Laconic Deep Learning Inference Acceleration

Corresponding author

Sayeh Sharify, Alberto Delmas Lascorz, Mostafa Mahmoud, Milos Nikolic, Kevin Siu, Dylan Malone Stuart, Zissis Poulos, Andreas Moshovos, University of Toronto

Keywords

Laconic Processing Element (LPE), inference, deep learning, bit sparsity, activation and weight, architecture

Summary

*Challenge*

The bulk of the energy and work in neural networks during inference is due to the data transfers and the computation needed to perform multiply-accumulate operations (MACs) of weight (W) and activation (A) values into partial sums (psum).

*Contribution*

1. Find that neural network exhibit high bit sparsity in their activations and weights.
2. Propose Laconic, a hardware accelerator design that exploits bit sparsity in activations and weights. Representing data with Booth Encoding.

*Innovation Points*

Propose a data-parallel histogram-based Laconic processing element (LPE) that is much smaller (5x in 8b) and energy efficient than a bit-parallel PE.

*Result*

By decomposing multiplications down to the bit level, the amount of work needed by multiplications during inference can be potentially reduced by at least 40× across a wide selection of neural networks (8b and 16b).