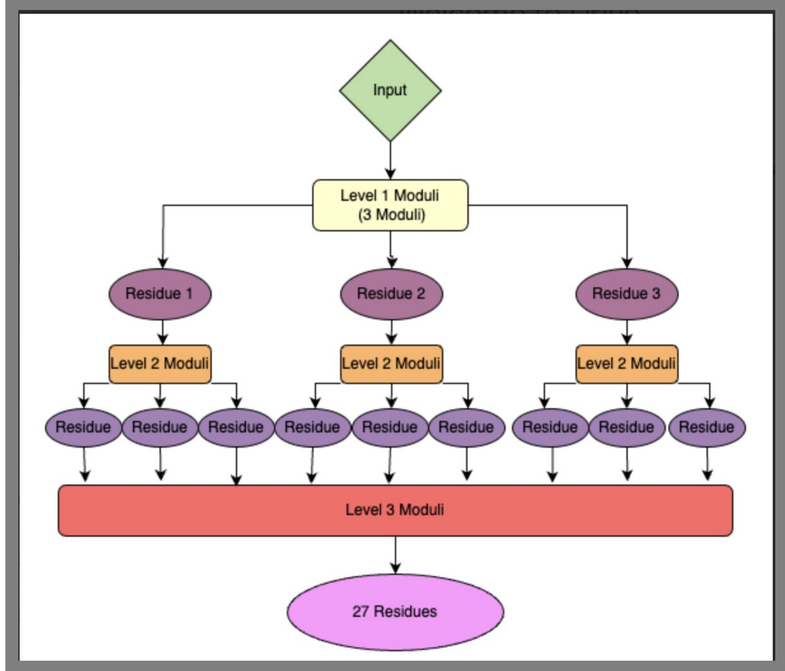
An Exploration of the Residue Number System in Neural Networks

Sonali Srikanth[1,2], Cansu Demikiran[2], Ajay Joshi[2]

Liberal Arts and Science Academy, 1012 Arthur Stiles Road Austin, TX 787211; Boston University, Department of Electrical and Computer Engineering, 8 St Mary's St, Boston, MA 022152

Abstract

- The Residue Number System (RNS) is based on moduli and remainders (called residues)
- Could significantly speed up emerging computation methods where numeric precision is limited
- A hierarchical structure (HRNS) can help speed computations up further by exploiting its parallelism



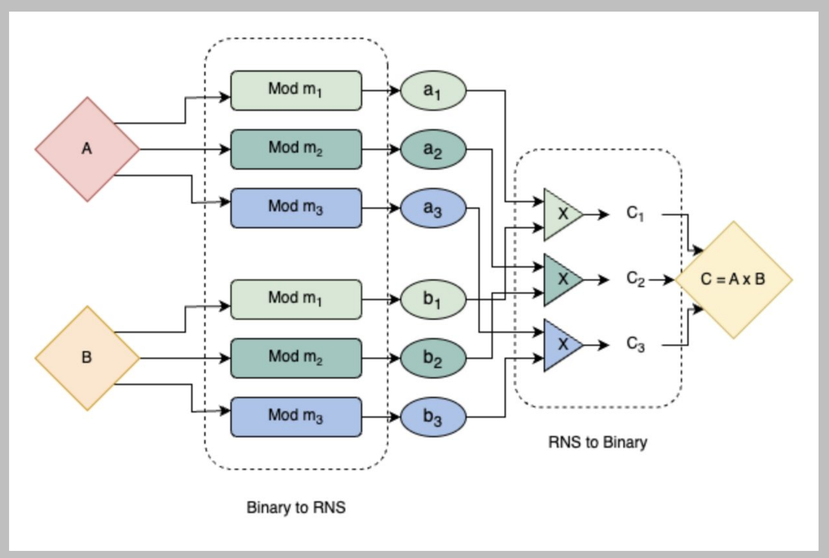
Hierarchical RNS (HRNS)

- Each residue from one "level" of RNS becomes a new "level" of RNS
- Same Moduli used for each level
- Reduces number of bits needed for operations
- Operations (addition, multiplication) happen at the lowest level
- After operations, each output is taken through the lowest level moduli to avoid overflow
- Example for 3 levels: Chinese Remainder Theorem (CRT) must be used 13 times for multiplication
  - 13 = 9 (level 3 to level 2) + 3 (level 2 to 1) +1 (level 1 to output)
  - Can have different number of levels or different number of moduli, changing CRT uses

Conclusion

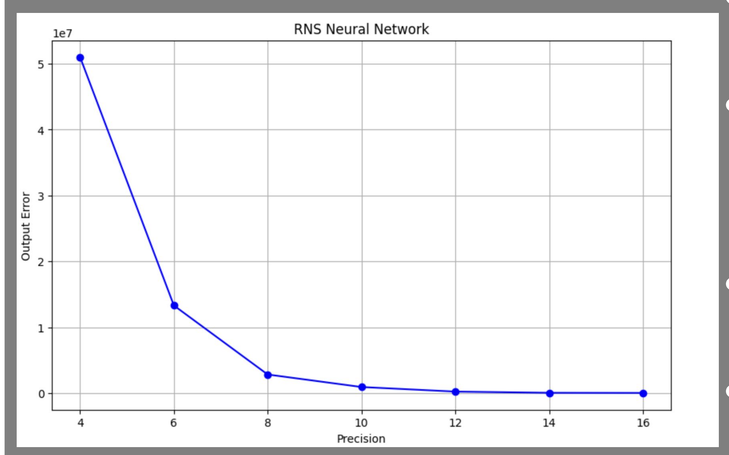
- RNS and HRNS can be used to perform computations for NN with in cases that require relatively low precision
- Can be applied to VLSI, Homomorphic encryption, and artificial intelligence
- More work can be done to understand and incorporate the HRNS and/or Recursive RNS in Neural Networks and test efficiency

Residue Number System (RNS)

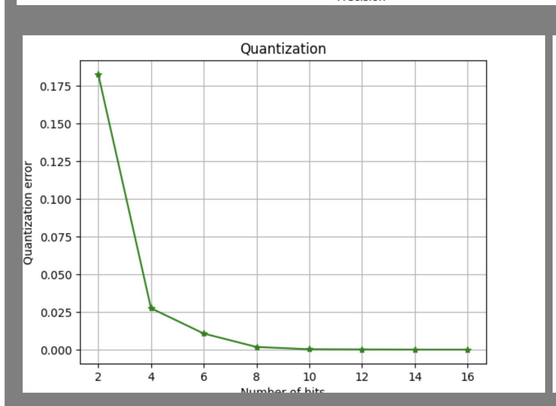


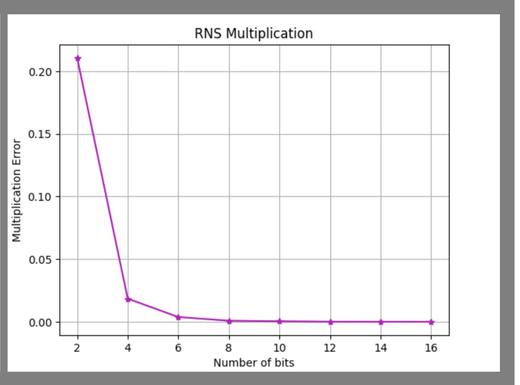
- Represents a set of integers using coprime moduli
- (m = m1, m2, m3)
- Multiple low precision operations can combine to form a high precision operation
- Closed under addition and multiplication
- Doesn't require the carrying of digits in these operations

Precision in Neural Networks



- Blue: RNS NN error in output vs # of bits used
- Green: Quantization and Dequantization error vs # of bits used
- Pink: RNS Multiplication error vs # of bits used
- Trend: More bits = less error





Moduli: [1023, 1024, 1025)

Acknowledgements

I would like to thank Dr. Ajay Joshi for this opportunity to work in he ICSG lab, and Dr. Cansu Demirkiran for the constant support and guidance this summer.

References