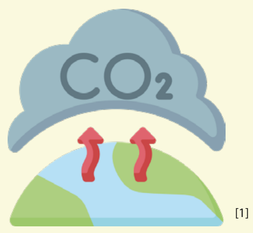
HAMM-LIPPSTADT
UNIVERSITY OF APPLIED SCIENCES

Energy Efficient AI

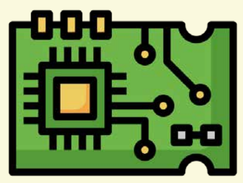
eki
ENERGY EFFICIENT AI
BY DNN APPROXIMATION
FOR FPGAs

Inference of an average DNN releases 2.840 t of CO2. [5]



[1]

How to optimize DNNs to be more efficient?


[1]

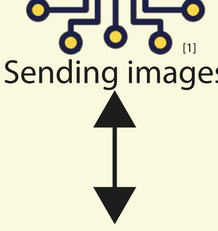
Scenarios

High speed cloud computing

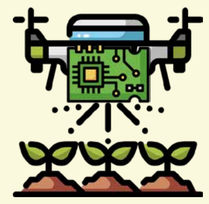

[1]

Sending images

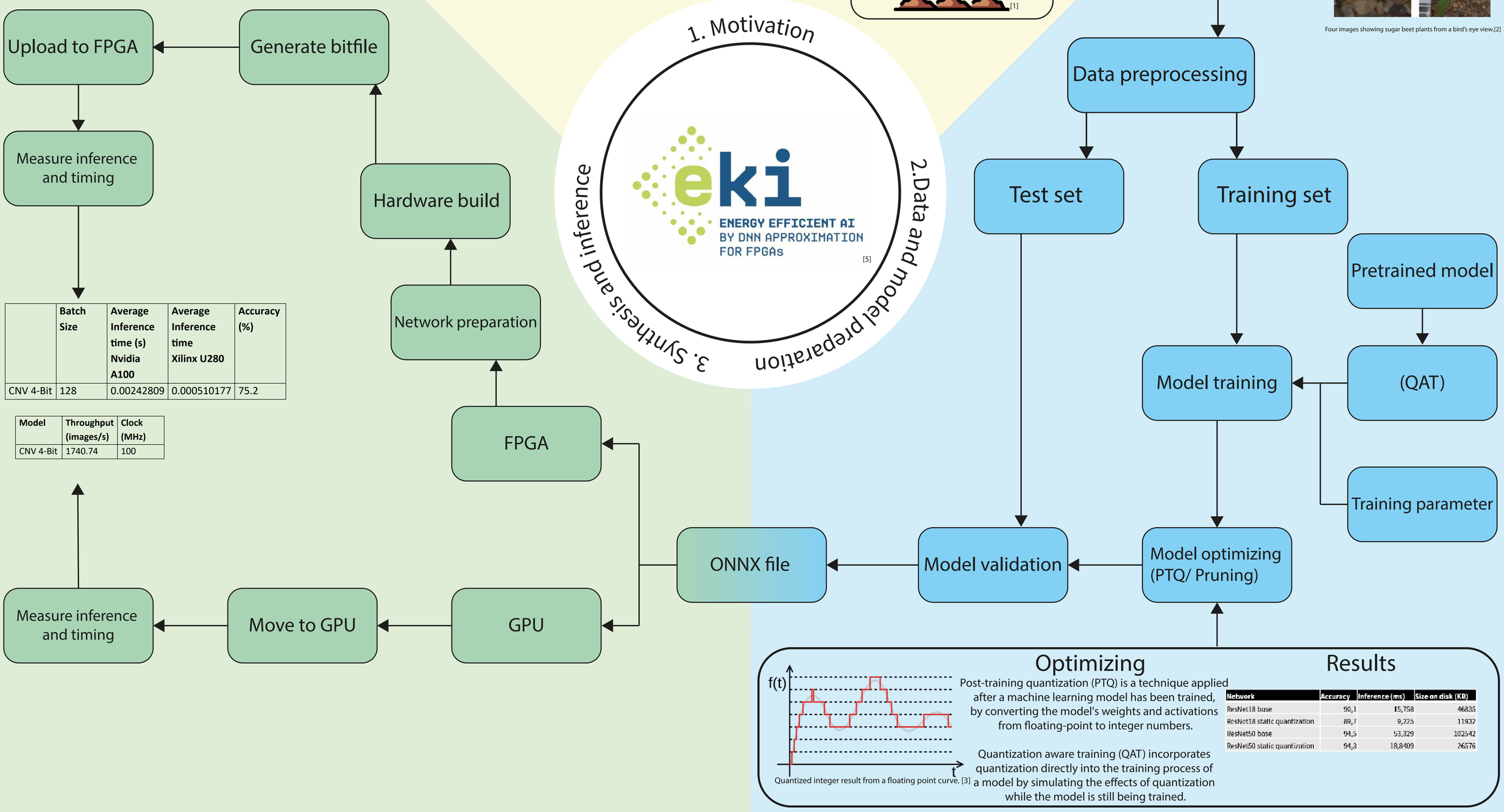
Receiving results


[1]

Efficient edge computing with FPGA as payload


[1]

Four images showing sugar beet plants from a bird's eye view.[2]



	Batch Size	Average Inference time (s) Nvidia A100	Average Inference time Xilinx U280	Accuracy (%)
CNV 4-Bit	128	0.00242809	0.000510177	75.2

Model	Throughput (images/s)	Clock (MHz)
CNV 4-Bit	1740.74	100

Quantized integer result from a floating point curve. [3]

Optimizing

Post-training quantization (PTQ) is a technique applied after a machine learning model has been trained, by converting the model's weights and activations from floating-point to integer numbers.

Quantization aware training (QAT) incorporates quantization directly into the training process of a model by simulating the effects of quantization while the model is still being trained.

Results

Network	Accuracy	Inference (ms)	Size on disk (KB)
ResNet18 base	90,1	15,758	46835
ResNet18 static quantization	89,7	9,225	11932
ResNet50 base	94,5	53,329	102542
ResNet50 static quantization	94,3	18,8409	26576

References:

[1] https://www.flaticon.com
[2] R. Sharma L. Fink. V2 plant seedling dataset.[online]. Available: https://www.kaggle.com/datasets/vbookshelf/v2-plant-seedlings-dataset, 26.09.2023.
[3] C. Rice. Gain, range, and quantization. [online]. Available: https://community.sw.siemens.com/s/article/gain-range-and-quantization, 26.09.2023.
[4] https://www.eki-project.tech/project

Fachhochschule Südwestfalen
University of Applied Sciences

PADERBORN UNIVERSITY

Paderborn Center for Parallel Computing

AMD

MEGWARE

Supported by:

Federal Ministry
for the Environment, Nature Conservation,
Nuclear Safety and Consumer Protection

based on a decision of
the German Bundestag

Hochschule Hamm-Lippstadt
Marker Allee 76-78
59063 Hamm

Contact
Prof. Dr. Stefan Henkler stefan.henkler@hshl.de
Prof. Dr. Achim Rettberg achim.rettberg@hshl.de
Domenic Drechsel domenic.drechsel@hshl.de

Grant number: 67KI32004A
Runtime: 01.2023 - 12.2025