

IWES 2021

6th Italian Workshop on Embedded Systems
Rome, 9-10 December 2021

HW/SW Inference-time Optimizations for Reliable Embedded ConvNets



Roberto Giorgio Rizzo

EDA Group - PoliTo

- **Introduction**

- ConvNets in embedded real-life scenarios: Quality and Performance challenges

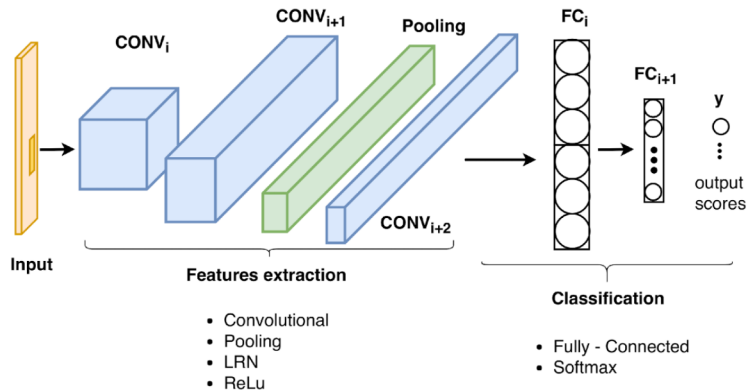
- **Inference-time optimizations for reliable embedded ConvNets**

- AdapTTA: Adaptive Test-Time Augmentation 
 - Improve Quality (i.e., Accuracy) of embedded ConvNets in real-life use
- TVFS: Topology Voltage Frequency Scaling 
 - Thermal-aware performance management technique for continuous inference of embedded ConvNets on low-power CPUs, under latency constraints

Introduction

ConvNets in embedded real-life scenarios

- **ConvNets** state-of-the-arts for several tasks and apps in **Computer Vision** (also NLP, Time-Series)

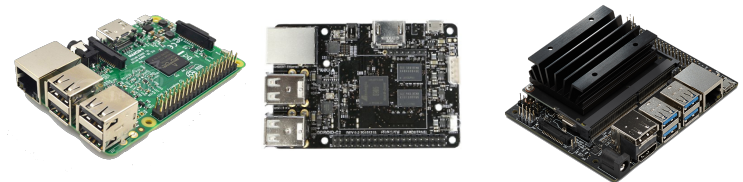


Training on the cloud



Inference on the Edge

Quality & Performance Challenges



on the field -> where the data is collected

AdapTTA: Adaptive Test-Time Augmentation

ConvNets in embedded real-life scenario: Quality challenges

- Input patterns collected in harsh environment might differ from those used at training time:
 - Size & orientation of the objects
 - Background
 - Lights conditions & contrast
- Model generalization capability at training-time w/ data augmentation is not always sufficient
- No ConvNet fine-tuning (i.e., re-training) w/ data collected on the field

Training Data

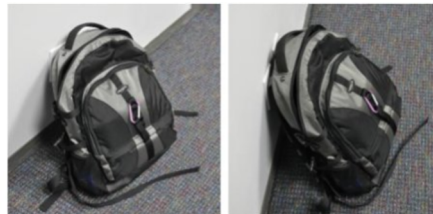


No background, vertical orientation

VS.

Misled Prediction

Inference real-life collected Data

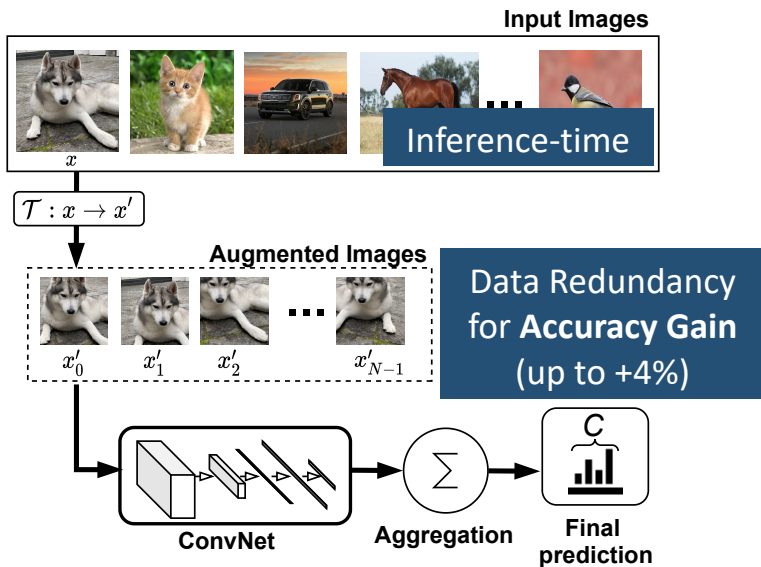


Complex background, different orientation

AdapTTA: Adaptive Test-Time Augmentation

Test-Time Augmentation (TTA): main features

- Improve accuracy with multiple inferences on a set of N input images altered through transforms: *Geometric, Luminosity, Contrast, Blur, Channel shuffle, etc.*
- Final prediction through a consensus of the aggregated predictions



TTA conceived for **GPUs**:
exploit batch inference

Embedded **CPUs**:
batch $\approx N$ proportional

Inference latency (ms)

ConvNet	NVIDIA Titan Xp			ARM Cortex-A53		
	1	5	10	1	5	10
MobileNetV1	18.2	18.6	18.7	53.1	290.6	569.9
MobileNetV2	12.1	12.4	12.9	44.2	261.8	513.5
EfficientNet-B0	21.3	22.4	22.6	68.5	358.9	682.3
EfficientNet-B1	31.9	33.4	33.9	103.4	536.4	1290.2

$\times 1$

$\times 5$

$\times 10$

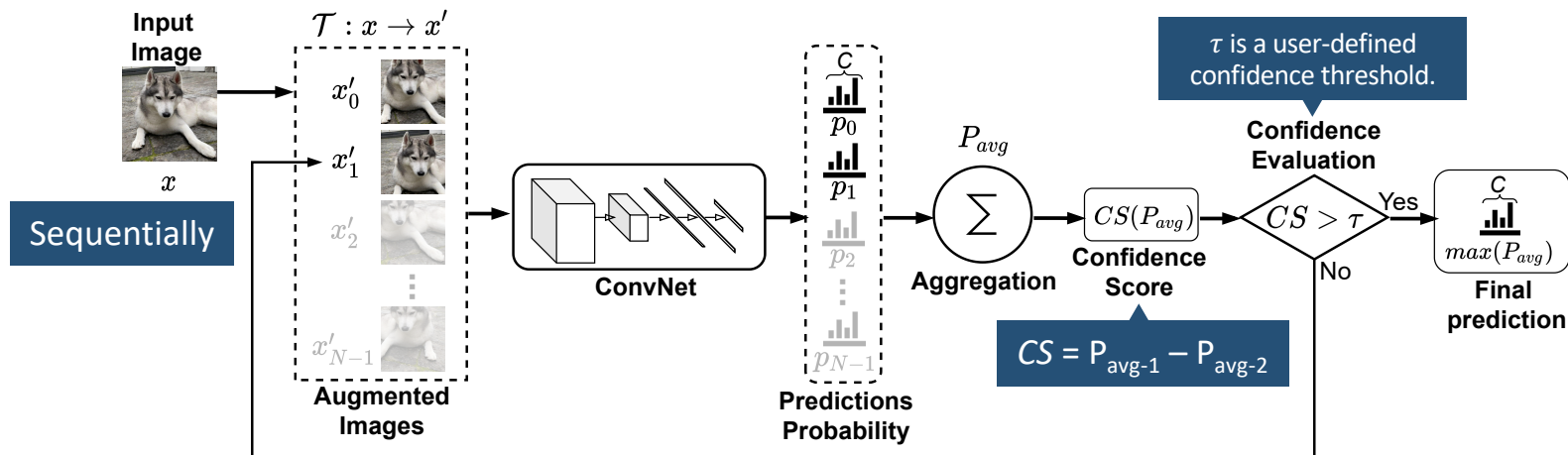
Minimize this cost



AdapTTA: Adaptive Test-Time Augmentation

From TTA to AdapTTA for Embedded ConvNets: main intuition

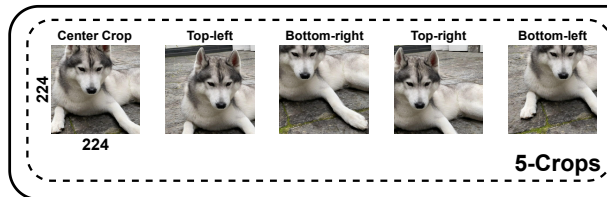
- Fixed N transformed images might be too conservative
- For certain inputs, the key features are well exposed and easy to be detected.
- AdapTTA: *a more flexible TTA mechanism exploiting intermediate results*



AdapTTA: Adaptive Test-Time Augmentation

Experimental Setup

- Augmentation Policies: 5C e 10C (sota)



More redundancy, higher accuracy



- Hardware Platforms & Compiler Toolchain

- Odroid-C2 [1]:
 - 4 Arm Cortex-A53 @ 1.5 GHz
 - 1 GB RAM
- TFLite v1.14
- GNU Toolchain v6.5



[1] <https://wiki.odroid.com/odroid-c2/odroid-c2>

- Image Classification ConvNet Benchmarks

	ConvNet [ImageNet]	Storage [MB]	Top-1 [%]	L_{nom} [ms]
[2]	MobileNetV1	4.3	70.0	53.1
	MobileNetV2	3.4	70.8	44.2
[3]	EfficientNet-B0	5.4	74.4	68.5
	EfficientNet-B1	6.4	75.9	103.4

[2] *Tensorflow lite hosted models* - https://www.tensorflow.org/lite/guide/hosted_models

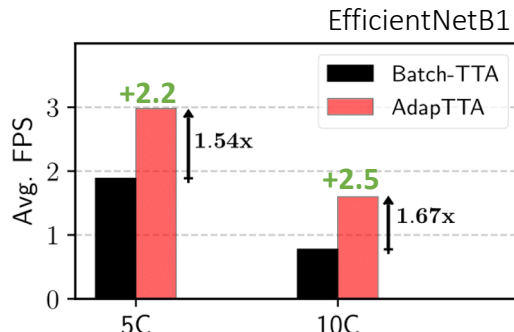
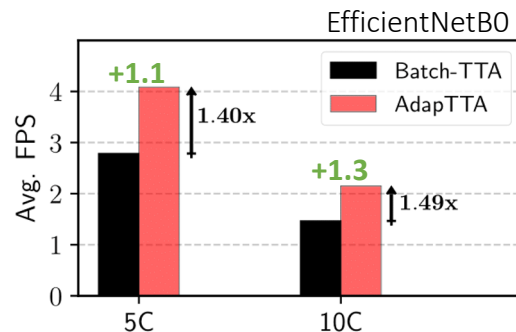
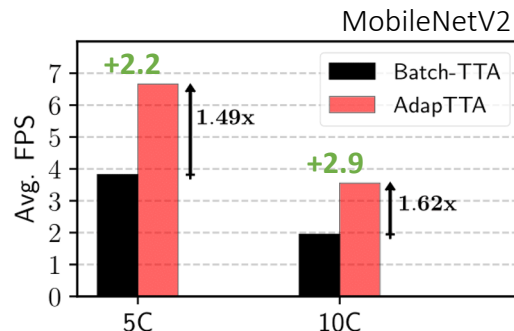
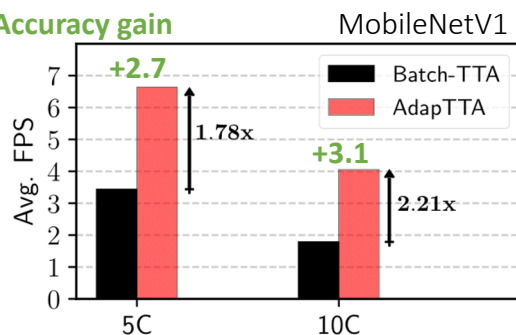
[3] *Tensorflow hub* - <https://tfhub.dev>

AdapTTA: Adaptive Test-Time Augmentation

Experimental Results

- ImageNet validation set (50k images)
- Confidence Threshold $\tau = 0.8$

Accuracy gain



Number of inferences (avg)

ConvNet	5C	10C
MobileNetV1	2.81	4.57
MobileNetV2	3.37	6.26
EfficientNet-B0	3.57	6.75
EfficientNet-B1	3.24	6.02

More efficient Embedded TTA

1. Accuracy gain at inference-time
2. Contained latency overhead

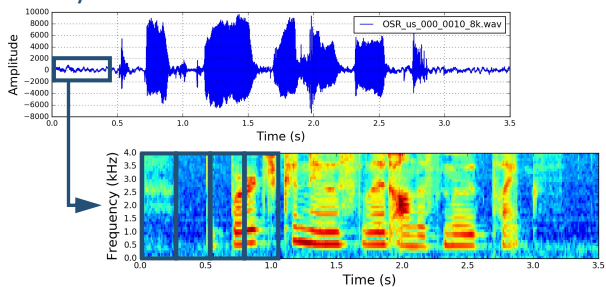


TVFS: Topology Voltage Frequency Scaling

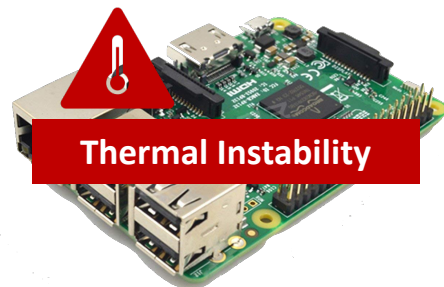
ConvNets in embedded real-life scenarios: Performance challenges

- Deploy **continuous inference** regime with latency constraints:
 - **Power demanding task**: intensive workload at max. frequency
 - Embedded systems with **limited Thermal Design Power (TDP)**, no room for heat spreader or active cooling

audio/time-series classification



Continuous Inference



on the field

Image classification w/ consensus (e.g., TTA)

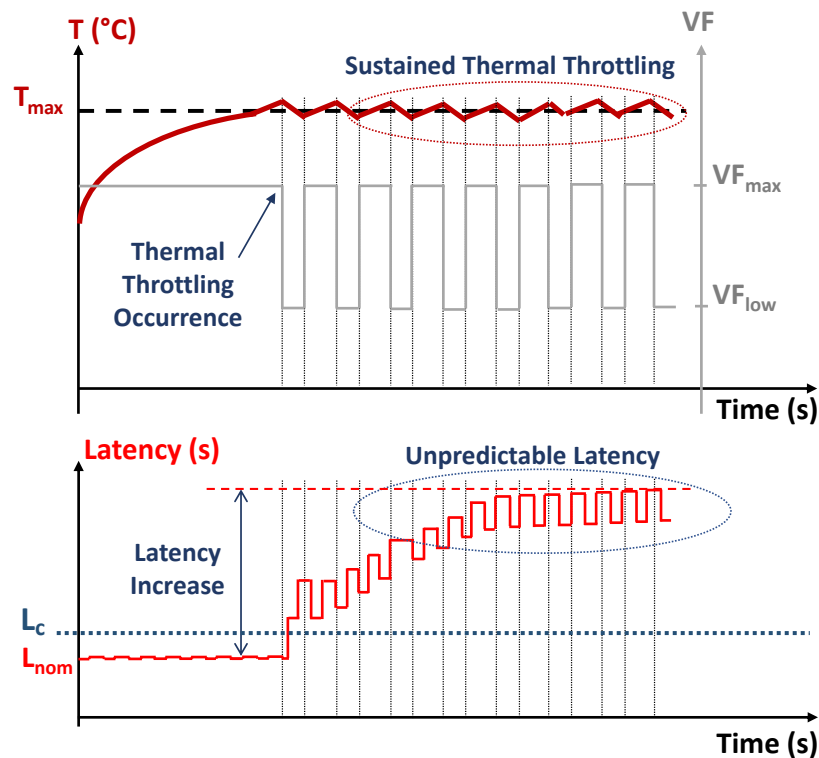


Horse

TVFS: Topology Voltage Frequency Scaling

Thermal-induced performance loss

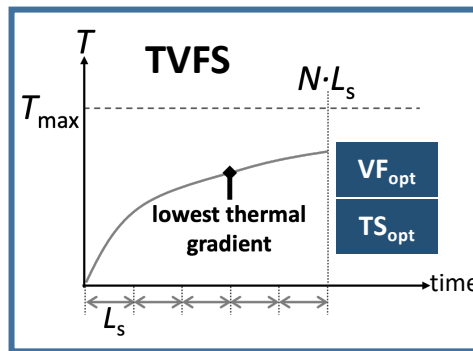
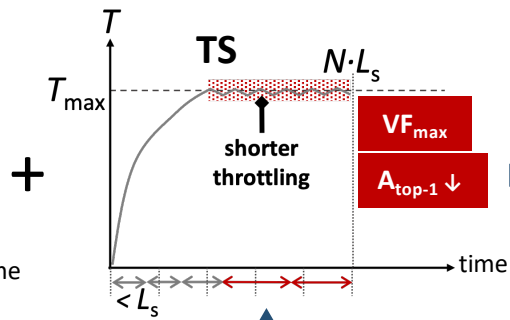
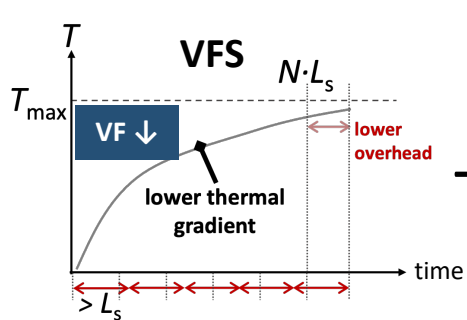
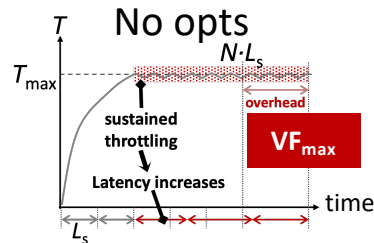
- **Continuous inference** of embedded ConvNets on CPUs w/ limited TDP under latency constraint (L_c)
- Intensive workload at max. frequency (VF_{max}) affects the thermal stability
 - On-chip temperature oversteps safety threshold (T_{max}) – Thermal Throttling
 - To avoid irreversible damages, OS policy lowers core's VF (VF_{low}) until temperature is back to safety
- Repeated thermal throttling leads to performance penalty, thus, latency constraint mismatch



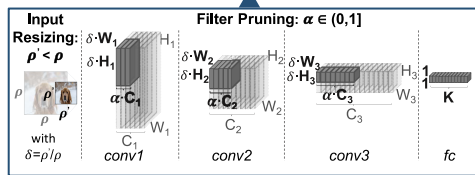
TVFS: Topology Voltage Frequency Scaling

Main Idea

- Thermal-aware performance management through the cooperation of:
 - power-reduction** techniques -> Voltage-Frequency Scaling (**VFS**)
 - algorithmic** optimizations -> ConvNet Topology Scaling (**TS**)



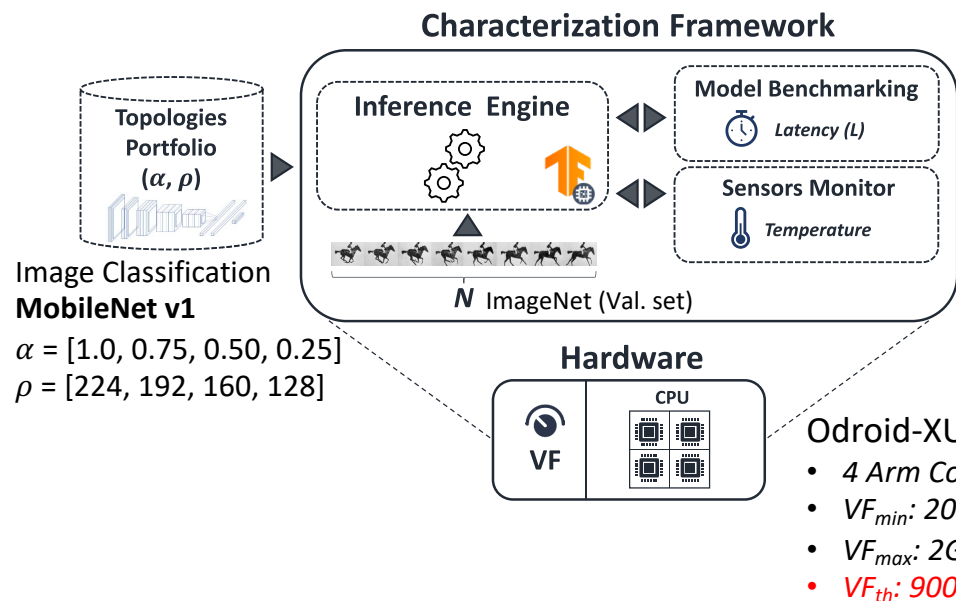
- No Thermal Throttling
- Latency $\leq N \cdot L_s$
- Min. Accuracy Loss



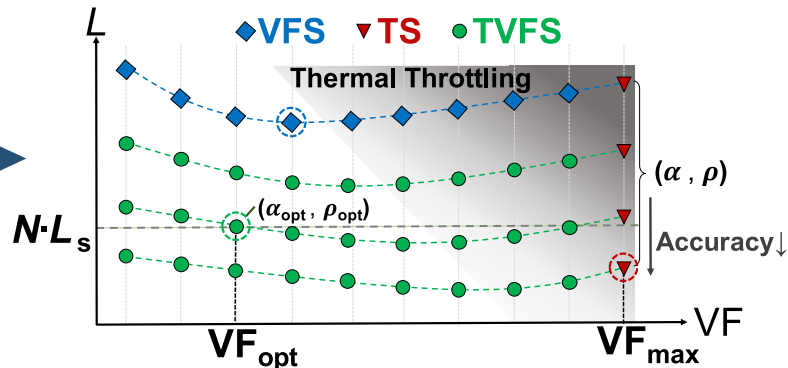
TVFS: Topology Voltage Frequency Scaling

Problem Formulation and Experimental Setup

Static workload of ConvNets Inference



Performance Trade-Off Profiling



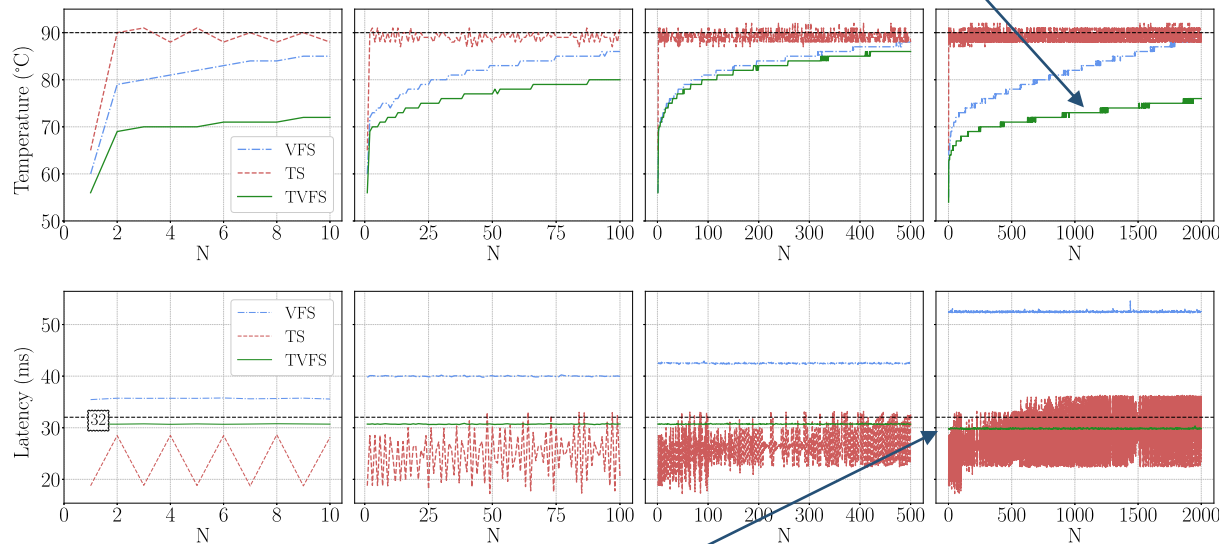
Odroid-XU4:

- 4 Arm Cortex-A15 w/ 19 VF-levels
- VF_{min} : 200MHz @ 0.85V
- VF_{max} : 2GHz @ 1.3625V
- VF_{th} : 900MHz @ 0.8875V

TVFS: Topology Voltage Frequency Scaling

Experimental Results

Preserves On-chip Thermal Stability



Meets Performance Constraint



$$A_{TVFS} \geq A_{TS}$$

N	Tech.	α	ρ	VF (GHz)	Top-1 (%)	L_{avg} (ms)	Th (%)	T_{avg} (°C)
10	NS	1.0	224	2.0	70.0	44.1	46.3	88.8
	VFS	1.0	224	1.8	70.0	35.7	0.0	82.4
	TS	1.0	160	2.0	66.9	23.1	38.5	89.4
	TVFS	1.0	192	1.5	69.1	30.7	0.0	70.7
100	NS	1.0	224	2.0	70.0	48.6	60.4	89.1
	VFS	1.0	224	1.6	70.0	40.0	0.0	82.1
	TS	1.0	160	2.0	66.9	24.9	53.7	89.1
	TVFS	1.0	192	1.5	69.1	30.7	0.0	76.8
500	NS	1.0	224	2.0	70.0	51.9	69.9	89.3
	VFS	1.0	224	1.5	70.0	42.5	0.0	83.5
	TS	1.0	160	2.0	66.9	26.3	63.3	89.2
	TVFS	1.0	192	1.5	69.1	30.7	0.0	82.1
2000	NS	1.0	224	2.0	70.0	58.6	82.1	89.5
	VFS	1.0	224	1.2	70.0	52.5	0.0	81.4
	TS	1.0	160	2.0	66.9	28.6	73.7	89.3
	TVFS	1.0	160	1.1	66.9	29.8	0.0	73.6



Contains Accuracy Loss

Questions



Contacts:

robertogiorgio.rizzo@polito.it

andrea.calimera@polito.it

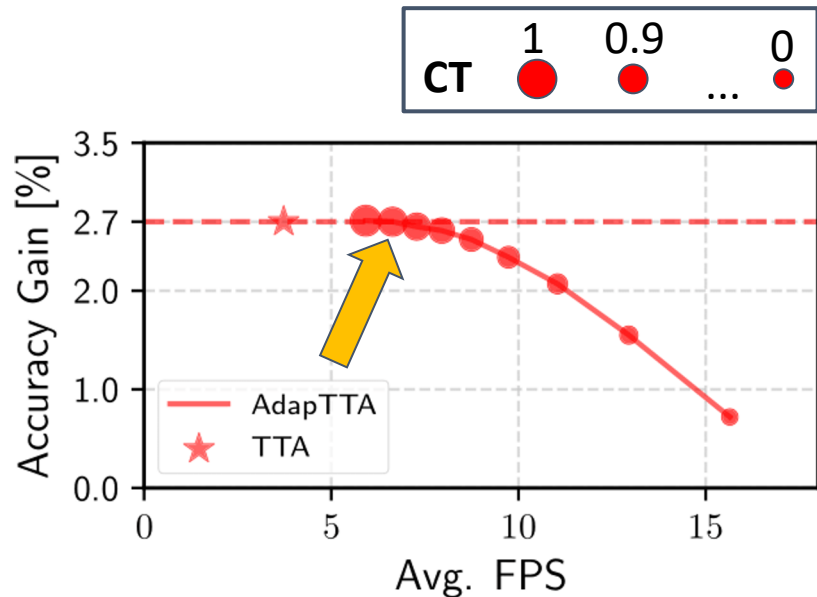
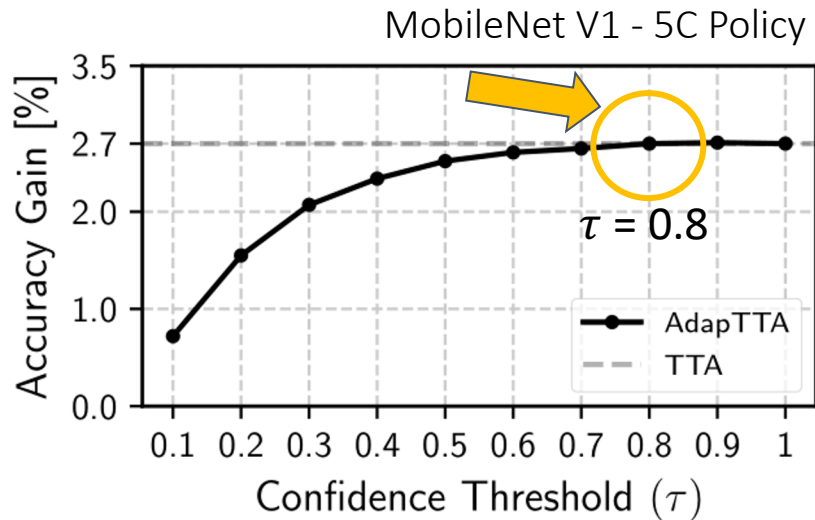
References

- “*AdapTTA: Adaptive Test-Time Augmentation for Reliable Embedded ConvNets*”. L. Mocerino, **R. G. Rizzo**, V. Peluso, A. Calimera, E. Macii; In: VLSI-SoC 2021.
- “*TVFS: Topology Voltage Frequency Scaling for Reliable Embedded ConvNets*”. **R. G. Rizzo**, V. Peluso, A. Calimera. In: IEEE TCAS-II 68 (2), 672-676 (2020)
- “*Performance profiling of embedded convnets under thermal-aware DVFS*”. V. Peluso, **R. G. Rizzo**, A. Calimera. In: Electronics 8 (12), 1423 (2020)
- “*Efficacy of topology scaling for temperature and latency constrained embedded convnets*”. V. Peluso, **R. G. Rizzo**, A. Calimera. In: Journal of Low Power Electronics and Applications 10 (1), 10 (2020).

AdapTTA: Adaptive Test-Time Augmentation

Experimental Results

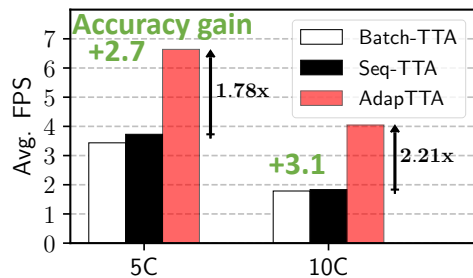
- Calibration set: ImageNet val (1k images)



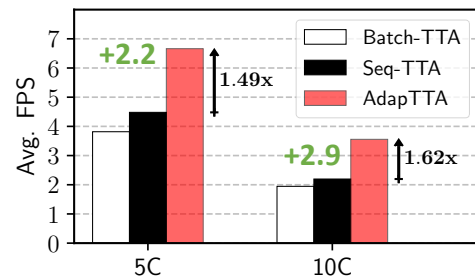
AdapTTA: Adaptive Test-Time Augmentation

Experimental Results

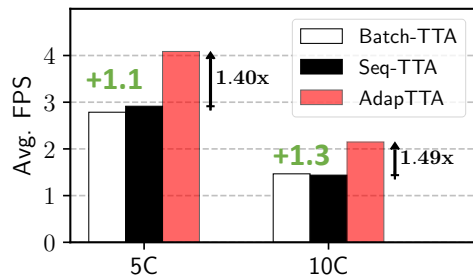
- ImageNet validation set (50k images)
- Confidence Threshold $\tau = 0.8$



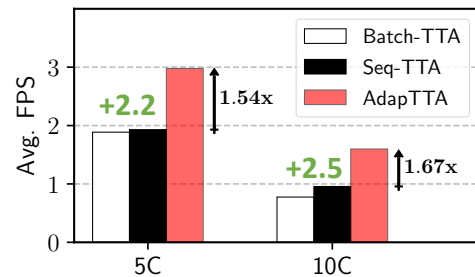
(a) MobileNetV1



(b) MobileNetV2



(c) EfficientNet-B0

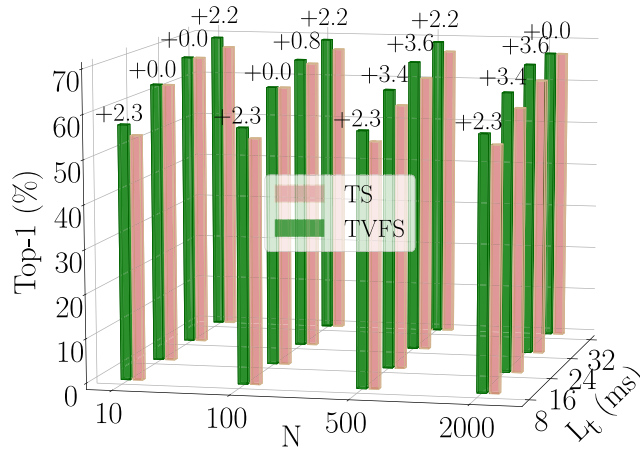


(d) EfficientNet-B1

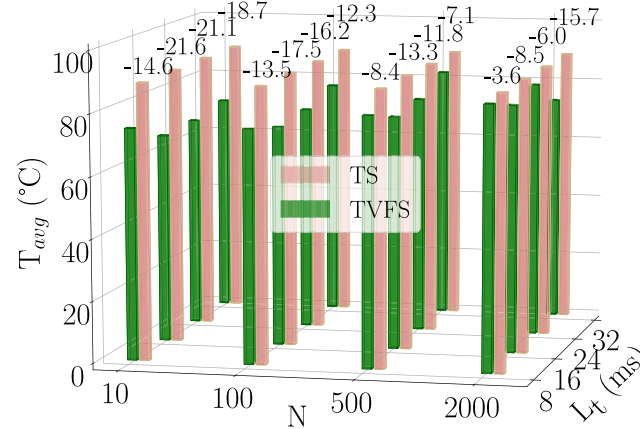
TVFS: Topology Voltage Frequency Scaling

Experimental Results

- More stringent latency constraints



$$A_{TVFS} \geq A_{TS}$$



$$T_{TVFS} < T_{TS}$$