TABLE I: Comparison of different quantization methods. $\hat{w} \approx \hat{w}_q$

Method	w	$\hat{w}$	$\hat{w}_q$	Parameter of quantizer	layer-wise
DoReFa	$w \in \mathbb{R}$	$\hat{w} = \frac{\tanh(w)}{\max(\lvert\tanh(w)\rvert)} \in [-1, 1]$	$\hat{w}_q = \lfloor \frac{(\hat{w}+\beta)}{\alpha} \rceil \alpha - \beta$ $q_{\hat{w}} \in [0, \cdots, 2^k - 1]$	$\alpha = \frac{2}{2^k - 1}; \beta = 1$	
WRPN	$w \in \mathbb{R}$	$\hat{w} = w$ $\hat{w} \in \mathbb{R}$	$\hat{w}_q = \lfloor \text{clip}(\frac{\hat{w}}{\alpha}, -2^{k-1} - 1, 2^{k-1} - 1) \rceil \alpha$ $q_{\hat{w}} \in [-(2^{k-1} - 1), \ldots, 0, \ldots, 2^{k-1} - 1]$	$\alpha = \frac{1}{2^{k-1} - 1}$	
QIL	$w \in \mathbb{R}$	$\hat{w} = \begin{cases} 0, & \text{if } \lvert w\rvert < c - d \\ \text{sign}(w), & \text{if } \lvert w\rvert > c + d \\ (\frac{d}{2}\lvert w\rvert + \frac{d-c}{2d})^{\gamma} \cdot \text{sign}(w), & \text{otherwise} \end{cases}$ $\hat{w} \in [-1, 1]$	$\hat{w}_q = \lfloor \frac{\hat{w}}{\alpha} \rceil \alpha$ $q_{\hat{w}} \in [-(2^{k-1} - 1), \ldots, 0, \ldots, 2^{k-1} - 1]$	$\alpha = \frac{1}{2^{k-1} - 1};\ \gamma; c; d;$	
LSQ	$w \in \mathbb{R}$	$\hat{w} = w$ $\hat{w} \in \mathbb{R}$	$\hat{w}_q = \lfloor \text{clip}(\frac{\hat{w}}{\alpha}, -2^{k-1}, 2^{k-1} - 1) \rceil \alpha$ $q_{\hat{w}} \in [-2^{k-1}, \ldots, 0, \ldots, 2^{k-1} - 1]$	α	