tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

“tinyML: Designing Efficient Neural Architectures and Scaling Strategies for Edge Computing”

Francesco Paissan – Junior Researcher, Fondazione Bruno Kessler (FBK)

November 28, 2023

TINY ML

www.tinyML.org

TINY ML TALKS webcast

Thank you, **tinyML Strategic Partners**, for committing to take tinyML to the next Level, together

Executive Strategic Partners

Qualcomm
AI research

Advancing AI research to make efficient AI ubiquitous

Power efficiency
Model design, compression, quantization, algorithms, efficient hardware, software tool

Personalization
Continuous learning, contextual, always-on, privacy-preserved, distributed learning

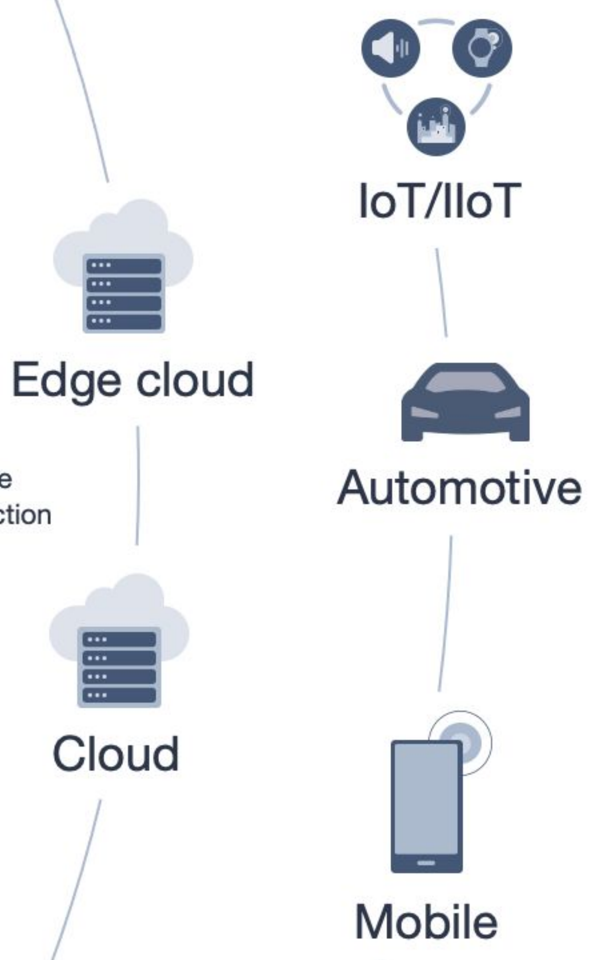
Efficient learning
Robust learning through minimal data, unsupervised learning, on-device learning

A platform to scale AI across the industry

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making



Accelerate Your Edge Compute

SYNTIANT

Making Edge AI A Reality

www.syntiant.com

Platinum Strategic Partner

AITRIOS

DEPLOY VISION AI
AT THE EDGE AT SCALE

Gold Strategic Partners

Build the Future of tinyML

on arm

TINY ML TALKS webcast

EDGE IMPULSE

The Leading Development Platform for Edge ML

edgeimpulse.com

Driving decarbonization and digitalization. Together.

Infineon serving all target markets as
Leader in Power Systems and IoT

www.infineon.com

innatera

NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE

www.innatera.com

Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.

renesas.com

life.augmented

STMicroelectronics provides extensive solutions to make tiny Machine Learning easy

www.st.com/ai

TINY ML TALKS webcast

Silver Strategic Partners

AIZIP

ARDUINO

brainchip

EFFICIENT

GREENWAVES TECHNOLOGIES

Grovety Inc.

Himax

imagimob

Nota AI

NXP

P&G

PIMIC

POLYN TECHNOLOGY

Schneider Electric

SensiML

SILICON LABS

TDK

Join Growing tinyML Communities:

17.7k members in
49 Groups in 41 Countries

tinyML - Enabling ultra-low Power ML at the Edge

https://www.meetup.com/tinyML-Enabling-ultra-low-Power-ML-at-the-Edge/

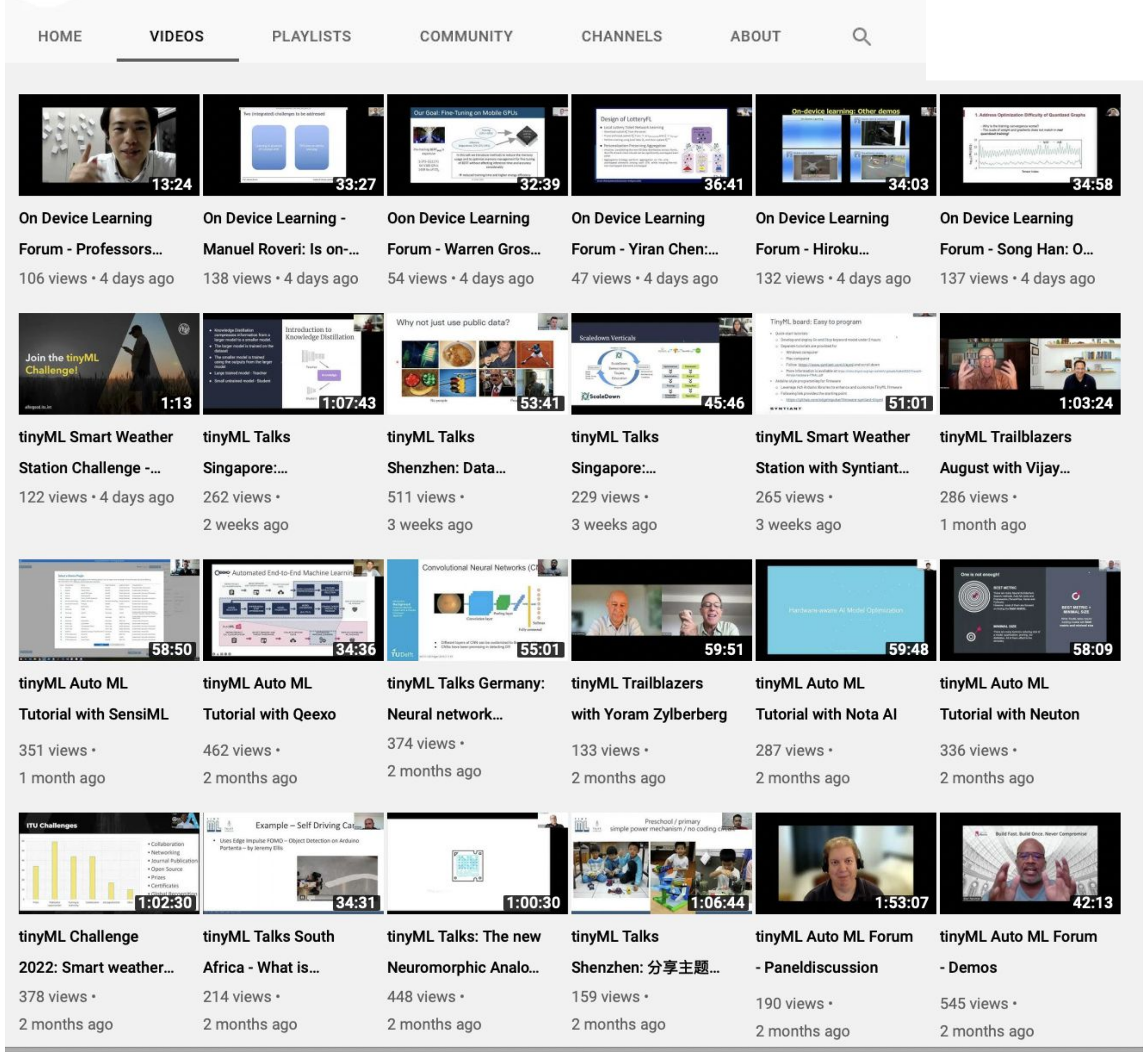
4k members
&
13k followers

The tinyML Community

https://www.linkedin.com/groups/13694488/

TINY ML TALKS webcast

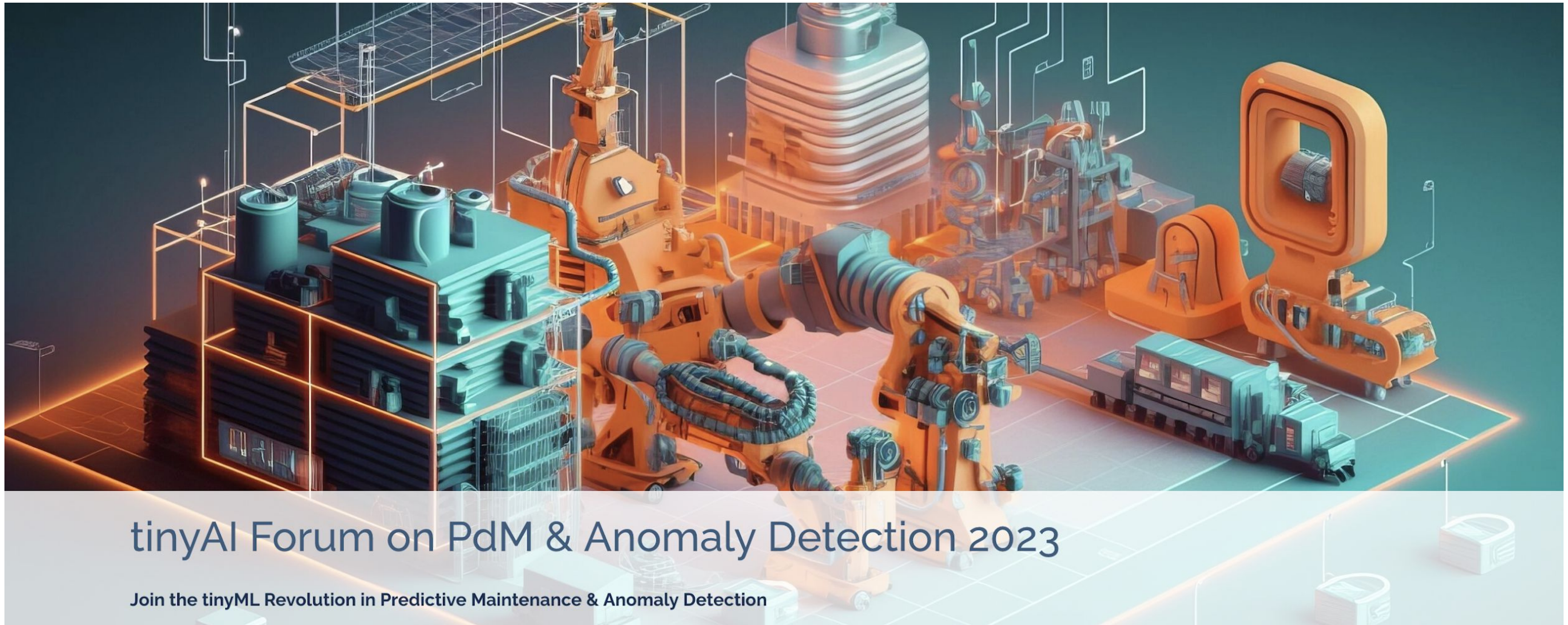
11k subscribers, 633 videos with 391k views



Subscribe to
tinyML YouTube Channel
for updates and notifications
(including this video)

www.youtube.com/tinyML

tinyAI Forum on PdM & Anomaly Detection 2023



Interactive live webinar December 5, 2023 at 8AM Pacific Time
Registration is free of charge

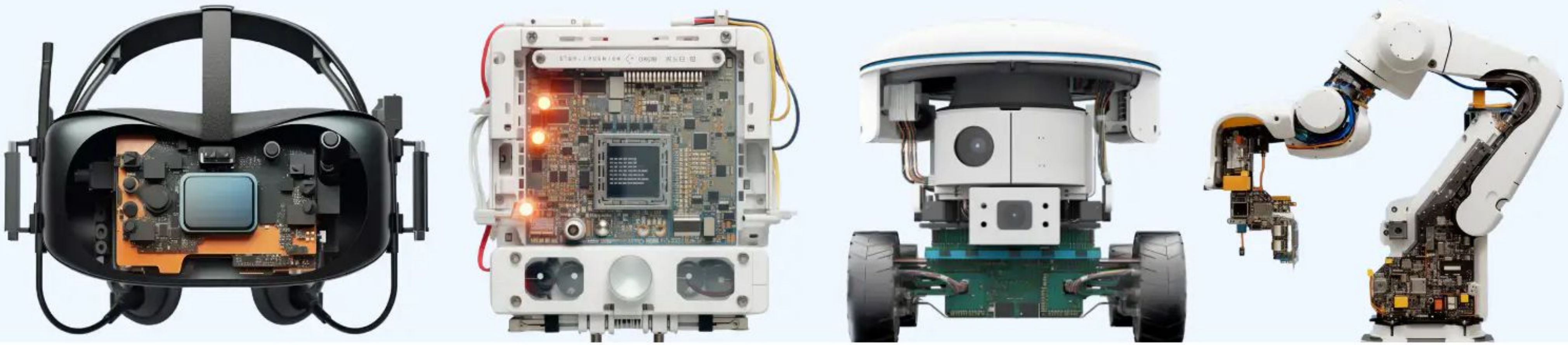
tinyML Research Symposium
April 22, 2023
Call for Papers

tinyML Summit April 23-24, 2024
Call for Presentations and Posters

https://www.wevolver.com/article/2023-edge-ai-technology-report

2023 Edge AI Technology Report

The guide to understanding the state of the art in hardware & software in Edge AI.



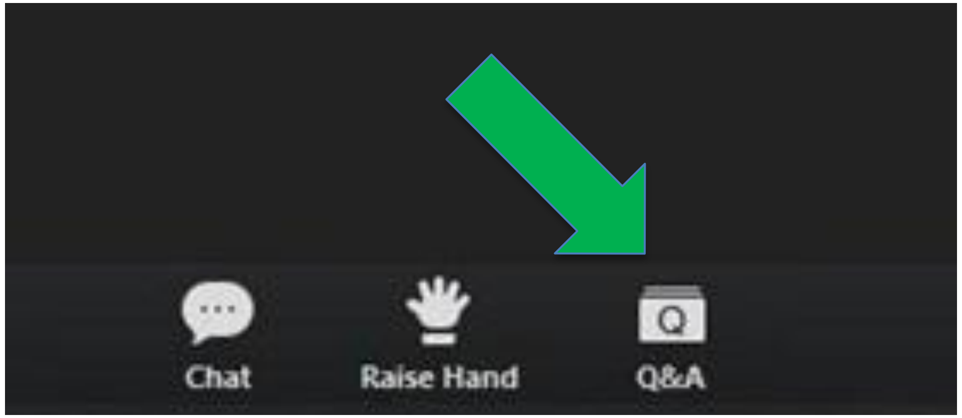
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions



Francesco Paissan

Francesco Paissan has been a Junior Researcher in the Energy Efficient Embedded Digital Architectures (E3DA) unit in Fondazione Bruno Kessler (FBK) since 2018. His research interests include diverse topics, from developing and modelling scalable neural architectures for multimedia analytics to bio-signals analysis with deep learning architectures. In 2021, Francesco joined the LEGEND experiment for the design of novel physics-inspired ML algorithms (e.g. learning-based triggering logics for cosmogenic rejection in the experiment's veto). Francesco was a research intern at the Montreal Institute of Learning Algorithms (Mila) in Montreal, where he worked on post-hoc interpretability techniques for neural networks. WWW speaker: https://francescopaissan.it/

tinyML: Designing Efficient Neural Architectures and Scaling Strategies for Edge Computing

Francesco Paissan

Energy Efficient Embedded Digital Architectures
Fondazione Bruno Kessler

fpaissan@fbk.eu

November 28, 2023

Presentation Overview

Outline

What?

- a fast-growing subfield of machine learning targeting **on-device** and **near-sensor processing**;

The Five (-1) Ws of tinyML

What?

- a fast-growing subfield of machine learning targeting **on-device** and **near-sensor processing**;

Why?

- many practical **benefits** (e.g. bandwidth reduction, infrastructure sustainability, scalability);
- **privacy** by design: enable processing on-device, thus sensitive data is never leaked;

The Five (-1) Ws of tinyML

What?

- a fast-growing subfield of machine learning targeting **on-device** and **near-sensor processing**;

Why?

- many practical **benefits** (e.g. bandwidth reduction, infrastructure sustainability, scalability);
- **privacy** by design: enable processing on-device, thus sensitive data is never leaked;

When?

- not clear, it was a continuous process, sometimes driven by necessity...

Who?

(tiny)AI researchers:

- come up with novel ML algorithms to compress and simplify NN model;
- generally approach tinyML as a ML problem;

Who?

(tiny)AI researchers:

- come up with novel ML algorithms to compress and simplify NN model;
- generally approach tinyML as a ML problem;

(AI)Embedded engineers:

- design custom NN accelerator and neuromorphic processors to speed up NN inference;
- approach tinyML as an engineering problem;

Who?

(tiny)AI researchers:

- come up with novel ML algorithms to compress and simplify NN model;
- generally approach tinyML as a ML problem;

(AI)Embedded engineers:

- design custom NN accelerator and neuromorphic processors to speed up NN inference;
- approach tinyML as an engineering problem;

But there's stuff also in the gray area...

Challenges of tinyML?

WORKSTATION

RAM: 10-100 GB
Storage: 10s of TB
Speed: 100 Billions of ops/s

PC/SBC

RAM: 1-10 GB
Storage: 10-100 GB
Speed: 1-10 Billions of ops/s

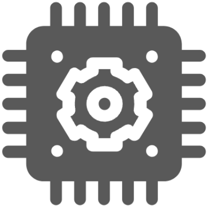
MCU

RAM: 10s - 100s of KBs
Storage: KBs - MBs
Speed: Millions of ops/s

÷10

÷10 000

Target platforms



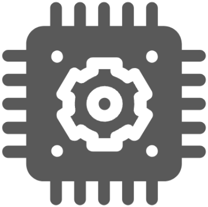
microcontrollers, SBC,

neuromorphic processors, ...

Target platforms

small parameter memory available

(kB - MB)



microcontrollers, SBC,

neuromorphic processors, ...

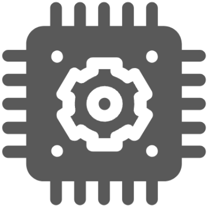
Target platforms

small parameter memory available

(kB - MB)

few operations per second

(million ops/s)



microcontrollers, SBC,

neuromorphic processors, ...

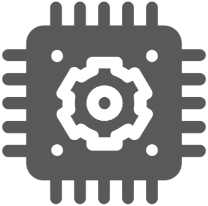
Target platforms

small parameter memory available

(kB - MB)

few operations per second

(million ops/s)



microcontrollers, SBC,

neuromorphic processors, ...

small working memory

(kB - MB)

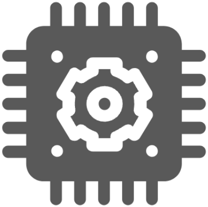
Target platforms

small parameter memory available

(kB - MB)

few operations per second

(million ops/s)



microcontrollers, SBC,

neuromorphic processors, ...

small working memory

(kB - MB)

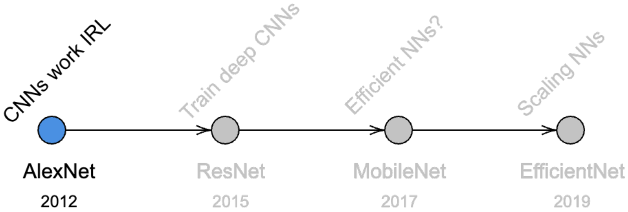
limited operations support

(generally optimized for CNNs)

Outline

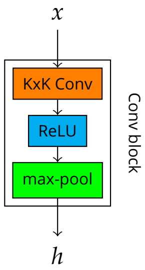
1. Introduction
   - The Five (-1) Ws of tinyML
   - Challenges of tinyML
2. Neural Network design
   - Rise and development of CNNs
   - tinyML-first CNNs
   - Hardware-Aware Scaling
3. Some applications...
   - YOLO-based
   - Zero-shot audio classification
   - micromind

A quick peek at the literature



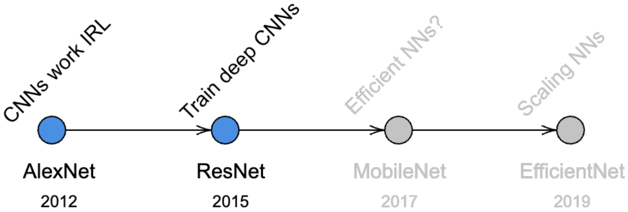
AlexNet

- ground-breaking CNN from 2012 was the first one to get good results on ImageNet;
- composed by a **sequence of convolutional blocks**, with varying configurations;



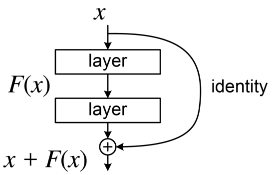
Krizhevsky, Sutskever, and Hinton, "ImageNet classification with deep convolutional neural networks".

A quick peek at the literature



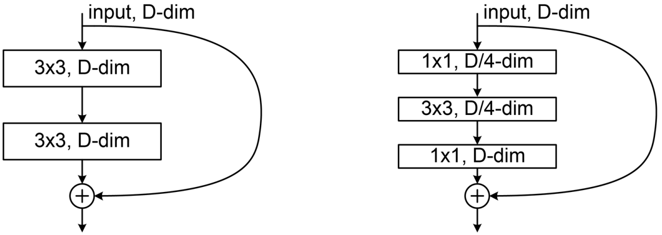
ResNet

- improves the performance by enabling deeper networks via **skip connections**;
- again, is composed by a **sequence of convolutional blocks**, called residual blocks;
- residual blocks follow a wide/narrow/wide structure in the number of channels;

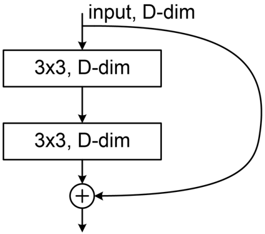


Wightman, Touvron, and J'egou, "ResNet strikes back: An improved training procedure in timm".

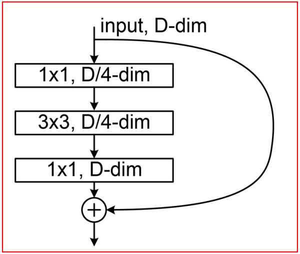
ResBlock variants



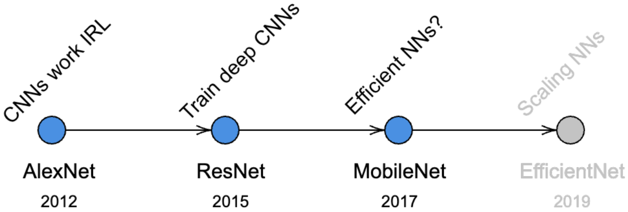
ResBlock variants



Wide-narrow-wide channel structure



A quick peek at the literature



MobileNet

- tries to improve CNN efficiency by proposing the **inverted residual block**;

Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications".

MobileNet

- tries to improve CNN efficiency by proposing the **inverted residual block**;
- differently from a ResBlock, this uses a narrow/wide/narrow structure in the number of channels;

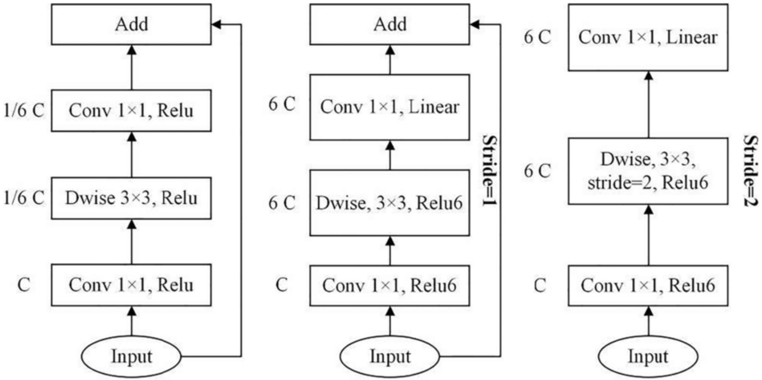
Howard et al., “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”.

MobileNet

- tries to improve CNN efficiency by proposing the **inverted residual block**;
- differently from a ResBlock, this uses a narrow/wide/narrow structure in the number of channels;
- additionally, groups are used inside the convolutions to reduce the computational complexity (depthwise convolutions);

Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications".

Inverted Convolutional Block

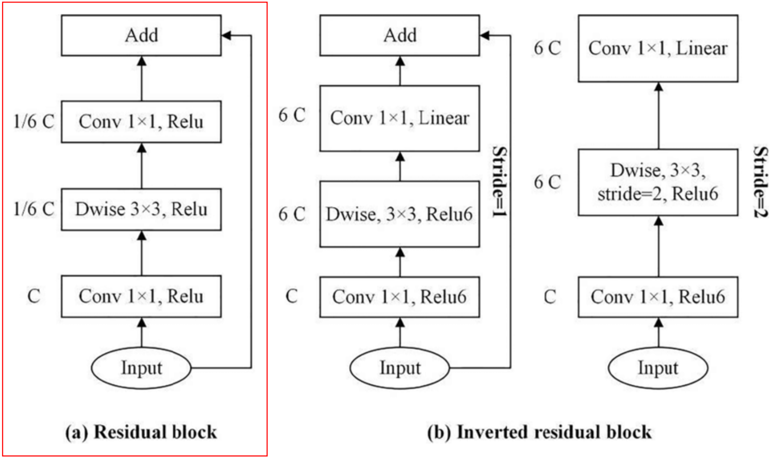


(a) Residual block

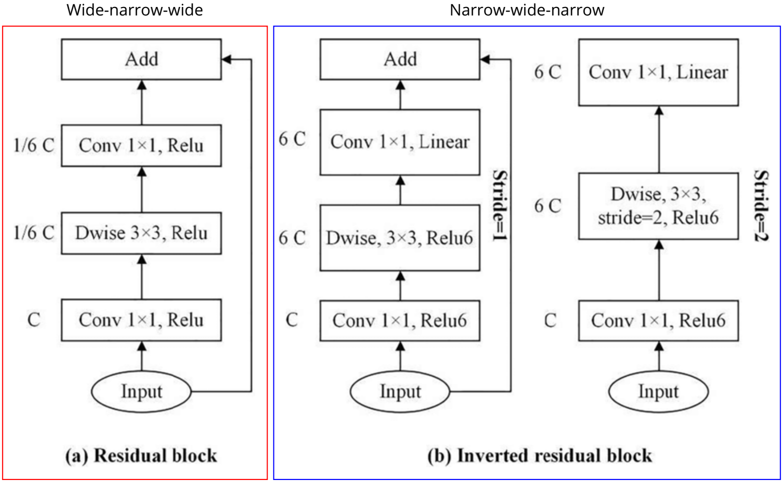
(b) Inverted residual block

Inverted Convolutional Block

Wide-narrow-wide

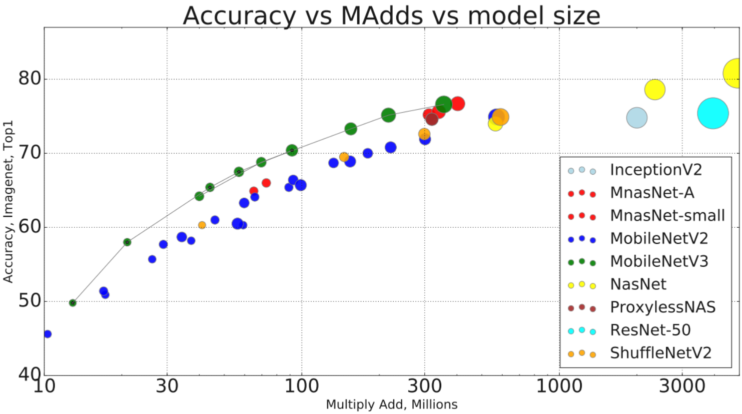


Inverted Convolutional Block

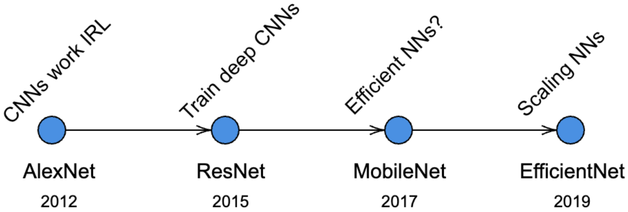


Just for comparison...

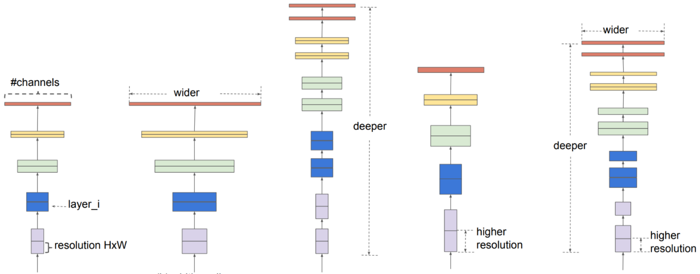
As of MobilNetv3 (Nov. 2019)...



A quick peek at the literature



- focuses on how we ‘should’ be scaling CNNs to obtain optimal performance;
- introduces the concept of compound scaling (i.e. scaling all dimensions is better than one dimension at a time);



Tan and Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”.

Shortcomings of mainstream CNNs

- these neural networks are **too demanding** to run on edge devices and/or compromise performance too much trying to fit;

Shortcomings of mainstream CNNs

- these neural networks are **too demanding** to run on edge devices and/or compromise performance too much trying to fit;
- edge devices have different capabilities conf blocks **cannot exploit**;

Shortcomings of mainstream CNNs

- these neural networks are **too demanding** to run on edge devices and/or compromise performance too much trying to fit;
- edge devices have different capabilities conf blocks **cannot exploit**;
- compound scaling changes all the computational complexities in a **coupled** way;

Ideal CNN for edge processing

- a neural network that can **scale to low computational complexity** ($\leq 1\,\mathrm{MB}$ of FLASH, $\leq 1\,\mathrm{MB}$ of RAM);
- a convolutional block that is designed to **exploit the available resources** maximally;
- a scaling strategy that allows fitting neural networks on **different edge platforms** based on the applications scenarios;

PhiNets

- based on **inverted residual blocks**, modified to decouple the computational resources;

Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

PhiNets

- based on **inverted residual blocks**, modified to decouple the computational resources;
- designed and optimized for **multimedia analytics** at the edge (audio-video);

Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

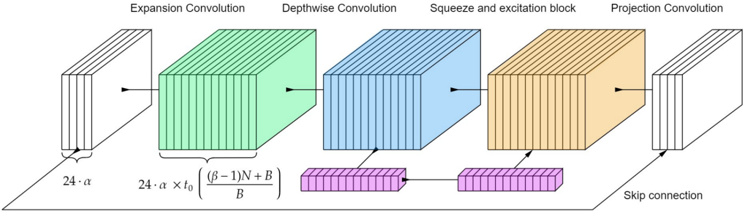
PhiNets

- based on **inverted residual blocks**, modified to decouple the computational resources;
- designed and optimized for **multimedia analytics** at the edge (audio-video);
- controls RAM (t_0), FLASH (β) and operations (α) using three hyperparamters;

Paissan, Ancilotto, and Farella, “PhiNets: A Scalable Backbone for Low-power AI at the Edge”.

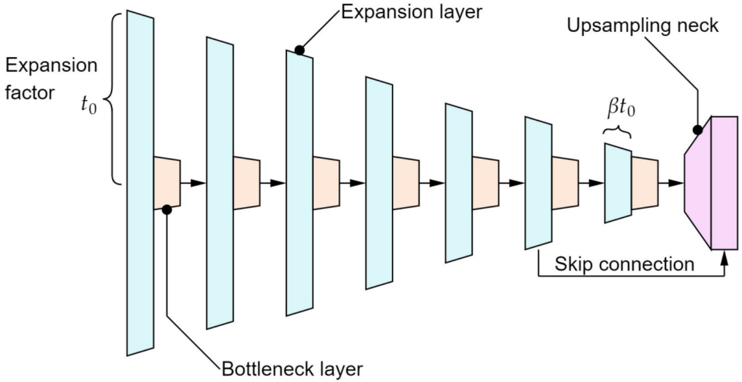
PhiNets convolutional block

Narrow-wide-narrow structure for the number of channels...



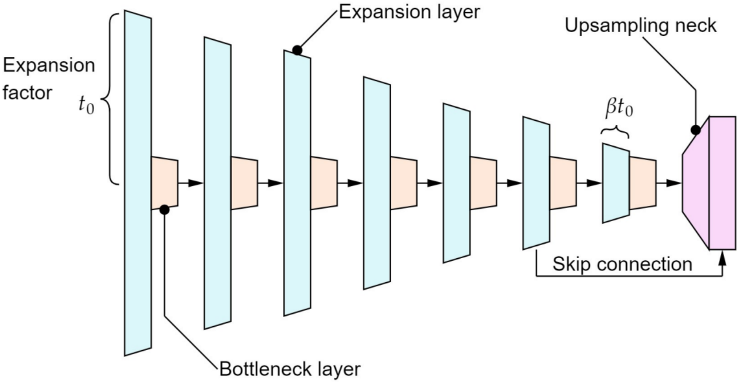
Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

The sequence of PhiNets conv blocks



Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

The sequence of PhiNets conv blocks




```
from micromind.networks import PhiNet
```

Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

Designing an optimized convolutional block

- PhiNets are designed based on **indirect efficiency metrics**, thus could be an ideal version of edge CNNs;

Designing an optimized convolutional block

- PhiNets are designed based on **indirect efficiency metrics**, thus could be an ideal version of edge CNNs;
- what happens if we try to break free of the common standards for convolutional block design and investigate from first principles?

Designing an optimized convolutional block

- PhiNets are designed based on **indirect efficiency metrics**, thus could be an ideal version of edge CNNs;
- what happens if we try to break free of the common standards for convolutional block design and investigate from first principles?

Let's see...

Formal definition of efficiency

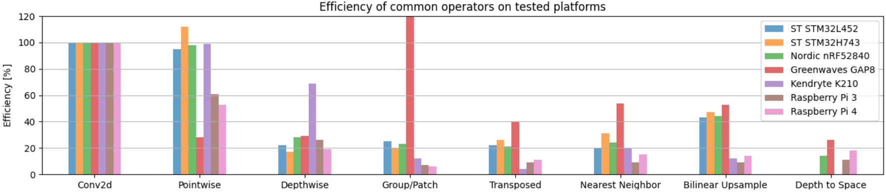
Definition 2.1

We assessed the actual efficiency of each operator (η_{op}) by calculating the ratio between the energy needed for a standard convolution (E_S) and the energy of the chosen operator (E_{op}) to perform an equivalent number of MACs.

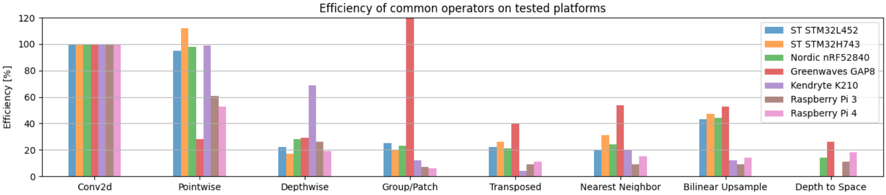
$$\eta_{op} = \frac{E_S}{E_{op}}$$

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

Empirical evaluation of CNN operators...

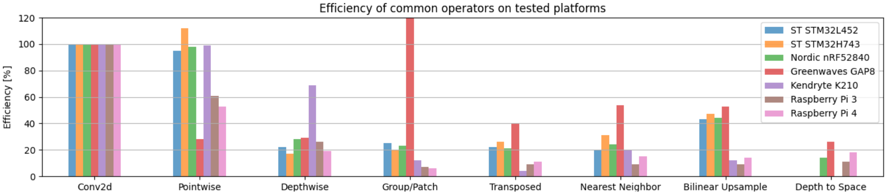


Empirical evaluation of CNN operators...



Ancilotto, Paissan, and Farella, “XiNet: Efficient Neural Networks for tinyML”.

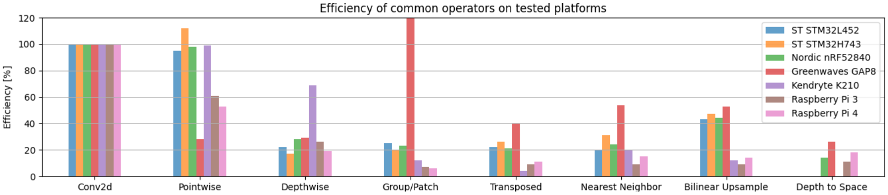
Empirical evaluation of CNN operators...



- this suggests that standard convolutions (AlexNet-style) are, on average, more efficient than other variants;

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

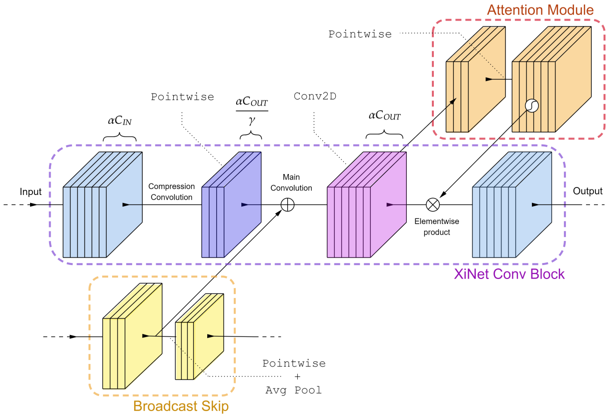
Empirical evaluation of CNN operators...



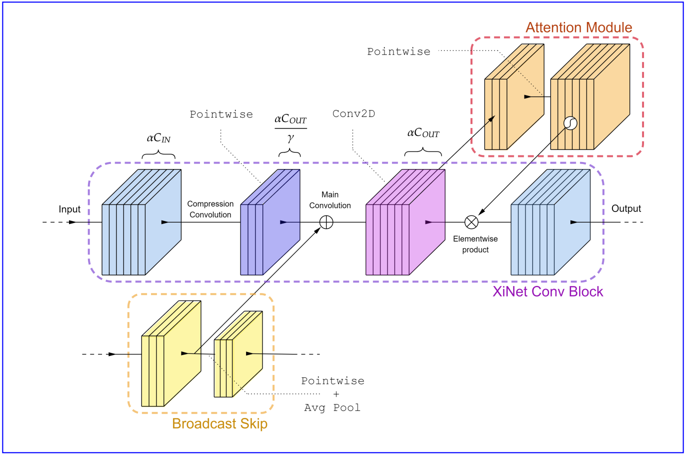
- this suggests that standard convolutions (AlexNet-style) are, on average, more efficient than other variants;
- but how do we exploit them with low parameter memory?

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

XiNet convolutional block

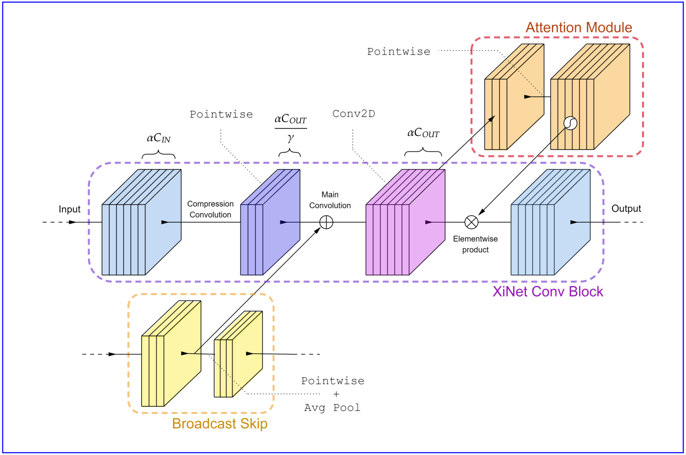


XiNet convolutional block



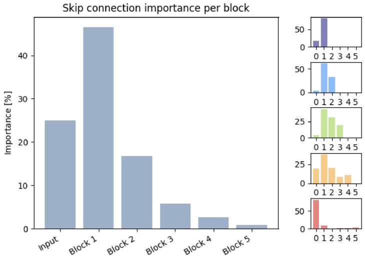
XiNet convolutional block

Wide-narrow-wide structure for channels, and much more...



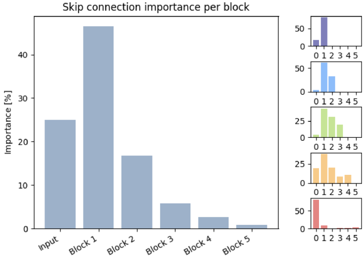
Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

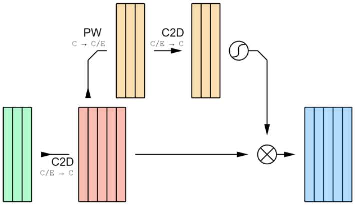
Skip connections and attention block



Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

Skip connections and attention block





Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

XiNet

- composed by a sequence of XiNet convolutional blocks;

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

XiNet

- composed by a sequence of XiNet convolutional blocks;
- similarly to PhiNets, its computational complexity is controlled using **three hyperparameters** (α, γ, β);

Ancilotto, Paissan, and Farella, “XiNet: Efficient Neural Networks for tinyML”.

XiNet

- composed by a sequence of XiNet convolutional blocks;
- similarly to PhiNets, its computational complexity is controlled using **three hyperparameters** (α, γ, β);
- designed based on the **empirical benchmark** of the different operators to be very efficient;

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

XiNet

- composed by a sequence of XiNet convolutional blocks;
- similarly to PhiNets, its computational complexity is controlled using **three hyperparameters** (α, γ, β);
- designed based on the **empirical benchmark** of the different operators to be very efficient;

```
from micromind.networks import XiNet
```

Ancilotto, Paissan, and Farella, "XiNet: Efficient Neural Networks for tinyML".

Hardware-aware scaling

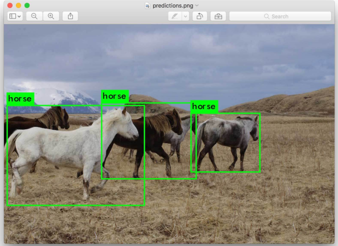
- **scaling strategy** that exploits the advanced PhiNets and XiNet architectures;
- helps deploy CNNs on a wide variety of edge platforms via its one-shot network optimization procedure;
- **inverts the mapping between computational complexity and hyperparameters** so that it can be solved with a mathematical programming toolkit for specific computational requirements;

Paissan, Ancilotto, and Farella, "PhiNets: A Scalable Backbone for Low-power AI at the Edge".

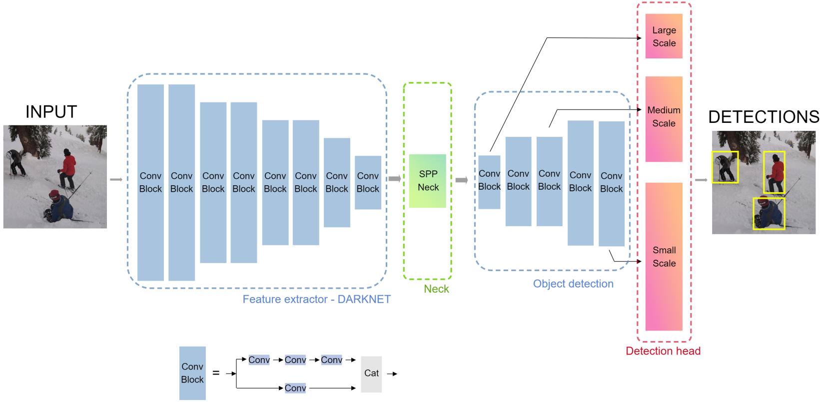
Outline

You Only Look Once (YOLO)

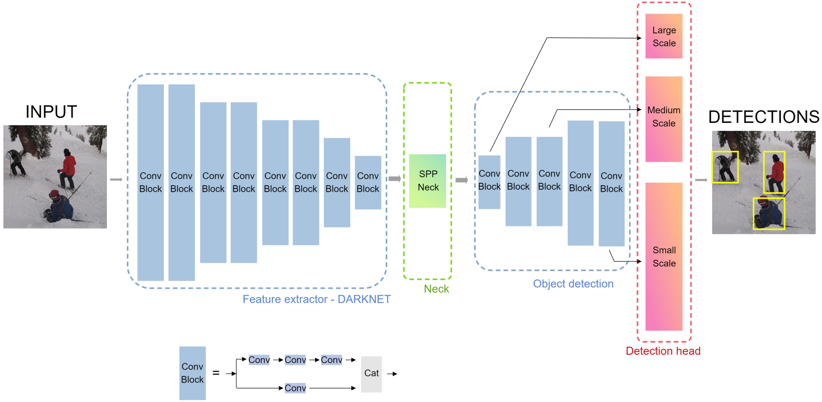
- originally proposed as an object detection pipeline;
- well known for its **good performance/complexity tradeoff**;
- mainly related to its ability to detect objects using **only one inference step** (no region proposal networks, etc...);
- recently extended to support image segmentation, keypoint detection/pose estimation;



YOLO Architecture

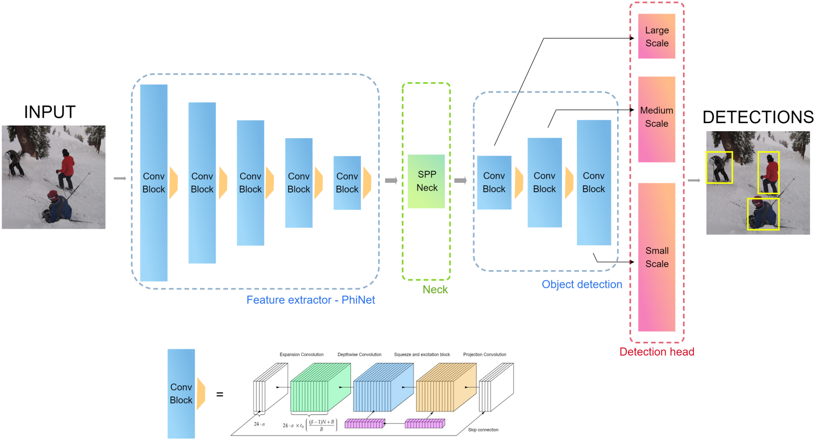


YOLO Architecture

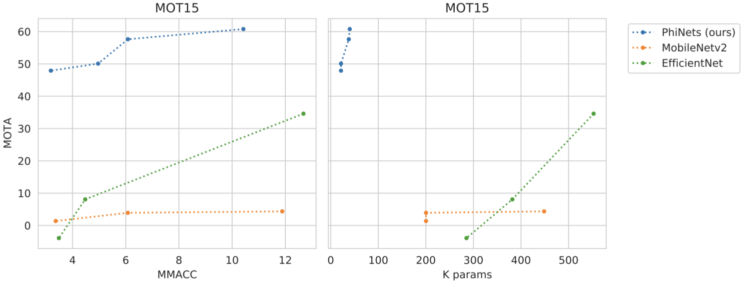


In the literature, some works propose to solve a simplified version of the object detection task; thus, reducing computational complexity... but here is what we do:

YOLOPhiNet


INPUT
Conv Block
Conv Block
Conv Block
Conv Block
Conv Block
Feature extractor - PhiNet
SPP Neck
Neck
Conv Block
Conv Block
Conv Block
Object detection
Large Scale
Medium Scale
Small Scale
Detection head
DETECTIONS
Conv Block
=
Expansion Convolution
Depthwise Convolution
Squeeze and excitation block
Projection Convolution
Skip connection

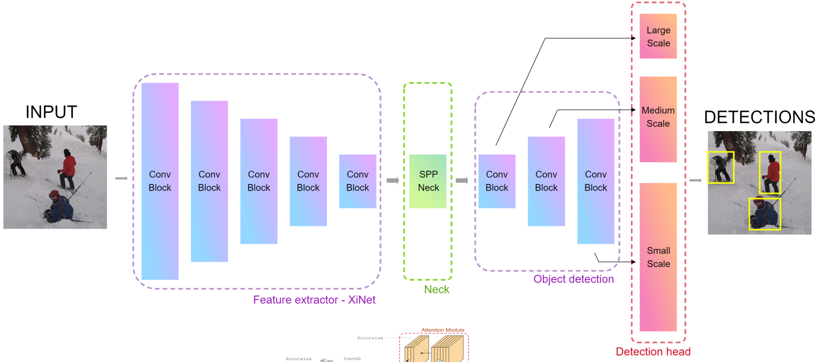
YOLOPhinet

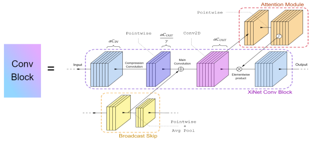


Deployed on an Arm-Cortex M7 MCU with $2\,\mathrm{MB}$ of internal Flash and $1\,\mathrm{MB}$ of RAM; achieves **power requirements in the order of** $10\,\mathrm{mW}$ @ 52% mAP on VOC2012.

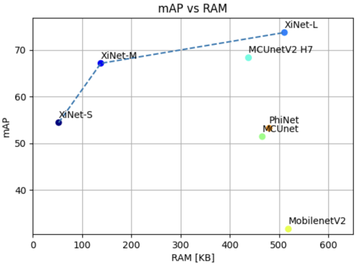
`micromind/recipes/object_detection`

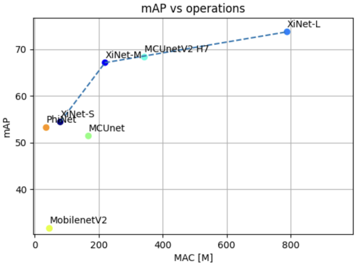
YOLOXiNet


INPUT
Conv Block
Conv Block
Conv Block
Conv Block
Conv Block
Feature extractor - XiNet
SPP Neck
Neck
Conv Block
Conv Block
Conv Block
Object detection
Large Scale
Medium Scale
Small Scale
Detection head
DETECTIONS


Conv Block
=
Attention Module
Pointwise
Conv2D
Input
Compression Convolution
Main Convolution
Elementwise product
Output
XiNet Conv Block
Broadcast Skip
Pointwise
+
Avg Pool

YOLOXiNet





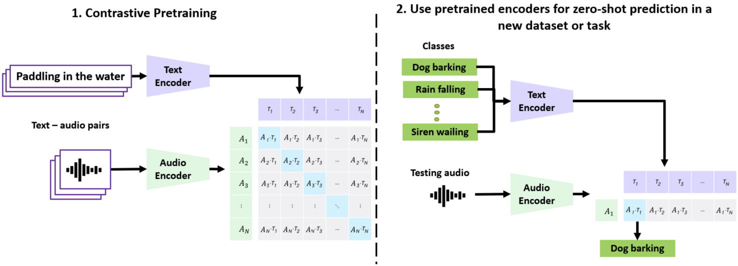
Deployed on an Arm-Cortex M7 MCU with $2\,\mathrm{MB}$ of internal Flash and $1\,\mathrm{MB}$ of RAM; Achieves a reduction in the **number of operations of 2**$\times$ and a reduction in **RAM usage of 9**$\times$ with respect to MCUNet, with the same performance. Achieves a **power consumption of around** $20\,\mathrm{mW}$ @ 67% mAP on VOC2012.

micromind/recipes/object_detection

Contrastive Language-Audio pretraining

- learns a **similarity score** between two modalities (audio and text);
- can be exploited for **zero-shot** classification;
- makes the network very **flexible** wrt the applications scenario they can be deployed to;

Zero-shot classification



- exploits the learned similarity score to learn a **more efficient audio network** (via a distillation process);

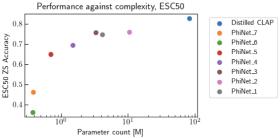
- exploits the learned similarity score to learn a **more efficient audio network** (via a distillation process);
- assumes the pre-trained **text encoder** does **not** need to be **deployed**;

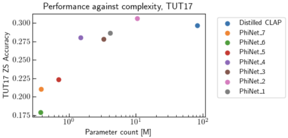
tinyCLAP

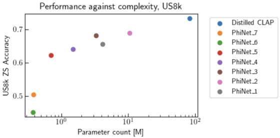
- exploits the learned similarity score to learn a **more efficient audio network** (via a distillation process);
- assumes the pre-trained **text encoder** does **not** need to be **deployed**;
- achieves good performance-complexity tradeoff for ZS classification, and state-of-the-art for a benchmark;

`micromind/recipes/tinyCLAP`

tinyCLAP: performance



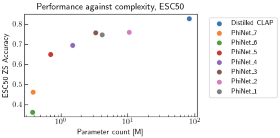


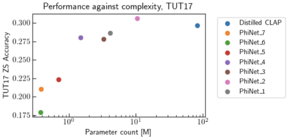


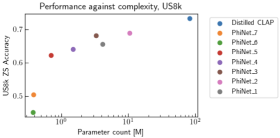
- follows a common power-law scaling behaviour;

Paissan and Farella, "tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models".

tinyCLAP: performance



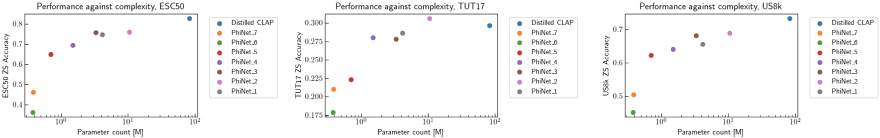




- follows a common power-law scaling behaviour;
- was not yet deployed on edge platforms (WIP);

Paissan and Farella, "tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models".

tinyCLAP: performance



- follows a common power-law scaling behaviour;
- was not yet deployed on edge platforms (WIP);
- **94% reduction in parameter count** wrt to original CLAP (from 82M to 4M), with a minor ZS accuracy drop (4% averaged on all benchmarks);

Paissan and Farella, "tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models".

micromind: tinyML research made simple

- not a startup or a research project, just an **open-source project** for tinyML research;
- tries to provide the **full research pipeline** for model design, development, and deployment;

Checkout the project on GitHub and leave a star!

Follow me on X `@fpaissan_` for updates.

Additional references to our works

Following is a list of references to works related to the topics discussed in the presentation:

- Video processing: Ancilotto, Paissan, and Farella, “On the Role of Smart Vision Sensors in Energy-Efficient Computer Vision at the Edge”; Paissan, Ancilotto, and Farella, “PhiNets: A Scalable Backbone for Low-power AI at the Edge”; Ancilotto, Paissan, and Farella, “XiNet: Efficient Neural Networks for tinyML”
- Generative modeling: Ancilotto, Paissan, and Farella, “PhiNet-GAN: Bringing real-time face swapping to embedded devices”; Ancilotto, Paissan, and Farella, “XimSwap: many-to-many face swapping for TinyML”
- Audio processing: Paissan et al., “Scalable Neural Architectures for End-to-End Environmental Sound Classification”; Brutti et al., “Optimizing PhiNet architectures for the detection of urban sounds on low-end devices”; Ali et al., “Scaling strategies for on-device low-complexity source separation with Conv-Tasnet”; Paissan et al., “Improving latency performance trade-off in keyword spotting applications at the edge”
- Multimodal processing: Paissan and Farella, “tinyCLAP: Distilling Constrastive Language-Audio Pretrained Models”

Ali, Mohamed Nabih et al. "Scaling strategies for on-device low-complexity source separation with Conv-Tasnet". In: *ArXiv* abs/2303.03005 (2023). URL: https://api.semanticscholar.org/CorpusID:257364800.

Ancilotto, A., F. Paissan, and Elisabetta Farella. "XiNet: Efficient Neural Networks for tinyML". In: *ICCV2023* (2023). URL: https://openaccess.thecvf.com/content/ICCV2023/papers/Ancilotto_XiNet_Efficient_Neural_Networks_for_tinyML_ICCV_2023_paper.pdf.

Ancilotto, Alberto, Francesco Paissan, and Elisabetta Farella. "On the Role of Smart Vision Sensors in Energy-Efficient Computer Vision at the Edge". In: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (2022), pp. 497–502. URL: https://api.semanticscholar.org/CorpusID:248546511.

— ."PhiNet-GAN: Bringing real-time face swapping to embedded devices". In: *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other*

Copyright Notice

www.tinyml.org

Copyright Notice

www.tinyml.org