EFFICIENT AND INTELLIGENT COMPUTING

HW-NAS-Bench: Hardware-aware Neural Architecture Search Benchmark

Chaojian Li[1], **Zhongzhi Yu[1]**, Yonggan Fu[1], Yongan Zhang[1], Yang Zhao[1], Haoran You[1], Qixuan Yu[2], Yue Wang[2], Yingyan (Celine) Lin[1]

[1] Georgia Institute of Technology, SCS [2] Rice University, ECE

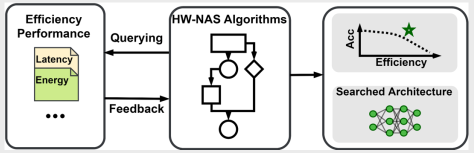
Georgia Tech College of Computing
Center for Research into
Novel Computing Hierarchies

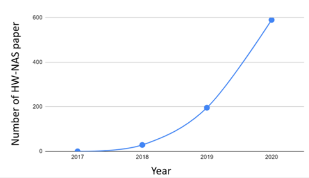
ICLR'21 Spotlight

Background & Motivation

HW-NAS as an Automation Tool is on Growing Demand

- HardWare-aware Neural Architecture Search (HW-NAS) **automatically** searches optimal architectures for **a target application and device**
- The number of HW-NAS research increases rapidly

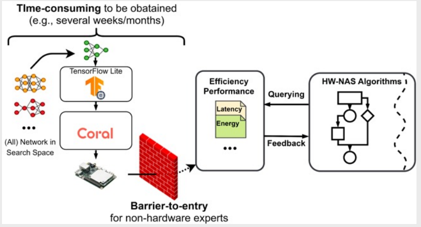




Challenge 1: Non-trivial to Obtain Hardware-cost

HW-NAS requires **hardware-cost of (all) networks** in the search space

- Existing methods: hardware-cost look-up tables/device-specific estimator
- Limitations:
  - ☹ **Time-consuming**
  - ☹ **Barrier-to-entry** for non-hardware experts

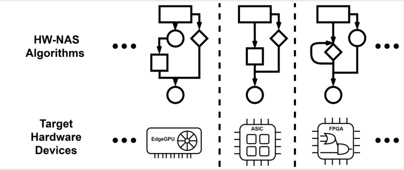


Challenge 2: Difficult to Benchmark HW-NAS

Difficult to benchmark different HW-NAS algorithms because of the **different adopted devices**

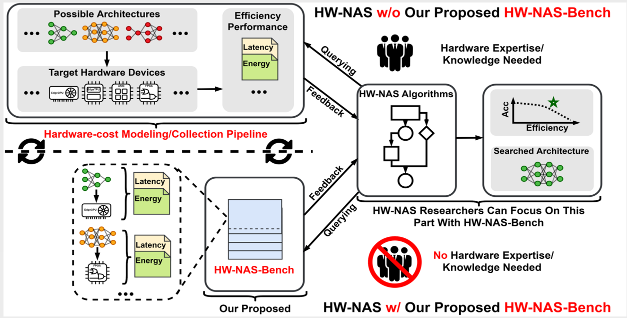
- Existing methods: Benchmarks focusing on accuracy/FLOPs/#Params/ server-level hardware-cost
- Limitations:
  - ☹ **No/Limited** hardware-cost
  - ☹ **FLOPs/#Params does not correlate well** with hardware-cost

Which one is better?



The Proposed HW-NAS-Bench

HW-NAS-Bench: Overview


Possible Architectures
Target Hardware Devices
Efficiency Performance
Latency
Energy
Hardware-cost Modeling/Collection Pipeline
HW-NAS w/o Our Proposed HW-NAS-Bench
Hardware Expertise/ Knowledge Needed
Querying
Feedback
HW-NAS Algorithms
Acc
Efficiency
Searched Architecture
HW-NAS Researchers Can Focus On This Part With HW-NAS-Bench
HW-NAS-Bench
Our Proposed
No Hardware Expertise/ Knowledge Needed
HW-NAS w/ Our Proposed HW-NAS-Bench

HW-NAS-Bench: Highlighted Features

- **Six** devices, **three** categories
- **Two** search spaces: NAS-Bench-201 and FBNet search spaces
- **Both energy and latency**

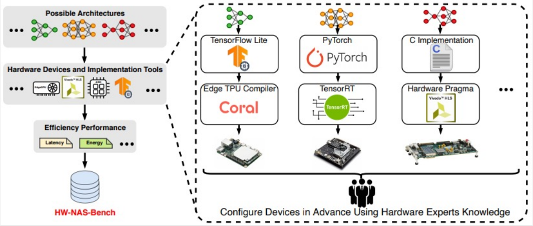
Devices	Edge GPU	Raspi 4	Edge TPU	Pixel 3	ASIC-Eyeriss	FPGA
Collected Metrics	Latency (ms) Energy (mJ)	Latency (ms)	Latency (ms)	Latency (ms)	Latency (ms) Energy (mJ)	Latency (ms) Energy (mJ)
Collecting Method	Measured	Measured	Measured	Measured	Estimated	Estimated
Runtime Environment	TensorRT	TensorFlow Lite	Edge TPU Runtime	TensorFlow Lite	Accelergy+Timeloop / DNN-Chip Predictor	Vivado HLS
Customizing Hardware?	✗	✗	✗	✗	✓	✓
Category	Commercial Edge Devices				ASIC	FPGA

☺ **Democratize** HW-NAS research to **non-hardware experts**
- Solved Challenge 1: Non-trivial to Obtain Hardware-cost

☺ Facilitate **a unified benchmark for HW-NAS**
- Solved Challenge 2: Difficult to Benchmark HW-NAS

HW-NAS-Bench: Hardware-cost Collection Pipeline

- Implementation in each device is **optimized to make sure the best hardware performance** of networks in the search space



HW-NAS-Bench: Analysis

Analysis 1: Real Hardware-cost Is Necessary

Theoretical metrics (e.g., FLOPs/#Params) do NOT correlate well with real-measured/estimated hardware-cost

- Coefficient can be **as low as 0.36**

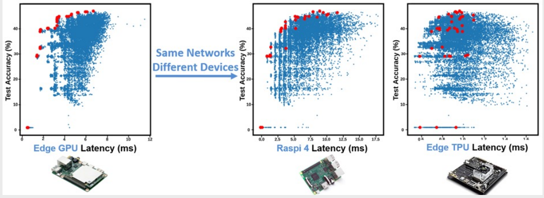
Analysis 2: Device-specific Cost Is Necessary

Hardware-cost on one device do NOT correlate well with hardware-cost on other devices

- Coefficient can be **as low as 0**

Analysis 3: Device-specific HW-NAS is Necessary

DNN architectures with the optimal accuracy vs. hardware-cost trade-offs in one device may not perform well in another device



HW-NAS-Bench: Easy-to-use APIs

- Create API

```
from hw_nas_bench_api import HWNASBenchAPI as HWAPI
hw_api = HWAPI("HW-NAS-Bench-v1_0.pickle", search_space="nasbench201")
```

- Get the real-measured/estimated hardware-cost

```
HW_metrics = hw_api.query_by_index(0, "cifar10")
```

- Example output:

```
===> Example to get use the hardware metrics
edgegpu_latency: 5.807418537139893 (ms)
edgegpu_energy: 24.226614330768584 (mJ)
raspi4_latency: 10.481976820010459 (ms)
edgetpu_latency: 0.9571811309997429 (ms)
pixel3_latency: 3.6058499999999998 (ms)
eyeriss_latency: 3.645620000000001 (ms)
eyeriss_energy: 0.6872827644999999 (mJ)
fpga_latency: 2.57296 (ms)
fpga_energy: 18.01072 (mJ)
```

Try it now:

https://github.com/GATECH-EIC/HW-NAS-Bench