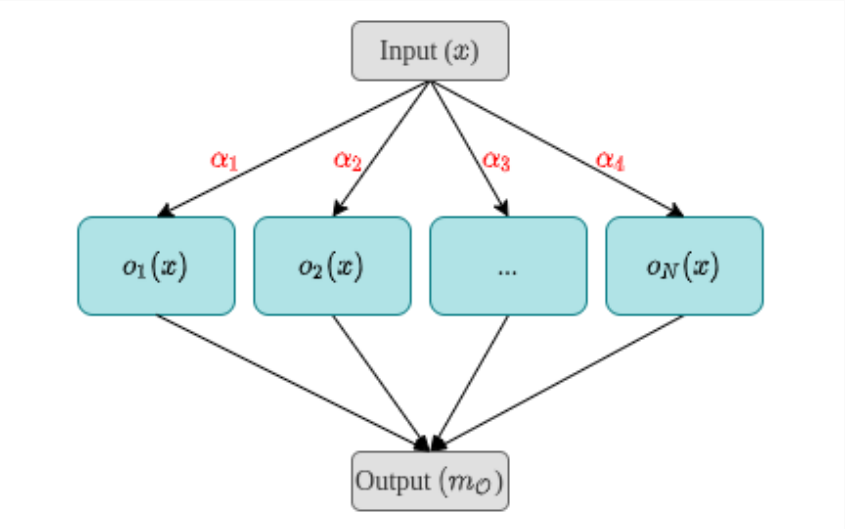
# VPU Specific CNNs through Neural Architecture Search

Ciarán Donegan, Hamza Yous, Saksham Sinha, Jonathan Byrne

Intel R&D Ireland Ltd, Trinity College Dublin.

# Motivation

- Complex CNNs are hard to deploy onto edge devices.
- Edge AI hardware accelerators - Intel Movidius Vision Processing Unit (VPU).
- NAS can optimize accuracy and efficiency.
- Differentiable NAS methods have a low search cost.



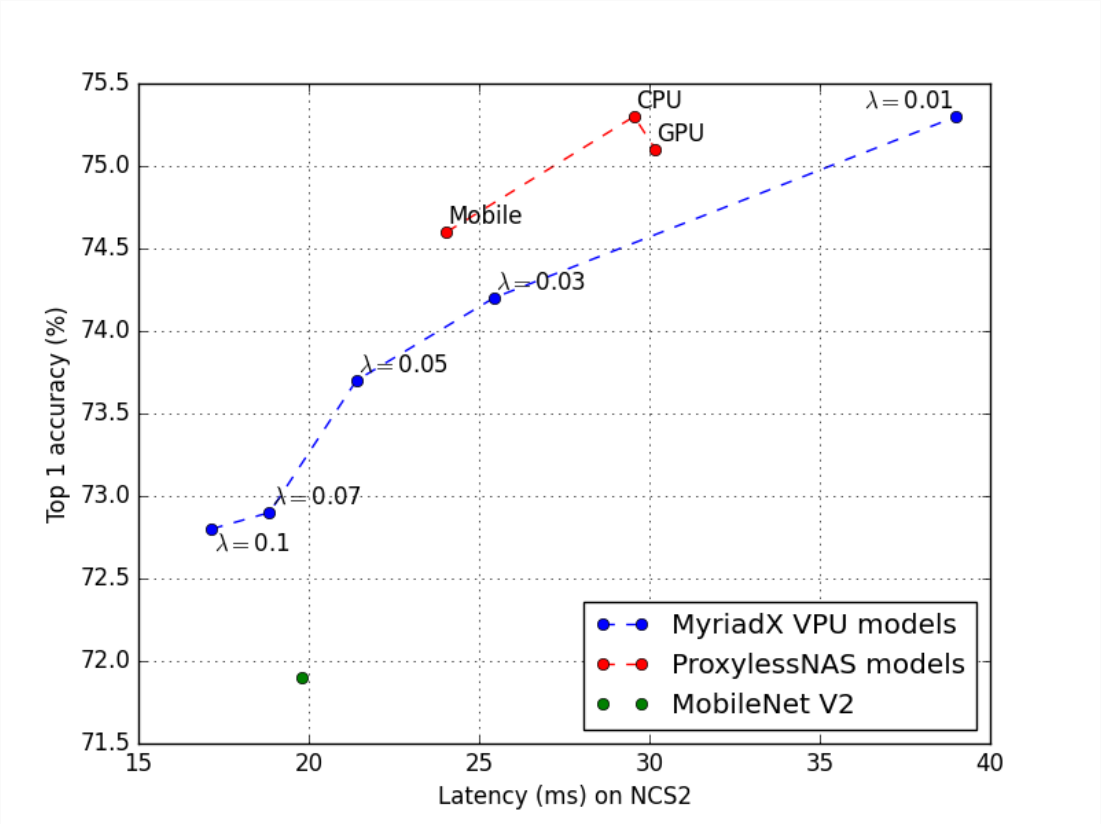
# Method

- Search space of MBConv blocks - kernel size {3, 5, 7}, expansion ratio {3,6}
- Build a latency look-up-table (LUT) by profiling operations on MyriadX VPU.
- Incorporate the network latency into the loss function.

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

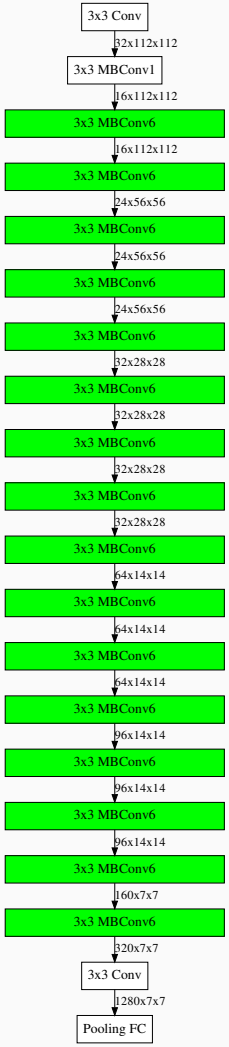
- Search for networks using the ProxylessNAS search algorithm.

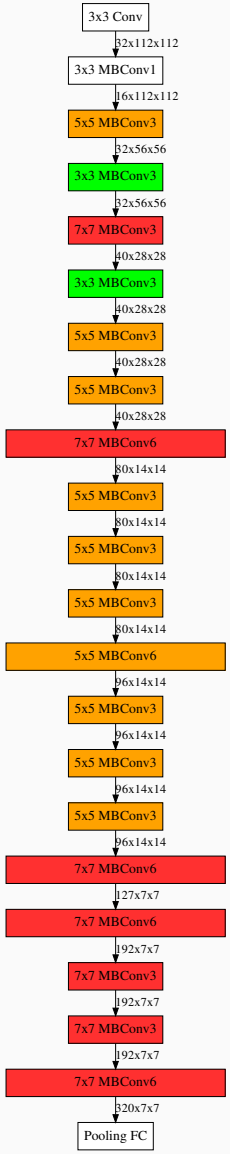
# Results




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

# Comparison of network architectures



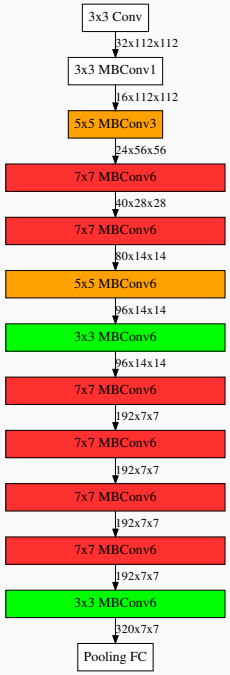




**Figure 1:** MobileNetV2 **Figure 2:** ProxylessNAS **Figure 3:** MyriadX VPU

# Conclusions

- NAS designed network outperforms MobileNetV2 on MyriadX VPU.
  - 1% more accurate
  - 13% faster on VPU.
- We demonstrate the use of differentiable NAS methods to design state-of-the-art networks for a specific hardware.
- Hardware specific CNNs are the future.