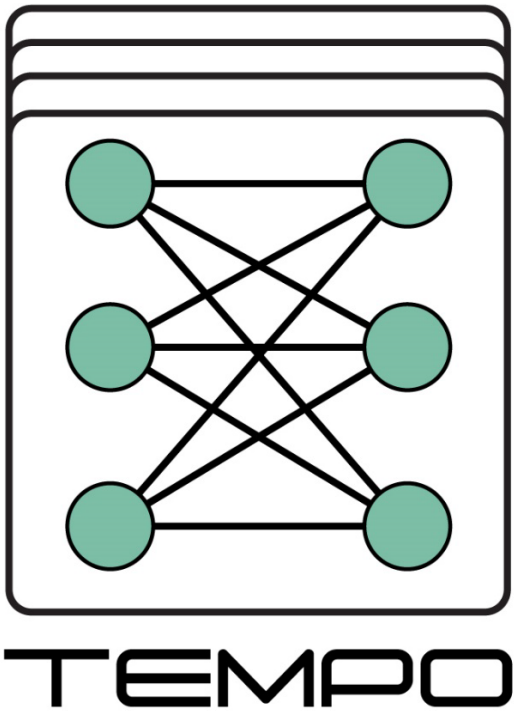
Technology & Hardware for nEuromorphic coMPuting

- ECSEL Research and Innovation Actions (RIA*) –




# Deliverable 4.1
# – Flexible precision blocks for DNN –

| | |
|---|---|
| Work Package | WP N° 4 – Design and Architecture |
| Document Date | 29.04.2020 |
| Revision N° | 0.4 |
| Status | Final |
| Dissemination Level | Confidential |
| Responsible Partner | FhG |
| Name | Michael Rothe |
| Contact Information | michael.rothe@iis.fraunhofer.de |

© Copyright 2019 TEMPO Project. All rights reserved
This document and its contents are the property of the TEMPO Partners. All rights relevant to this document are determined by the applicable laws. This document is furnished on the following conditions: no right or license in respect to this document or its content is given or waived in supplying this document to you. This document or its contents are not be used or treated in any manner inconsistent with the rights or interests of TEMPO Partners or to its detriment and are not be disclosed to others without prior written consent from TEMPO Partners. Each TEMPO Partner may use this document according to the TEMPO Consortium Agreement.

* This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826655. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, Netherlands, Switzerland".

ECSEL Joint Undertaking

# Publishable Summary

While DNNs are compute intensive they especially benefit in embedded systems from dedicated accelerators and model compression approaches. Reducing the complexity of computations using fixed point arithmetic is one of the key parts to achieve a small form factor and reasonable power consumption for these accelerators. However, fixed point arithmetic has to address its own challenges like numerical resolution, overflow and saturation handling. Another key part is to reduce the huge amount of network-parameters to efficiently use chip internal memory and to reduce external memory access. Here, compression techniques come into play.

Therefore, this deliverable first provides an overview regarding the topic of quantization and training, that makes possible to build flexible synapsis and activation functions. Bosch focuses here on the transition of pre-trained DNNs from Float32 to fixed-point variants. Here, various quantization strategies are evaluated and elimination of using standard multiplication is investigated to further reduce die size and power consumption of the computation. In addition, videantis describes their approach to take a trained floating-point (32fp) model of a complete DNN as input, perform an analysis step with representative input data, and quantize weights and activations per each layer to either 16 bit or 8 bit fixed point resolution, depending on given accuracy requirements. Fraunhofer IIS is on the one hand concentrating on different compression methods to reduce the external memory access and on the other hand on quantization-aware training using a custom activation function for an analog accelerator. In this context Fraunhofer EMFT also presents their results for 1 bit quantization.

The work on quantization and training provides the basis for the second part of this deliverable. Here the architecture for a digital deep learning inference accelerator consisting of the accelerator subsystem developed by videantis and the decompression core developed by Fraunhofer IIS is presented as well as the architecture for an analog deep learning inference accelerator developed by Fraunhofer IIS and Fraunhofer EMFT.

The digital accelerator subsystem consists of a number of videantis v-MP+ accelerator cores connected to an OCP bus fabric and to a multi-banked Local SRAM. The developed decompression core also is connected to this bus fabric and will work as coprocessing unit to the v-MP+ processor units.

Afterwards, the architecture for the fully connected layers to be implemented with an analog deep learning inference accelerator is described. Here, the top-level architecture and interfaces for the analog accelerator, the connection between layers as well as an overview of the crossbar circuit for a fully connected layer is presented. Furthermore, Fraunhofer IIS and EMFT present possible solutions linked to the quantization and training sections for implementing the crossbar array architecture either with 3 bit weights or with 1 bit weights. For the analog accelerator the usage of the FeFET technology is planned.

In the further project course, the presented digital and analog accelerator together will be used to run the neuronal networks of the use case partners InnoSent and Valeo.