Laconic deep learning inference acceleration

Björn Bebensee (2019–21343)
Topics in Artificial Intelligence

October 10, 2019

As hardware is energy-constraint, an important research direction in hardware acceleration is increasing energy efficiency. Sharify et al. [1] show that there is a large amount of ineffectual computations during in inference in deep learning models. They propose a new hardware accelerator which they call *Laconic* which exploits this fact to achieve high-energy efficiency.

The authors demonstrate that inference computations in (deep) neural networks exhibit high bit sparsity in their activations and weights. They observe bit sparsity in different neural network architectures such as GoogleNet, Resnet50 and MobileNet and show high potential speed-ups by skipping both zero activation and weight terms during inference. While most previous work focused on removing ineffectual activation and weight multiplication terms by exploiting a combination of activation or weight sparsity with activation bit sparsity, this work exploits both activation bit sparsity and weight bit sparsity allowing for a potential speedup of a magnitude higher than exploiting activation bit sparsity only. *Laconic* implements this policy of removing zero terms in booth-encoded activation and weight terms and thus aims to only process the necessary bits.

Sharify et al. newly propose a *Laconic* Processing Element (LPE) and describe its architecture in detail. Instead of processing the product $A \times W$ of a weight W and an activation A in a single cycle, it processes a single term of A and a single term of W individually. Here, a term refers to a signed power of two from the booth-encoding of A, W. In order to implement this PE architecture efficiently, *Laconic* utilizes a histogram-based front-end with a modified adder tree back-end. They further explain how these LPEs can be organized into tiles to allow reuse of activations and weights to achieve greater efficiency. The authors hint at the future possibility of further optimizing *Laconic* by exploiting intra-value bit-level parallelism (as opposed to the inter-value bit-value parallelism that they already exploit) in a way that is not possible in bit-parallel hardware.

Their proposed implementation is evaluated in terms of performance, energy and area against a number of other state-of-the-art hardware accelerators: DaDianNao++, SCNN, Eyeriss, Pragmatic and BitFusion. They find that *Laconic* provides a speedup against other hardware accelerators while achieving higher energy efficiency.

References

[1] Sharify, Sayeh, et al. "Laconic deep learning inference acceleration." *Proceedings of the 46th International Symposium on Computer Architecture*. ACM, 2019.