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

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

November 28, 2023
Thank you, tinyML Strategic Partners, for committing to take tinyML to the next Level, together.
Executive Strategic Partners
Advancing AI research to make efficient AI ubiquitous

A platform to scale AI across the industry
Accelerate Your Edge Compute
SYNTIANT
Making Edge AI A Reality
www.syntiant.com
Platinum Strategic Partner
DEPLOY VISION AI AT THE EDGE AT SCALE
Gold Strategic Partners
Build the Future of tinyML on ARM
The Leading Development Platform for Edge ML

dgeimpulse.com
Driving decarbonization and digitalization. Together.

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

www.infineon.com
NEUROMORPHIC INTELLIGENCE FOR THE SENSOR-EDGE
Renesas is enabling the next generation of AI-powered solutions that will revolutionize every industry sector.
STMicroelectronics provides extensive solutions to make tiny Machine Learning easy
ENGINEERING EXCEPTIONAL EXPERIENCES

We engineer exceptional experiences for consumers in the home, at work, in the car, or on the go.

www.synaptics.com
Join Growing tinyML Communities:

tinyML - Enabling ultra-low Power ML at the Edge

The tinyML Community
https://www.linkedin.com/groups/13694488/

17.7k members in
49 Groups in 41 Countries

4k members
&
13k followers
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

The tinyML research symposium serves as a flagship venue for related research at the intersection of machine learning applications, algorithms, software, and hardware in deeply embedded machine learning systems.

Call for Papers
tinyML Summit April 23-24, 2024
Call for Presentations and Posters
2023 Edge AI Technology Report

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

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

Slides & Videos will be posted tomorrow

tinyml.org/forums  youtube.com/tinyml

Please use the Q&A window for your questions
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
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
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
The Five (-1) Ws of tinyML

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...
(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 in the gray area...
(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;
(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, ...

Francesco Paissan (FBK)  
November 28, 2023
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, ...

Francesco Paissan (FBK) November 28, 2023
Target platforms

small parameter memory available
(kB - MB)

few operations per second
(million ops/s)

small working memory
(kB - MB)

microcontrollers, SBC,
neuromorphic processors, ...
Target platforms

- Microcontrollers, SBC, neuromorphic processors, ...
- Small parameter memory available (kB - MB)
- Few operations per second (million ops/s)
- 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
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;


Francesco Paissan (FBK)  
November 28, 2023
A quick peek at the literature

- CNNs work IRL
- Train deep CNNs
- Efficient NNs?
- Scaling NNs

- AlexNet: 2012
- ResNet: 2015
- MobileNet: 2017
- EfficientNet: 2019
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

Wide-narrow-wide channel structure
ResBlock variants

Wide-narrow-wide channel structure
A quick peek at the literature

- CNNs work IRL
- Train deep CNNs
- Efficient NNs?
- Scaling NNs

AlexNet 2012
ResNet 2015
MobileNet 2017
EfficientNet 2019
MobileNet

• tries to improve CNN efficiency by proposing the **inverted residual block**;
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”.
• 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”.

Francesco Paissan (FBK)
Inverted Convolutional Block

(a) Residual block

(b) Inverted residual block
Inverted Convolutional Block

Wide-narrow-wide

(a) Residual block

(b) Inverted residual block
Inverted Convolutional Block

Wide-narrow-wide

Input

\[ \text{Add} \]

\[ \frac{1}{6} \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Relu} \]

\[ \frac{1}{6} \text{C} \]

\[ \text{Dwise} \ 3 \times 3, \ \text{Relu} \]

\[ \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Relu} \]

(a) Residual block

Narrow-wide-narrow

Input

\[ \text{Add} \]

\[ 6 \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Linear} \]

\[ 6 \text{C} \]

\[ \text{Dwise} \ 3 \times 3, \ \text{Relu}6 \]

\[ 6 \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Relu}6 \]

\[ 6 \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Linear} \]

\[ 6 \text{C} \]

\[ \text{Dwise} \ 3 \times 3, \ \text{stride}=2, \ \text{Relu}6 \]

\[ \text{C} \]

\[ \text{Conv} \ 1 \times 1, \ \text{Relu}6 \]

(b) Inverted residual block
Just for comparison...

As of MobilNetv3 (Nov. 2019)...

Accuracy vs MAdds vs model size

<table>
<thead>
<tr>
<th>Model</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>InceptionV2</td>
<td></td>
</tr>
<tr>
<td>MnasNet-A</td>
<td></td>
</tr>
<tr>
<td>MnasNet-small</td>
<td></td>
</tr>
<tr>
<td>MobileNetV2</td>
<td></td>
</tr>
<tr>
<td>MobileNetV3</td>
<td></td>
</tr>
<tr>
<td>NasNet</td>
<td></td>
</tr>
<tr>
<td>ProxylessNAS</td>
<td></td>
</tr>
<tr>
<td>ResNet-50</td>
<td></td>
</tr>
<tr>
<td>ShuffleNetV2</td>
<td></td>
</tr>
</tbody>
</table>
A quick peek at the literature

- CNNs work IRL
- Train deep CNNs
- Efficient NNs?
- Scaling NNs

- AlexNet (2012)
- MobileNet (2017)
- EfficientNet (2019)
EfficientNet

- 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”.

Francesco Paissan (FBK)
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** (≤ 1 MB of FLASH, ≤ 1 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;
PhiNets

- based on **inverted residual blocks**, modified to decouple the computational resources;
- designed and optimized for **multimedia analytics** at the edge (audio-video);
• 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 ($\beta$) and operations ($\alpha$) using three hyperparameters;
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”.

Francesco Paissan (FBK)
The sequence of PhiNets conv blocks

Paissan, Ancilotto, and Farella, “PhiNets: A Scalable Backbone for Low-power AI at the Edge”. Francesco Paissan (FBK) November 28, 2023 24/47
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”.

Francesco Paissan (FBK)
Designing an optimized convolutional block

- PhiNets are designed based on **indirect efficiency metrics**, thus could be an ideal version of edge CNNs;
• 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?
• 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 ($\eta_{op}$) 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.

$$\eta_{op} = \frac{E_S}{E_{op}}$$
Empirical evaluation of CNN operators...

Efficiency of common operators on tested platforms

- ST STM32L452
- ST STM32H743
- Nordic nRF52840
- Greenwaves GAP8
- Kendryte K210
- Raspberry Pi 3
- Raspberry Pi 4
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”.

Francesco Paissan (FBK)  November 28, 2023  28/47
• this suggests that standard convolutions (AlexNet-style) are, on average, more efficient than other variants;
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”. Francesco Paissan (FBK) November 28, 2023 28 / 47
XiNet convolutional block

![Diagram of XiNet convolutional block]

- Input
- Compression Convolution
- Pointwise
- Conv2D
- Elementwise product
- Output
- Attention Module
- Broadcast Skip
- Avg Pool

XiNet Conv Block

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

Francesco Paissan (FBK)
November 28, 2023
29 / 47
XiNet convolutional block

[Diagram of XiNet convolutional block]

- **Input**
- **Compression Convolution**
- **Pointwise**
- **Conv2D**
- **Pointwise**
- **Elementwise product**
- **Output**

- **Attention Module**
- **Broadcast Skip**
- **Avg Pool**

---

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

Francesco Paissan (FBK)

November 28, 2023

29 / 47
XiNet convolutional block

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

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

Francesco Paissan (FBK)
Skip connections and attention block

Ancilotto, Paissan, and Farella, “XiNet: Efficient Neural Networks for tinyML”. Francesco Paissan (FBK)
Skip connections and attention block

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

Francesco Paissan (FBK)
• composed by a sequence of XiNet convolutional blocks;
• composed by a sequence of XiNet convolutional blocks;
• similarly to PhiNets, its computational complexity is controlled using **three hyperparameters** \((\alpha, \gamma, \beta)\);
• composed by a sequence of XiNet convolutional blocks;
• similarly to PhiNets, its computational complexity is controlled using three hyperparameters \((\alpha, \gamma, \beta)\);
• designed based on the empirical benchmark of the different operators to be very efficient;
XiNet

- composed by a sequence of XiNet convolutional blocks;
- similarly to PhiNets, its computational complexity is controlled using three hyperparameters $(\alpha, \gamma, \beta)$;
- designed based on the empirical benchmark of the different operators to be very efficient;

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

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

Francesco Paissan (FBK)
November 28, 2023
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:
Deployed on an Arm-Cortex M7 MCU with 2 MB of internal Flash and 1 MB of RAM; achieves **power requirements in the order of** 10 mW @ 52% mAP on VOC2012.
Deployed on an Arm-Cortex M7 MCU with 2 MB of internal Flash and 1 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 $20 \text{mW} @ 67\% \text{mAP}$ on VOC2012.**
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

1. Contrastive Pretraining

Text – audio pairs

2. Use pretrained encoders for zero-shot prediction in a new dataset or task

Classes

- Dog barking
- Rain falling
- Siren wailing

Testing audio

Audio Encoder

Dog barking
• 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**;
• 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;
tinyCLAP: performance

- follows a common power-law scaling behaviour;
tinyCLAP: performance

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

Paissan and Farella, “tinyCLAP: Distilling Contrastive Language-Audio Pretrained Models”.

Francesco Paissan (FBK)
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:


- **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”


—. “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

This multimedia file is copyright © 2023 by tinyML Foundation. All rights reserved. It may not be duplicated or distributed in any form without prior written approval.

tinyML® is a registered trademark of the tinyML Foundation.

www.tinymce.org
Copyright Notice

This presentation in this publication was presented as a tinyML® Talks webcast. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org