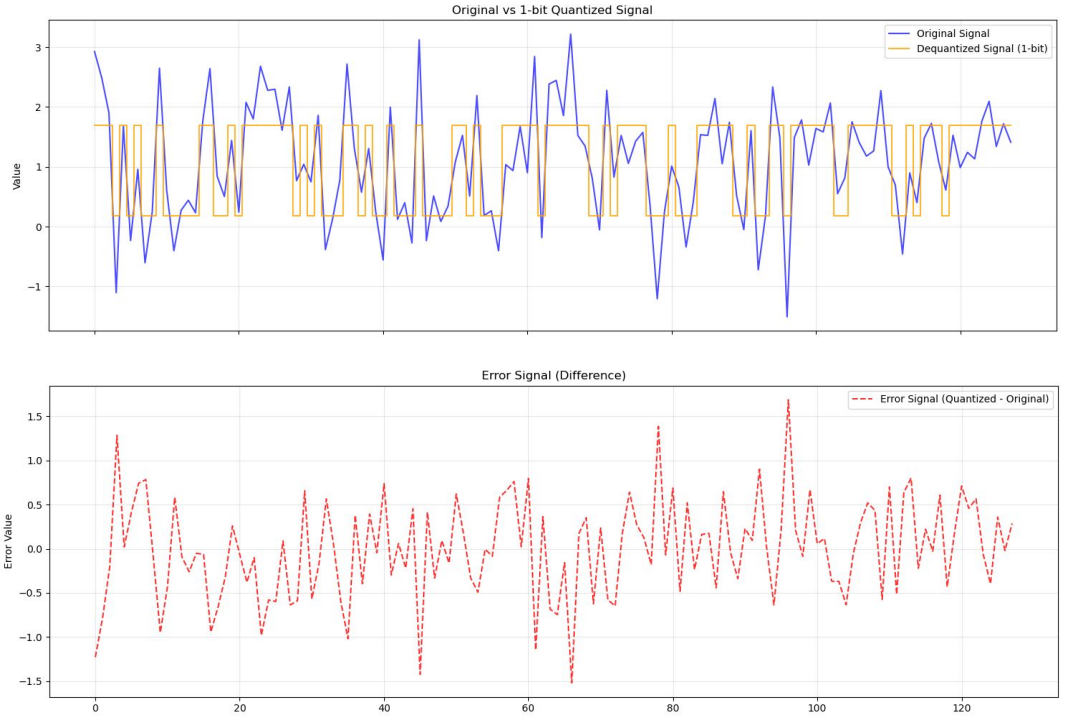
Robust Training of Neural Networks at Arbitrary Precision and Sparsity

Chengxi Ye, Grace Chu, Yanfeng Liu, Yichi Zhang,
Lukasz Lew, Li Zhang, Mark Sandler, Andrew Howard

Quantized Signal = Unquantized Signal + Error Signal

- $y = x + n$
- NNs with 0,1 signals
- Hodgkin-Huxley Model



Classic Formulation of Quantization

- Quantization + Dequantization: $$y = s \cdot \text{round}\left(\frac{x}{s}\right)$$
- round() is not differentiable.
- The straight-through estimator approach: $$\frac{dy}{dx} = 1$$

Our Error Injection Formulation

- Quantization Error: $\delta = \text{round}(\frac{x}{s}) - \frac{x}{s}$
- Error Injection: $y = s \cdot (\frac{x}{s} + \delta) = x + s \cdot \delta$
- The gradient is well defined: $\frac{dy}{dx} = 1 + s'(x) \cdot \delta$

Comparing the Two Formulations

- Input signal gradient: $\frac{dL}{dx} = \frac{dL}{dy} \cdot \frac{dy}{dx}$
- Straight-through Estimator: $\frac{dy}{dx} = 1$
- The true gradient is: $\frac{dy}{dx} = 1 + s'(x) \cdot \delta$

Novelties in Our Work

- Sparsification is a special case of quantization
  - Quantize small values to 0
- Minimize the Quantization Error

$$\min_{s_g, b_g} \frac{1}{2N} \left\| s_g \cdot \boldsymbol{q} + b_g \cdot \mathbf{1} - \boldsymbol{x} \right\|^2 + \frac{\lambda}{2} s_g^2$$

- Efficient Quantized Matrix Multiplication

Pareto Frontier Curves for Quantization