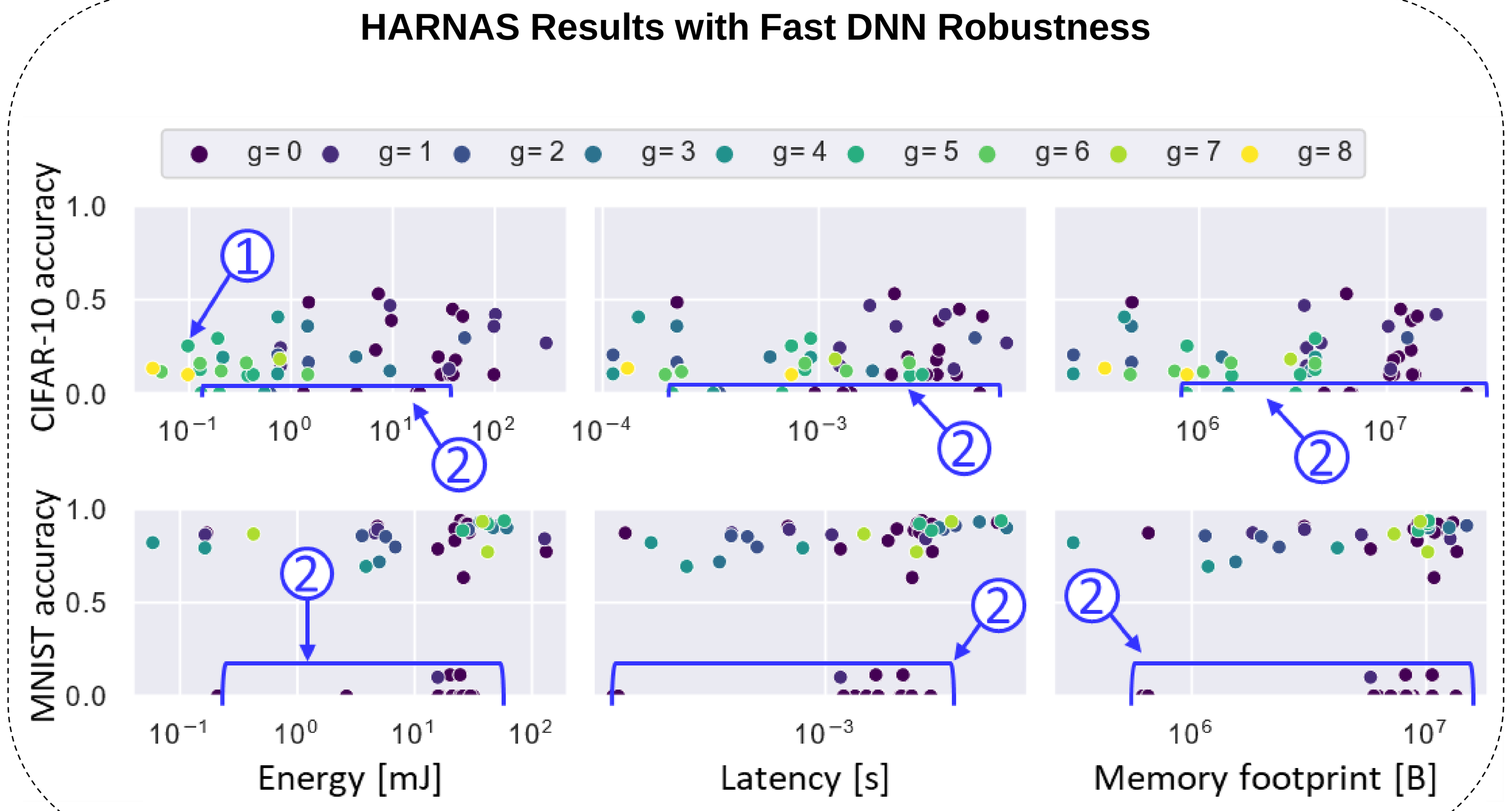
Selection of Adversarial Perturbation for the NAS



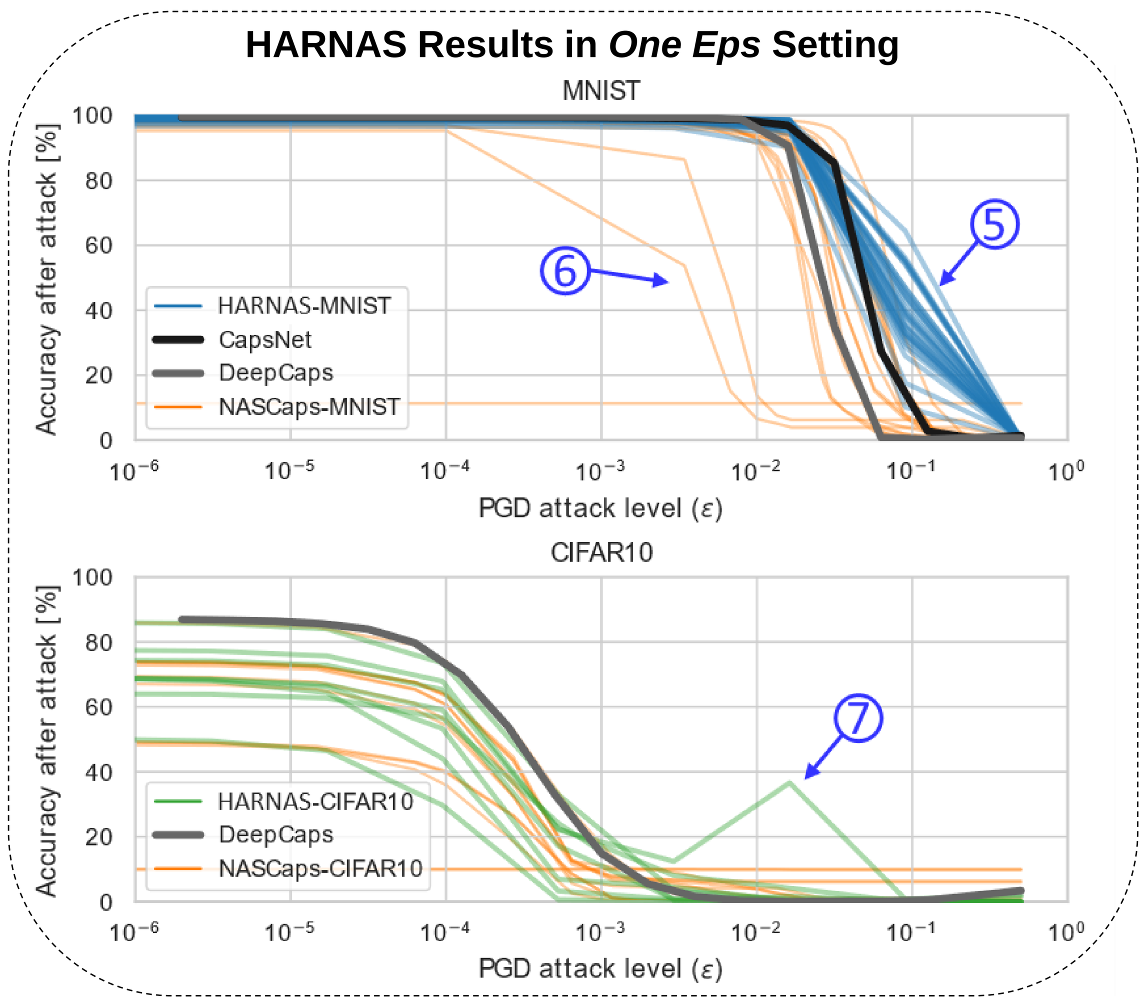
	Two EPS ε_{low}	One EPS ε	Two EPS ε_{high}
MNIST	3e-3	3e-2	1e-1
CIFAR10	3e-5	3e-4	1e-3

Evaluation and Related Work Comparison

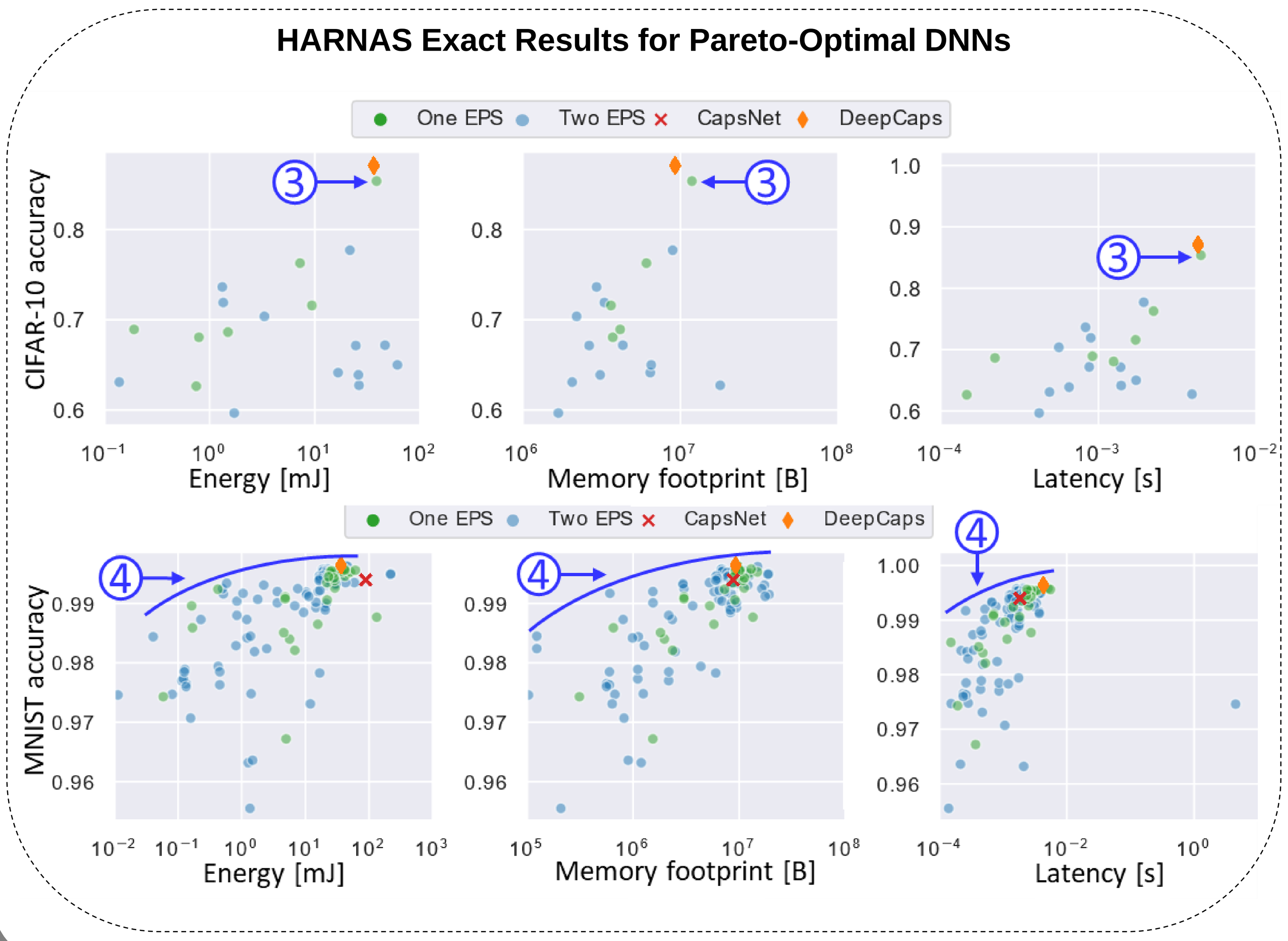
HARNAS Results with Fast DNN Robustness



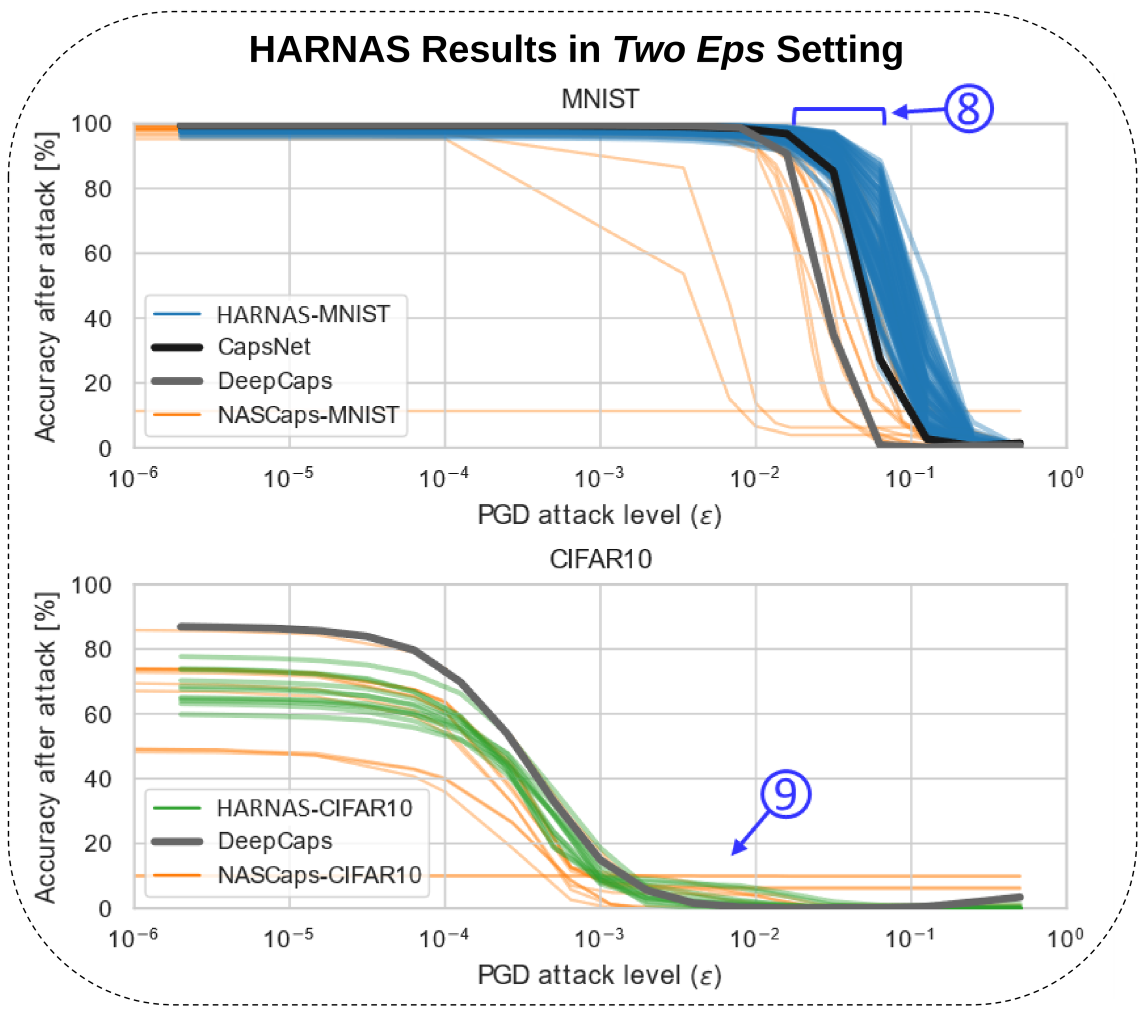
HARNAS Results in *One Eps* Setting



HARNAS Exact Results for Pareto-Optimal DNNs



HARNAS Results in *Two Eps* Setting



Key Observations

1) For the HARNAS evaluated on the CIFAR10 dataset, the latest generations find DNNs that are less robust to the PGD attack, but still belong to the Pareto-frontier due to the low energy consumption.
2) Several candidate DNNs found in the earliest generations are automatically discarded by the Pareto-frontier selection, since they are highly vulnerable to the PGD attack.
3) A Pareto-optimal solution found by the HARNAS framework for the CIFAR10 dataset achieves 86.07% accuracy while having an energy consumption of 38.63 mJ, a memory footprint of 11.85 MiB, and a latency of 4.47 ms.
4) The Pareto-optimal DNN search for MNIST covers a wider range of values, leveraging tradeoffs between different objectives.
5) For the MNIST dataset, the Pareto-optimal solutions obtained with the HARNAS framework are particularly robust for a high range of perturbation ε.
6) The accuracy starts dropping at around one order of magnitude higher ε than NASCaps.
7) For the CIFAR10 dataset, the HARNAS DNNs' behavior is similar to the DeepCaps for low values of ε, while a Pareto-optimal HARNAS solution offer a respectable robustness also with higher adversarial perturbation.
8) The HARNAS framework with the Two EPS setting, compared to the One EPS setting, produces different levels of robustness w.r.t. ε for the MNIST dataset.
9) For the CIFAR10 dataset, the Two EPS search leads to worse results than the One EPS counterpart.

This work has been supported in part by the Doctoral College Resilient Embedded Systems, which is run jointly by the TU Wien's Faculty of Informatics and the UAS Technikum Wien. This work was also supported in part by Czech Science Foundation grant GA22-02067S. The computational resources were supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140).

ICML Workshop on Dynamic Neural Networks
Baltimore, MD, USA, July 22, 2022