Paper critique on

Planaria: Dynamic Architecture Fission for Spatial Multi-Tenant Acceleration of Deep Neural Networks

By **Brandon Saldanha**

Summary

With Dennard scaling slowing down in recent years, the 'free' hardware boost that the silicon industry has enjoyed is waning. To keep up with the additional requirement in computing, companies have moved towards accelerating machine learning operations. To this end, companies have invested resources into Accelerated Hardware, developing computing resources such as Edge TPU, NVIDIA Jetson, Apple Bionic Engine, etc for commercial use and Google TPU, NVIDIA T4, Microsoft Brainwave, Facebook DeepRecSys for datacenter use.

The current DNN accelerator designs are focused on running a single neural network model as fast as possible. This model can be improved by using a multi-tenant approach as implemented with cloud-computing for multiple years. This possibility is discussed with Planaria - where a single DNN accelerator's resources can be 'rented' by different tasks so multiple models can be deployed at the same time.

Planaria suggests changes to the following components in the current DNN accelerator design: 1) Dynamic architecture splitting of compute resources 2) New and flexible communication routes between compute resource sections 3) A new SIMD Vector Unit to handle data transfer to multiple compute resources.

This new architecture is compared with previous existing architectures with respect to Throughput, Compute Resource Utilization, QoS requirements and found improvements in these areas.

Strengths

Planaria can adapt the available resources for a wide range of use cases. Because of the way Planaria is designed and configured, an NxN compute unit can be used in any formation right from $1xN^2$ to N^2x1. This flexibility allow Planaria to run a variety neural nets more efficiently than a monolithic systolic array.

The papers does a good job of explaining why and how the current industry design is not efficient enough and how new modifications are necessary to better utilize available resources. The paper shows how the new modifications improve on efficiency, power utilization and fairness.

The paper considers a wide range of resource use case and shows improvements in all these cases. Planaria considers the priority and QoS requirements of task while executing them - something that the current architecture does not consider.

Weaknesses

Planaria's architecture limits the granularity of fission. It is not possible to achieve greater number of pods without an exponential increase in memory pods and communication resources. As such there is more scope of efficient utilization of compute resources.

The new architecture described in Planaria requires separation of the Unified Multi-Bank Activation Buffer and SIMD Vector Unit into multiple blocks instead of a single block which increases resources required to run Planaria. For situations where we need to use the architecture as a monolithic systolic array (when we have only one task at a time) it would benefit to have a way to turn off fission instead of using the forwarding routes to transfer data.

Future scope

We could work towards more aggressive fission, where even a single processing element can be used as a complete subarray. This is currently not possible with Planari's architecture due to the very costly power and area overload required to have NxN fission pods.

The current architecture of fission pods includes multiple 'pod memories' to be present for the architecture's working. A different approach that utilizes a big bank of data with some form of forwarding could be researched to mitigate the costs of splitting memory.