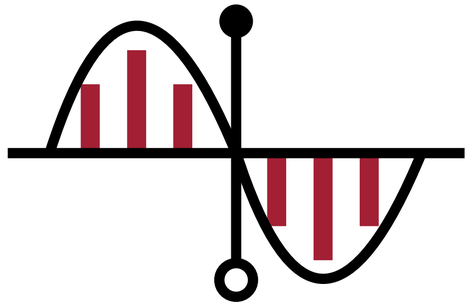
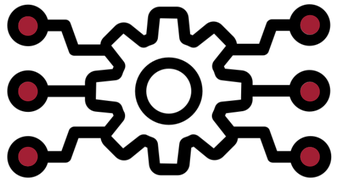






1. Quantization-aware scaling

$$\tilde{\mathbf{G}}_{\bar{\mathbf{W}}} = \mathbf{G}_{\bar{\mathbf{W}}} \cdot s_{\mathbf{W}}^{-2}$$

2. Sparse layer/tensor update

3. Tiny Training Engine

