Hardware-Aware Neural Architecture Search : Survey & Taxonomy

Manually building deep learning architectures requires expertise and time. Hardware-aware Neural Architecture Search are methods to automatically create efficient architectures.

Authors
Hadjer Benmeziane, Kaoutar El Maghraoui, Hamza Ouarnoughi, Smail Niar, Martin Wistuba and Naigang Wang

Affiliation
--> Université Polytechnique Hauts-de-France, LAMIH/CNRS, Valenciennes, France
--> IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
--> IBM Research AI, IBM Technology Campus, Dublin, Ireland

01 Motivation

- Making AI mainstream by bringing powerful, power hungry Deep Neural Networks (DNNs) to resource-constrained devices requires an efficient co-design of algorithms, hardware and software.
- Increased popularity of DNN applications deployed on a wide variety of platforms,
- From tiny microcontrollers to data centers: Multiple questions and challenges in constraints introduced by the hardware.
- Surveys on conventional NAS exists [1]: ours is the 1st survey dedicated to HW-NAS.

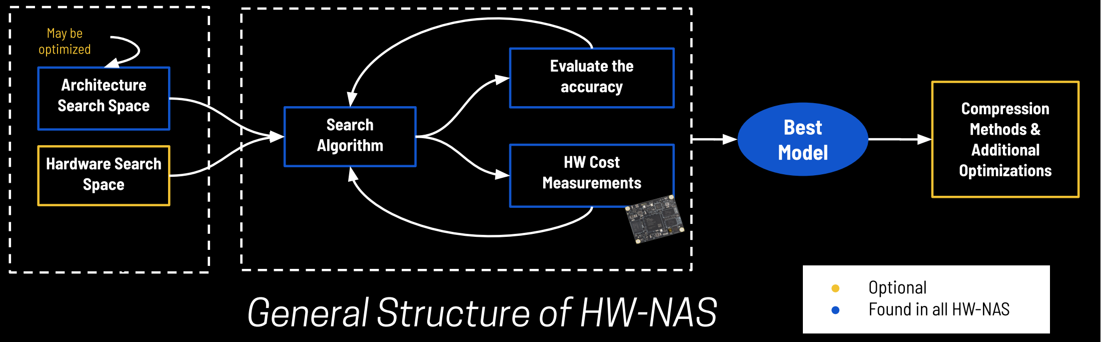
02 Goal

Study Hardware-aware neural architecture (HW-NAS), understand its main components and explore the hardware friendly design options.

03 General HW-NAS Structure

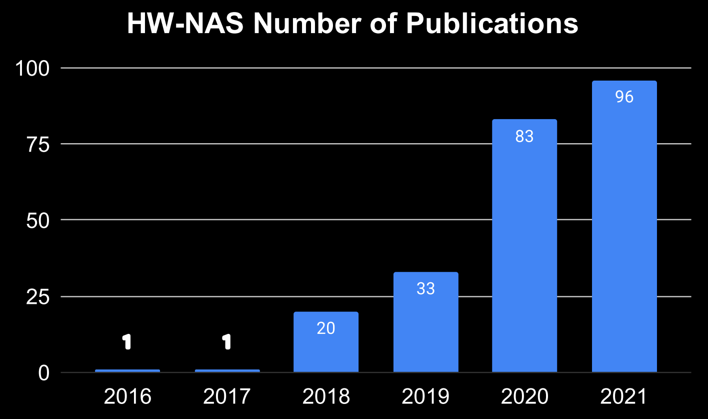
The general structure of HW-NAS different than the conventional NAS process. We still find the three main components: **Architecture search space, Search algorithm** and **Evaluation methods.**

- The Architecture Search Space can be optimized:
  - Remove unoptimized operators
  - Remove too large architectures
- Hardware search space used -> **joint optimization** between DL architecture and HW configurations.

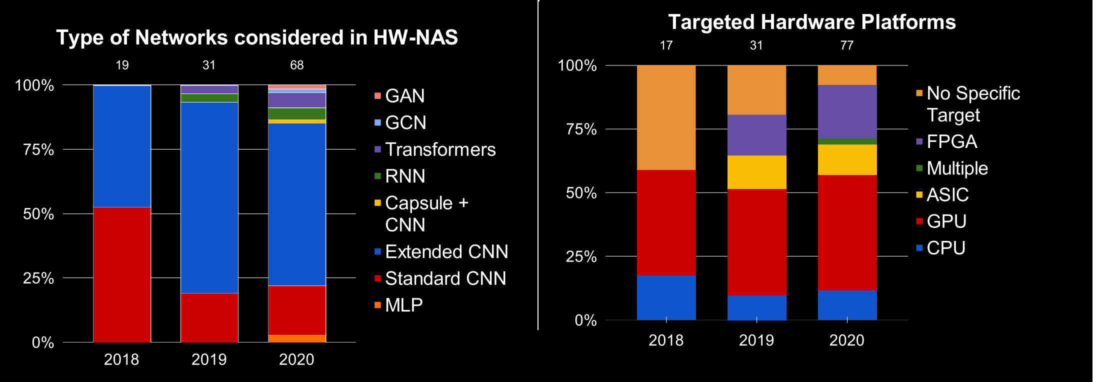


General Structure of HW-NAS

04 HW-NAS... A Trending Topic

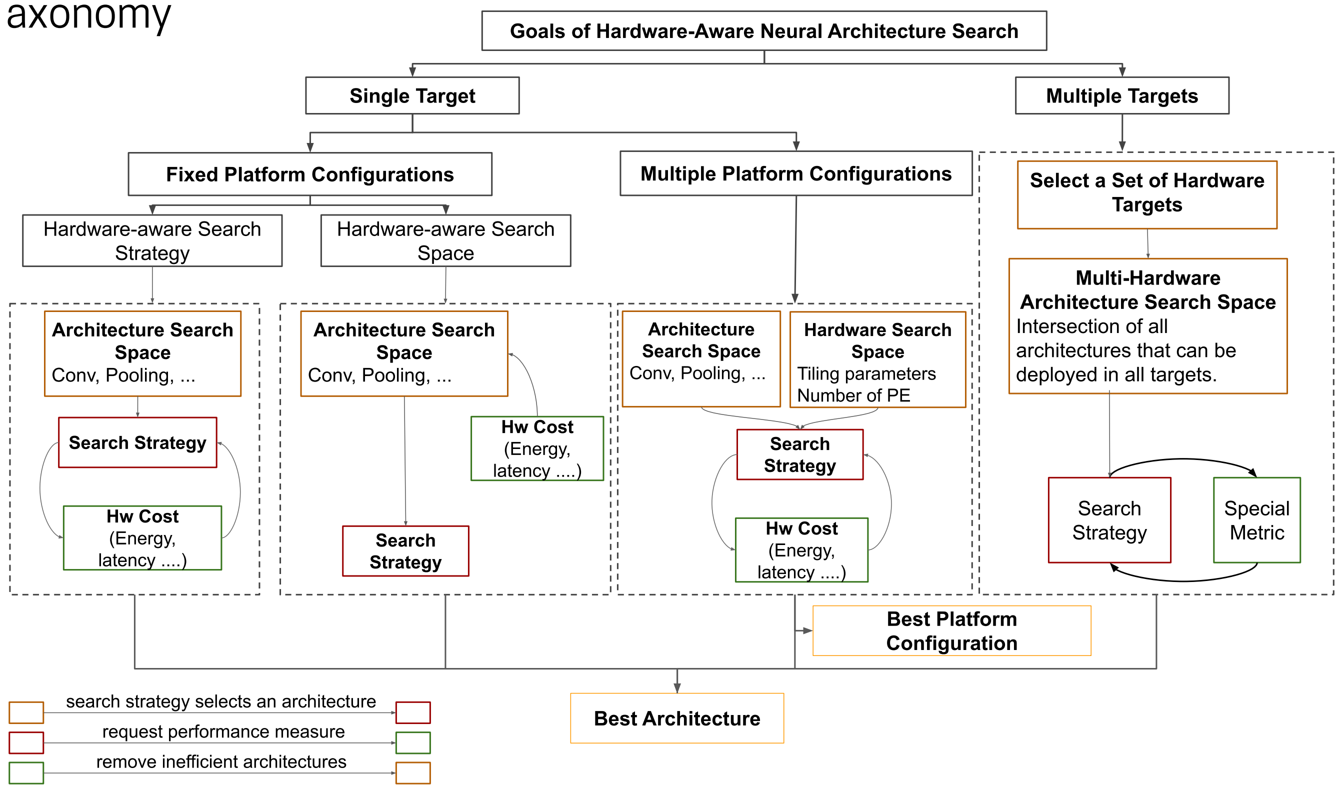


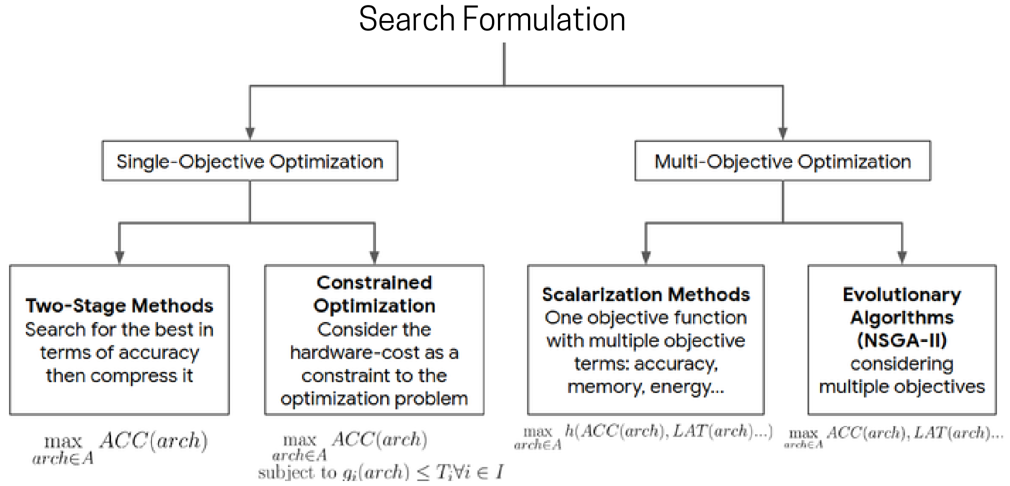
(updated by June 2021)



Proportion over 139 HW-NAS

05 Taxonomy


Goals of Hardware-Aware Neural Architecture Search
Single Target
Multiple Targets
Fixed Platform Configurations
Multiple Platform Configurations
Hardware-aware Search Strategy
Hardware-aware Search Space
Architecture Search Space
Conv, Pooling, ...
Search Strategy
Hw Cost
(Energy, latency ....)
Architecture Search Space
Conv, Pooling, ...
Hw Cost
(Energy, latency ....)
Search Strategy
Architecture Search Space
Conv, Pooling, ...
Hardware Search Space
Tiling parameters
Number of PE
Search Strategy
Hw Cost
(Energy, latency ....)
Select a Set of Hardware Targets
Multi-Hardware Architecture Search Space
Intersection of all architectures that can be deployed in all targets.
Search Strategy
Special Metric
Best Platform Configuration
Best Architecture
search strategy selects an architecture
request performance measure
remove inefficient architectures


Search Formulation
Single-Objective Optimization
Multi-Objective Optimization
Two-Stage Methods
Search for the best in terms of accuracy then compress it
Constrained Optimization
Consider the hardware-cost as a constraint to the optimization problem
Scalarization Methods
One objective function with multiple objective terms: accuracy, memory, energy...
Evolutionary Algorithms (NSGA-II)
considering multiple objectives
$\max_{arch \in A} ACC(arch)$
$\max_{arch \in A} ACC(arch)$
subject to $g_i(arch) \leq T_i \forall i \in I$
$\max_{arch \in A} h(ACC(arch), LAT(arch)...)$
$\max_{arch \in A} ACC(arch), LAT(arch)...$

06 HW-NAS Components

■ Search Space

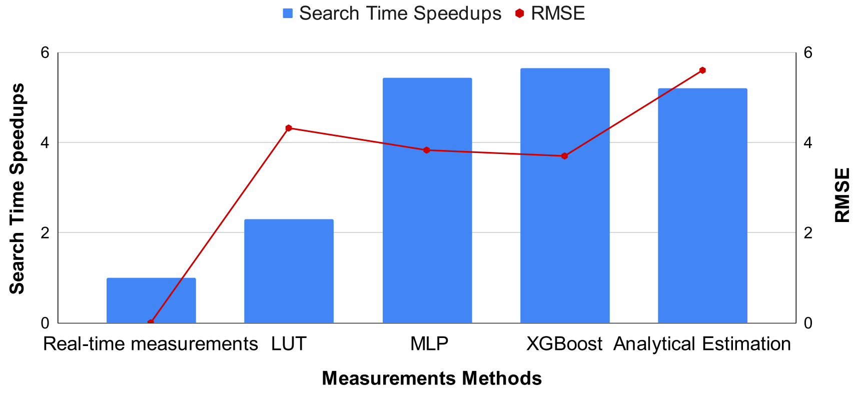
Architecture Search Space: same space used by conventional NAS.
In HW-NAS, it can be optimized by either removing some operators or some architectures.
Hardware Search Space: space of different configurations of one platform (e.g, frequency scaling)

Observations

- Layer-wise search space is more hardware friendly than cell-based search space [2].

■ HW Cost Techniques

- Real-world measurements
- Lookup tables (LUT)
- Analytical Estimation
- **Prediction Models**



Comparison of different evaluation methods for the latency on NAS-Bench-201

07 Conclusion & Key Takeaways

- HW-NAS an important to tool to find efficient architectures.
- No method that beats every other strategy.
- HW-NAS works focus on computer vision and CNN.
- Benchmarking a big challenge in HW-NAS,
  - HW-NAS-Bench [3] extend NAS-Bench-201 and FBNet with HW metrics

08 References

[1] M. Wistuba, et Al. A survey on neural architecture search. CoRR, abs/1905.01392, 2019
[2] B. Wu, et Al. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search., CVPR. 2019
[3] C. Li, et Al. HW-NAS-bench: Hardware-aware neural architecture search benchmark. In ICLR 2021.

Université Polytechnique HAUTS-DE-FRANCE
LAMIH
IBM Research