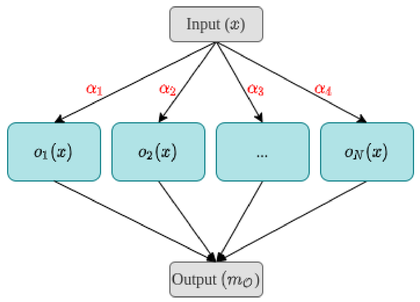
VPU specific CNNs through Neural Architecture Search

Ciarán Donegan, Hamza Yous, Saksham Sinha, Jonathan Byrne

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Overview

- Neural Architecture Search (NAS) methods can automate the design of neural networks and optimize both accuracy and efficiency.
- We use a differentiable NAS method to design CNNs optimized to run on Intel Movidius Vision Processing Unit (VPU).



- We profile CNN operations on MyriadX VPU to gather latency metrics.
- We use the ProxlessNAS method and incorporate the latency of the network on VPU hardware into the loss function.

$$\mathcal{L} = \mathcal{L}_{\mathsf{CE}} + \lambda \cdot \frac{\mathsf{LAT} - \mathsf{LAT}_T}{\mathsf{LAT}_T}$$

- Our NAS designed CNN outperforms MobileNetV2, being 1% more accurate and 13% faster on MyriadX VPU.
- We demonstrate that CNNs designed specifically for MyriadX VPU perfrom better than CNNs designed for mobile devices.

Comparison of Network Architectures

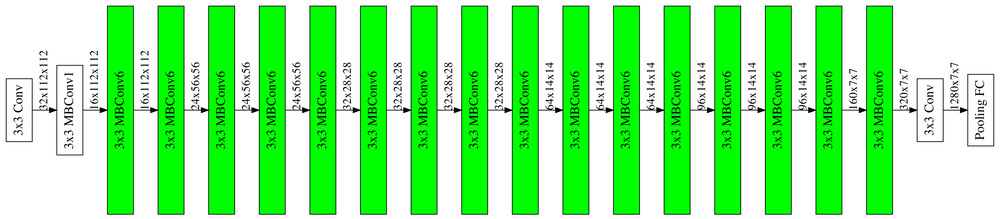


Figure 1: MobileNetV2

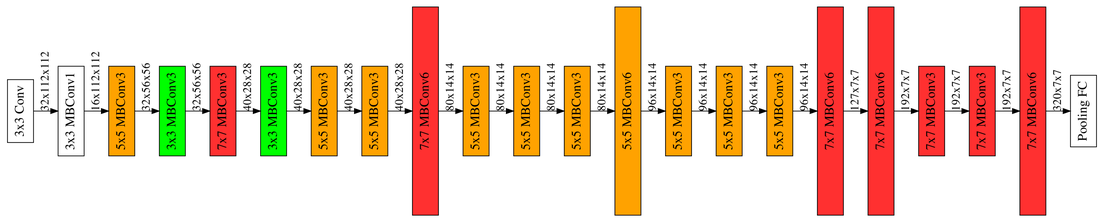


Figure 2: ProxylessNAS mobile

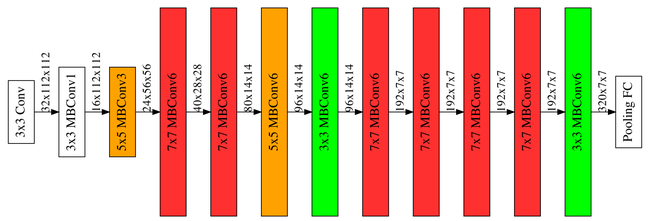


Figure 3: MyriadX VPU

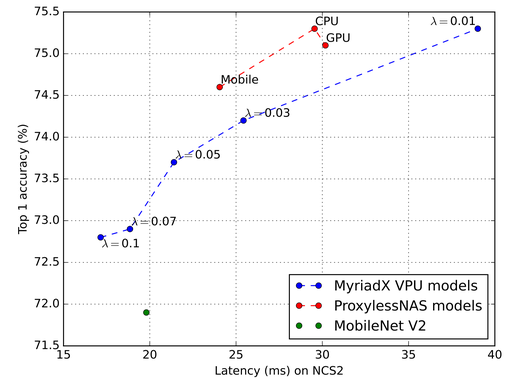
The colour of each block denotes the kernel size (Green = 3 x 3, Orange = 5 x 5 and Red = 7 x 7), the white blocks are fixed layers and the width of the block indicates the expansion ratio (wide block is expansion 6 and narrow block is expansion 3)

The NAS designed MyriadX VPU network reveals some the hardware's affinities with regard to network design.

- VPU network is far shallower.
- VPU network makes use of wider expansion blocks (in MBConv blocks) with expansion ratio 6 almost always chosen.
- VPU network uses larger kernels (7×7 kernel) more often than the other networks because the VPU hardware is designed to perform a high degreee of parallelism and so can efficiently process these large kernels.

Results

Performance of networks on ImageNet 2012 classification dataset. Latency is measured on Intel Neural Compute Stick 2 (NCS2).



	Top 1 acc (%)	NCS2 Latency (ms)
MobileNet V2	71.9	19.8
ProxylessNAS mobile	**74.6**	24.1
MyriadX VPU	72.8	**17.2**

Conclusion

- Our VPU-specific NAS designed network achieves impressive performance on ImageNet and outperforms MobileNetV2 in terms of both accuracy and latency.
- We show the usefulness of differentiable NAS and in particular hardware aware NAS methods like ProxylessNAS to design state-of-the-art CNNS.