TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

Institut für Computertechnik
Institute of Computer Technology

Embedded Machine Learning

AVL Open Networking Day

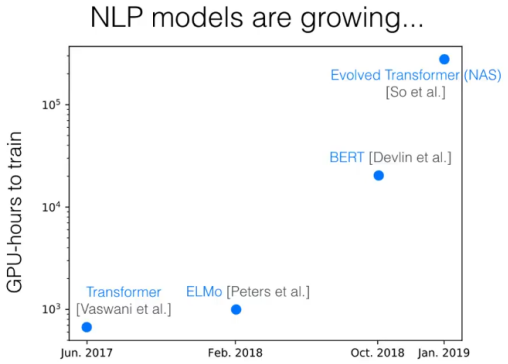
Axel Jantsch

October 4, 2021

Challenge and Motivation

ML is Resource demanding

- NAS based training is beyond the reach of most organizations

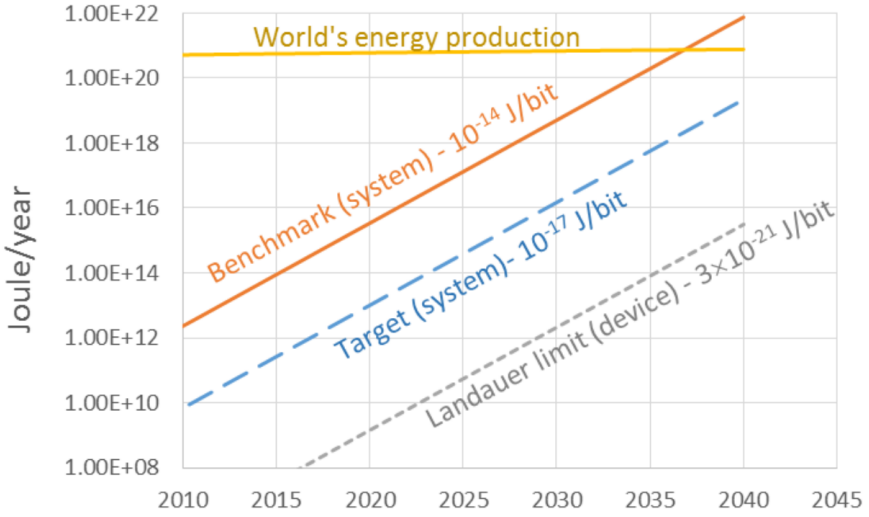


Full architecture search for a big transformer model requires

- 979M training steps and
- 32,623hours of TPU or 274,120 hours on 8 P100 GPUs,
- carbon footprint equivalent to the **lifetime of 5 US cars**.

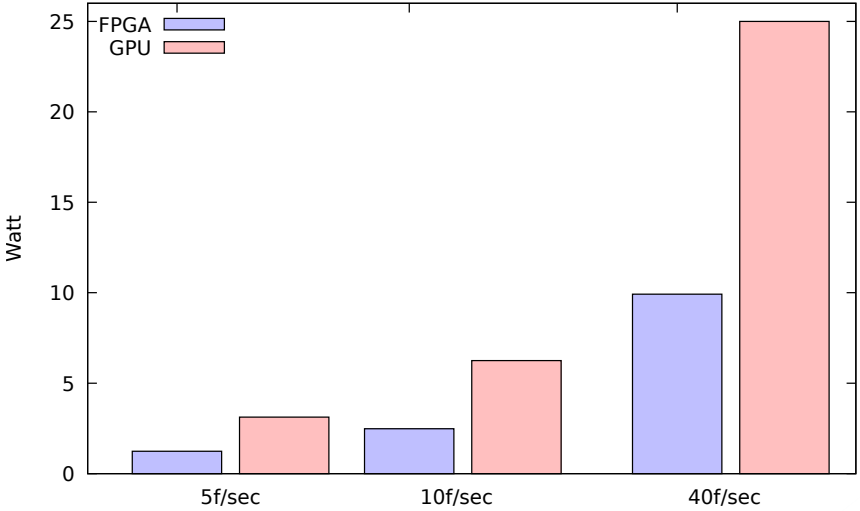
Emma Strubell, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pages 3645–3650

ML is Resource Usage is Unsustainable



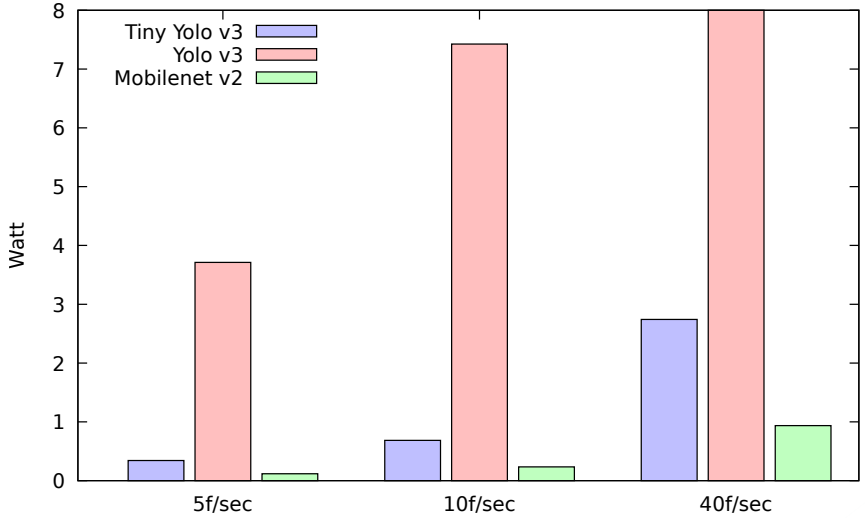
SIA - SRC. *Rebooting the IT Revolution: A Call to Action*. Technical report. Semiconductor Industry Association and Semiconductor Research Corporation, Sept. 2015

Power Consumption in Inference



VGG16 applied to the ImageNet data set based on published papers.

Power Consumption in Inference

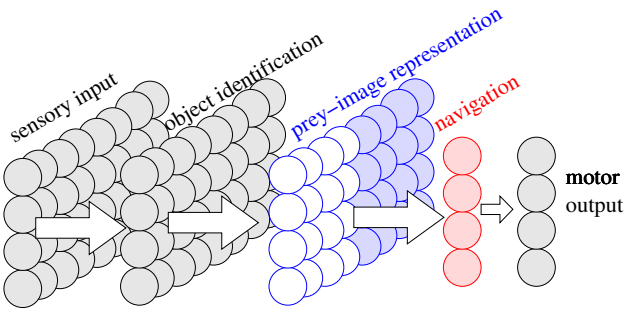


Object detection on the NCS2 platform; own measurements.

Dragonfly

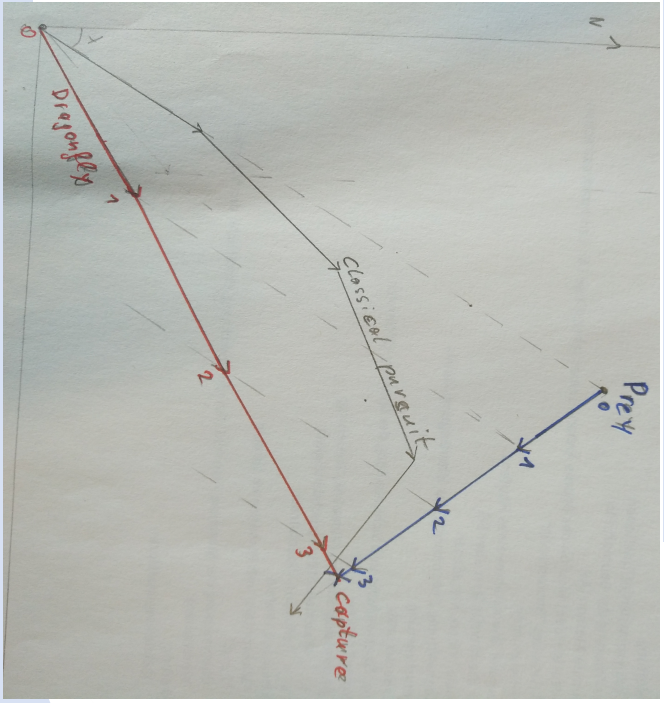
- Brain volume: 1 mm^3
- Weight: 1 mg
- Number of neurons: 1 Million
- Power consumption: 2–8 mW
- 200 frames/second
- 95 % hunting success rate
- Reaction time: 50 ms

Dragonfly



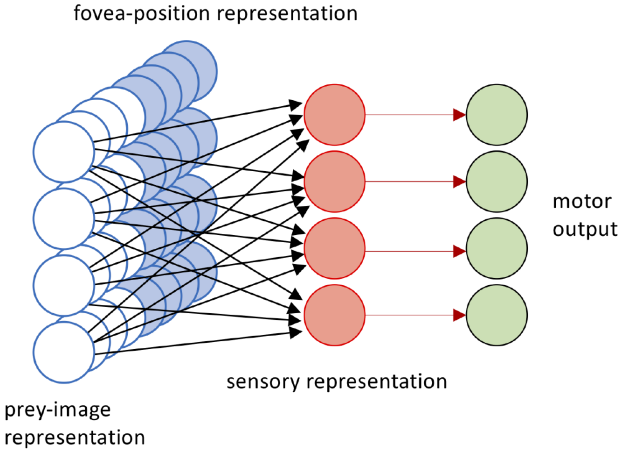
- 50 ms reaction time
- One neuron needs 10 ms to integrate inputs
- 10 ms for the photo detectors and the prey identification
- 5 ms for the muscles to produce force
- leaves 35 ms for route planning
- $\Rightarrow$ maximum 5 layer NN

Proportional Navigation



Dragonfly

- Proportional navigation has been implemented in a 3-layer NN;

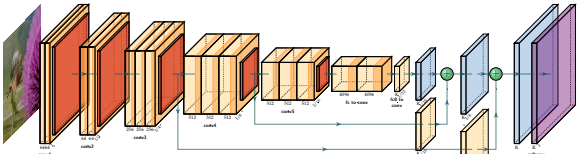


Frances S. Chance. "Interception from a Dragonfly Neural Network Model". In: *International Conference on Neuromorphic Systems 2020 (ICONS)*. Oak Ridge, TN, USA. ACM, New York, NY, USA, July 2020

In terms of energy efficiency we are about 2-3 orders of magnitude from what is possible and feasible.

Activities and Results

Design Space



DNN Choices

Convolutional layers
Filter kernels
Number of filters
Pooling layers
Filter shape
Stride
Fully connected layer
Number of layers
Regularization
etc.

Mapping Choices

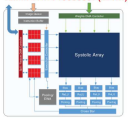
Neuron pruning
Data type selection
Approximation
Retraining
Connection pruning
Weight sparsifying
Regularization
etc.

Platform Choices

Platform Selection
Reconfiguration
Batch processing
Deep pipelining
Resource reuse
Hierarchical control
Processing unit selection
Memory allocation
Memory reuse
etc.

Intel® Vision Products with Intel® Arria® 10 FPGA

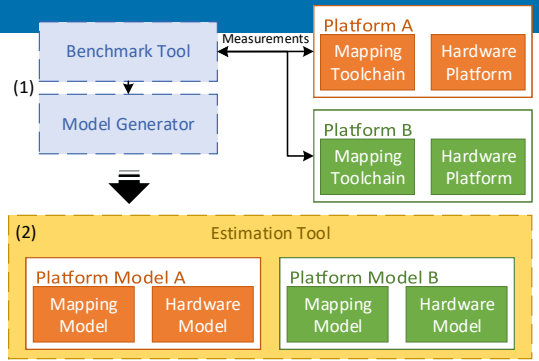
ARM NN



Nvidia Turing

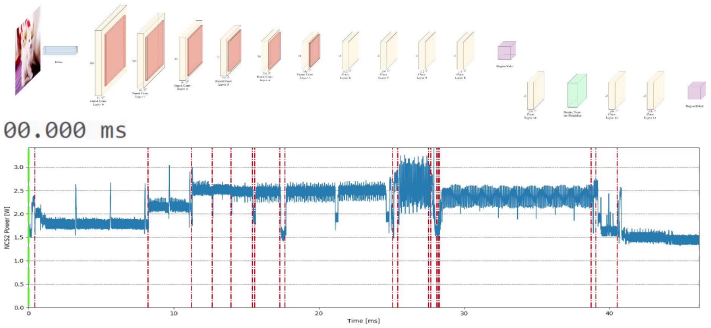
Estimation

- Two leading performance estimation tools: ANNETTE and Blackthorn
- For NCS2, Xilinx FPGA, and Jetson
- Combine analytic, statistical model and partial measurements



Network	Estimation Error [%]			
	NCS2	ZCU102	Jetson Nano	Jetson TX2
YoloV3	4.1	3.2	-	-
MobileNetV2	4.3	4.2	3.6	4.2
ResNet50	8.2	1.2	2.4	2.8
FPN Net	9.3	7.5	-	-
AlexNet	5.2	4.8	5.5	6.6
VGG16	11.3	6.2	0.5	1.4

Yolov3-tiny power profile on NCS2

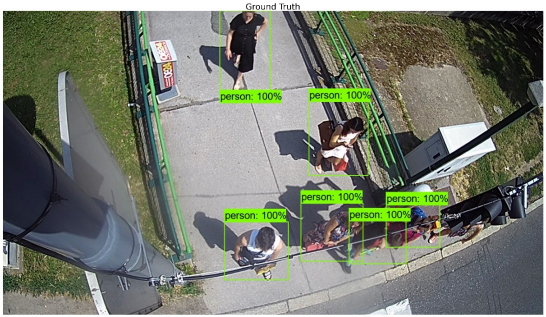


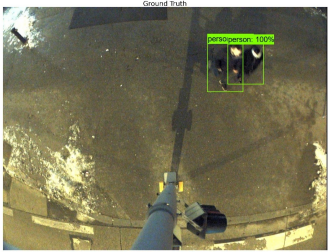
- NCS2 and Nvidia platforms
- Detailed, per layer latency and power profiling
- Hardware setting have significant influence
- 100 kHz sampling frequency is required for 5 % accuracy
- Number of operations is a poor predictor for energy consumption

Traffic Light Controller

Data set:

- training: 19087 images
- positive examples 47%
- validation: 13184
- positive examples 26%
- Resolution: 1280x720
- Issue: Validation 4h/network
  $\rightarrow$ validation set: 1319

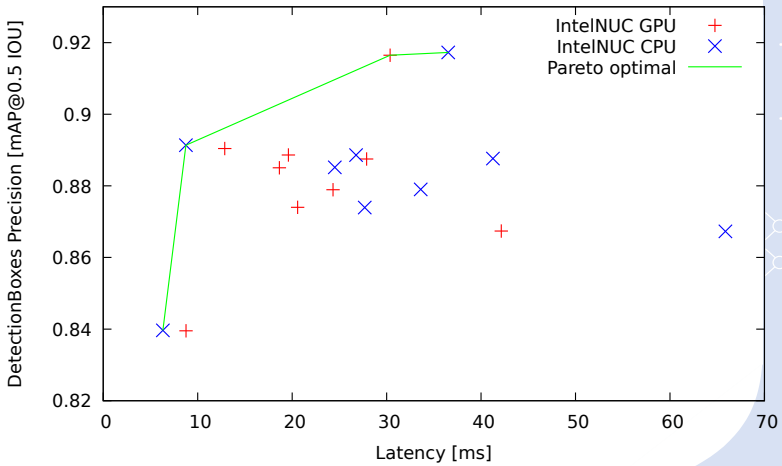




Traffic Light Controller

Intel NUC CPU vs. GPU

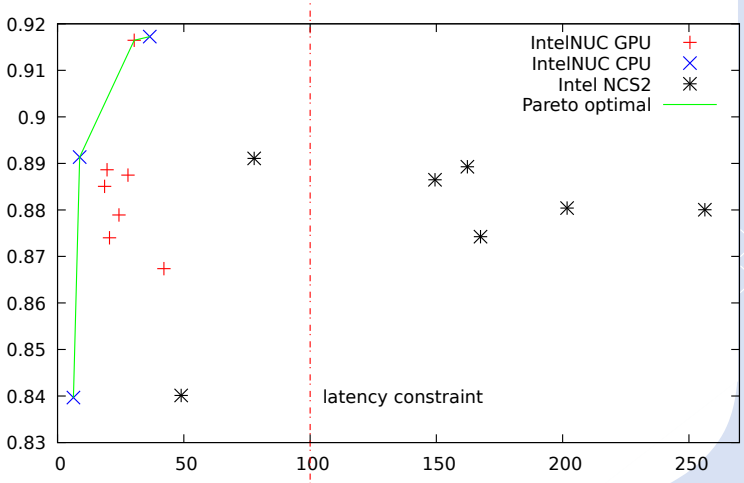
- SSD MobileNet V2
- image resolution: 320x320, 768x768, 1280x720
- Pareto optimal:
  - CPU 320x320
  - CPU 320x320 FPN Lite
  - GPU 640x640
  - CPU 640x640
- For low resolution images, CPU is preferable



Traffic Light Controller

Intel NUC CPU, GPU and NCS2

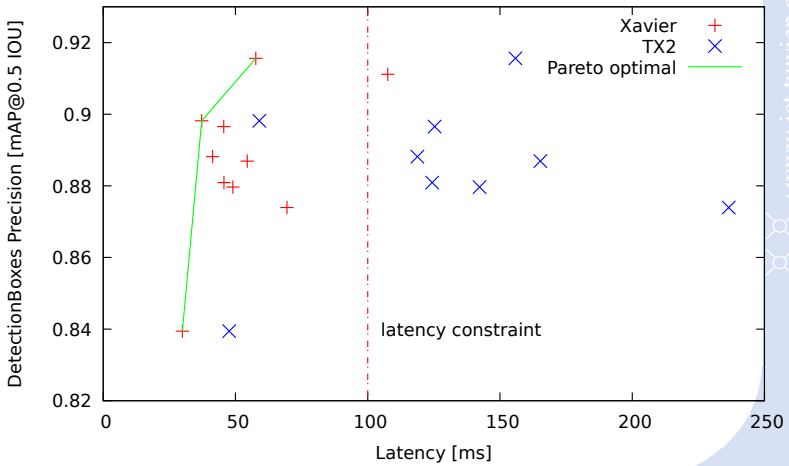
- SSD MobileNet V2
- image resolution: 320x320, 768x768, 1280x720
- Pareto optimal:
  - CPU 320x320
  - CPU 320x320 FPN Lite
  - GPU 640x640
  - CPU 640x640
- Latency constraint leaves many options



Traffic Light Controller

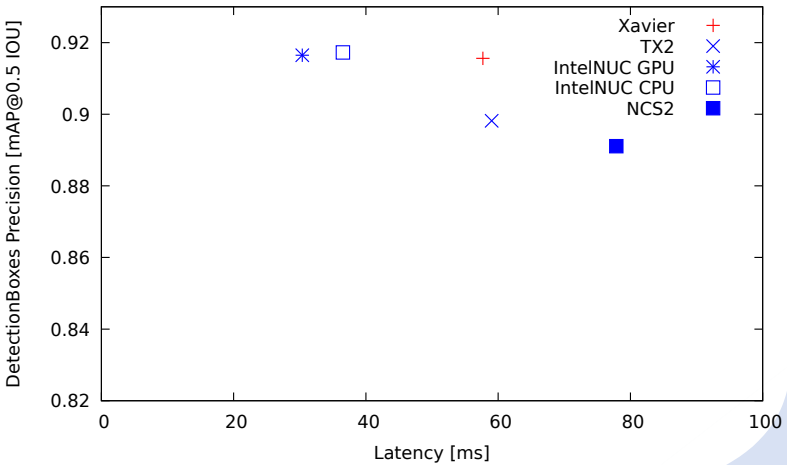
Nvidia Xavier and TX2

- Xavier is too expensive
- Best TX2: 320x320 with 59ms



Traffic Light Controller

Best platform solutions



Results, publications, demos, code on

eml.ict.tuwien.ac.at

CD-Lab for Embedded Machine Learning

Duration	7 years, Oct 2019 - Sept 2026
Partner	TU Wien, TU Graz, AVL, Mission Embedded, Siemens
3 WPs	WP1 Embedded Platforms (TUW, Mission Embedded) WP2 DNN Architecture and Optimization (TUW, Siemens) WP3 Continuous Learning (TUG, AVL)
Budget	2.8 M€, 400 k€/year
People	Funded: 2 Postdocs, 5 PhD Students, 3 MSc Students Total: 2 Postdocs, 5 PhD Students, 14 MSc+BSc Students

¿ Questions ?