“Advancing Medical Imaging Analysis with Multi-task and Hardware-Efficient Neural Architecture Search”

Hadjer Benmeziane – Visiting Researcher, IBM Research Europe

January 23, 2024
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

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

Perception
Object detection, speech recognition, contextual fusion

Reasoning
Scene understanding, language understanding, behavior prediction

Action
Reinforcement learning for decision making

A platform to scale AI across the industry
Accelerate Your Edge Compute

Syntiant

Making Edge AI A Reality

www.syntiant.com
Platinum Strategic Partners
DEPLOY VISION AI
AT THE EDGE AT SCALE
Gold Strategic Partners
Build the Future of tinyML on ARM
EDGE IMPULSE

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 EXCELLENT 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:

**Meetup**

*tinyML - Enabling ultra-low Power ML at the Edge*

**LinkedIn**

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

18.9k members in 49 Groups in 41 Countries
4.2k members & 14.5k followers
Subscribe to tinyML YouTube Channel for updates and notifications (including this video)
www.youtube.com/tinyML
tinyML Research Symposium
April 22, 2024
Call for Papers

Research Symposium - April 22, 2024

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
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
Dr. Hadjer Benmeziane is a visiting researcher at IBM Research Europe, specializing in hardware-aware neural architecture search for emerging AI accelerators such as analog in-memory computing. She received her PhD from Université Polytechnique des Hauts-de-France in August 2023, following her Master's and Engineering degree in Computer Science from Ecole Supérieure d'Informatique, Algiers, Algeria. Her work on Analog Neural Architecture Search received the prestigious IEEE open source science award and best paper award at IEEE Services Computing 2023 Symposium. Her research focuses on making hardware-aware neural architecture search more efficient, flexible and practical.
Advancing Medical Imaging Analysis with Multi-task and Hardware-Efficient Neural Architecture Search

Hadjer Benmeziane

23/01/2024
Advancing Medical Imaging Analysis with Multi-task and Hardware-efficient Neural Architecture Search

1. Context & Motivation

2. Background:
   - Medical Imaging Deep learning Architectures
   - Hardware-aware Neural Architecture Search

1. MED-NAS-Bench:
   - a. Benchmark design
   - b. Performance distribution
   - c. Cross-dataset performance analysis

2. Multi-task HW-NAS:
   - a. Evaluating multi-task ability
   - b. Initial Results

3. Other Hardware Platforms Potential

4. Conclusion and Future Perspectives
Context & Motivation

Medical Imaging Segmentation

(Isensee, Fabian, et al. 2021)

Generative AI for Medical Tasks

(Zhang, Peng, et al. 2023)

Medical Tasks on Wearable devices

(Banerjee, Amit, et al. 2020)
Context & Motivation

- A high number of deep learning algorithms are now FDA-approved.
- Many are targeting the fields of Radiology, Cardiology and Internal Medicine/General.
- Growing need in automating the design the these algorithms to enable multi-task benefits.

List of FDA-approved Medical Algorithms

(Benjamens, Stan, et al. 2020)
Context & motivation

★ Deep Learning models are more accurate and automate many radiologist tasks.

☐ Urgent need for efficient AI solutions in fast-paced hospital environments.

☐ Current U-net like architectures are slow and large.

☐ Other constraints are more critical in medical tasks such as robustness, certainty, interpretability.
Background: Deep Learning Architectures

U-Net Architecture
(Long, Shelhamer, et al. 2014)
Background: Hardware-aware Neural Architecture Search

- **Architecture Search Space**
  - Conv
  - Pooling
  - Activations
  - ...

- **Hardware Search Space**
  - Tiling parameters
  - Number of PE Templates
  - ...

- **Search Space**
  - [May be optimized]

- **Search Algorithm**
  - RL
  - EA
  - Gradient-based
  - ...

- **Search Strategy**

- **Evaluate the accuracy**
  - Proxy
  - Early stopping
  - Accuracy Predictors
  - ...

- **HW Cost Measurements**
  - Real-time measurements
  - Performance Predictor
  - ...

- **Additional Optimization**
  - [Optional]
  - Found in all HW-NAS

- **Best Model**
  - Transfer the model
  - Quantization
  - Pruning
Medical Imaging Analysis with HW-NAS
MED-NAS-Bench
Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

- The absence of standardized benchmarks for HW-NAS in medical imaging makes it difficult to compare and validate the effectiveness of different approaches.
- Medical imaging data often requires specialized preprocessing, such as noise reduction or contrast enhancement, which can influence the effectiveness of the derived architectures.
- Ensuring generalization across diverse patient populations and imaging conditions is challenging, given the high variability in medical images.

Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

We propose a NAS benchmark for medical AI imaging analysis.

- The benchmark targets 11 tasks.
  - Medical Segmentation Decathlon\(^1\)
  - NIH Chest X-Ray Dataset\(^2\)

- The benchmark architectures are represented within a *supernetwork*.

- The supernetwork is a *U-Net like* representation.

- Weight-sharing methodology is used to estimate the performance scores.

---


Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

- We create a supernetwork per task
- Each supernetwork is based on the U-Net architecture
- The supernetwork is created with a recursive implementation of U-Net.
Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

- Multiple blocks are possible per task.
- Each operation is associated with its recursive upsampling option.
- We use different 2D and 3D convolutions.

<table>
<thead>
<tr>
<th>Block</th>
<th>Operations</th>
<th>Respective Upsampling block</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Identity</td>
<td>IdentityLayer</td>
<td>IdentityLayer</td>
</tr>
<tr>
<td>A</td>
<td>LinearLayer(in, out)</td>
<td>LinearLayer(out, in)</td>
</tr>
<tr>
<td>B</td>
<td>2DConv(in, out, k=3, use_bn=false, act=relu)</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>C</td>
<td>2DConv(in, out, k=3, use_bn=true, act=relu)</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>D</td>
<td>2DConv(in, out, k=3, use_bn=false, act=leakyrelu)</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>E</td>
<td>2DConv(in, out, k=3, use_bn=true, act=leakyrelu)</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>F</td>
<td>[B, C]</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>G</td>
<td>[D, E]</td>
<td>T2DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>H</td>
<td>3DConv(in, out, k=3, use_bn=false, act=relu)</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>I</td>
<td>3DConv(in, out, k=3, use_bn=true, act=relu)</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>J</td>
<td>3DConv(in, out, k=3, use_bn=false, act=leakyrelu)</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>K</td>
<td>3DConv(in, out, k=3, use_bn=true, act=leakyrelu)</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>L</td>
<td>[H, I]</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
<tr>
<td>M</td>
<td>[J, K]</td>
<td>T3DConv(out, in, k=2, s=2)</td>
</tr>
</tbody>
</table>
Except for lung tumor segmentation, weight-sharing method gave a better approximation of the ranking with an average of 0.85 kendall tau-b correlation.
Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

Hardware Efficiency Evaluation

- We deployed each model on two different hardware devices: Raspberry Pi3 and Laptop.
- Except for 10K architectures in each task, the latency and energy consumption are estimated using a lookup table.

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Processor</th>
<th>RAM</th>
<th>Storage</th>
<th>Operating System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raspberry Pi 3</td>
<td>quad-core ARM Cortex-A53 CPU</td>
<td>1GB</td>
<td>32G</td>
<td>Raspbian</td>
</tr>
<tr>
<td>Laptop</td>
<td>AMD Ryzen 7 6800H</td>
<td>16GB</td>
<td>1T</td>
<td>Microsoft Windows 11</td>
</tr>
</tbody>
</table>
## Medical Imaging Analysis with HW-NAS: MED-NAS-Bench

### Cross datasets ranking evaluation

Is there a single architecture that can be used for all medical imaging tasks?

![Similar Top Architectures](image)

### Similar Top Architectures

<table>
<thead>
<tr>
<th></th>
<th>brain</th>
<th>hippocampus</th>
<th>lung</th>
<th>liver</th>
<th>prostate</th>
<th>heart</th>
<th>hepatic vessels</th>
<th>spleen</th>
<th>chest</th>
</tr>
</thead>
<tbody>
<tr>
<td>brain</td>
<td>1.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.4</td>
<td>0.6</td>
</tr>
<tr>
<td>hippocampus</td>
<td>0.8</td>
<td>1.0</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>lung</td>
<td>0.6</td>
<td>0.6</td>
<td>1.0</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
<td>0.4</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>liver</td>
<td>0.4</td>
<td>0.5</td>
<td>0.9</td>
<td>1.0</td>
<td>0.6</td>
<td>0.9</td>
<td>0.4</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>prostate</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.6</td>
<td>1.0</td>
<td>0.5</td>
<td>0.9</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>heart</td>
<td>0.7</td>
<td>0.7</td>
<td>0.9</td>
<td>0.8</td>
<td>0.5</td>
<td>1.0</td>
<td>0.7</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>pancreas</td>
<td>0.5</td>
<td>0.3</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>1.0</td>
<td>0.5</td>
<td>0.8</td>
<td>0.5</td>
</tr>
<tr>
<td>colon</td>
<td>0.4</td>
<td>0.6</td>
<td>0.4</td>
<td>0.5</td>
<td>0.4</td>
<td>0.8</td>
<td>0.5</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>hepatic vessels</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>1.0</td>
</tr>
<tr>
<td>spleen</td>
<td>0.6</td>
<td>0.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>1.0</td>
</tr>
<tr>
<td>chest</td>
<td>0.4</td>
<td>0.4</td>
<td>0.7</td>
<td>0.8</td>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
<td>0.6</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Medical Imaging Analysis with Multi-task and Hardware-efficient Neural Architecture Search

Advancing Medical Imaging Analysis with Multi-task and Hardware-efficient Neural Architecture Search

A two-stage differentiable neural architecture search, allowing for simultaneous cell-level and network-level optimization.

Evolutionary search on two stages: fine (operation) and coarse (connections or topology)

A two-phase search includes a differentiable NAS to narrow down the search space, followed by a novel progressive evolutionary search.
Medical Imaging Analysis with HW-NAS MT-MIAS
Medical Imaging Analysis with HW-NAS: MT-MIAS

Once the supernetwork is trained, we evaluate each block (layer in this case) from bottom to top.

**Block Importance Score**

\[ s_{l,o} = \| w^{(l+1)} \|^T s_{l+1} \]

* The Pareto front is obtained by relaxing the number of selected blocks

---

Medical Imaging Analysis with HW-NAS: MT-MIAS

General Search Objective

\[
\max_{o \in l} \ s_{l,o} \cdot \frac{\text{Avg}(\alpha(A_{o \in A}))}{\text{Lat}(o)} + \sigma(s_{t,l,o} \cdot w_o, s_{t,l,o} \cdot w_o)
\]

- alpha is the validation accuracy obtained using weight-sharing
- sigma is the cosine similarity between the weights of the same operation used in different tasks
Medical Imaging Analysis with HW-NAS: MT-MIAS

We validate MT-MIAS using three scenarios:

➔ **MIAS Scenario**: Classical HW-NAS

\[
\max_{o \in l} s_{l,o} \times \frac{\text{Avg}(\alpha(A_{o \in A}))}{\text{Lat}(o)}
\]

➔ **MT-MIAS Scenario**: Classical HW-NAS + Generalization

\[
\max_{o \in l} s_{l,o} \times \frac{\text{Avg}(\alpha(A_{o \in A}))}{\text{Lat}(o)} + \sigma(s_{t,l,o} \times w_o, s_{t,l,o} \times w_o)
\]

➔ **MT-MIAS-C Scenario**: Relax the final number of blocks.
The goal is to construct the smallest supernetwork that is completely deployable, but each sub-network is trained on a different task.
MIAS generally outperforms all other methodologies
MT-MIAS induces a performance reduction due to generalization
MT-MIAS-C relaxes the generalization and outperforms other SOTA methods
## Medical Imaging Analysis with HW-NAS: MT-MIAS

<table>
<thead>
<tr>
<th>Method</th>
<th>Colon Dice (%)</th>
<th>Hepatic Vessels Dice (%)</th>
<th>Spleen Dice (%)</th>
<th>Chest F1-score</th>
<th>Hardware aware</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net [301]</td>
<td>54.32</td>
<td>38.5</td>
<td>89.54</td>
<td>95.32</td>
<td>-</td>
</tr>
<tr>
<td>U-Net++ [302]</td>
<td>59.82</td>
<td>48.93</td>
<td>88.95</td>
<td>95.38</td>
<td>-</td>
</tr>
<tr>
<td>Att. U-Net [303]</td>
<td>45.7</td>
<td>56.73</td>
<td>90.56</td>
<td>95.78</td>
<td>-</td>
</tr>
<tr>
<td>nnU-Net [308]</td>
<td>56</td>
<td>66.08</td>
<td>96</td>
<td>96.8</td>
<td>-</td>
</tr>
<tr>
<td>C2FNAS [305]</td>
<td>58.9</td>
<td>67.65</td>
<td>96.28</td>
<td>96.34</td>
<td>No</td>
</tr>
<tr>
<td>EA</td>
<td>50.8</td>
<td>55.78</td>
<td>89.76</td>
<td>95.78</td>
<td>No</td>
</tr>
<tr>
<td>RS</td>
<td>46.7</td>
<td>39.76</td>
<td>80.56</td>
<td>89.56</td>
<td>No</td>
</tr>
<tr>
<td>MixSearch [304]</td>
<td>57.8</td>
<td>71.65</td>
<td>96.75</td>
<td>97.54</td>
<td>No</td>
</tr>
<tr>
<td>C2FNAS [305]</td>
<td>60.67</td>
<td>65.42</td>
<td>97.34</td>
<td>98.68</td>
<td>No</td>
</tr>
<tr>
<td>BiX-NAS [306]</td>
<td>57.59</td>
<td>66.78</td>
<td>96.76</td>
<td>96.83</td>
<td>No</td>
</tr>
<tr>
<td>MT-MIAS</td>
<td>55.46</td>
<td>60.98</td>
<td>88.56</td>
<td>87.45</td>
<td>Yes</td>
</tr>
<tr>
<td>MT-MIAS-C</td>
<td>56.35</td>
<td>60.45</td>
<td>93.61</td>
<td>86.77</td>
<td>Yes</td>
</tr>
<tr>
<td>MT-MIAS (T)</td>
<td>63.45</td>
<td>70.54</td>
<td>95.66</td>
<td>93.45</td>
<td>Yes</td>
</tr>
<tr>
<td>MT-MIAS-C (T)</td>
<td>64.5</td>
<td>68.7</td>
<td>97.65</td>
<td>98.65</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Other hardware platforms potential

Analog in-memory computing
Analog In-memory Computing Potential

- Compute Vector-Matrix multiplication directly in the memory.
- **Less data movement.**
- **More energy efficient.**

**BUT!**
- **Noise** and **Drift** inherent characteristics
- Accuracy drops over time
- Robust & Noise resilient architectures

**AnalogNAS: A Neural Network Design Framework for Accurate Inference with Analog In-Memory Computing (IEEE Edge)**

Hadjer Benmeziane, Corey Lammie, Irem Boybat, Malte Rasch, Manuel Le Gallo, Hsinyu Tsai, Ramachandran Muralidhar, Smail Niar, Ouarnoughi Hamza, Vijay Narayanan, Abu Sebastian, Kaoutar El Maghraoui

---

1. **The 1-day accuracy** measures the performance of an architecture on a given dataset.

2. **The Accuracy Variation over one Month (AVM)** computes the difference between the 1-month and 1-sec accuracy.

3. **The 1-day accuracy standard deviation** measures the variation of the architecture's performance across experiments.

---

Evaluate

- Resnet-like Search Space
- Optimized Evolutionary
- Surrogate Models

---

robustness, accuracy ??

---

23/01/2024

Advancing Medical Imaging Analysis with Multi-task and Hardware-efficient Neural Architecture Search
Analog In-Memory Computing with Uncertainty Quantification for Efficient Edge-based Medical Imaging Segmentation

Hadjer Benmeziane, Imane Hamzaoui, Zayneb Cherif, Kaoutar El Maghraoui

We experiment multiple medical segmentation models on different tasks, comparing analog to digital inference.

Key insights:
- ★ Transformer based models are more robust to noise injection and analog training than pyramidal architectures.
- ★ Due to additional hardware-aware training, analog inference is more reliable and certain, despite the noise.
Conclusion & Perspectives

➔ We're proud to present **MED-NAS-Bench**, a trailblazing benchmark that bridges the gap between NAS and the intricate world of medical imaging analysis.

The **MED-NAS-Bench** API represents a significant contribution to the field of medical imaging research. [https://github.com/IHIaadj/med_nas_bench](https://github.com/IHIaadj/med_nas_bench)

➔ **MT-MIAS** encapsulates a methodology that seeks architectures optimized for holistic medical analysis, ensuring adaptability across diverse medical imaging tasks.

➔ Analog in-memory computing presents an efficient alternative to medical imaging analysis.

**Perspectives:**
- Increase the collected objectives and targeted hardware platforms in MED-NAS-Bench.
- Improve the API access to a sub-network in the benchmark.
- Develop a search process for automatically designing robust and noise-resilient medical architectures on analog in-memory computing.
MED-NAS-Bench

https://github.com/IHIaadj/med_nas_bench
https://theses.hal.science/tel-04224035v1/file/Benmeziane_Hadjer2.pdf
haadjer.benmeziane@gmail.com
Copyright Notice

This multimedia file is copyright © 2024 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