QNN

1 Deep Neural Networks(DNN)

A DNN consists of (1) an input layer, (2) multiple hidden layers, and (3) an output layer.

A DNN with d layers is a non-linear multivariate function $\mathcal{N}: \mathbb{R}^n \to \mathbb{R}^s$.

- Input: $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = \mathbf{x}^1$.
- Hidden layer: $\mathbf{x}^i = \phi(\mathbf{W}^i \mathbf{x}^{i-1} + \mathbf{b}^i)$.
  - Activation function: ϕ, e.g. $\text{ReLU}(x) = \max(x, 0)$.
  - Weight matrix: $\mathbf{W}^i$, $2 \leq i \leq d$.
  - Bias vector: $\mathbf{b}^i$, $2 \leq i \leq d$.
- Output: $\mathcal{N}(\mathbf{x}) = \mathbf{x}^d$.
- Notations:
  - $n_1 = n, n_2, \ldots, n_d = s$

2 Quantization

Symmetric uniform quantization is considered here.

Quantization Configuration. A quantization configuration $\mathcal{C}$ is a tuple $\langle \tau, Q, F \rangle$, where Q and F are the total bit size and the fractional bit size allocated to a value, respectively, and $\tau \in \{+, \pm\}$ indicates if the quantized value is unsigned or signed.

Example. Given a real number $x \in \mathbb{R}$ and a quantization configuration $\mathcal{C} = \langle \tau, Q, F \rangle$, its quantized integer counterpart $\hat{x}$ and the fixed-point counterpart $\tilde{x}$ under the symmetric uniform quantization scheme are: $\hat{x} = \text{clamp}(\lfloor 2^F \cdot x \rceil, \mathcal{C}^{\text{ub}}, \mathcal{C}^{\text{lb}})$ and $\tilde{x} = \hat{x}/2^F$, where:

- $\mathcal{C}^{\text{ub}}, \mathcal{C}^{\text{lb}}$

$$\mathcal{C}^{\text{lb}} = \begin{cases} 0, & \tau = + \\ -2^{Q-1}, & \text{otherwise} \end{cases}$$

$$\mathcal{C}^{\text{ub}} = \begin{cases} 2^Q - 1, & \tau = + \\ 2^{Q-1} - 1, & \text{otherwise} \end{cases}$$

- $\lfloor \cdot \rceil$ is the round-to-nearest integer operator
- The clamping function $\text{clamp}(x, a, b)$ with a lower bound a and an upper bound b

$$\text{clamp}(x, a, b) = \begin{cases} a, & \text{if } x < a; \\ x, & \text{if } a \leq x \leq b; \\ b, & \text{if } x > b. \end{cases}$$

Definition (Quantized Neural Network). Given quantization configurations for the weights, biases, output of the input layer and each hidden layer as $\mathcal{C}_w = \langle \tau_w, Q_w, F_w \rangle$, $\mathcal{C}_b = \langle \tau_b, Q_b, F_b \rangle$, $\mathcal{C}_{in} = \langle \tau_{in}, Q_{in}, F_{in} \rangle$, $\mathcal{C}_h = \langle \tau_h, Q_h, F_h \rangle$, the quantized version (i.e., QNN) of a DNN $\mathcal{N}$ with d layers is a function $\widehat{\mathcal{N}}: \mathbb{Z}^n \to \mathbb{R}^s$ such that $\widehat{\mathcal{N}} = \hat{l}_d \circ \hat{l}_{d-1} \circ \cdots \circ \hat{l}_1$. Then, given a quantized input $\hat{\mathbf{x}} \in \mathbb{Z}^n$, the output of the QNN $\hat{\mathbf{y}} = \mathcal{N}(\hat{\mathbf{x}})$ can be obtained by the following recursive

computation:

- Input layer $\hat{l}_1 : \mathbb{Z}^n \to \mathbb{Z}^{n_1}$ is the identity function;
- Hidden layer $\hat{l}_i : \mathbb{Z}^{n_{i-1}} \to \mathbb{Z}^{n_i}$ for $2 \leq i \leq d-1$ is the function such that for each $j \in [n_i]$,

$$\hat{\mathbf{x}}_j^i = \text{clamp}(\lfloor 2^{F_i} \widehat{\mathbf{W}}_{j,:}^i \cdot \hat{\mathbf{x}}^{i-1} + 2^{F_h - F_b} \hat{\mathbf{b}}_j^i \rceil, 0, \mathcal{C}_h^{\text{ub}}),$$

  where F_i is $F_h - F_w - F_{in}$ if $i = 2$, and $-F_w$ otherwise;
- Output layer $\hat{l}_d : \mathbb{Z}^{n_{d-1}} \to \mathbb{R}^s$ is the function such that

$$\hat{\mathbf{y}} = \hat{\mathbf{x}}^d = \hat{l}_d(\hat{\mathbf{x}}^{d-1}) = 2^{-F_w} \widehat{\mathbf{W}}^d \hat{\mathbf{x}}^{d-1} + 2^{F_h - F_b} \hat{\mathbf{b}}^d.$$

  where for every $2 \leq i \leq d$ and $k \in [n_{i-1}]$, $\hat{\mathbf{W}}_{j,k}^i = \text{clamp}(\lfloor 2^{F_w} \hat{\mathbf{W}}_{j,k}^i, \mathcal{C}_w^{\text{ub}}, \mathcal{C}_w^{\text{lb}})$ is the quantized weight and $\hat{\mathbf{b}}_j^i = \text{clamp}(\lfloor 2^{F_b} \hat{\mathbf{b}}_j^i, \mathcal{C}_b^{\text{ub}}, \mathcal{C}_b^{\text{lb}})$ is the quantized bias.

Definition (Quantization Error Bound). Given a DNN $\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^s$, the corresponding QNN $\widehat{\mathcal{N}} : \mathbb{Z}^n \to \mathbb{R}^s$, a quantized input $\hat{\mathbf{x}} \in \mathbb{Z}^n$, a radius $r \in \mathbb{N}$ and an error bound $\epsilon \in \mathbb{R}$. The QNN $\widehat{\mathcal{N}}$ has a quantization error bound of ϵ w.r.t. the input region $R(\hat{\mathbf{x}}, r) = \{\hat{\mathbf{x}}' \in \mathbb{Z}^n \mid \|\hat{\mathbf{x}}' - \hat{\mathbf{x}}\|_\infty \leq r\}$ if for every $\hat{\mathbf{x}}' \in R(\hat{\mathbf{x}}, r)$, we have $||2^{-F_h} \mathcal{N}(\hat{\mathbf{x}}') - \mathcal{N}(\mathbf{x}')\|_\infty < \epsilon$, where $\mathbf{x}' = \hat{\mathbf{x}}' / (\mathcal{C}_{in}^{\text{ub}} - \mathcal{C}_{in}^{\text{lb}})$.