Bio-Inspired Edge Learning on the Akida Event-Based Neural Processor

Sasskia Brüers, **Kristofor D. Carlson,** Marco Cheng, Sébastien Crouzet, Mahendran Devarajlu, Hussein Makki, Douglas McLelland, Nicolas Oros, Charles Wilson, and Kenneth Wu, BrainChip

The Akida event-based neural processor is a high-performance, low-power SoC targeting edge applications, distinguishing itself from traditional deep learning accelerators (DLAs) through 2 key features.

First, multiple design choices maximise efficiency, and allow for a highly configurable accuracy/power trade-off. On the algorithmic side, these include aggressive 1 to 4-bit quantization of weights and activations, along with event-based implementation of computations. On the hardware side, CNN layers are distributed across many (~80) small neural processing units (NPUs), each with its own collocated processing and memory. As a result, Akida requires 40-60% fewer computations to process a given CNN when compared to a DLA, often without the need for off-chip memory access or host CPU communication even for relatively large CNNs, such as MobileNet v1. For example, using the Akida Execution Engine (AEE) we simulated the performance of object classification on the ImageNet data set using MobileNet v1 and observed an average activity sparsity of 42% with 30 inferences per second at ~160 mW and very little loss in accuracy. Similarly, we used a six-layer CNN to classify the Google Speech Commands data set with ~93% accuracy, an average activity sparsity of 72%, and 7 inferences per second at ~200 uW.

Second, Akida incorporates a bio-inspired learning algorithm, adapted from spike timing-dependent plasticity (STDP). Combined with pre-trained feature extractor networks, this allows us to perform learning directly on the chip. For example, building on the two networks described above, we present state-of-the-art few-shot learning in both visual (MobileNet on mini-imagenet) and auditory (6-layer CNN on Google Speech Commands) domains.