**Status:**

In our 2016 continuing research proposal we wrote:

*An open-ended pruning framework is a happy compromise between the needs of our SRC unbundled programming model research and the needs of sponsors working on their own hardware and in need of a domain specific compiler frontend, as well as those exploring memory bandwidth and energy footprint of accelerator implementations.*

The tool, codenamed WeedWacker, has been successfully completed and committed to the BDD repository (bdd2/isg/weedwacker) along with documentation and a number of examples demonstrating its use. The tool is pruning strategy agnostic and enables easy experimentation and evaluation of different pruning constraints. It is built on top of TensorFlow framework.

Using the tool we have developed a number of pruning strategies that result in GPU friendly computation patterns. Our results are reported in our paper “Structured Deep Neural Network Pruning via Matrix Pivoting” (bdd2/isg/weedwacker).

**Publications:**

Structured Deep Neural Network Pruning via Matrix Pivoting

Abstract:

Deep Neural Networks (DNNs) are the key to the state-of-the-art machine vision, sensor fusion and audio/video signal processing. Unfortunately, their computation complexity and tight resource constraints on the Edge make them hard to leverage on mobile, embedded and IoT devices. Due to great diversity of Edge devices, DNN designers have to take into account the hardware platform and application requirements during network training. In this work we introduce pruning via matrix pivoting as a way to improve network pruning by compromising between the design flexibility of architecture-oblivious and performance efficiency of architecture-aware pruning, the two dominant techniques for obtaining resource-efficient DNNs. We also describe local and global network optimization techniques for efficient implementation of the resulting pruned networks. In combination, the proposed pruning and implementation result in close to linear speed up with the reduction of network coefficients during pruning.
The paper is currently under peer review for publication in the special issue of Computer Magazine. An early-access copy has been provided for the BDD sponsors in the repository (bdd2/isg/papers).

**Current efforts:**

1. We are actively developing WeedWacker to allow working with generalized training constraints, such as quantization of network coefficients during training.
2. We are actively using the tool to explore different pruning strategies that allow efficient implementation of DNN models in resource-constrained environments.