A APPENDIX

A.1 Architectures optimized with wiNAS

Our framework wiNAS, takes a given macro-architecture and optimizes each $3 \times 3$ convolutional layer by choosing from direct convolution or different Winograd configurations. For the search, all $1 \times 1$ convolutions were fixed to use $im2row$.

For wiNAS$\text{WA}$ in FP32, the resulting architecture only substituted the last convolution layer with $im2row$ instead of $F2$. The rest of the layers remained unchanged from the WA$F4$ configuration (which was described in Section 5.1). The same micro-architecture was used in CIFAR-10 and CIFAR-100.

For wiNAS$\text{WA}$ with 8-bit quantization and CIFAR-10, wiNAS replaced the 5th and second last convolutional layers with $im2row$, instead of $F4$ and $F2$ respectively. For CIFAR-100, more optimization was compared to WA$F4$.

The resulting micro-architecture optimization is shown in Figure 9 (left).

When introducing quantization in the search space, wiNAS$\text{WA-Q}$, the resulting architectures are shown in Figure 9 for both CIFAR-10 (middle) and CIFAR-100 (right).
Figure 9. Resulting architectures after optimizing a ResNet-18 macro-architecture using WiNAS. For WiNAS\textsubscript{WA} and CIFAR-100, the architecture resulted is shown on the left. With WiNAS\textsubscript{WA-Q}, that introduces quantization in the search space, the optimization resulted in different architectures for CIFAR-10 (middle) and CIFAR-100 (right), evidencing the difference in complexity of the later.