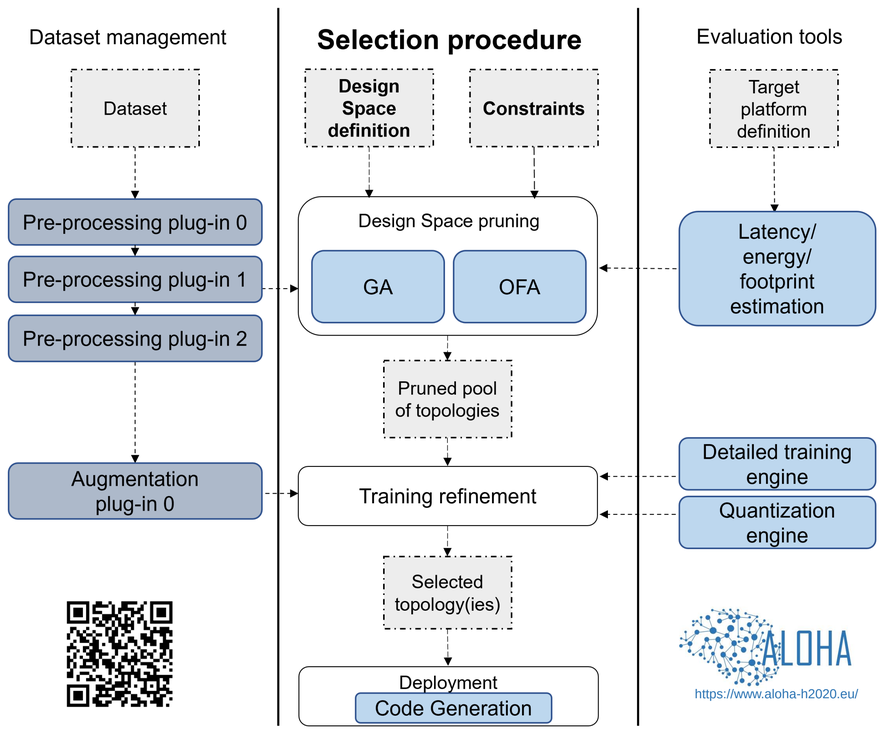
Optimizing Neural Networks for Embedded Edge-Processing Platforms

Paola Busia, Paolo Meloni

DIEE, Università degli Studi di Cagliari, Italy

INTRODUCTION : The design of network models targeting the **edge-computing** domain requires a tight trade-off between **accuracy** and **efficiency**. Evaluating the different design alternatives within a reasonable **time** requires efficient exploration strategies, enabling a certain degree of **platform awareness**. This work aims to contribute to these challenges, proposing an **efficient exploration strategy** to comprehensively consider the impact of multiple design parameters and an accurate estimation method to pre-evaluate the **on-hardware efficiency** of candidate design points. Finally, we apply hardware-aware design concepts to the emerging **transformer** model.

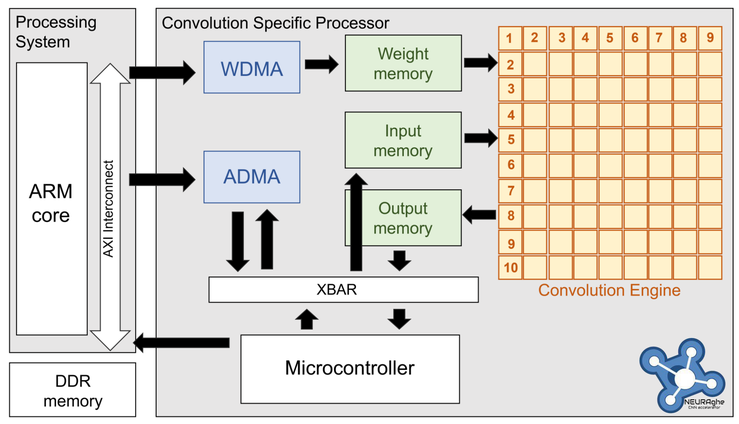
Design Space Exploration



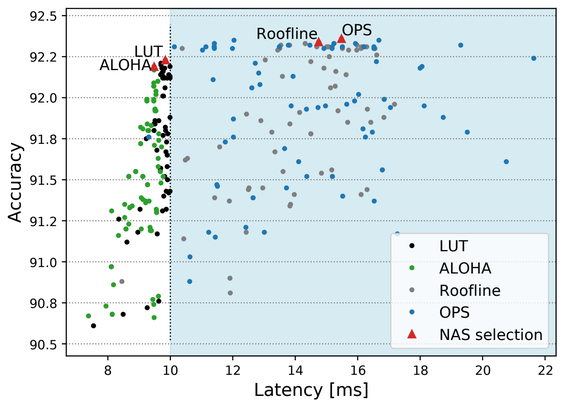
Selection procedure for target-oriented CNN design exploiting the ALOHA and OFA one-shot train [1], enabling **flexible customization of data pre-processing**.

[1] H.Cai et al, Once for all: Train one Network and Specialize it for Efficient Deployment, 2019

Hardware Performance Estimation



- Available parallelism
  - par0 : 9 Input Features
  - par1: 10 Output Features
  - par2: 4 Width pixels
- Storage resources
  - id0: Input Features
  - id1: Output Features
  - id2: Weights
- IO channels
  - id0: Input Features
  - id1: Output Features
  - id2: Weights



Abstract model considering the resources available on the platform to capture details impacting performance, such as the **efficiency in exploiting the available parallelism**, or the need to **repeat some data transfers** during inference execution. When exploited in NAS, targeting CIFAR classification on Neuraghe, it allows obtaining **near-optimal** results, very similar to the ones obtained by accessing direct measurements stored in Look Up Tables.

Optimal Transformer Design

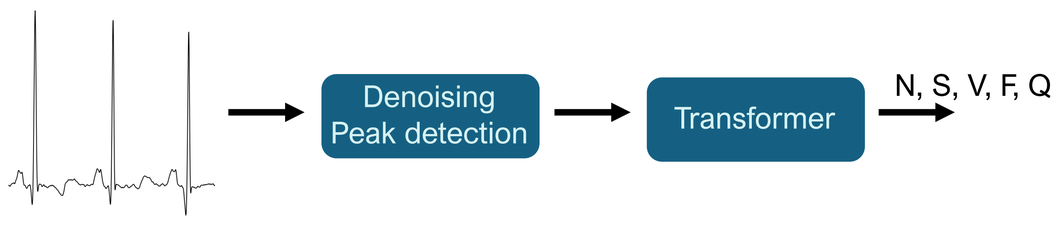
Use case: long-term electroencephalography monitoring for seizure detection

Unobtrusive signal acquisition limited to the temporal channels
Reference dataset CHB-MIT Scalp EEG database

Model	Segment-level		Event-level		
	Sensitivity	Specificity	Sensitivity	FP/h	Detection Latency
This Work	**65.5**	**99.9**	**73 (32/44)**	**0.15 (38s)**	**15.2 s**
AdaBoost [2]	72	99.9	86 (38/44)	0.5 (4s)	19 s

[2] Ingolfsson et al, Towards Long-Term Non-invasive Monitoring of Epilespy via Wearable EEG devices, 2019

Use case: long-term electrocardiography monitoring for arrhythmia recognition



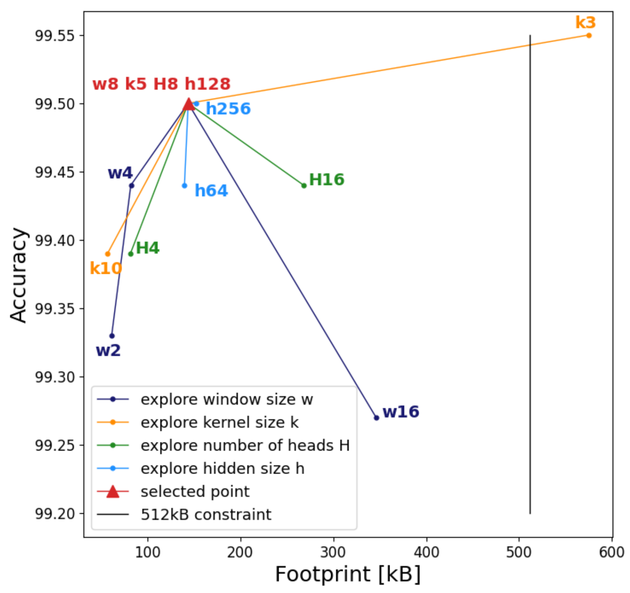
Reference dataset MIT-BIH Arrhythmia database

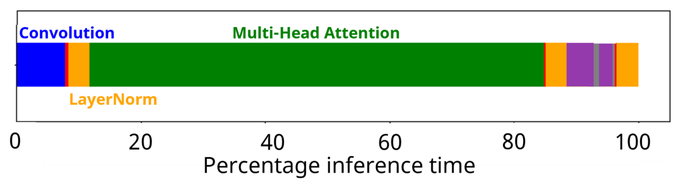
[3] Yan et al, Fusing transformer model with temporal features for ecg heartbeat classification, 2019

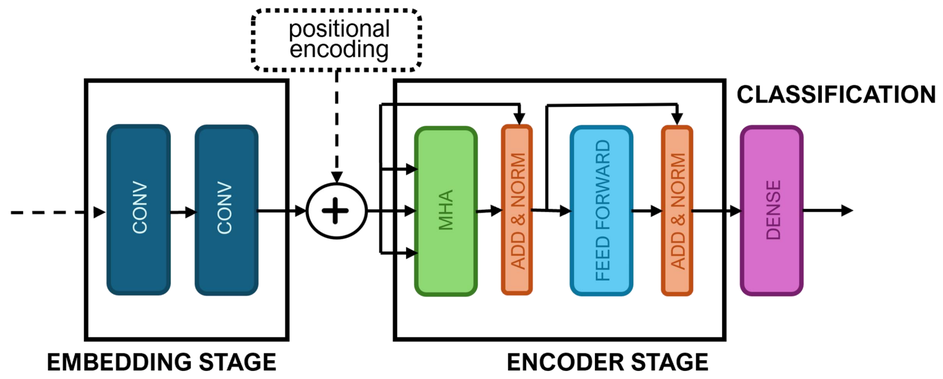
Model	Accuracy		Noisy Test		Parameters	Footprint
	32-bit	8-bit	32-bit	8-bit		
This Work	**99.05**	**98.97**	**98.65**	**98.36**	**6649**	**49 kB**
Yan [3]	99.62	-	-	-	405711**	3 MB**

** *Estimated from the paper.*

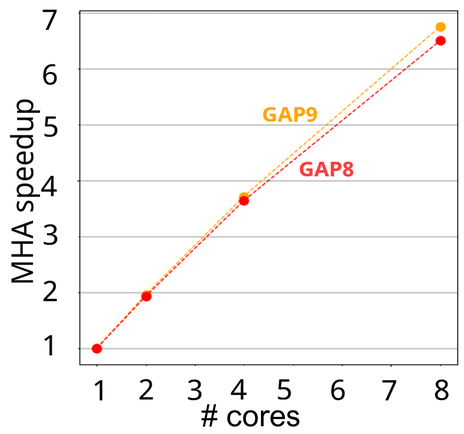
The design exploration focused on the parameters impacting the most on the **workload** in the Multi-Head Attention layer. We considered a **memory constraint** of **512 kB**, targeting **real-time** execution on **low-power** microcontrollers and considering the typical size of their working memories.







	GAP8		GAP9	
	1 core	8 cores	1 core	8 cores
Frequency	65 MHz	65 MHz	240 MHz	**240 MHz**
Time/ Inference	283.9 ms	62.2 ms	72 ms	**13.7 ms**
Power	10 mW	18.1 mW	12.9 mW	**22.9 mW**
Energy/ Inference	2.9 mJ	1.2 mJ	0.93 mJ	**0.31 mJ**



Conclusions

We presented our research effort to provide guidelines and methodologies for efficiently addressing three of the main challenges of edge-oriented network design, resulting in an efficient **design exploration strategy** and an accurate estimation method to enable **platform awareness**. We finally applied hardware-aware design concepts to the **transformer** model.

Acknowledgments

The ALOHA framework provided the context and reference for the research activity. The findings presented in this work are the results of different cooperations, with the research team of Santer Reply SpA, the Leiden Institute of Advanced Computer Science (LIACS), and Zurich's ETH.