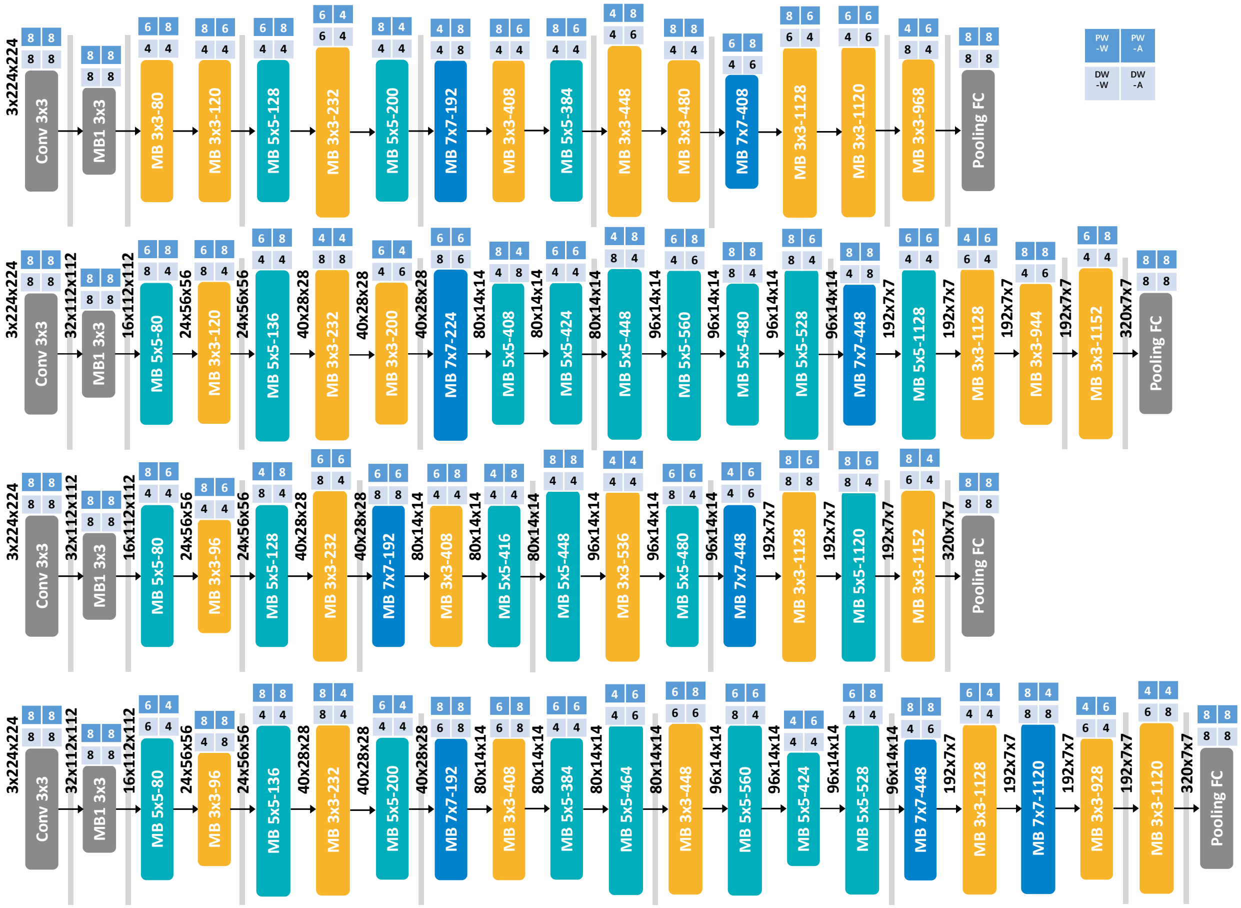

Figure 1: Illustration of the network architecture searched by APQ. PW/DW-W/A denotes the bit width for pointwise/depthwise convolution layer's weight/activation. The models are searched under 4-bit MobileNetV2 latency constraint, 6-bit MobileNetV2 latency constraint, 4-bit MobileNetV2 energy constraint, 6-bit MobileNetV2 energy constraint, respectively. One can see that shallow layer models are searched under a tight resource budget while deeper models are searched when the budget is loose, which is in line with intuition. The implementation for APQ can be found at https://drive.google.com/drive/folders/1wbPQSPC-rfivH8f0c80s8hKBwemq8c9N?usp=sharing