To reduce memory footprint and run-time latency, techniques such as neural net-work pruning and binarization have been explored separately. However, it is u