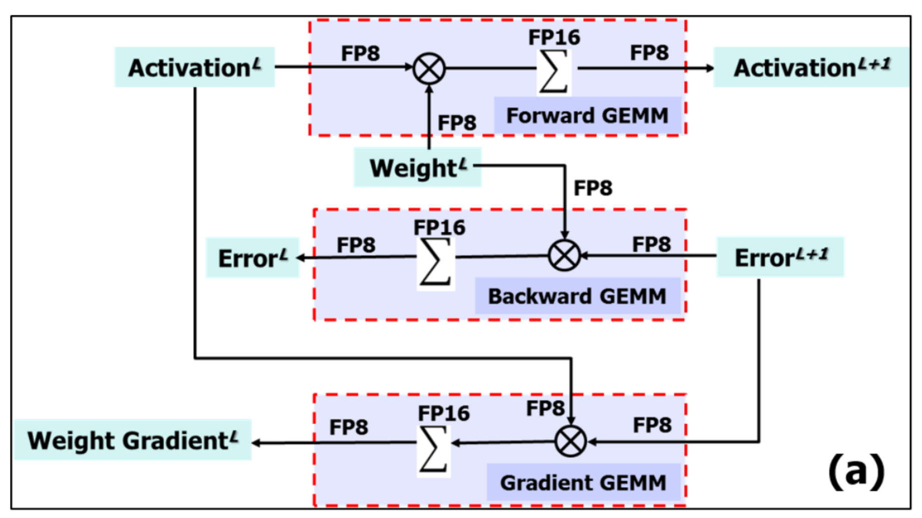
Accumulation Bit-Width Scaling for Ultra-Low Precision Training of Deep Neural Networks

Charbel Sakr[1], Naigang Wang[2], Chia-Yu Chen[2], Jungwook Choi[2], Ankur Agrawal[2], Naresh Shanbhag[1], Kailash Gopalakrishnan[2]

[1]University of Illinois at Urbana-Champaign, [2]IBM T.J. Watson Research Center

Reduced Precision FP Training

Representation & Accumulation Quantization



[Wang et al., NeurIPS 2018]

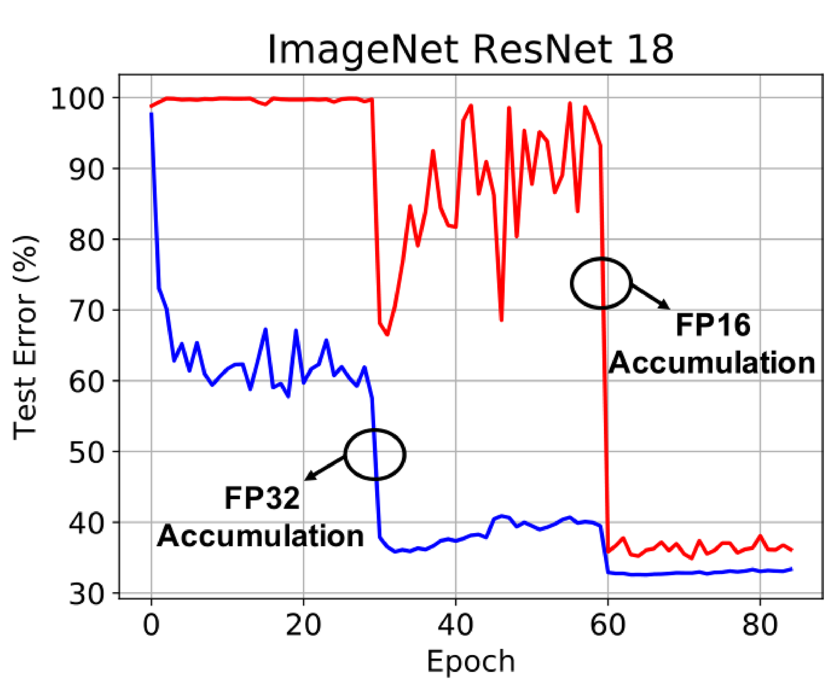
Chunk-Based Accumulation

```
Input: {x_n}_{n=1:N}, {y_n}_{n=1:N} (FP_mult),
Parameter: chunk size CL
Output: sum  (FP_acc)
sum = 0.0; idx = 0; num_ch = N/CL
for n=1:num_ch {
  sum_ch = 0.0
  for i=1:CL {
    idx++
    tmp = x_idx · y_idx  (in FP_mult)
    sum_ch += tmp  (in FP_acc) }
  sum += sum_ch  (in FP_acc) }        (a)
```

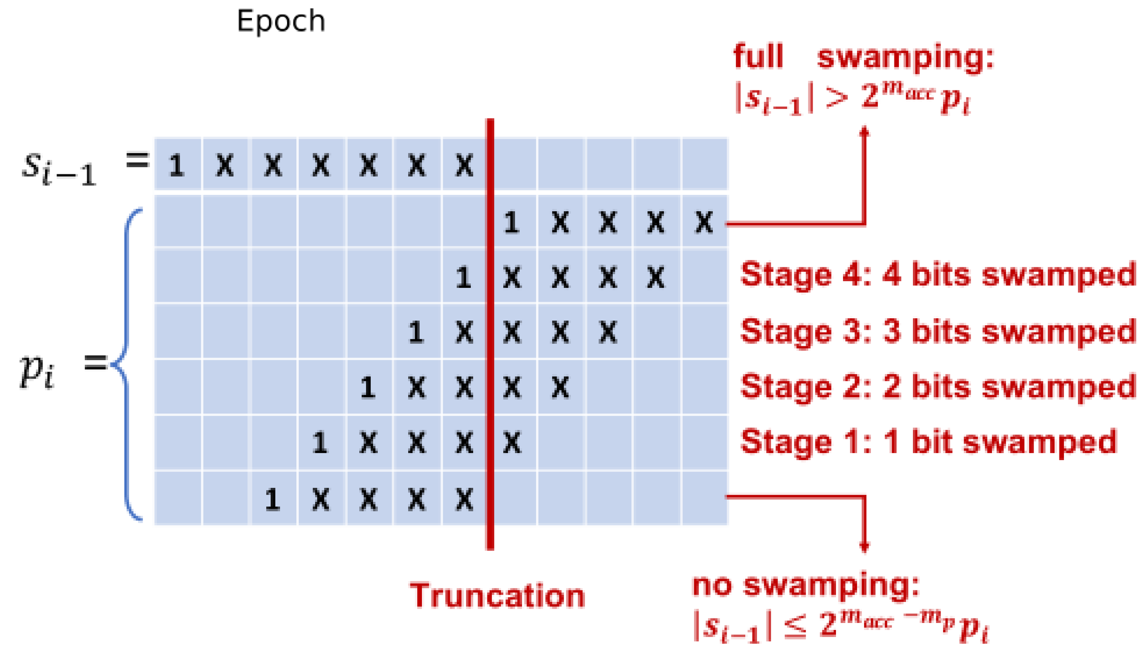
[Wang et al., NeurIPS 2018]

- tensors are quantized to FP8 = (1,5,2)
- accumulations are in FP16 = (1,6,9) but require chunking
- what is the accumulation precision required?

Problem Setup



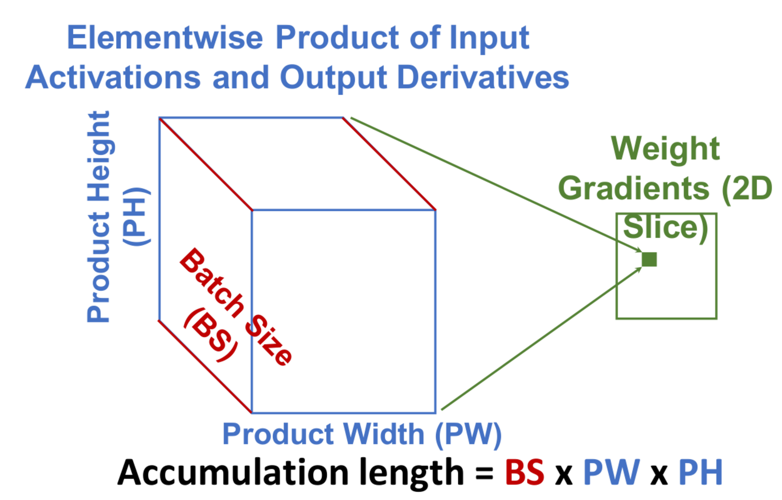
- reducing representation precision in FP format is well studied [Wang et al., NeurIPS'2018]
- problem of reducing accumulation precision largely overlooked



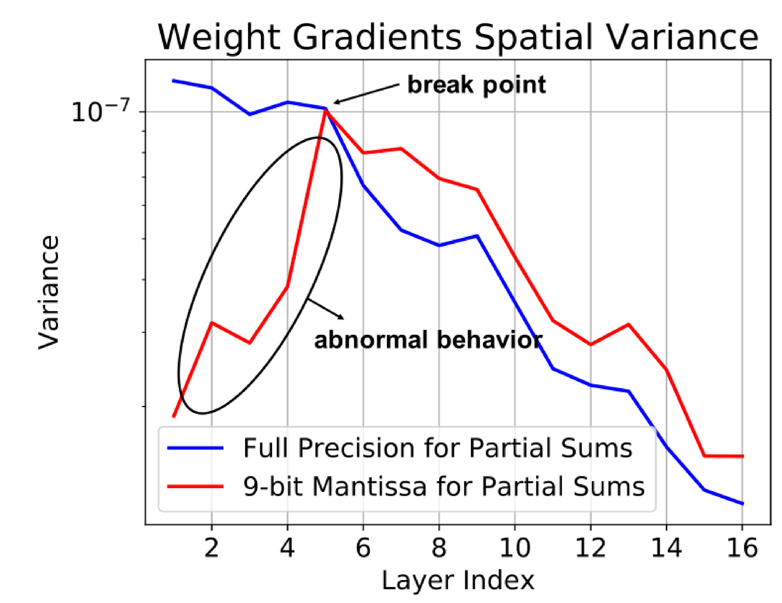
- accumulators typically designed conservatively because swamping effects are very destructive and intractable (hard to analyze)

Accumulation Variance

Gradient Accumulation

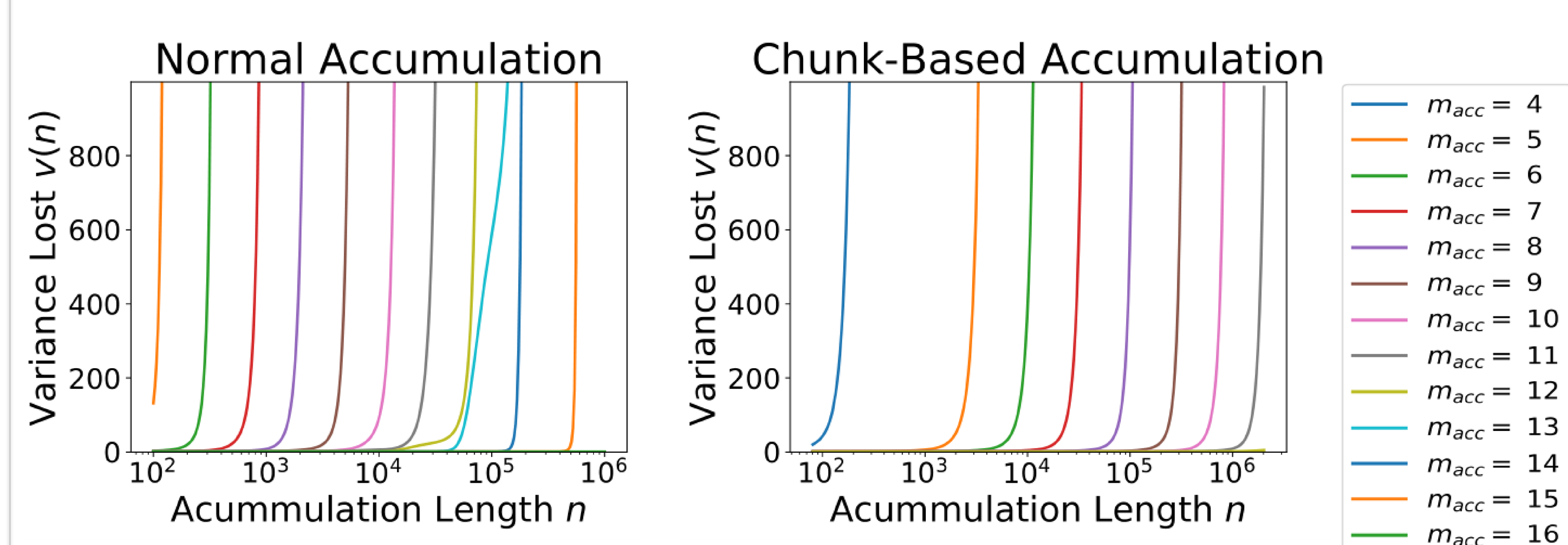


Variance Lost



- analysis reveals correlation between accumulated variance and convergence behavior
- break point corresponds to increase in accumulation length
- hypothesis: there is a relationship between precision, accumulation variance, and accumulation length

Variance Retention Ratio



$$VRR = \frac{\sum_{i=2}^{n-1}(i-\alpha)_+ q_i \mathbf{1}_{\{i>\alpha\}} + \sum_{j_r=2}^{m_p}(n-\alpha_{j_r})_+ q_i' \mathbf{1}_{\{n>\alpha_{j_r}\}} + nk_3}{kn}$$

$$where\ (x)_+ = \begin{cases} x \text{ if } x>0 \\ 0 \text{ otherwise}\end{cases},\ \mathbf{1}_A = \begin{cases} 1 \text{ if } A \text{ is true} \\ 0 \text{ otherwise}\end{cases},$$

$$\alpha = \frac{2^{m_{acc}-3m_p}}{3}\sum_{j=1}^{m_p} 2^j(2^j-1)(2^{j+1}-1),\ q_i = 2Q\left(\frac{2^{m_{acc}}}{\sqrt{i}}\right)\left(1-2Q\left(\frac{2^{m_{acc}}}{\sqrt{i-1}}\right)\right),$$

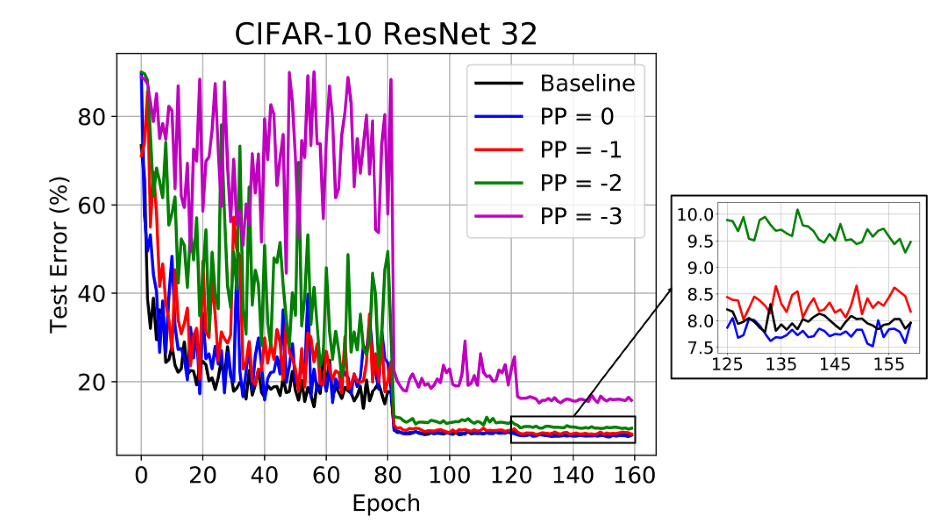
$$\alpha_{j_r} = \frac{2^{m_{acc}-3m_p}}{3}\sum_{j=1}^{j_r-1} 2^j(2^j-1)(2^{j+1}-1),$$

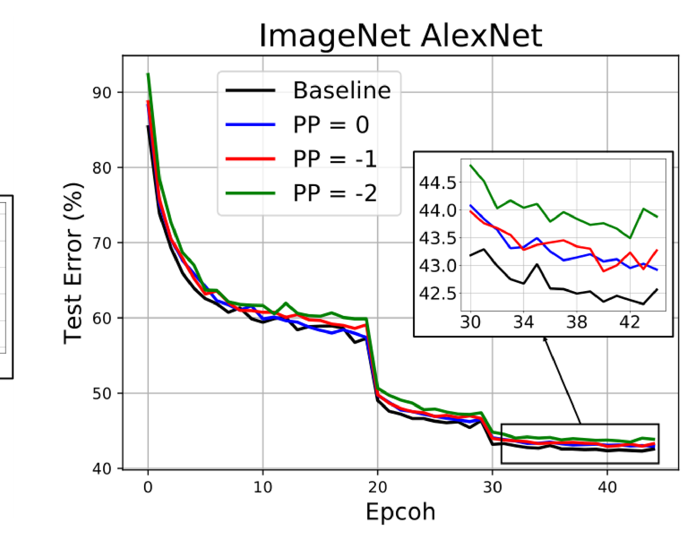
$$q_{j_r}' = N_{j_r-1}2Q\left(\frac{2^{m_{acc}-m_p+j_r-1}}{\sqrt{n}}\right)\left(1-2Q\left(\frac{2^{m_{acc}-m_p+j_r}}{\sqrt{n}}\right)\right), k = k_1+k_2+k_3,$$

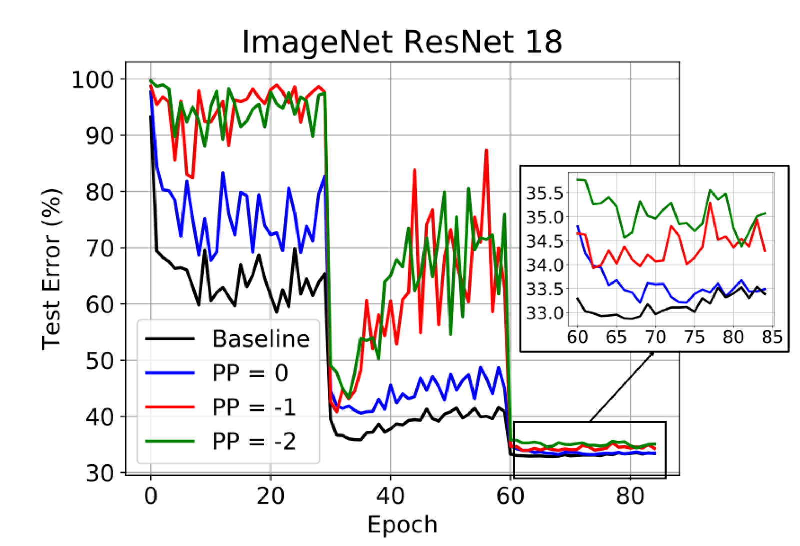
$$k_1 = \sum_{i=2}^{n-1} q_i \mathbf{1}_{\{i>\alpha\}},\ k_2 = \sum_{j_r=2}^{m_p} q_i' \mathbf{1}_{\{n>\alpha_{j_r}\}},\ and\ k_3 = 1-2Q\left(\frac{2^{m_{acc}-m_p+1}}{\sqrt{n}}\right).$$

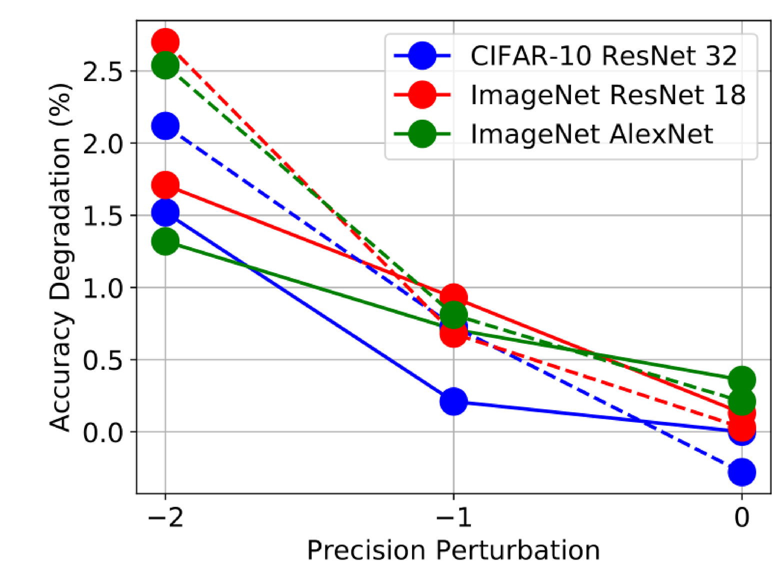
- swamping causes loss of accumulated variance in partial sums
- variance retention ratio (VRR) is derived analytically as a function of precision and accumulation length
- VRR 'knee' corresponds to the maximum accumulation length allowed for a given precision

Convergence with Low-Precision Accumulation



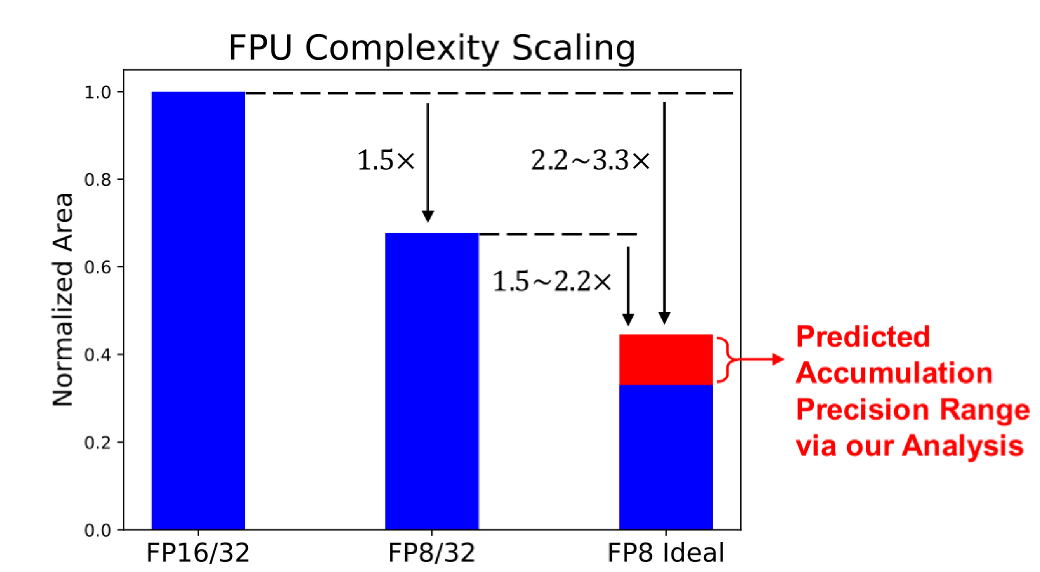






- VRR-based analysis enables convergence with low-precision accumulation and is tight

Hardware Benefits



- low-precision accumulation reduces hardware cost over by ~2× compared to representation quantization

Acknowledgement

This work is supported in part by IBM Research; IBM Soft Layer; IBM Cognitive Computing Cluster (CCC); IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM AI Horizons Network; and C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.