# Efficient Deep Learning Inference: A Digital Hardware Perspective

Evaluating and improving performance and efficiency of artificial and spiking neural networks hardware accelerators

Fabrizio Ottati

February 2024

From smartphones to televisions and cars, deep learning (DL) has become pervasive in our daily lives. Many modern DL models, especially large language models (LLMs), recommender systems, and vision transformers (ViTs) require huge amounts of power and energy for both training and inference. For instance, ViT-G/14, a top performing transformer model in object recognition, requires 2.86 GFLOP for a single inference on ImageNet, and around 159 MW h of energy to train on specialized tensor processing units (TPUs) running on 220 W.

The human brain, instead, has been trained during the evolution of humankind. Since the brain power budget is around 20 W, one could say that using only this power it is able to perform multiple actions at once and model fine-tuning (*i.e.*, learn new things).

Beyond training, DL models inference costs prove to be a serious problem: for instance, processing an user prompt using OpenAI LLM ChatGPT costs 0.04 in terms of energy and hardware (*i.e.*, graphic processing units (GPUs) used for the inference.). Hence, efficient hardware implementations and neural network models should be considered also for when DLs models are deployed.

Given the extraordinary efficiency of the brain in cognitive tasks, researchers are trying to take inspiration from biology when designing new artificial intelligence (AI) models and hardware; in this context, spiking neural networks (SNNs) are arising as a possible alternative to reach energy efficient AI. In this thesis, the inference of artificial neural networks (ANNs) and SNN models on digital hardware is investigated, analyzing state-of-the-

art (SOTA) digital hardware accelerators targeting deep neural networks (DNNs) from both the spiking and artificial domains, on vision and audio workloads.

A C++ library for the deployment of spiking convolutional neural networks (CNNs) to field-programmable gate arrays (FPGAs) using high level synthesis (HLS) tools, is presented. The hardware architecture proposed is a dataflow one, and SOTA techniques for activations buffering are exploited to minimize the memory footprint of the neural network model, in order to target edge-class FPGAs with limited computational resources.

The library targets event-based vision tasks, in which new vision sensors inspired by the human retina are exploited to retrieve visual information from a scene. These sensors provide luminance variation events in output instead of full-frame images, contrarily to conventional RGB sensors. This data format naturally fits the spike-based processing of SNNs. The choice of this application is justified in the analysis of DL models digital hardware acceleration, in which it is shown that SNNs are not competitive with conventional DL accelerators on static frame-based vision tasks, such as object recognition, when it comes to energy consumption per inference and accuracy reached on the selected task. Using CNNs implemented in hardware by the library, one can target multiple event-based vision tasks, such as object recognition, object detection, optical flow estimation etc.

An FPGA accelerator produced using the proposed HLS library is used as controller for autonomous small drones equipped with event cameras, in collaboration with Delft University of Technology. In the application considered, a CNN is interfaced with an event-based camera to compute optical flow from raw events, and its output used by the motor controller to drive the drone in some actions autonomously. The hardware accelerator shows a $10\times$ reduction in the power consumption when compared to a neuromorphic processor baseline on the same task, with no performance loss in terms of accuracy using the same spiking neural network model.