Rice IP Applied to Neural Networks

(An Overview)

There are two areas of Rice IP technology relating to neural networks. As referenced at this site, these are building blocks for;

- Advanced Processing (especially Frequency Domain Processing)
- Unique Data Compression

Their relevance to Convolutional Neural Networks (CNNs) is summarized below.

Processing IP (Frequency Domain Processing)

CNNs expend the bulk of their computations and power on the convolution process itself. This poses extreme computational requirements for either programmable or specialized processors. Modern CNNs often restrict convolution filter (kernel) sizes and weight resolution to compensate for this basic problem. This can reduce flexibility and complicate "training".

A concept addressing these issues is convolution in the "frequency domain". Mathematically, this provides a more efficient approach to performing convolution. However, the concept typically requires complex "DFT" or "FFT" structures. Addressing this issue, the Rice Processing IP provides structures of high efficiency for frequency domain convolution.

As discussed in the Projects of this site, this involves a "hierarchical" Processing Architecture which;

- is "multi-mode" in nature, capable of various operational functions (including FFTs), where

- said FFTs employ "DFT-like" building blocks, and

- said DFTs are based on specialized multiplier blocks consisting only of addition circuitry

The proprietary multipliers referenced above minimize complexity while maintaining performance. They enable DFT multiplications (of 4 to 20 bit operands) to be achieved with the equivalent of two to three fixed-point adder circuits, (while not using publicly known techniques such as "Cordic arithmetic" or "residue number systems").

The Processing Architecture can achieve a ten-fold reduction in complexity over traditional frequency domain approaches, and enables;

- Extended dimensions and weight precision of convolution filters
- Acceleration of the convolution process
- Substantial reduction of size/power consumption of the CNN processor

Use of frequency domain processing for CNN convolution layer(s) is well documented. Also, open literature has described the utility of "spectral pooling" (i.e. a pooling layer implemented in the frequency domain). Accordingly, the Company's IP has relevance to multiple steps (layers) in advanced CNN processing.

Data Compression IP

Rice IP includes novel Data Compression building blocks (as described at this site). Their processes are "lossy" in nature, but have the potential to alleviate storage and bandwidth limitations within advanced neural networks. The Compression building blocks may be efficiently implemented via software or hardware, being based heavily on LUT and primitive integer operations (e.g., scaling, addition/subtraction). Accordingly, they are conducive to a variety of hardware implementations, including custom, ASIC and FPGA circuits.

The IP is relevant to compression of filter weights in CNNs. Also, it may be applied to compress "multi-media" of various forms (e.g., imagery, speech). Of particular importance is its efficiency relevant to pseudo-random data sets, as referenced by the discussion in the "Projects" section of this site.

The Data Compression IP may render up to ≈70% reduction of "raw" multi-media data (e.g., imagery, speech), while minimizing "perceptual" distortion. As referenced at this site, it is also possible to trade fidelity of the recovered (decompressed) data for higher degrees of compression. By making such tradeoffs, an objective of the Company is to enable up to 10X reduction of raw media data.

Such compression can also be applied to convolution kernels, especially kernels having expanded dimensions and weight precision. Accordingly, the Compression IP and Processing IP

(referenced above) can be complementary in the enhancement of CNN processors. The potential advantages accruing to this synergistic IP approach include:

- Accelerated processing
- Reduction of size, weight and power consumption
- Enhanced functionality (improvements in training and recognition performance)

The IP can therefore be of significant utility in the development of advanced Neural Network processors.

Rice Electronics

ricetronics@gmail.com

Filename: Neural Net Note 3-2021