Lucerne University of
Applied Sciences and Arts

HOCHSCHULE
LUZERN

Technik & Architektur

# AI at the Edge –
# Building a Framework for Efficient CNN Inference

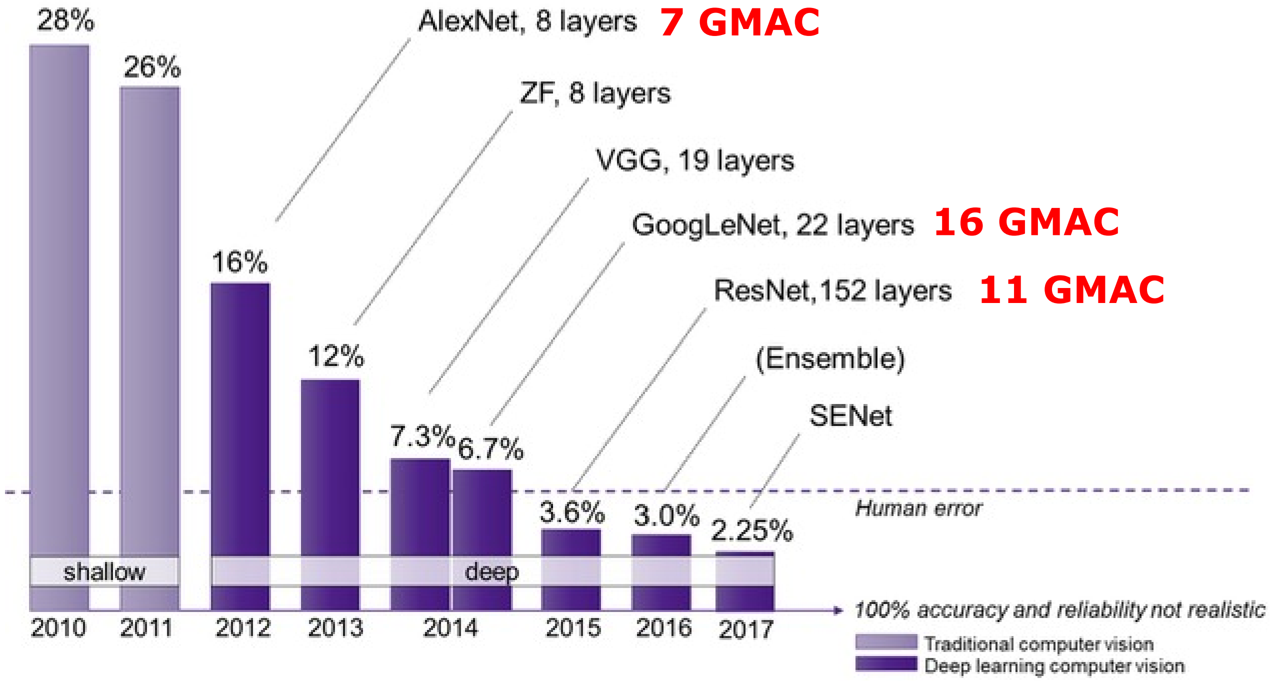

**Mario Fischer (MSE)**
**Silvio Emmenegger (MSE)**
**Prof. Dr. Jürgen Wassner**

IET TechLunch, 09. September 2020, 12:00-13:00

Fabio Johner(MSE), Josef Estermann (MSE), Mario von Flüe (BA), Michael Kurmann (MSE), Cyrill Durrer (BA)

# When it all began, again …

- Ignition spark from ImageNet object classification challenge
- 256 x 256 color pixel images in 1000 classes
- 2012 AlexNet: First **Deep** Neural Network Architecture
- 2017 better-than-human classification accuracy



# What’s wrong with a couple of GMACs?

- Multiply-Accumulate (MAC) operation is at the heart of any digital signal processing system

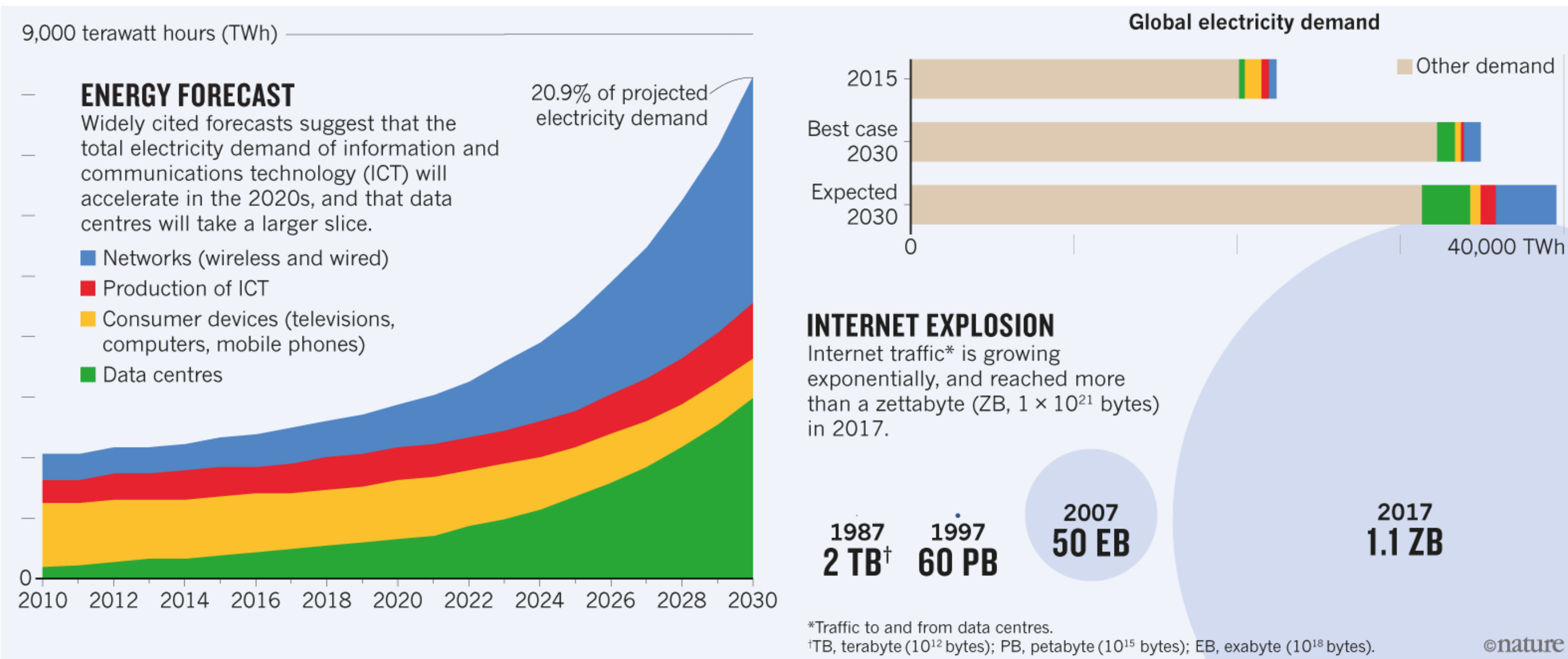
| | **2048-point FFT (OFDM Recvr)** | **ResNet-152 (better-than-human)** |  |
|---|---|---|---|
| Sample Interval | 50 us | 20 ms | |
| GMAC/s | 1.8 | 550 | |
| **Power** | **0.25 W** | **76 W** | **10 W** |

- Prototypes of self-driving cars
  - process 200 MB of sensor data per sec
  - dissipate **2.5 kW**


Source: Adapted from [1]

# It's a global problem …

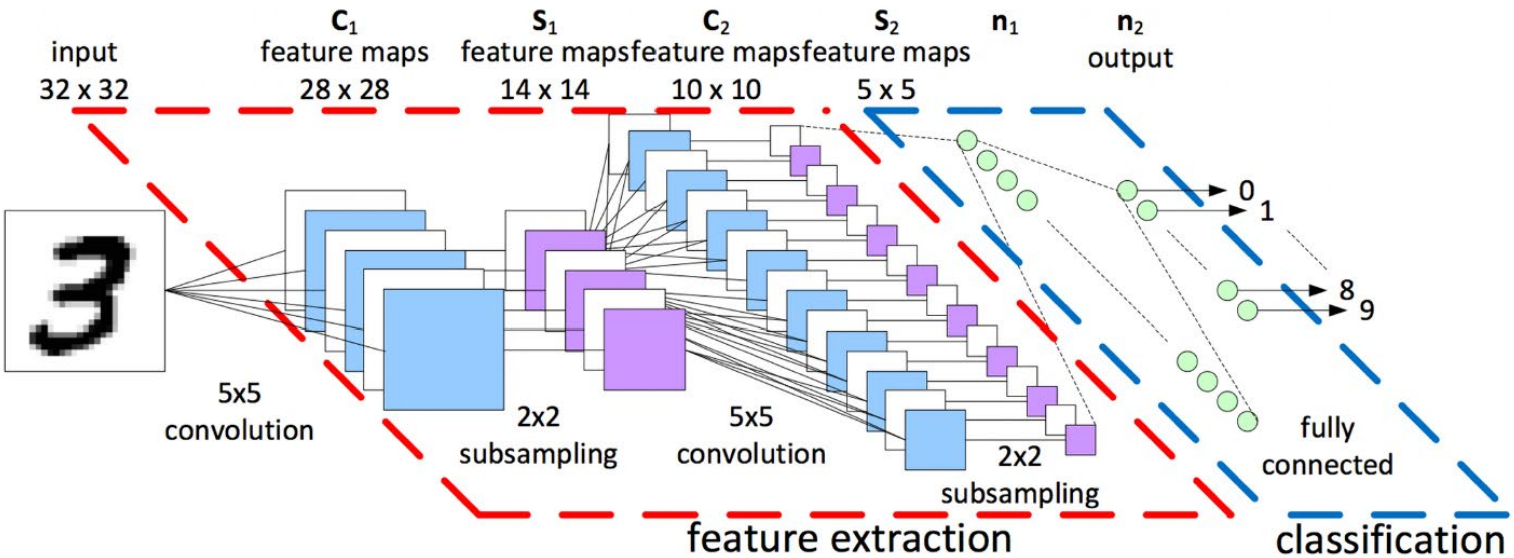
- ICT to account for 20% of total electricity in 2030
- Soon more CO2 emissions than world-wide traffic
- Energy consumption growing for **Network traffic ➔ Perform AI at the Edge**




Source: [2]

# Where do all the GMACs come from?

- Since AlexNet all ImageNet winners were **<u>Convolutional</u>** Neural Networks (CNN)

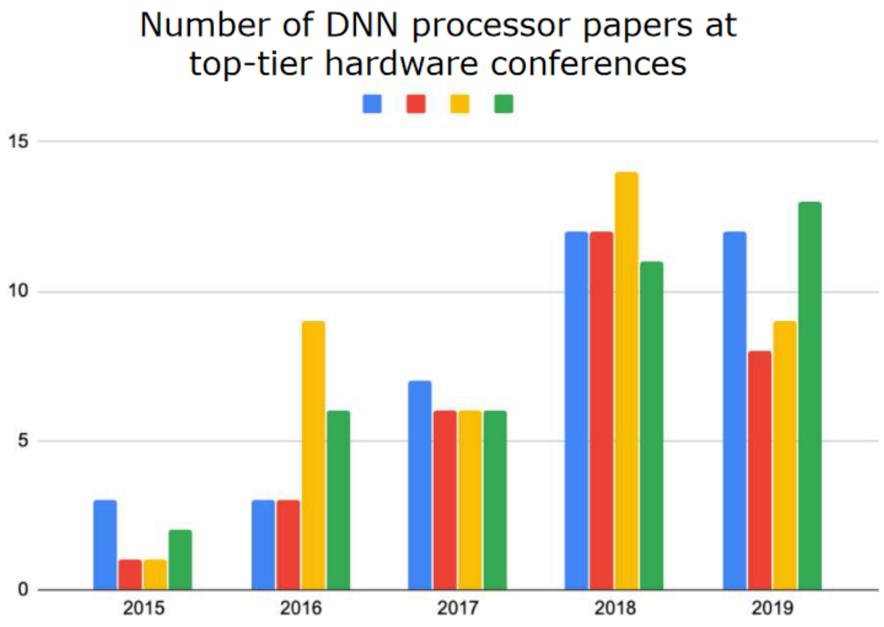



- No-Free-Lunch Theorem (X-th version):

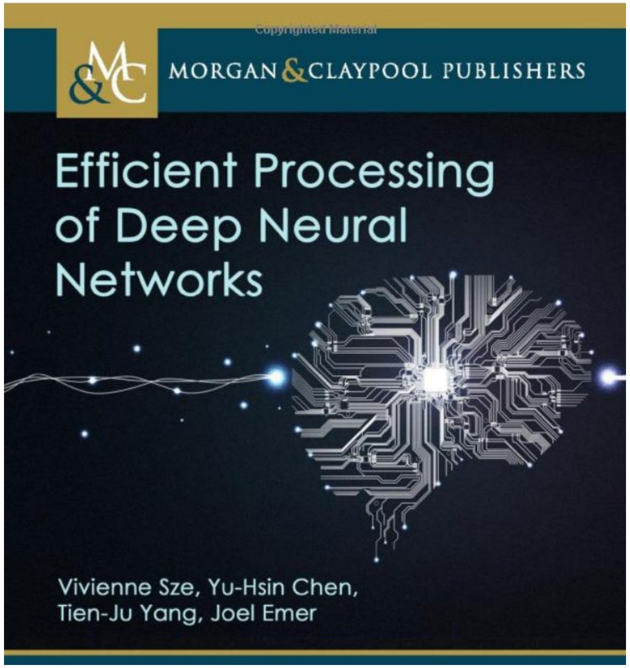
  **Who wants CNN-accuracy has to pay the power penalty. Really?**

# The second wave …

- … of AI research & development “pandemia”

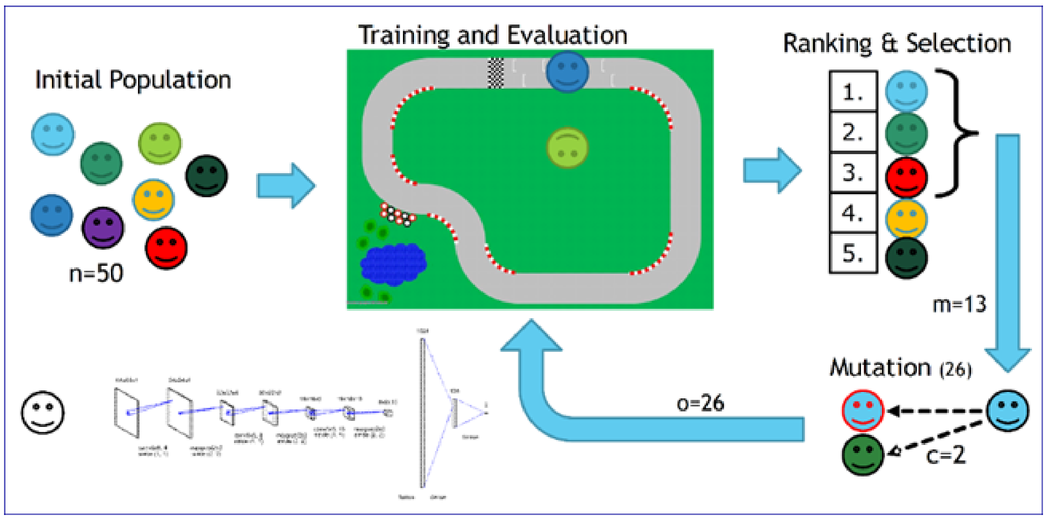



Source: [3]



# Breed your own …

- Optimization potential is larger at higher abstraction levels ➔ Try to reduce # of GMACs first
- Optimize neural networks by selective breeding

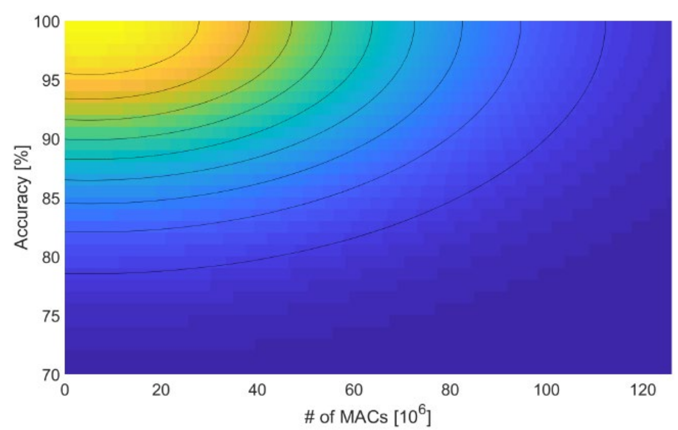



| Layer type | Layer Parameters | Parameter values |
|---|---|---|
| 2D Convolution | # of kernels<br>stride<br>kernel size<br>padding<br>activation<br>dropout usage<br>dropout rate<br>batch-norm usage | $\in \{1, 2, ..., 49, 50\}$<br>$\in \{1, 2, 4, 6, 8, 10\}$<br>Square, $\in \{3, 5, 7, 9, 11\}$<br>valid or same<br>linear or ReLU<br>true or false<br>$\in \{0.25, 0.3, ..., 0.7, 0.75\}$<br>true or false |
| Dense/ Fully-connected | # of neurons<br>activation<br>dropout usage<br>dropout rate | $\in \{20, 30, ..., 490, 500\}$<br>linear or ReLU<br>true or false<br>$\in \{0.25, 0.3, ..., 0.7, 0.75\}$ |
| Pooling | stride<br>kernel size<br>padding | $\in \{1, 2, 4, 6, 8, 10\}$<br>Square, $\in \{2, 3, ..., 10, 11\}$<br>valid or same |
| Dropout | dropout rate | $\in \{0.25, 0.5, 0.75\}$ |

Source: [4]

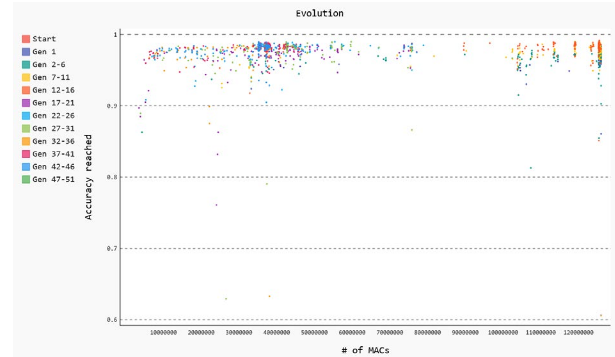
- 2-D ranking (cost) function

  ➔ Joint optimization of accuracy and # of GMACs



# … and get amazing Results …

- Traffic Sign Recognition Challenge




- Single Network

| Network | Accuracy [%] | MAC [$10^6$] | Param [$10^6$] |
|---|---|---|---|
| Original Winner | 98.47 | 126.0 (1x) | 1.54 (100%) |
| Opt 1 | 98.18 | 11.3 **(11x)** | 0.60 **(39%)** |
| Opt 2 | 98.03 | 10.9 **(11x)** | 0.17 **(11%)** |

- Network Ensemble

| Network | Accuracy [%] | MAC [$10^6$] | Param [$10^6$] |
|---|---|---|---|
| Original Winner | 99.46 | 3149.5 (1x) | 38.59 (100%) |
| Opt-Ensemble 1 | 99.15 | 22.4 **(140x)** | 1.28 **(39%)** |
| Opt-Ensemble 2 | 99.35 | 45.0 **(70x)** | 2.74 **(11%)** |

# ... that can be published

2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)

## Efficient Evolutionary Architecture Search for CNN Optimization on GTSRB

Fabio Marco Johner
*Competence Center Electronics (CCE)*
*Lucerne University of Applied Sciences and Arts (HSLU)*
Lucerne, Switzerland
fabio.johner@hslu.ch

Juergen Wassner
*Intelligent Sensors and Networks Laboratory (ISN)*
*Lucerne University of Applied Sciences and Arts (HSLU)*
Lucerne, Switzerland
juergen.wassner@hslu.ch

***Abstract*—Neural network inference on embedded devices has to meet accuracy and latency requirements under tight resource constraints. The design of suitable network architectures is a challenging and time-consuming task. Therefore, automatic discovery and optimization of neural networks is considered important for continuing the trend of moving classification tasks from cloud to edge computing.**

there is an increasing demand for neural networks that have low computational compl... eters. Such reduced netw... accuracy which in genera... tradeoff [13].

Our main goal is to aut...

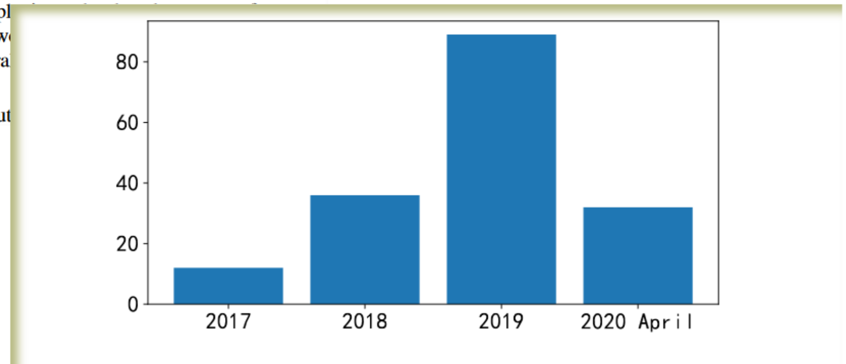


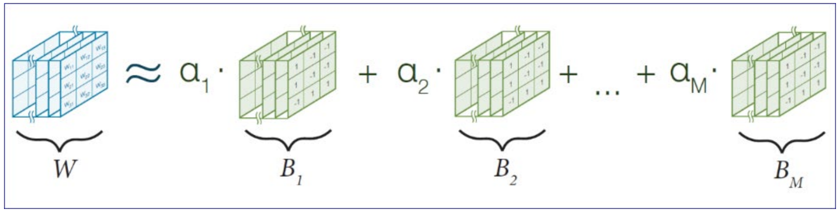

Fig. 1. The number of submissions focusing on evolutionary neural architecture search. The data is from Google Scholar with the keywords of "evolutionary"

Source: [5]

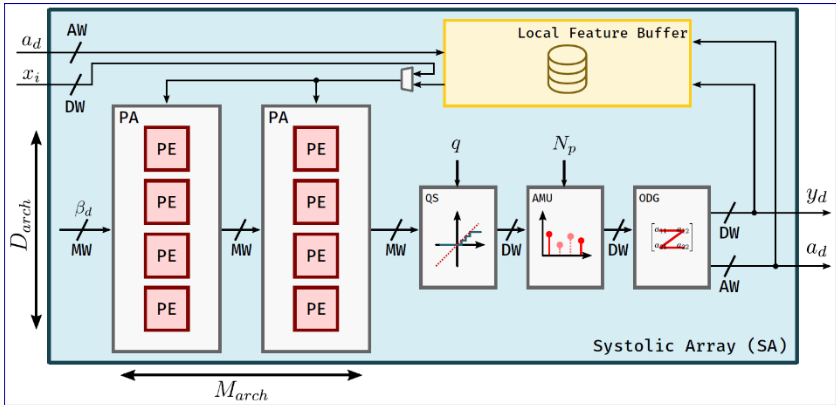
# Optimize (remaining) GMACs Implementation

- **<u>Mario Fischer (M.Sc. 2020)</u>**

  - Binary Weight Approximation [6]




  - Scalable HW Architecture [7]



# Techlunch September 9, 2020

## Techlunch September 9, 2020

Mario Fischer

Hochschule Luzern

*marioandrea.fischer@hslu.ch*

September 9, 2020

# Origins of Neural Networks

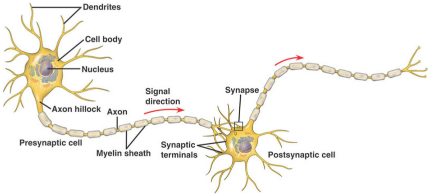



Figure 1: Brain Neuron[1]

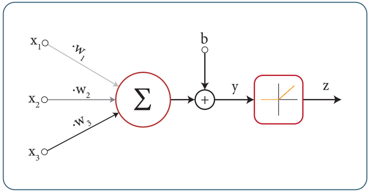



Figure 2: Mathematical Model of a Neuron

- *Dendrites*: Connected to $> 1000$ neighbouring neurons.
- *Soma*: Sums the number of exited neighbouring neurons.
- *Axon*: Fires, if the Soma exceeds a certain potential

$$z = \varphi(\sum_{i=0}^{n} x_i w_i + b) \tag{1}$$

- $\varphi$: Nonlinear function
  - e.g. Threshold, Logistic etc.

# Neural Networks for Classification

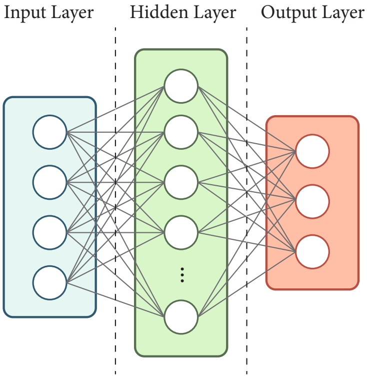



Figure 3: Fully Connected Neural Network

- Neuron in a layer connect to neurons in adjacent layers
- Creates a densly connectd network
  - Fully Connected Network
- Hidden layers are often cascaded

# Preprocessing in terms of Artifical Neural Networks

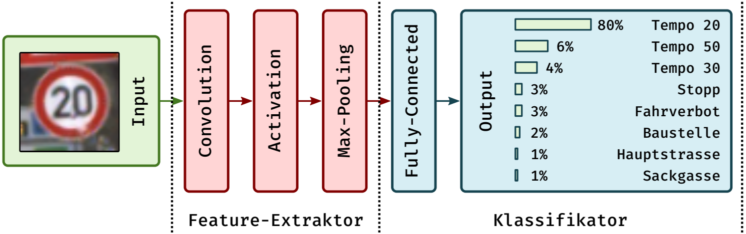



Figure 4: Complete Convolutional Neural Network (CNN)

- Use multiple filters for feature extraction
- Network learns the parameters of the kernel
- Cascade convolution filters (layers)
    - Low Level to High Level Features
    - $N_{Layers}$ between 1 to 100

# Challenge of Convolution Layers

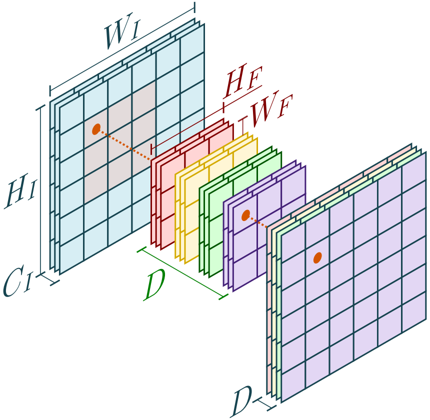



Figure 5: Convolution with Multiple Kernels

- Computationally intensive convolution
  - Each input pixel convolved with $D$ kernels
  - Large amounts of Multiply-Accumulates (MACs)
- Example Traffic Sign Recognition (IDSIA)
  - Total: 126 MMACs
  - 99.5% of MACs for feature extraction

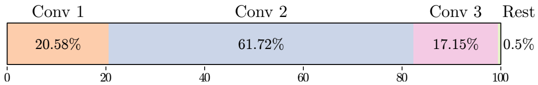



Figure 6: Percentage of total MACs per layer

# Challenge of Fully Connected Layers

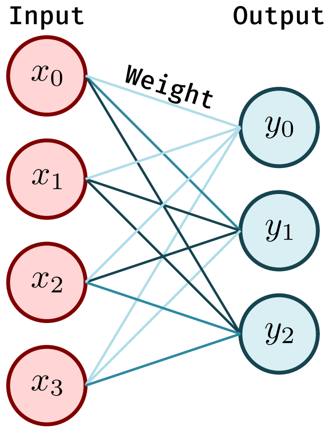



Figure 7: Fully Connected or Dense Layer

- Memory consumption of neurons
  - Each neuron has *weights* to all previous neurons
  - Usually $N_{Neurons} > 100$
- Example Traffic Sign Recognition (IDSIA)
  - Total: 1.54M parameters
  - FC1: nearly ½ of parameters
  - `float32`: 43.4Mb per frame

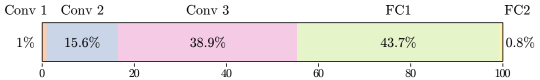



Figure 8: Memory Distribution IDSIA

# Binary Weight Approximation

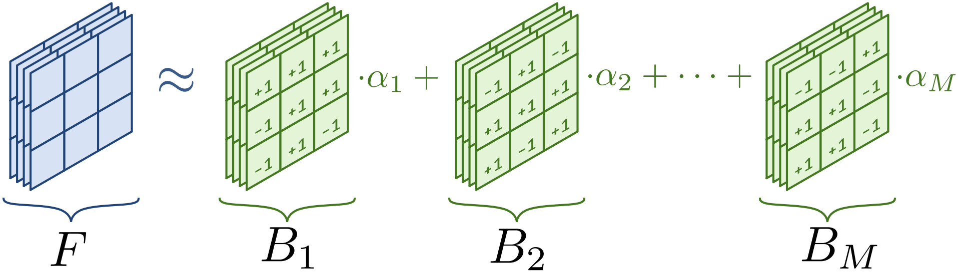



Figure 9: Binary Approximation with $M$ Binary Filters [2]

- Problem: Limited number of hardware multipliers
- Goal: Scalability in **throughput** und **ressource utilization**
- Idea: Replace point wise multiplication with mostly sign changes
  - Addition $\approx 7\times$ more energy efficient than multiplication [3]
- Compression of weights without loss in accuracy

# Hardware Architecture Design Paradigm

**Regular Processor**

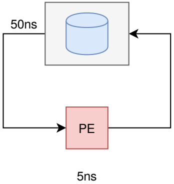



Figure 10: CPU

- $t_{\text{tot\_exec}} = 50\ ns$
- $N_{\text{OPS}} = 20\ MOPS$

**Systolic Array** (greek *systole*: contraction)

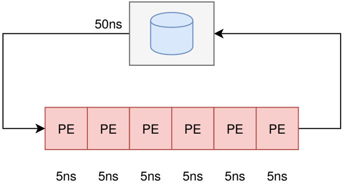
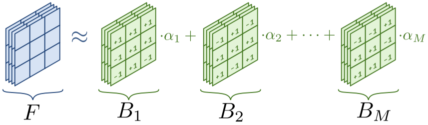



Figure 11: Proposed Paradigm for BinArray

- $t_{\text{tot\_exec}} = 50\ ns$
- $N_{\text{OPS}} = 120\ MOPS$

# Processing Element (PE)



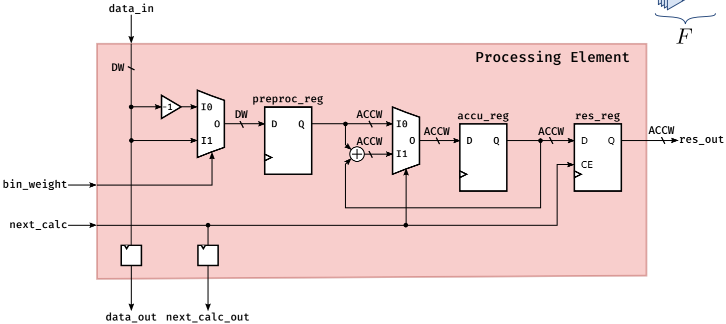




Figure 12: Processing Element performs Sign Changes according to Binary Weight

- Single accumulation per clock cycle

# Processing Array (PA)

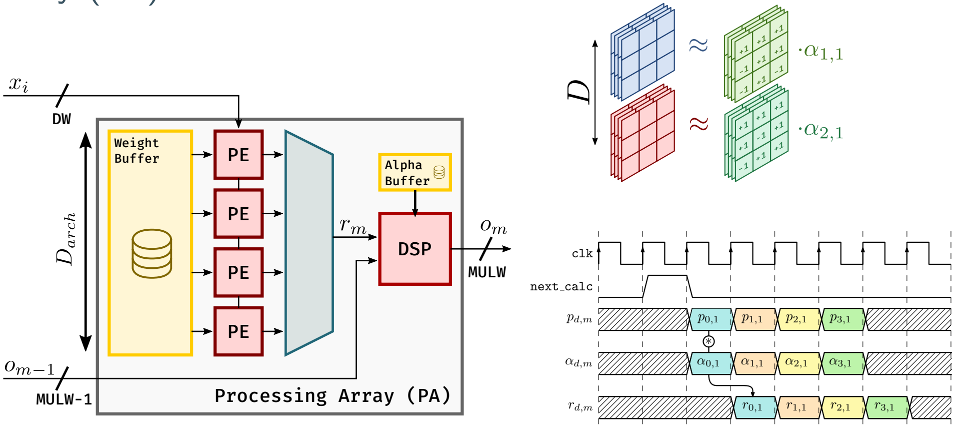



Figure 13: Parallel Computation of $D$ Channels by passing Input Features to PEs

- DSP slice for multiplication with $\alpha_m$

# Processing Arrays (PAs) and Systolic Array

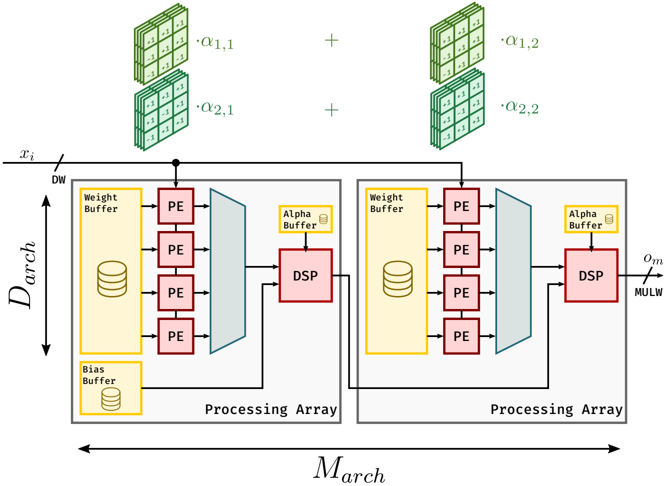



Figure 14: Parallel Computation of $D_{arch}$ Channels and $M_{arch}$ Base Filters

# Systolic Array

## Advantages of Systolic Arrays for Embedded Systeme

- Parameter $D_{arch}$ und $M_{arch}$
  - Scalability in throughput, accuracy and resource utilization
- Dataflow in a Systolic Array
  - Support for different kernel sizes $(3 \times 3,\ 5 \times 5,$ etc.)
  - Architecture convolution and fully connected layers

# Performance on GTSRB mit BinArray *Fast* $D_{arch} = 32$

| Network | MACs | Accuracy | | Throughput FPS | |
|---|---|---|---|---|---|
| | | CPU | BinArray | CPU | BinArray |
| IDSIA[4] | 126M | 97.2% | 96.8% | 7.9 | 92.1 |
| DNA821[5] | 9M | 97.8% | 97.0% | 111.8 | 819.8 |

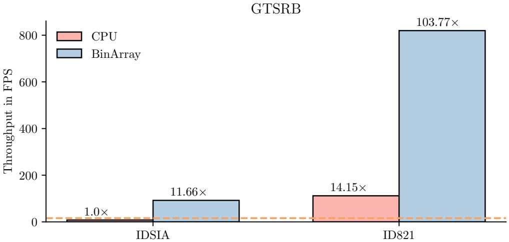



Figure 15: Accuracy and Throughput on GTSRB with $M = 2$

# Performance on GTSRB with BinArray small $D_{arch} = 8$

| Network | MACs | Throughput FPS | | $\Delta$ *Fast* |
|---|---|---|---|---|
| | | CPU | BinArray | |
| IDSIA[4] | 126M | 7.9 | 24.9 | 370% |
| DNA821[5] | 9M | 111.8 | 354.2 | 231% |

| Util. $D_{arch} = 8,\ M_{arch} = 2$ | | | |
|---|---|---|---|
| | Total | % of XC7Z045 | $\Delta$ *Fast* |
| LUTs | 1708 | 0.78% | 46% |
| FFs | 2311 | 0.53% | 43% |
| BRAM | 5.5 | 1.01% | 85% |
| DSPs | 2 | 0.22% | 100% |

# Summary

## BinArray: Design and implementation of a hardware accelerator for FPGAs

- Scalable accelerator for different network architectures
- Conserve precious hardware multipliers on the FPGA platform
- Limit the need for global communication with the Systolic Array paradigm

## Thank you for your attention!

- Back to Jürgen

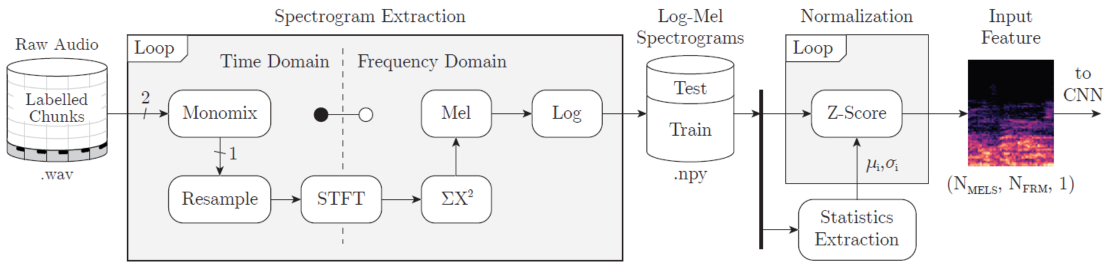
# Bibliographie


- M. Radice, *The Neuron: Simple, Yet Complex*.
  weebly.
- X. Lin, C. Zhao, and W. Pan, "Towards accurate binary convolutional neural network," *CoRR*, vol. abs/1711.11294, 2017.
- W. Dally, "High-performance hardware for machine learning," in *2016 Embedded Neural Network Summit*, (San Jose, CA), Cadence Design Systems, 2016.
- D. Cireșan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333 – 338, 2012.
  Selected Papers from IJCNN 2011.
- F. M. Johner and J. Wassner, "Efficient evolutionary architecture search for CNN optimization on GTSRB," *ICMLA*, 2019.
- Coral.ai, "Edge TPU performance benchmarks."

# Let's get really small ...

- Also 1-D signals ask for classification ➔ Microphone data
- Sometimes together with ultra-low power requirements

- **<u>Silvio Emmenegger (M.Sc. 2020)</u>**
  - Acoustic Scene and Room Classification for Real-Time Applications [8]



# Acoustic Scene and Room Classification for Real-Time Applications

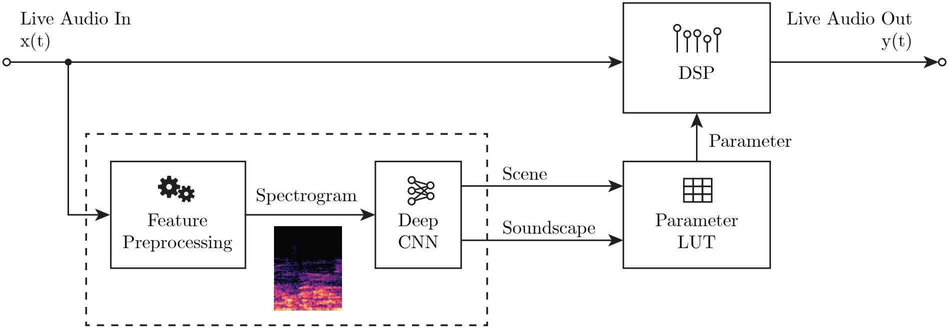
Silvio Emmenegger

Tech-Lunch, AI at the Edge

Sept 9, 2020

## Intention

- Hearing aids: acoustic classifier for scenes and soundscapes
- Use of latest AI techniques → Convolutional Neural Networks (CNN)



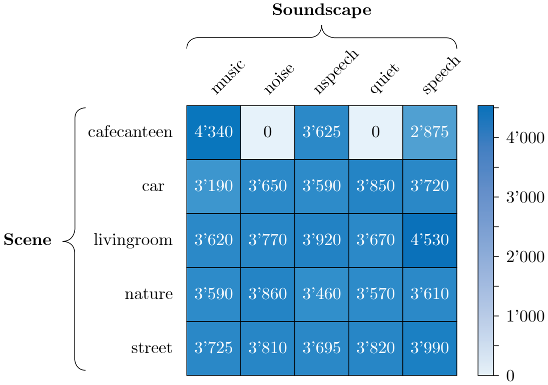
## Real-Time Specifications for Hearing Aids

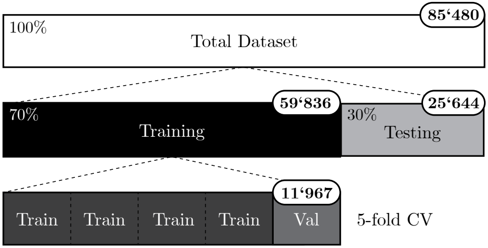

| Parameter | | Value | Unit |
|---|---|---|---|
| Audio | Sampling Freq. | 22.05 | kHz |
| | Quantization | 16 | bit |
| CPU | Clock Speed | 5 | MHz |
| Memory | Available | unknown | Mbit |
| | Total | 32-48 | Mbit |
| Battery | Capacity | 45 | mAh |
| | Voltage | 4 | V |
| | Lifetime | 3-7 | days |
| Inference Interval | | 1 | sec |

Source: https://bit.ly/388zFsc

## Recorded Dataset

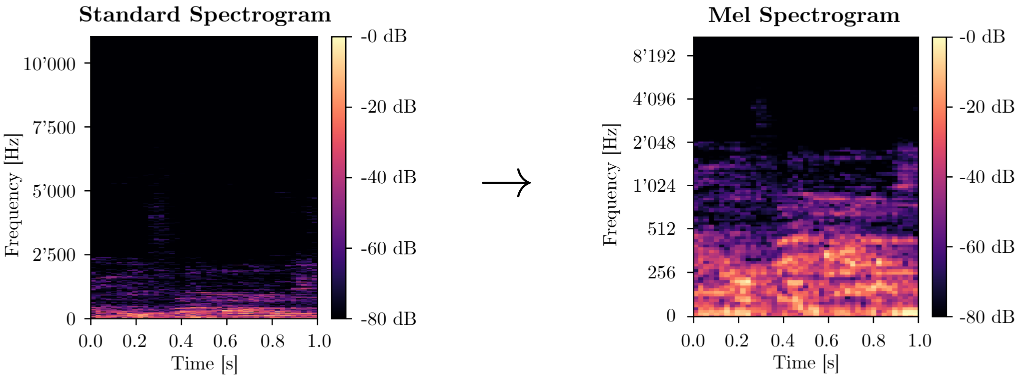
- Multi-class multi-output classification problem, supervised learning
- Total length: 23.8 hours → chunked to 1 sec



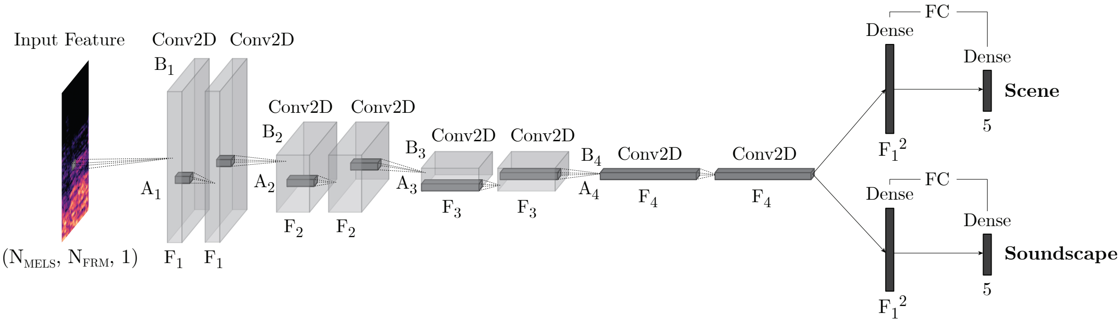
## Feature Preprocessing

- Convert audio chunks into frequency domain
- Short-Time Fourier Transform (STFT) → standard spectrogram
- Logarithmic compression of frequency axis → Mel spectrogram



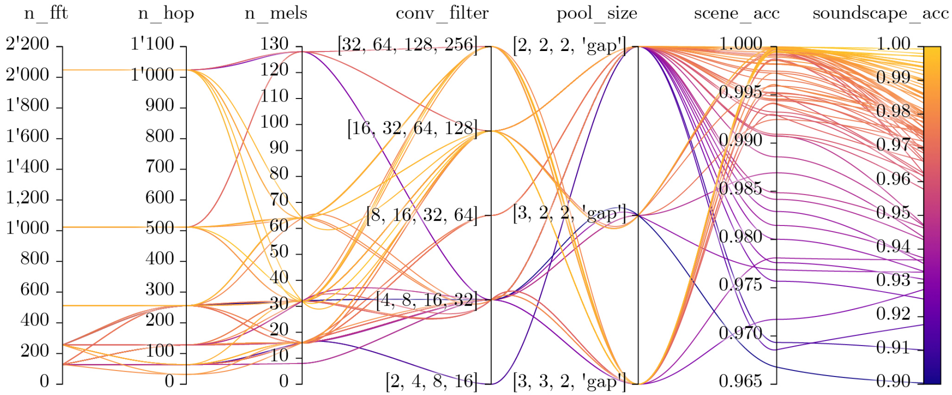
## CNN Architecture

- Inspired by VGGNet-16 (image classification)
- Two fully-connected (FC) outputs share same feature extraction



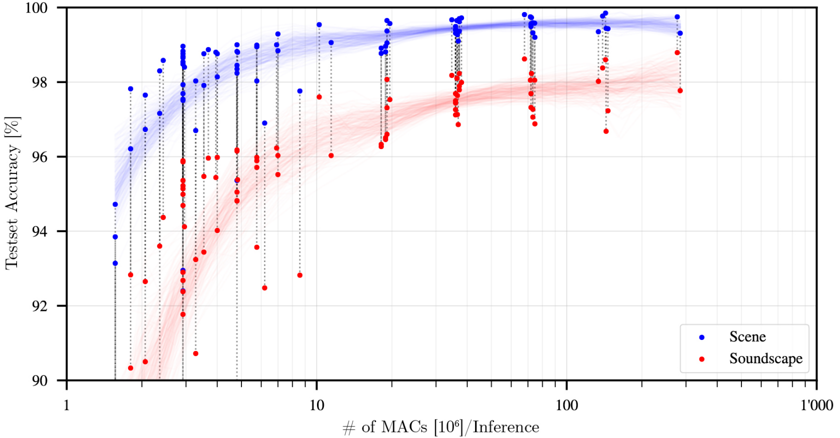
## Training metrics of model candidates

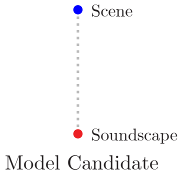
- Iterative training of ≈ 100 models
- Conclusion: soundscape always below scene training accuracy



# Inference Complexity

- Trend: testset accuracies increasing for models with more MACs[1]
- Model studies: selection of the three best models of each decade

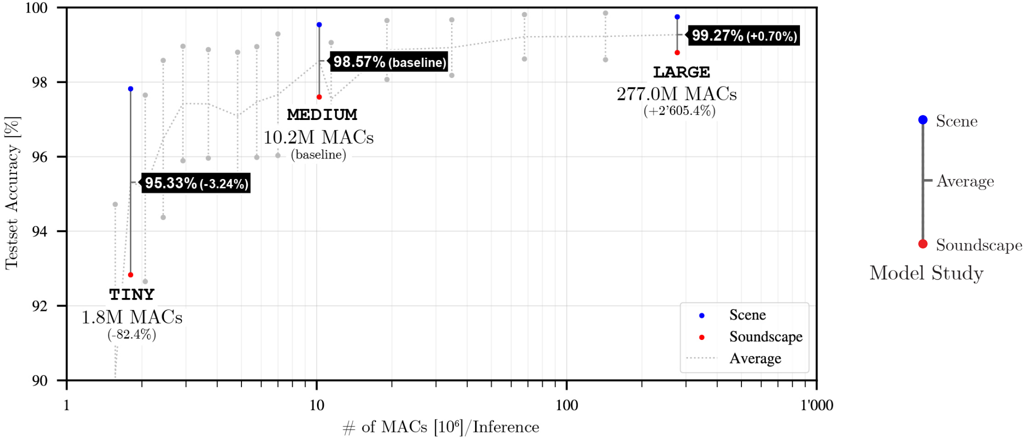



[1] Multiply-accumulate operations

# Inference Complexity

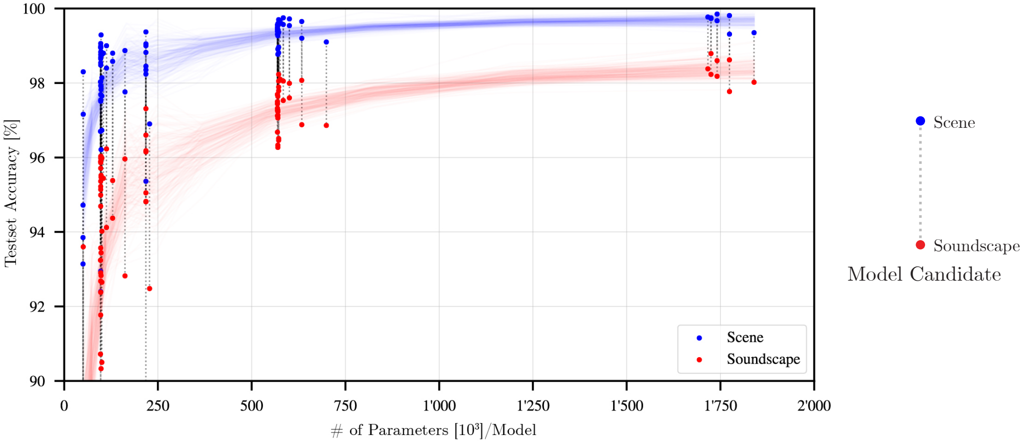
- Trend: testset accuracies increasing for models with more MACs[1]
- Model studies: selection of the three best models of each decade




[1]Multiply-accumulate operations

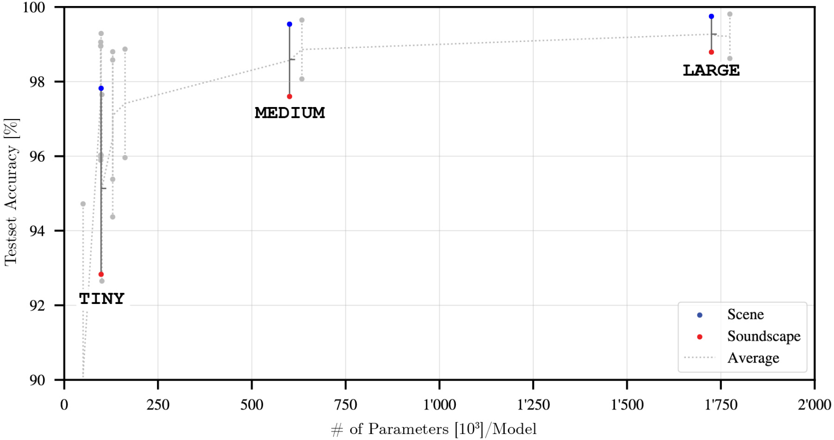
## Memory Complexity

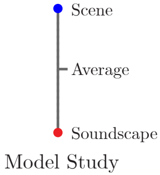
- Same trend: more parameters lead to higher testset accuracies
- Model studies: number of parameters scale on a linear basis



# Memory Complexity

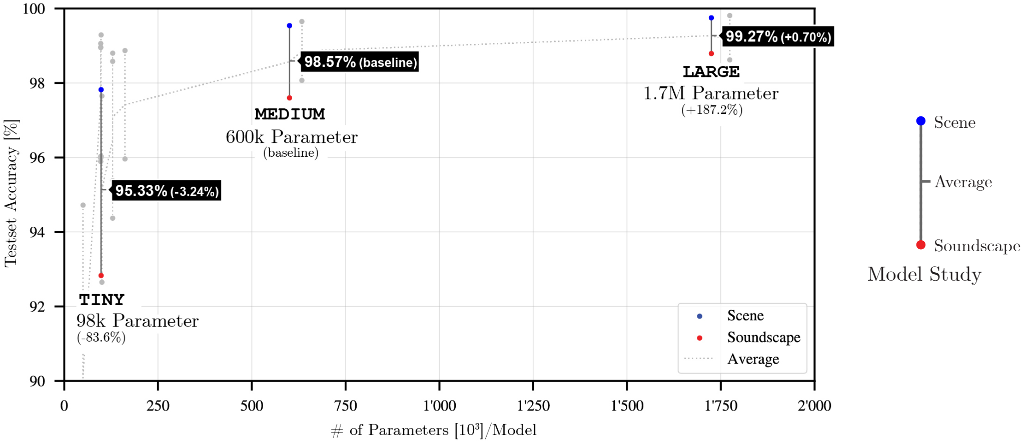
- Same trend: more parameters lead to higher testset accuracies
- Model studies: number of parameters scale on a linear basis



# Memory Complexity

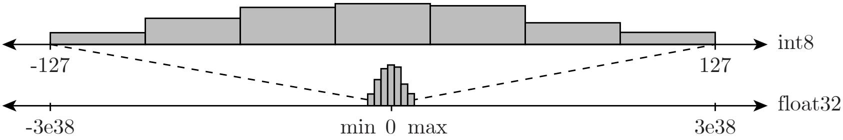
- Same trend: more parameters lead to higher testset accuracies
- Model studies: number of parameters scale on a linear basis



## Implementation Concept

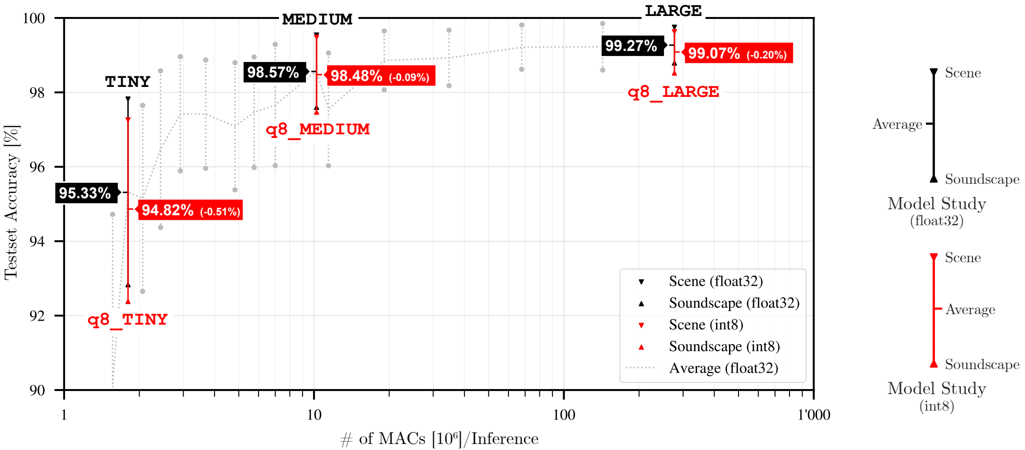
- Post-quantization of `float32` model into `int8` model with `TFLite`
- Approximation of `float32` values by rescaling and shifting:

$$real\_value = (int8\_value - zero\_point) \times scale$$



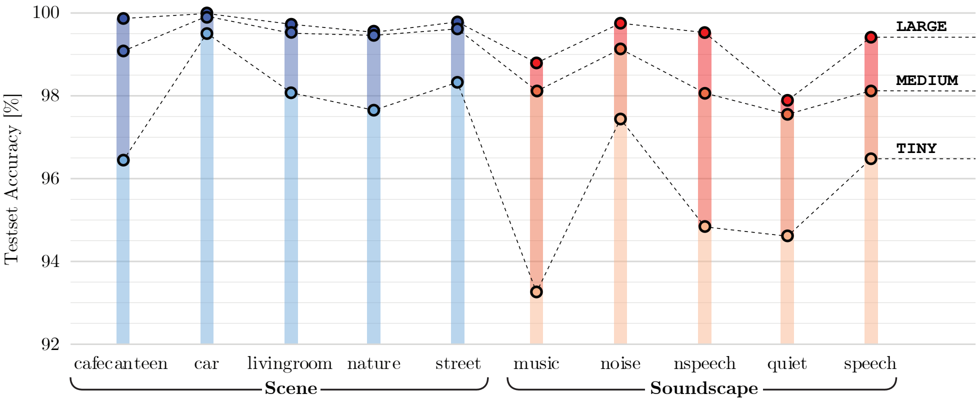
## Post-quantization

- 8-bit quantization applied to all three model studies
- Accuracy loss between [-0.56%, -0.06%] in worst/best case (label-avg)



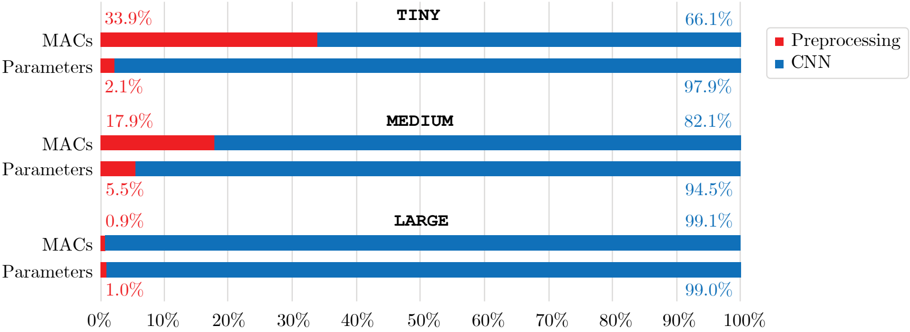
## Class-wise Results

- Larger models outperform smaller models in every class
- Not difficult to detect: `car` and `noise`
- Difficult to detect: `music`



## Partial Complexity Analysis

- CNN model requires the most computational effort
- Conclusion: preprocessing does not grow linearly to CNN size



## Inference Performance

| Study | Required Memory | Clock Speed (min) |
|---|---|---|
| `TINY` | 0.8 MBit | 7.2 MHz |
| `MEDIUM` | 4.8 MBit | 41.0 MHz |
| `LARGE` | 13.8 MBit | 1.1 GHz |

- Security margin of factor 2x for clock speed calculations
- Final conclusions:
  - `MEDIUM`: realistic compromise between accuracy and throughput
  - `TINY`: 5x less computational effort result in $\approx 3\%$ loss of accuracy
  - `LARGE`: 25x more computational effort result in $\approx 1\%$ gain of accuracy
  - Increase clock speed by factor 2x on hearing aids: 5 MHz $\rightarrow$ 10 MHz (`TINY`)

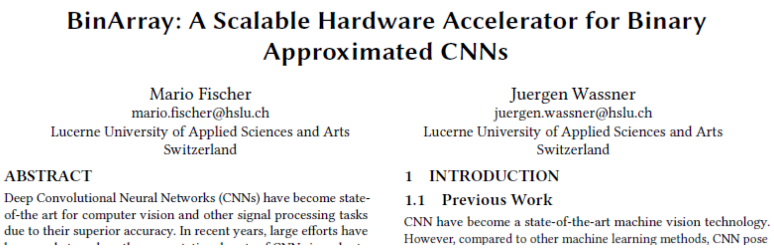
## Outlook

- Optimization: evolutionary algorithm, quantization-aware training
- Dataset: more labels/classes, new recordings with actual hearing aid (characteristic)
- Implementation: portation of model to hearing aid

# Demonstrator

# What's next ?

- Publication for International FPGA'2021 Symposium



**BinArray: A Scalable Hardware Accelerator for Binary Approximated CNNs**

Mario Fischer
mario.fischer@hslu.ch
Lucerne University of Applied Sciences and Arts
Switzerland

Juergen Wassner
juergen.wassner@hslu.ch
Lucerne University of Applied Sciences and Arts
Switzerland

ABSTRACT

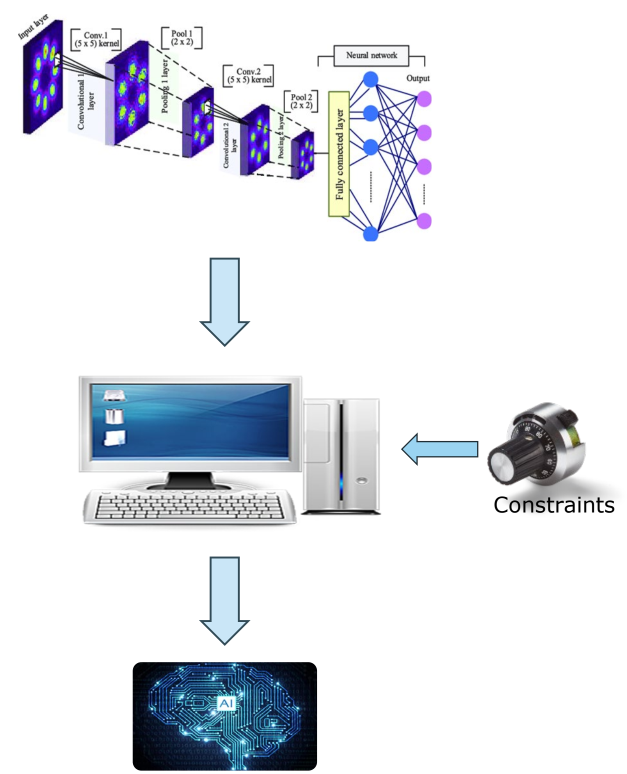
Deep Convolutional Neural Networks (CNNs) have become state-of-the art for computer vision and other signal processing tasks due to their superior accuracy. In recent years, large efforts have

1 INTRODUCTION

1.1 Previous Work

CNN have become a state-of-the-art machine vision technology. However, compared to other machine learning methods, CNN pose

- Edge-AI Framework
  - Hardware–aware network architecture optimization [5]
  - Compilation-based end-to-end work flow
  - Integration with heterogeneous HW/SW platform




- External R&D projects
  - 1-D Signals: Intelligent Nose
  - 2-D Signals: Vision in Space

# Sources

[1] J. Stewart. «Self-Driving Cars Use Crazy Amounts of Power, and It's Becoming a Problem».
Online (06.06.2020): https://www.wired.com/story/self-driving-cars-power-consumption-nvidia-chip/

[2] N. Jones. «How to stop data centres from gobbling up the world's electricity»
Online (06.06.2020): https://www.nature.com/articles/d41586-018-06610-y

[3] V. Sze. «Efficient Processing of Deep Neural Networks:from Algorithms to Hardware Architectures».
Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019).
Online (08.09.2020): http://eyeriss.mit.edu/2019_neurips_tutorial.pdf

[4] F. Johner, J. Wassner. «Efficient Evolutionary Architecture Search for CNN Optimization on GTSRB».
18th IEEE Conference on Machine Learning Applications (ICMLA), Dec 16-19 2019.
Online (06.06.2020): https://ieeexplore.ieee.org/document/8999305

[5] Liu, Y., Sun, Y., Xue, B., Zhang, M., & Yen, G.
A Survey on Evolutionary Neural Architecture Search. 2020.
Online (09.09.20) http://arxiv.org/abs/2008.10937

[5] M. Kurmann. «Optimization of neural networks for FPGA implementation».
MSE Project, Hochschule Luzern Technik & Architektur, Februar 2020.

[6] M. Fischer. «Hardware-friendly Weight Encoding of Convolutional Neural Networks».
MSE Project, Hochschule Luzern Technik & Architektur, Februar 2019.

[7] M. Fischer. «A Scalable Hardware Architecture for Binary Approximated Weights».
MSE Master Thesis, Hochschule Luzern Technik & Architektur, Februar 2020.
Online (06.06.2020): https://portfoliodb.hslu.ch/entries/34786555-7dec-4b37-9731-c8dc2318550c

[8] S. Emmenegger. «Acoustic Scene and Room Classification for Real-Time Applications».
MSE Master Thesis, Hochschule Luzern Technik & Architektur, July 2020 (in preparation)

[9] Xilinx Inc. «Adaptable and Real-Time AI Inference Acceleration»
Online (06.06.2020): https://www.xilinx.com/products/design-tools/vitis/vitis-ai.html