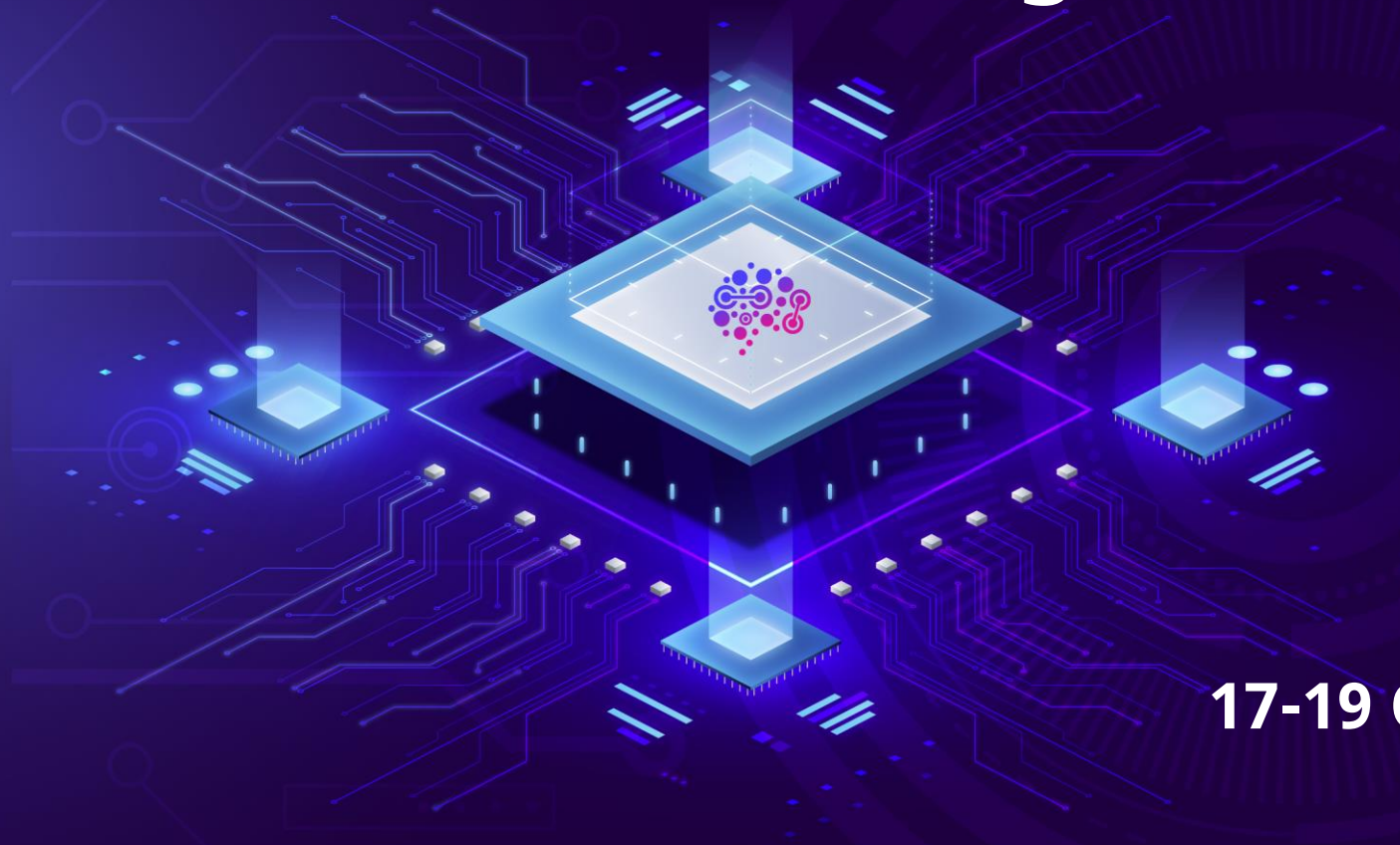
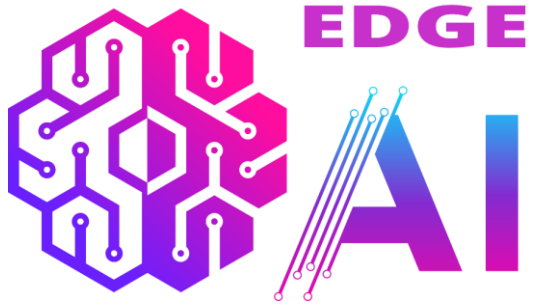
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



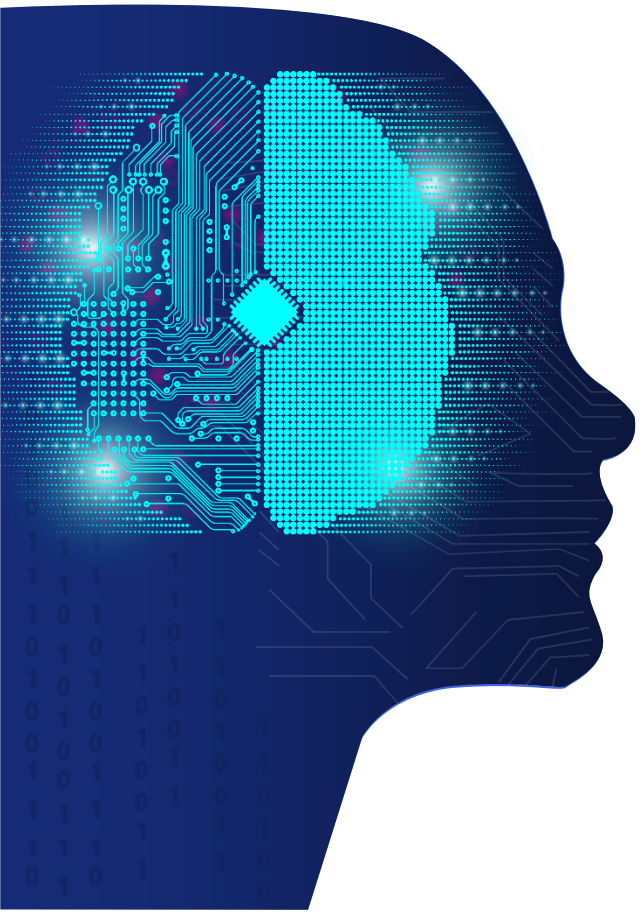
Trade-off Analysis of Pruning Methods for Compact Neural Networks on Embedded Devices

Presenter Duc V. Le, University of Twente, The Netherlands

17-19 October 2023 Athens, Greece

Copyright © 2023

Presentation Outline

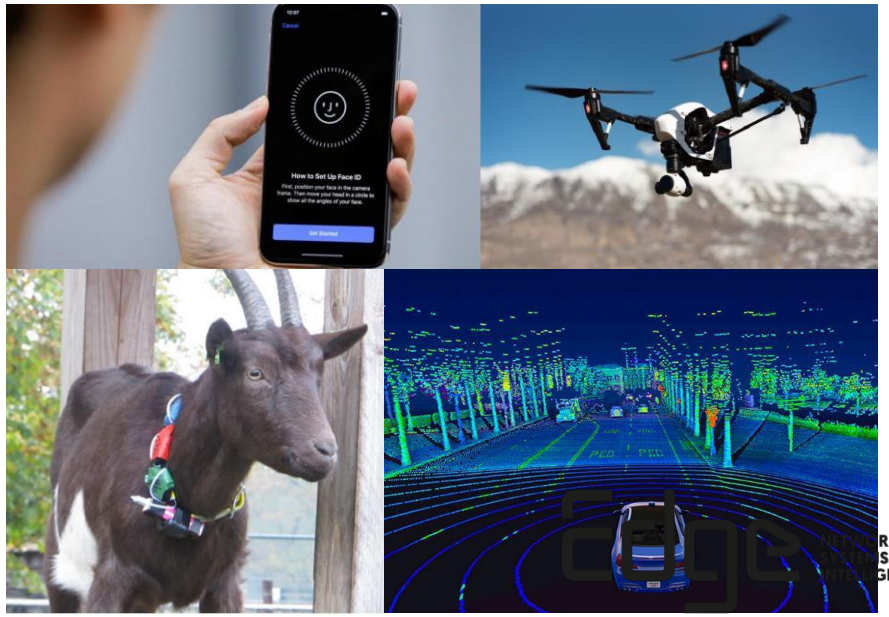


- Introduction
- State-of-the-art
- Research Goal
- Methodology
- Design and Implementation
- Experimental Results
- Future Research
- Conclusion

Copyright © 2023

Introduction

- Increasing demanding of AI on embedded devices
- Growing AI models' size



How To Lower Resource Usage?

- Model pruning techniques

In this work:

- Optimized neural network such as MobileNetV2 still can be 5 times smaller "without" sacrificing the accuracy
- Model size and power consumption are not linearly correlated.

Copyright © 2023

State-of-the-art

SoTA on Edge Deep Learning

- Neural Network Architecture Search (NAS) and autoML
- Neural Network Architecture Design: MobileNet, SqueezzNet
- Quantization: Tensorflow Lite
- Pruning: structured and unstructured

SoTA on comparision pruning techniques lack:

- pruning optimized neural networks such as MobileNet
- analyzing trade-off of power consumption and accuracy

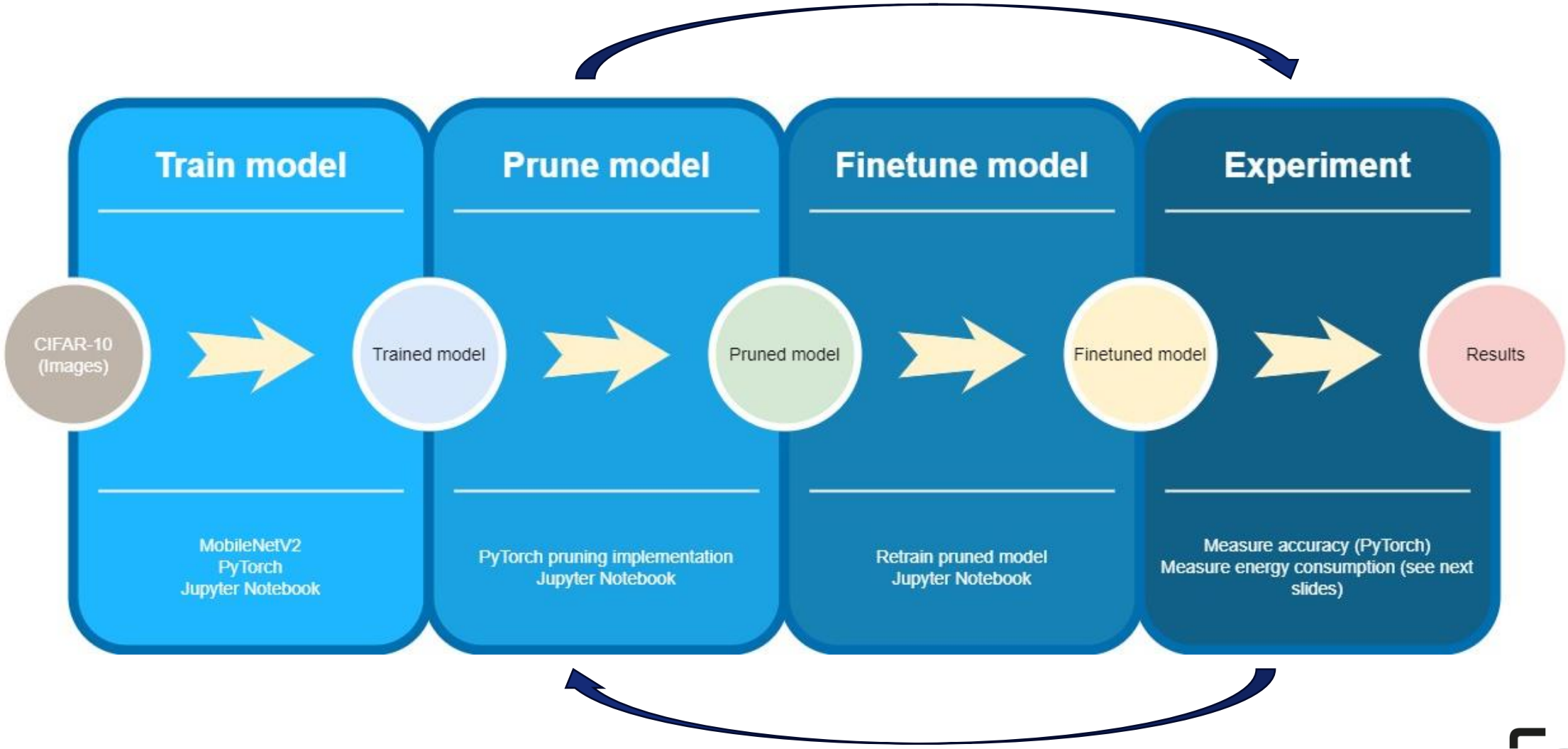
Copyright © 2023

Research Goal

- What would be the impact of applying pruning techniques on ready-optimized neural networks for embedded devices?
- What is the trade-off between accuracy and model size/power consumption?

Copyright © 2023

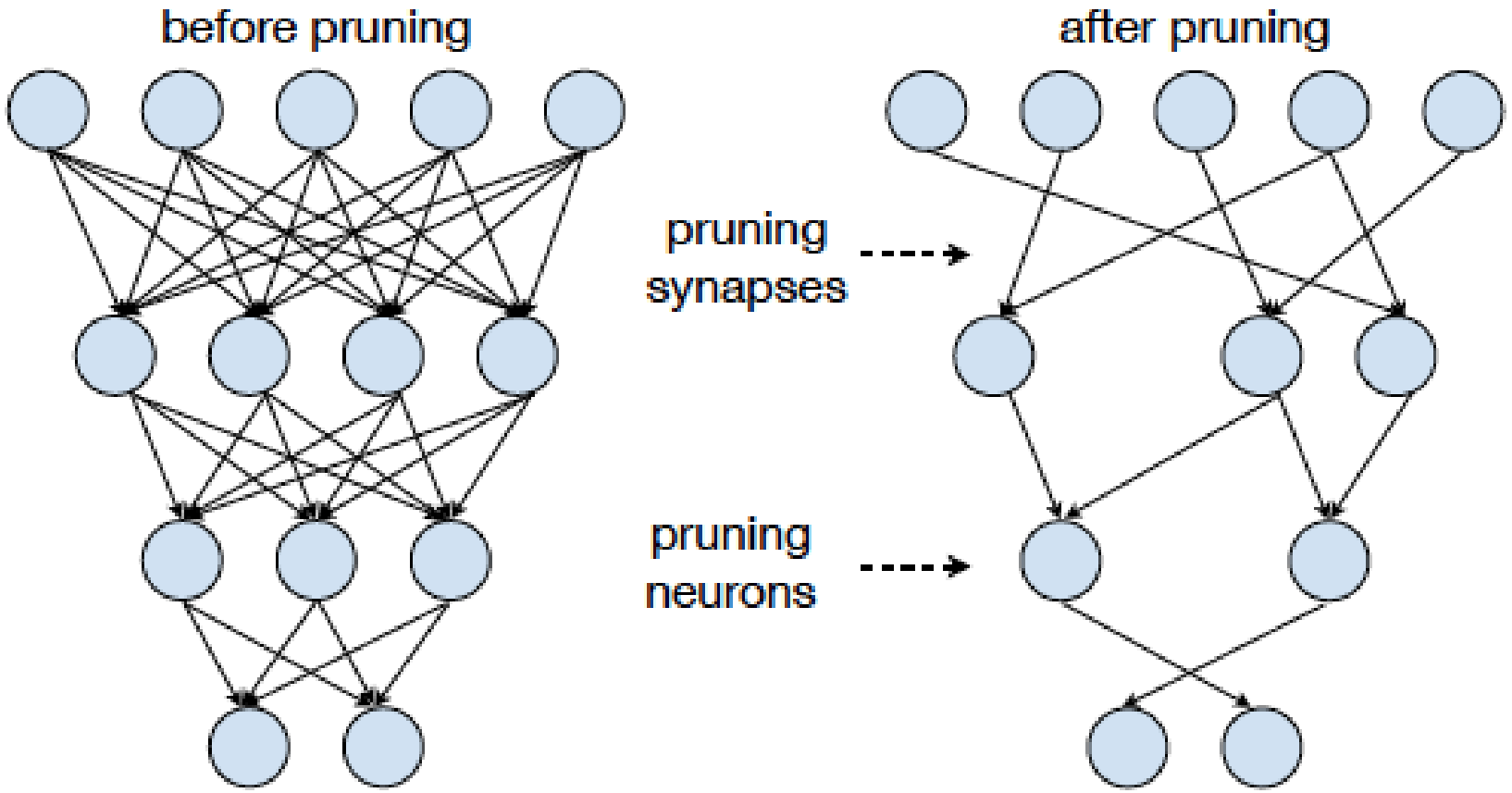
Methodology



Copyright © 2023

Design and Implementation

Network pruning: *Make neural network smaller by removing synapses and neurons*



Optimal Brain Damage [LeCun et al., NeurIPS 1989]

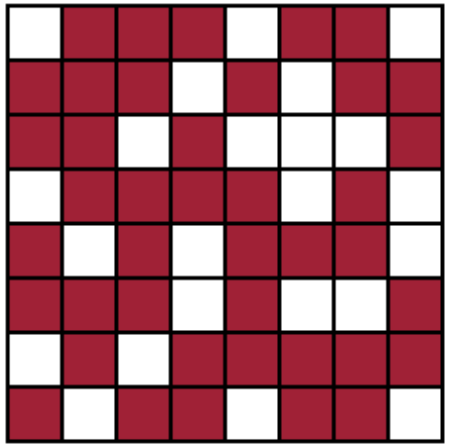
Copyright © 2023

Design and Implementation

UNIVERSITY OF TWENTE.

Network pruning

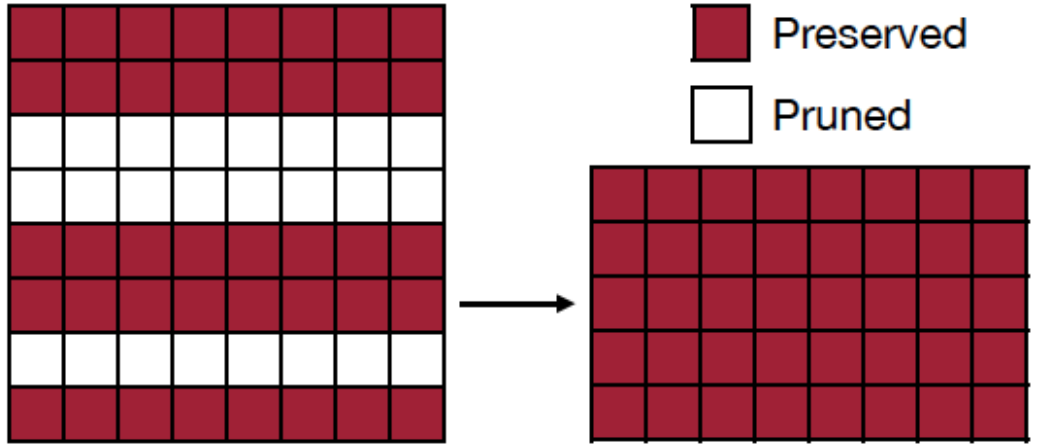
Fine-grained (unstructured)



- More flexible pruning index choice
- Hard to accelerate (irregular data expression)

- Global unstructured magnitude pruning [8]
- Global unstructured random pruning

Coarse-grained (structured)



- Less flexible pruning index choice (a subset of the fine-grained case)
- Easy to accelerate (just a smaller matrix!)

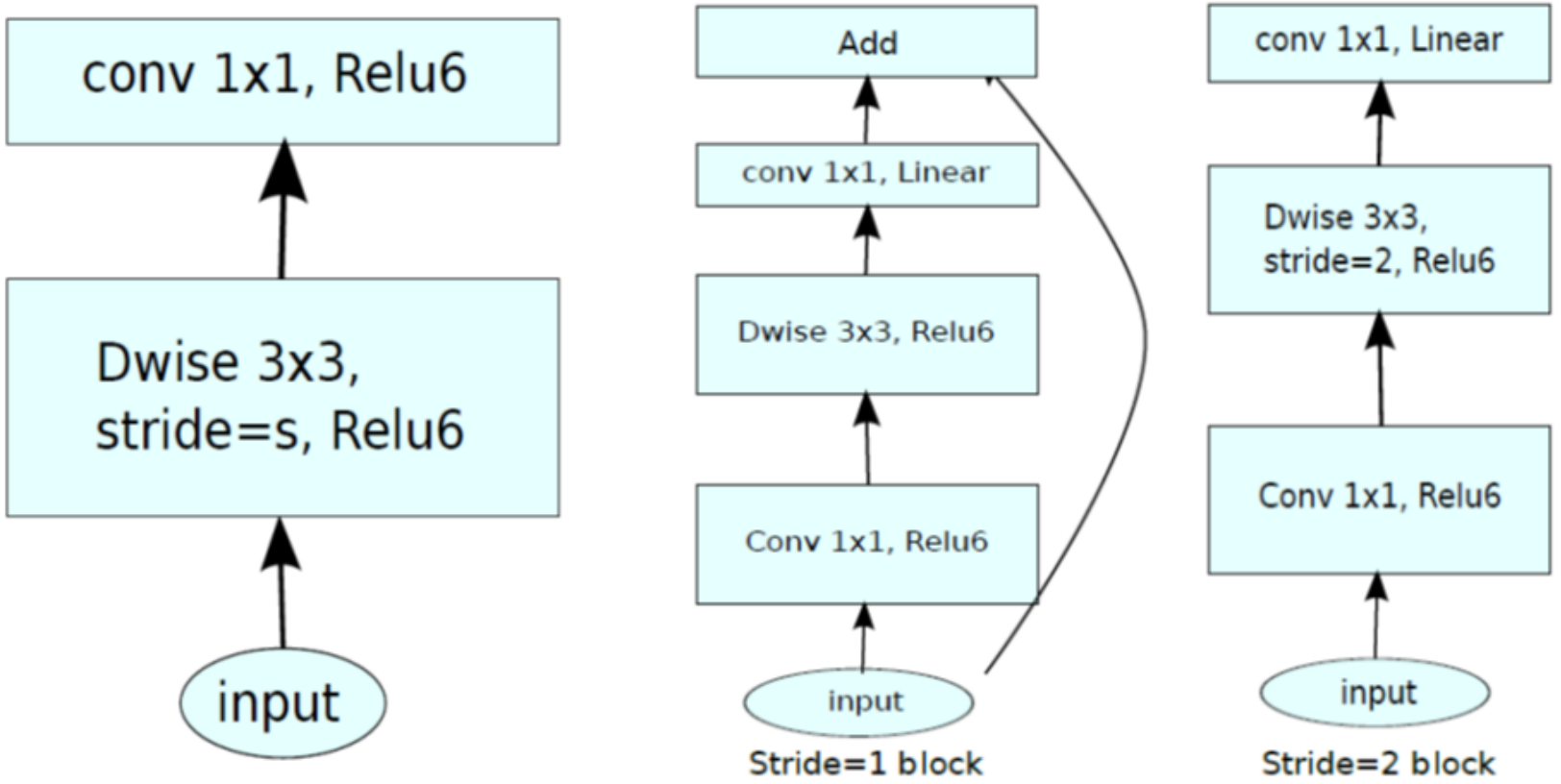
- L1-Norm filter pruning [15]
- Network Slimming [16]

Edge NETWORKS. SYSTEMS. INTELLIGENCE.

Copyright © 2023

Design and Implementation

Ready-optimized network: MobileNetV2



MobileNetV1

16 MB

MobileNetV2

14 MB

Copyright © 2023

Design and Implementation

UNIVERSITY OF TWENTE.

Dataset: CIFA-10

60000 32x32 colour images (50k training/10k testing)

For each combination (pruning method, sparsity):

2000 inferences to determine energy consumption

10000 inferences to determine model accuracy

Raspberry Pi 4B (running models)
INA219 chip (measuring power)

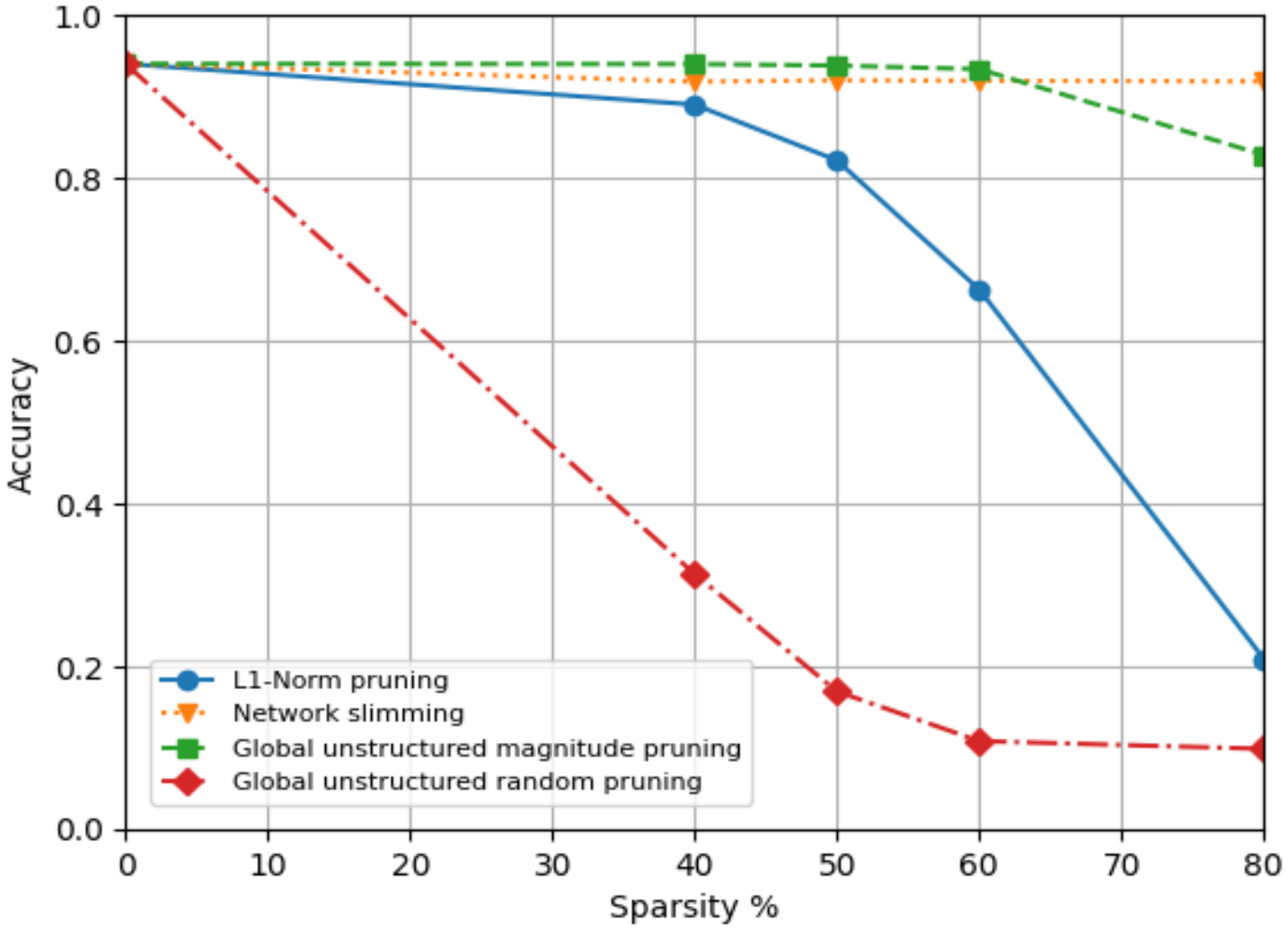
Edge NETWORKS. SYSTEMS. INTELLIGENCE.

Copyright © 2023

Experimental Results

UNIVERSITY OF TWENTE.

- Accuracy vs. model size



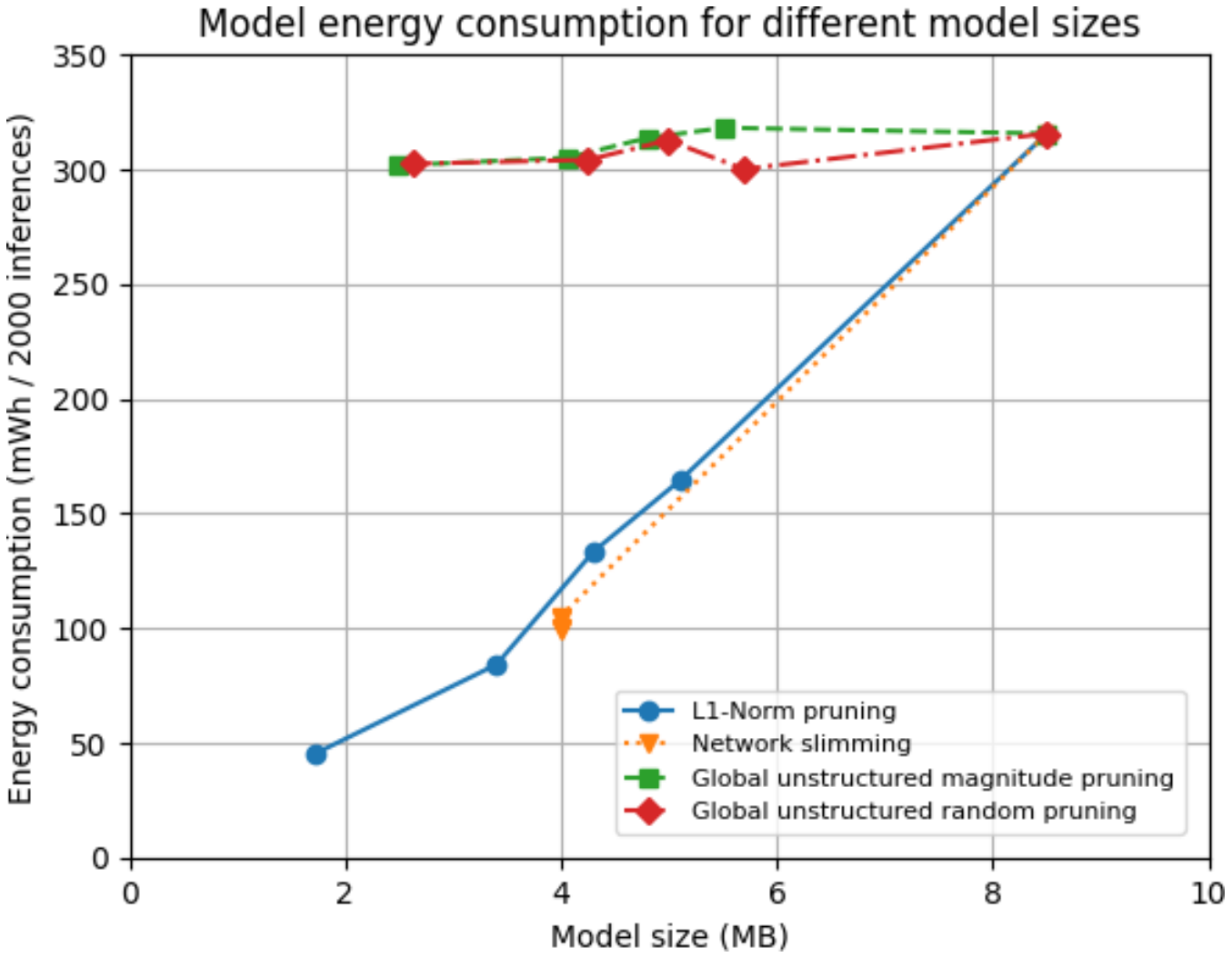
Edge NETWORKS. SYSTEMS. INTELLIGENCE.

Copyright © 2023

Experimental Results

UNIVERSITY OF TWENTE.

- Energy consumption vs. model size.



Edge NETWORKS. SYSTEMS. INTELLIGENCE.

Copyright © 2023

Future Research

- Investigate further the trade-off between *energy consumption /inferencing time/model size vs. accuracy* for real-time object detection

- Experiments on other network architectures and pruning methods

Copyright © 2023

Conclusion

In this work:

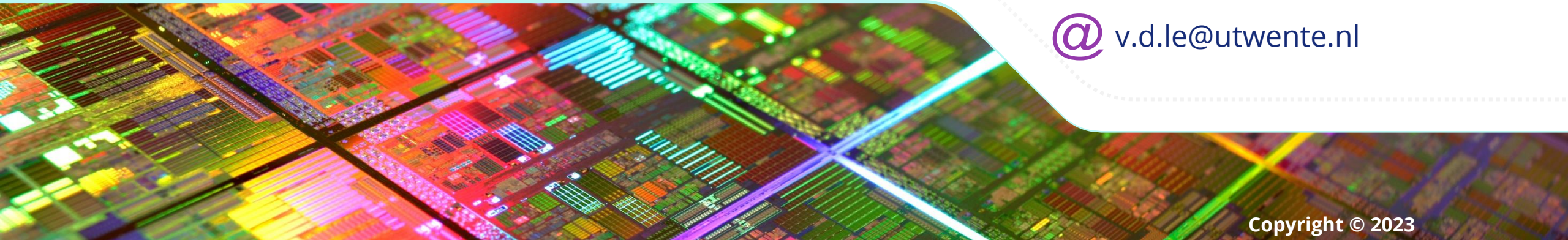
- Optimized neural network such as MobileNetV2 still can be 5 times smaller “without” sacrificing the accuracy
- Model size and power consumption are not linearly correlated.

Copyright © 2023

EDGE AI

Thank You

For your attention



@ v.d.le@utwente.nl

Copyright © 2023