Profiling Neural Blocks and Design Spaces for Mobile Neural Architecture Search

Keith G. Mills[1], Fred X. Han[2], Jialin Zhang[3], Seyed Saeed Changiz Rezaei[2], Fabian Chudak[2], Wei Lu[2], Shuo Lian[3], Shangling Jui[3] and Di Niu[1]

[1]University of Alberta [2]Huawei Technologies Canada Co., Ltd. [3]Huawei Kirin Solution, Shanghai, China

Open-source repository: https://github.com/Ascend-Research/BlockProfile

Neural Architecture Search (NAS) automates neural network design and has achieved state-of-the-art results in numerous deep learning applications.

Three primary components: Design Space, Search Algorithm, Performance Estimation Strategy. Bulk of research done for the latter two.

We propose a simple methodology for profiling Design Spaces across valuable performance metrics, e.g., accuracy and inference latency, across different target hardware devices.

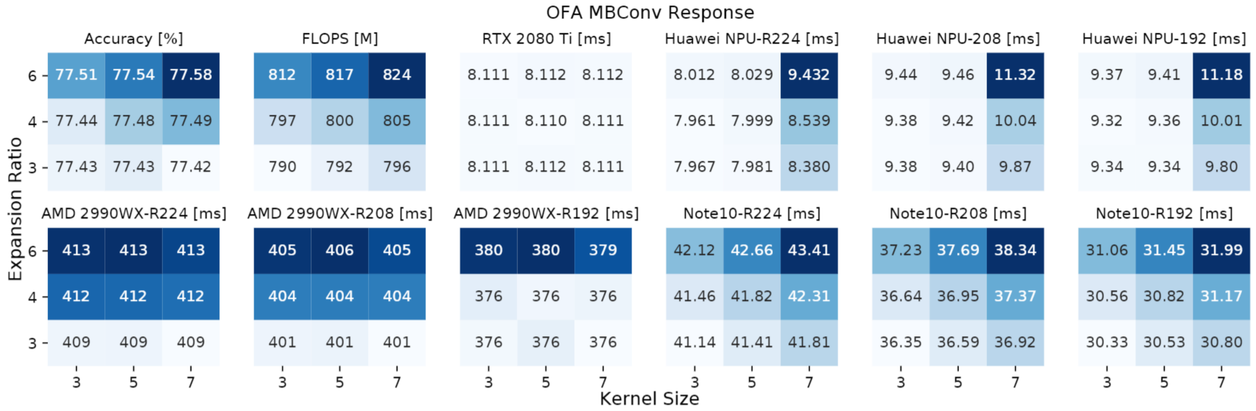
Devices considered: Huawei Kirin 9000 NPU, Nvidia RTX 2080 Ti GPU, AMD Threadripper 2990WX CPU and Samsung Note10.

Insights gleamed can be used to make pruned search spaces that outperform the original, including Once-for-All on MobileNetV3 (MBv3).

BLOCK-WISE SAMPLING AND PROFILING

Our method is rooted in the random sampling of architectures. To measure the impact of block *b* at layer *l* of unit *u* (denoted *(u,l,b)*), we sample many random architectures and affix *b* to *(u,l)* in each architecture, then measure the end-to-end response.

We can then iterate across all locations in the network, then average the results to get the general effect of block *b* on a desired metric.



RESULTS ON ONCE-FOR-ALL

Block-Wise Results on Once-for-All:

- Accuracy follows block size, FLOPS and visually correlated with Note10 latency.
- Latency on RTX 2080 Ti GPU close to constant.
- 2990WX CPU latency depends on expansion ratio (channels)
- NPU latency depends on kernel size.
- Lower resolution does not always mean lower latency.

EXPLOITING DISTRIBUTION DIFFERENCES

Pareto front performance depends on how different the accuracy and latency profiles are:

- NPU: Our insights outperform the original because Kernel Size 7 operations are not hardware friendly, but Kernel Size 5 is still very accuracy friendly.
- GPU: Negligible difference in block latency allows us to cut low-accuracy blocks with almost no additional latency.
- CPU: Latency differences exist but are very small.
- Note 10: Correlation means high latency operations are high accuracy operations. There is a trade-off.

WHAT IS A DESIGN SPACE?

The variable components of the network structure, primarily in the body. We abstract the Design Space to 3 levels:

- Units, *u*: Operate on unique tensor dimensions.
- Layers, *l*: Contain blocks, variable number within a unit.
- Blocks, *b*: The selectable operation sequences.
- Input image resolution size may also be considered.

We study three Design Spaces: Once-for-All, ProxylessNAS and ResNet50 and gain insights to block and layer sensitivity.

APPLICATION TO NAS

Derived insights can be used to prune a pre-defined space:

- Remove blocks from specific units or altogether.
- Limit the number of layers in a unit.
- Focus search on specific units.
- Optimize for one or more desired metrics.

Test insights using a simple random mutation algorithm.

- Compare pruned search space to the original.
- Accuracy-latency Pareto frontier optimization.
- Maximum accuracy search.

LATENCY CONSTRAINED PARETO FRONTIER SEARCH

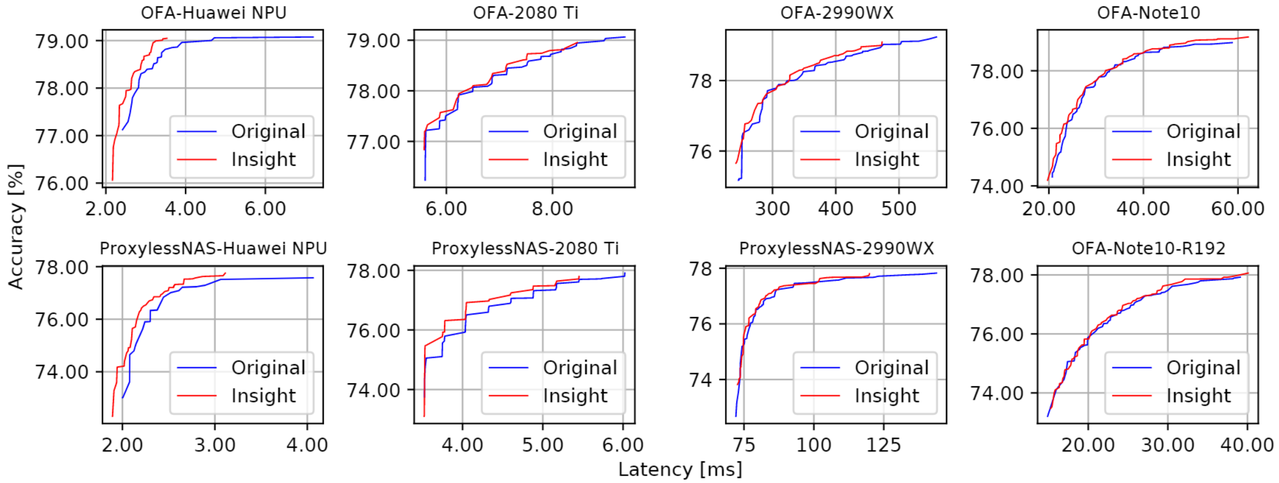


Figure 8: Pareto frontiers contrasting the original search spaces (blue) with our insight-based search spaces (red).

MAXIMUM ACCURACY SEARCH

Table 2: Maximum top-1 ImageNet accuracy search results on different design spaces, compared to existing works. We show averages over 5 random seeds for our experiments.

Model	Accuracy	MACs
MobileNetV2 [21]	72.0	300M
MobileNetV3-Large [10]	75.2	219M
OFA [2]	76.0	230M
OFA_{Large}	79.0	595M
OFA-insight	**79.2** ± 0.04	342M
OFA-base	78.9 ± 0.07	292M
ProxylessNAS-insight	**77.9** ± 0.04	417M
ProxylessNAS-base	77.6 ± 0.08	359M
ResNet50-insight	**80.0** ± 0.03	2.81B
ResNet50-base	79.9 ± 0.09	2.64B

Without considering latency and just optimizing for accuracy, our insights on Once-for-All allow us to find a pruned search space that not only outperforms the original version, but the original OFA_{Large}!