Reducing bit-widths of activations and weights of deep networks makes it efficient to compute and store them in memory, which is crucial in their deployments to