Note: The contributions described in the following sections are contributions of my thesis so far. More contributions will be added once the architecture is complete.

Contributions

This thesis describes a hardware architecture for accelerating the training of Convolutional Neural Networks (CNNs) on FPGAs. From the hardware perspective, convolution layers are particularly challenging because they have a multitude of configuration options which must be generalized in hardware to sustain a high computational throughput. For example, layer shapes, kernel sizes, padding, stride and batch size are configuration options that vary and depend on the architecture of the CNN. This thesis describes a runtime configurable data loader to address this issue in hardware. Specifically, the contributions of this thesis are:

- A generalized high-throughput implementation of the im2col transform on FPGAs for mapping convolution to matrix multiplication. This design is runtime configurable for different options including automatic zero padding, stride, kernel size, and input tensor dimensions and operates on streaming data without explicitly forming the large matrices involved.
- Quantitative analysis of the im2col data loader for different configuration options.