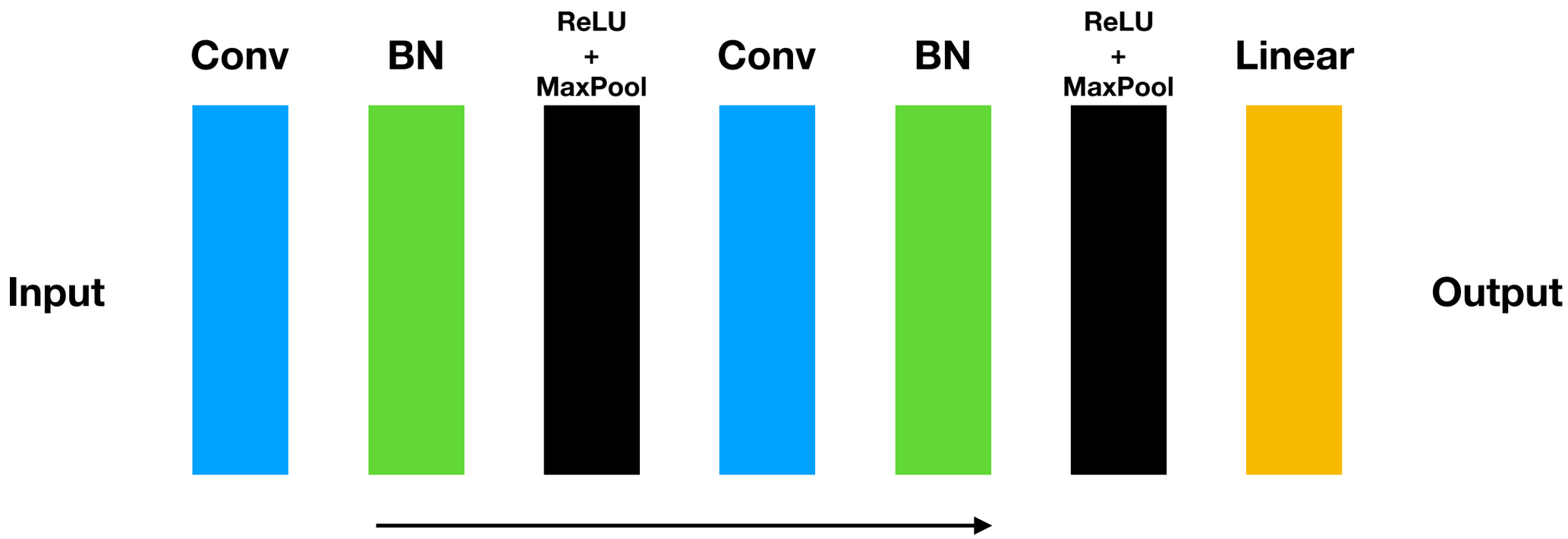
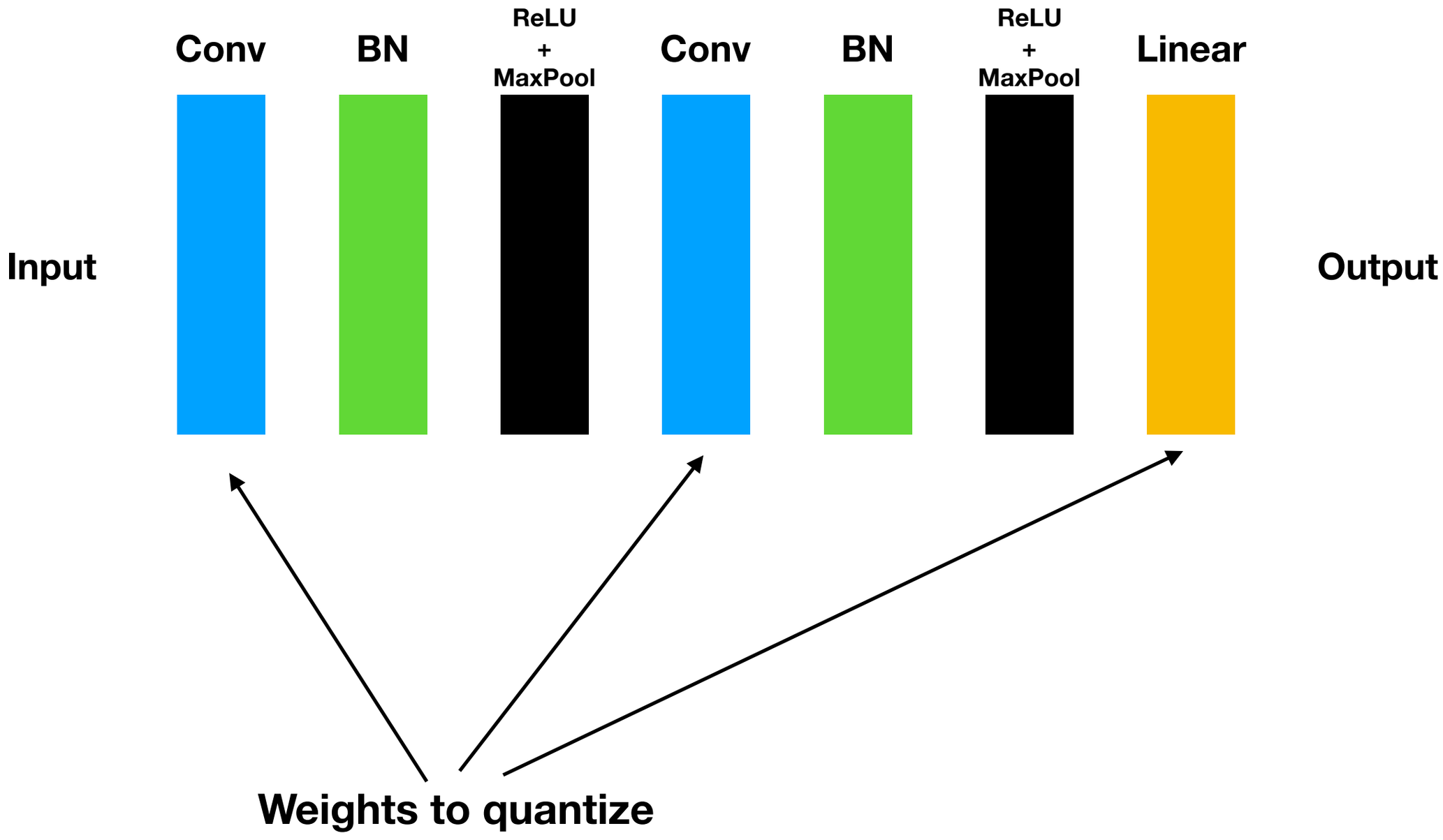
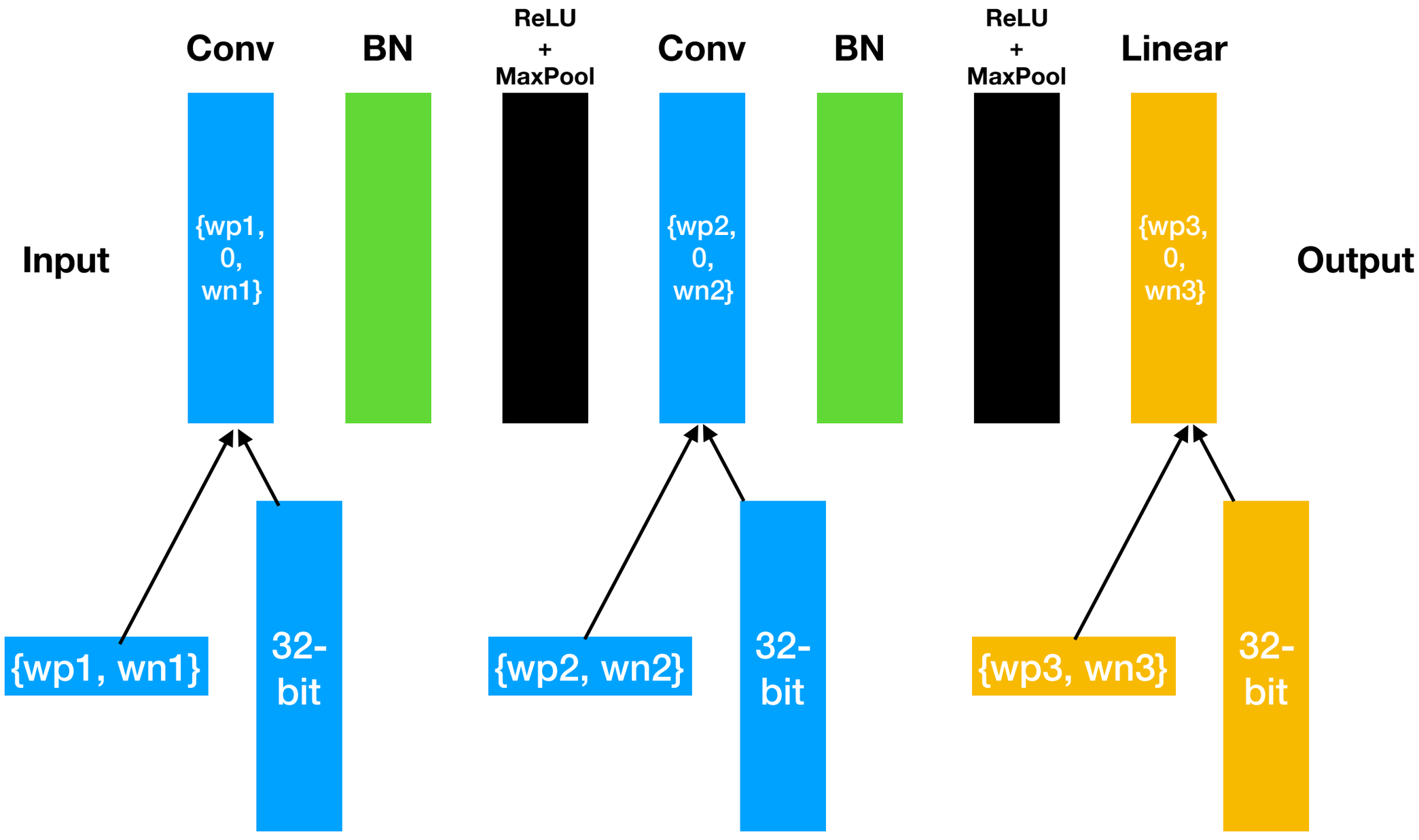
Original model: loaded with trained 32-bit weights

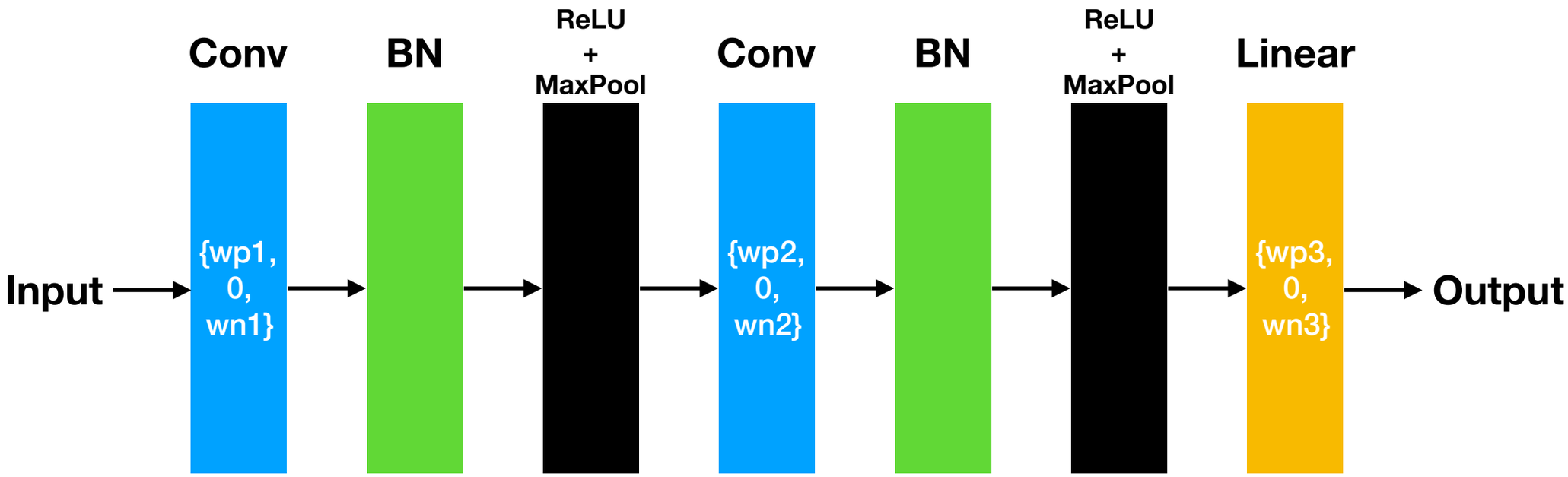


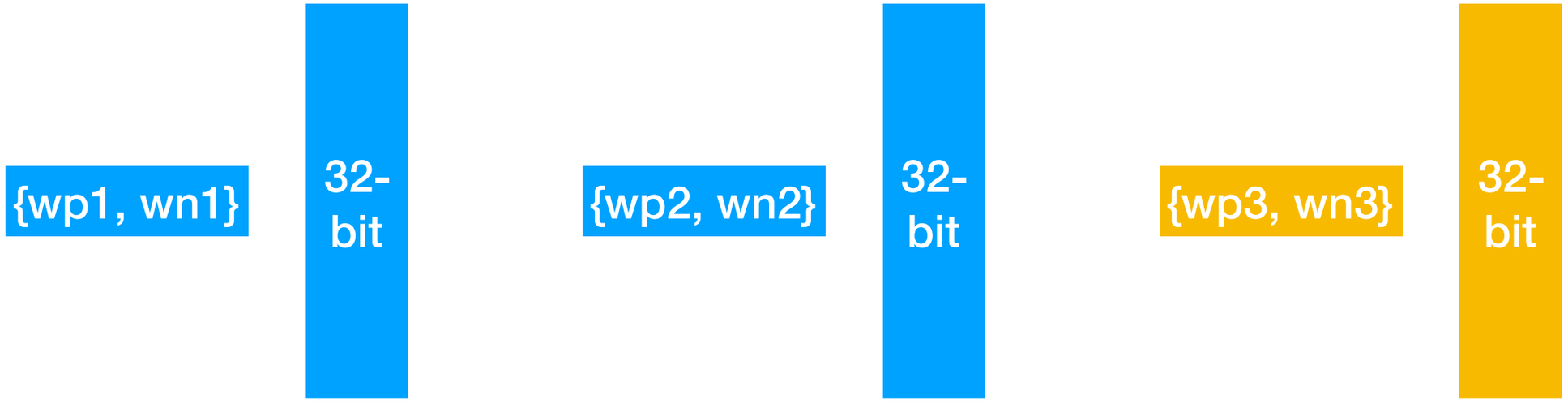

Input
Conv
BN
ReLU
+
MaxPool
Conv
BN
ReLU
+
MaxPool
Linear
Output
Weights to quantize

BEFORE forward propagation: Quantize weights

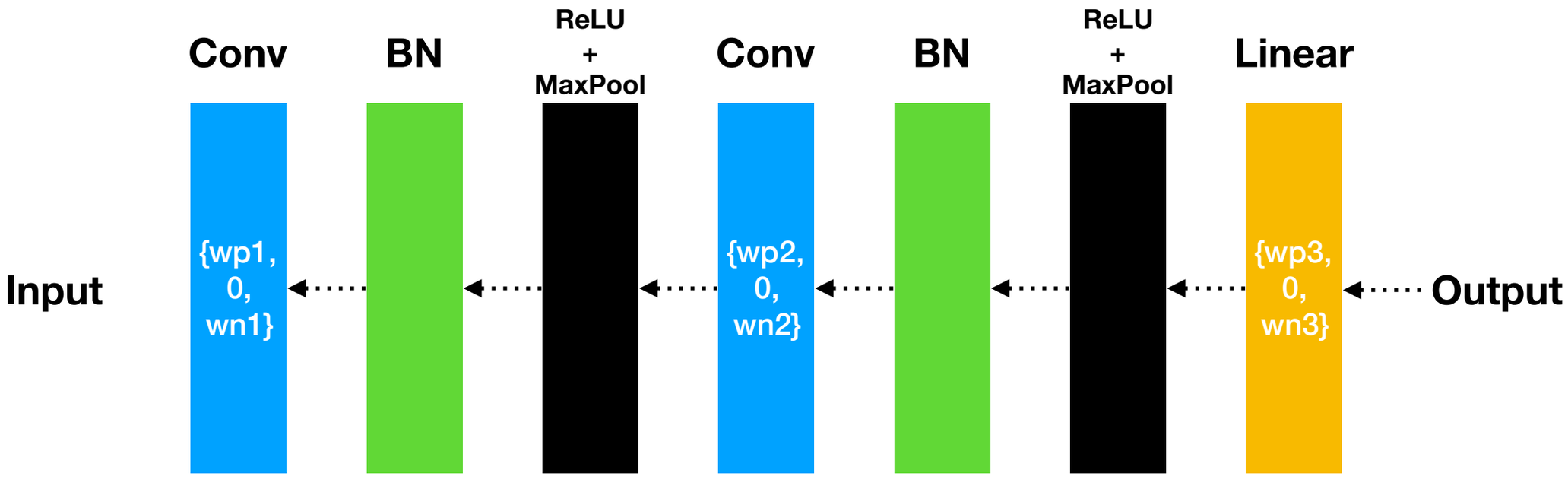


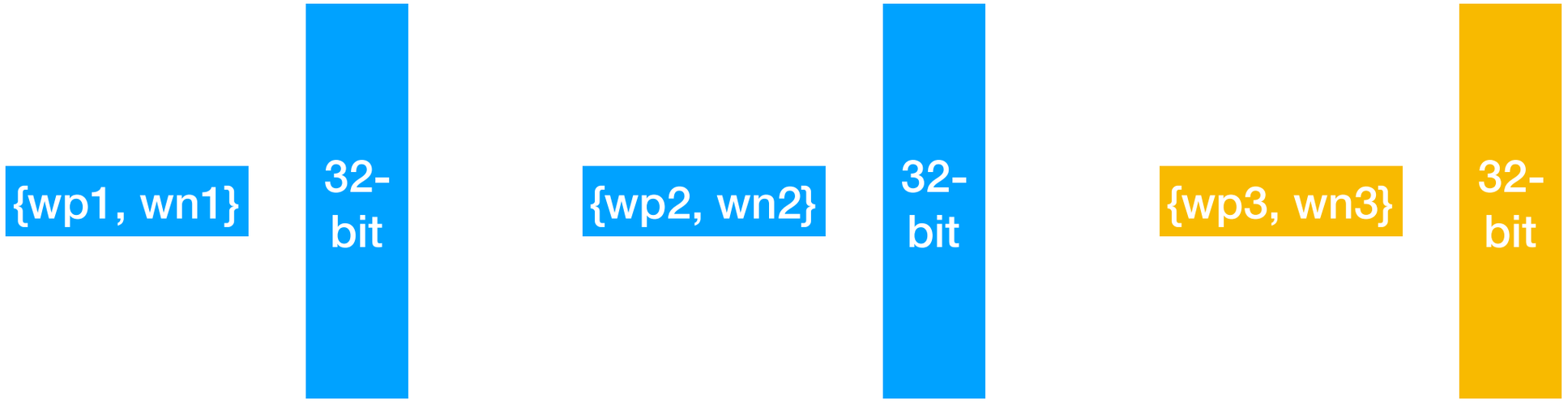
Forward propagation



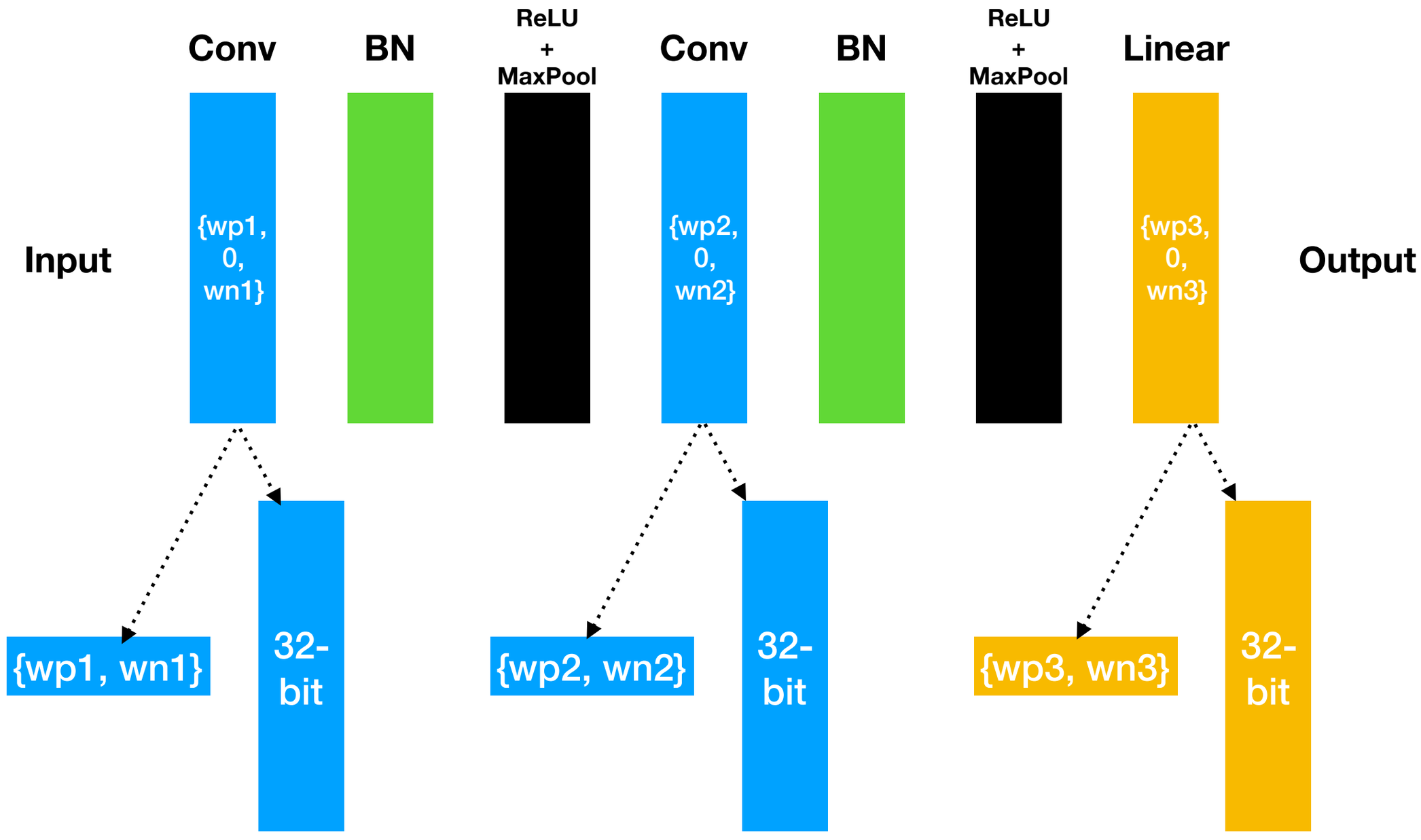


Back propagation 1: Gradients are backpropagated, only non-quantized weights are updated (BN layers in this case)





Back propagation 2: Manually set gradients for 32-bit weights and scaling factors (wp, wn) and updated them



BEFORE forward propagation: Quantize weights. Weights are now quantized from updated 32-bit weights