Analytical Guarantees on Numerical Precision of Deep Neural Networks
Charbel Sakr, Yongjune Kim, Naresh Shanbhag
University of Illinois at Urbana-Champaign

Motivation
Machine Learning ASICs

Current Approaches
- Stochastic Rounding [Gupta, ICML’15 – Bengio, NIPS’16]
- Trial-and-error approach [Sung, SIPS’14]
- SQNR based precision allocation [Lin, ICML’16]

Setup
Quantization Noise Model
\[ x_q = x + q \quad q \sim U\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right] \quad \Delta = 2^{-(B-1)} \]

Classification
\[ \hat{y} = \arg \max_{i=1, \ldots, M} z_i \]
\[ z_i = f \left( \{a_h, w_h\} \in A, \{w_h\} \in W \right) \]

Output Quantization
\[ z_{i} + q_{z_i} = f \left( \{a_h + q_{a_h}, w_h + q_{w_h}\} \in A, \{w_h\} \in W \right) \]
\[ q_{z_i} = \sum_{h \in A} q_{a_h} \frac{\partial z_i}{\partial a_h} + \sum_{h \in W} q_{w_h} \frac{\partial z_i}{\partial w_h} \]

Simplified but meaningful model of complexity
\[ \leq 1\% \]

Accelerator with height (\#neurons/layer)

1\% SQNR based precision allocation [Lin, 10 (MNIST) ; VGG = Other works considered + \( B \)]

MP2

Mismatch probability decreases
2048 on CIFAR

Data dependence (compute once and reuse):
\[ E_A \leq \mathbb{E} \left[ \sum_{i=1}^{M} \frac{\|z_i - z_{f_i}\|^2}{2^{4|Z_i - z_{f_i}|^2}} \right] \]
\[ E_W \leq \mathbb{E} \left[ \sum_{i=1}^{M} \frac{\|z_i - z_{f_i}\|^2}{2^{4|Z_i - z_{f_i}|^2}} \right] \]

\[ \Delta_A = 2^{-(B_A - 1)} \]
\[ \Delta_W = 2^{-(B_W - 1)} \]

Tighter Bound on \( p_m \)
\[ p_m \leq \mathbb{E} \left[ \sum_{i=1}^{M} e^{S(i, \hat{Y}_f)} P_1(i, \hat{Y}_f) P_2(i, \hat{Y}_f) \right] \]

\( M \): Number of Classes; \( S \): Signal to quantization noise ratio; \( P_1 \) & \( P_2 \): Correction factors

Second Order Bound on \( p_m \)
\[ p_m \leq \Delta_A^2 E_A + \Delta_W^2 E_W \]

- Input/Weight precision trade-off:
\[ B_A - B_W = \text{round} \left( \log_2 \frac{E_A}{E_W} \right) \]

- Computational cost
- Total number of FAs used assuming folded MACs
- Number of FAs per MAC:
\[ DB_A B_W + (D - 1)(B_A + B_W + \left\lfloor \log_2(D) \right\rfloor - 1) \]

- Represenational cost
- Total number of bits needed to represent weights and activations
- High level measure of area and communications cost (data movement):
\[ |A| B_A + |W| B_W \]

- Other works considered
- Stochastic quantization (SQ)
- 784 – 1000 – 1000 – 10 (MNIST)
- 64CS – MP2 – 64CS – MP2 – 64FC – 64FC – 10 (CIFAR10)
- BinaryNet (BN)
- 784 – 2048 – 2048 – 2048 – 10 (MNIST) ; VGG (CIFAR10)

Complexity Comparison

ConvNet on CIFAR-10
Architecture: 64C5 – 64C1 – 64C1 – MP2 – 64CS – 64C1 – 64C5 – MP2 – 64CS – 64FC – 64FC – 64FC – 10

A: \( B_W = B_A; p_m \leq 1\% \) (Theorem 1)
B: \( B_W = B_A; p_m \leq 1\% \) (Theorem 2)
C: \( B_W = B_A + 3; p_m \leq 1\% \) (Theorem 1)
D: \( B_W = B_A + 3; p_m \leq 1\% \) (Theorem 2)

MISMATCH PROBABILITY

\( \{ a_h \}_{h \in A} \rightarrow \hat{y} \)
\[ \{ w_h \}_{h \in W} \rightarrow \hat{f} \]
\[ B_N \]
\[ B_P \]
\[ \hat{y}_f \]
\[ \hat{f}_t \]
\[ p_m = \Pr \{ \hat{y}_f \neq \hat{f}_t \} \]

MLP on MNIST
Algorithm: 784 – 512 – 512 – 10

Architecture: 784 – 512 – 512 – 10 – 10

A: \( B_W = B_A; p_m \leq 1\% \) (Theorem 1)
B: \( B_W = B_A; p_m \leq 1\% \) (Theorem 2)
C: \( B_W = B_A + 3; p_m \leq 1\% \) (Theorem 1)
D: \( B_W = B_A + 3; p_m \leq 1\% \) (Theorem 2)

Simplification cost

No theoretical guarantees on accuracy

How are they choosing these precisions?
Why is it working?
Can it be determined analytically?

Acknowledgement
This work was supported in part by Systems on Nanoscale Information fabricS (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.