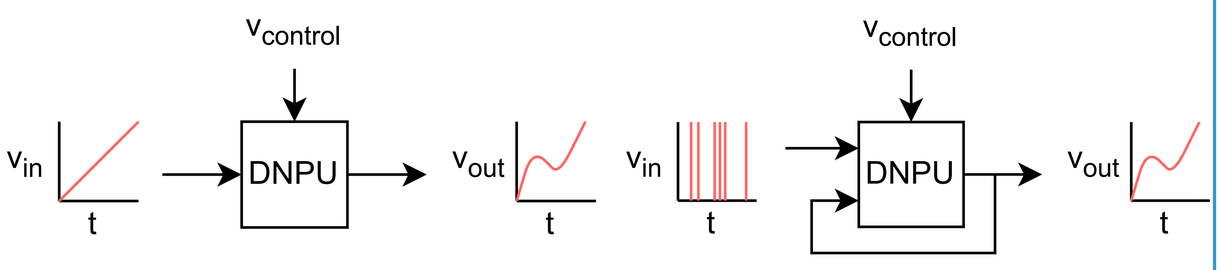
Analog Kolmogorov-Arnold networks for energy-efficient embedded AI

Sjoerd van den Belt - University of Twente

Introduction

Analog AI processing

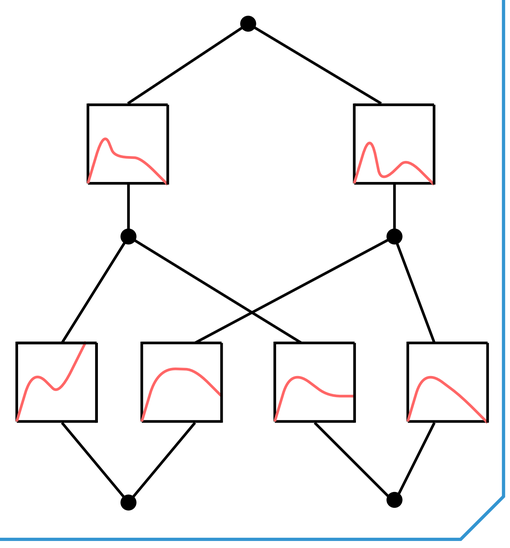
Dopant Network Processing Units (DNPUs) are nano-scale analog devices featuring multiple electrodes that can be used as input and output. Recent work has demonstrated that these devices exhibit non-linear transfer functions. By using some electrodes as control variables and others as input and output, DNPUs can be seen as tunable non-linear activation functions. Such devices can learn complex non-linear functions by tuning the control electrodes. In addition, by introducing feedback to the DNPU, a tunable analog dynamic system can learn to interpret time-dependent input data, such as modulated spike signals.



DNPU as tunable activation function **DNPU as tunable dynamic system**

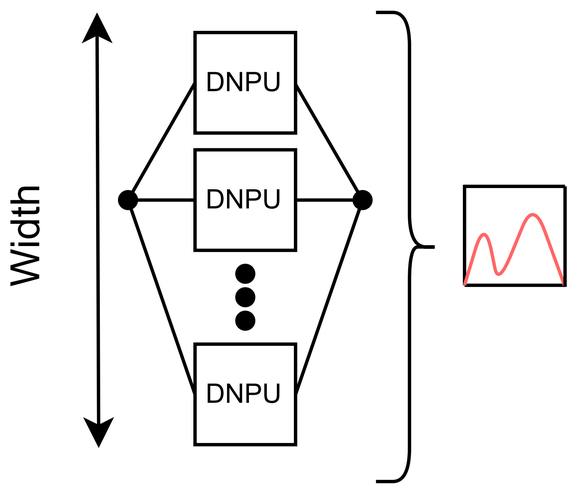
Kolmogorov-Arnold Networks

The performance of DNPUs as tunable activation functions to perform complex tasks is evaluated. Neural networks that learn through tunable activation functions instead of conventional weights are known as Kolmogorov-Arnold Networks (KANs). The edges of such a network apply a non-linear 1D transformation, and the nodes sum up their inputs.



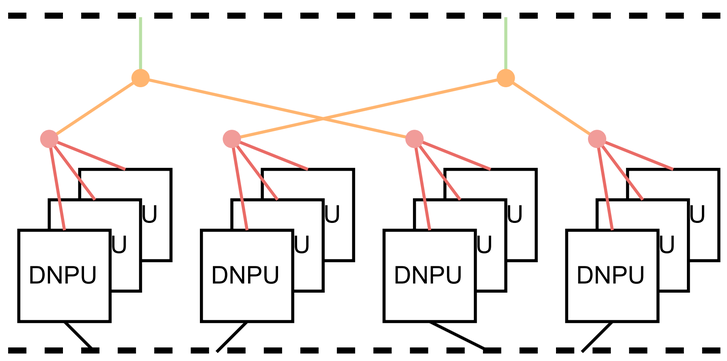
Method

Creating an energy-efficient KAN using DNPUs



To create a KAN network using DNPUs, each activation function in the KAN is substituted with an array of DNPUs. The width of the DNPU array is application dependent, a wider DNPU array can learn more complex activation functions.

The signals between network layers must be rescaled. To enable this, multipliers must be placed in the DNPU network. The number of required multipliers depends on whether scaling is applied before or after the KAN nodes (summations). This affects the scaling granularity. High granularity increases model flexibility at the cost of energy and resource efficiency.



Scaling granularities

DNPU-wise
Input channel-wise
Output channel-wise

Neural architecture search and evaluation

Using the DNPU-based KAN network architecture, models are trained to approximate the following function:

$$y = \frac{\sin(x_1 \cdot x_2)}{x_1}$$

The DNPU-based networks are compared to a full architecture search of baseline fully-connected MLP models. All MLP and DNPU architectures up to a given network depth and channel width are evaluated. The models are evaluated by their mean-square error (MSE) loss and their modeled energy consumption.

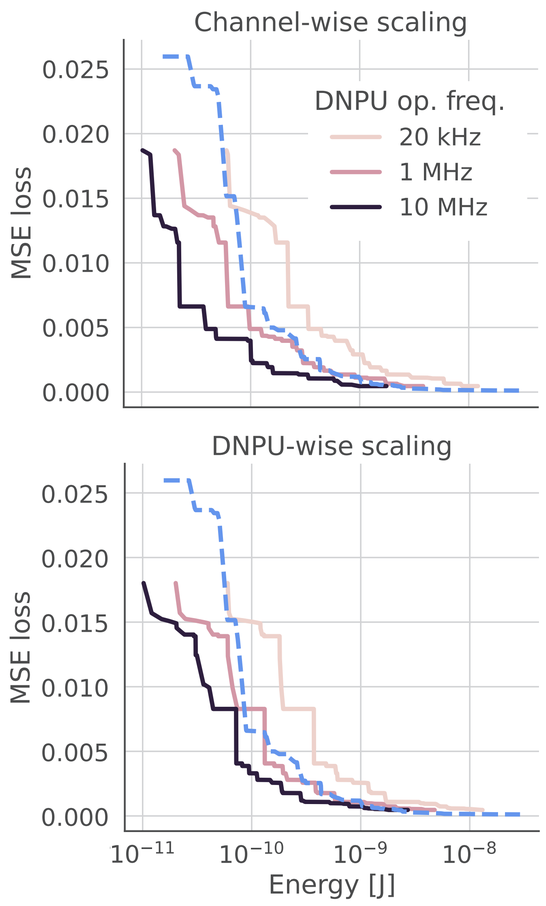
Energy consumption is calculated using a fixed cost per multiplication and addition. DNPU energy cost depends on the operating frequency of the DNPU. This leads to the following energy model:

$$\text{Energy} = C_{Add} \cdot N_{Add} + C_{Mult} \cdot N_{Mult} + \frac{P_{DNPU}}{f_{DNPU}} \cdot N_{DNPU}$$

Results

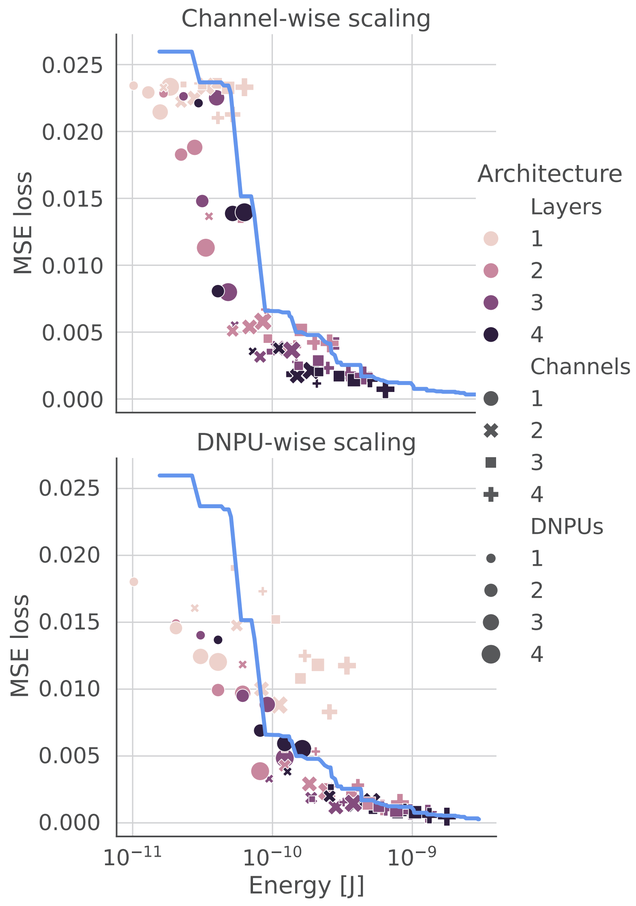
DNPU operating frequency

Below the best DNPU models are compared to the best MLP models (blue line) for three different DNPU operating frequencies.



DNPU model architectures

The scatter plot below shows the performance of all DNPU model architectures. The blue line indicates the best performance of the MLP model.



Identifying the best scaling granularity

To identify the optimal scaling granularity, the difference in loss for the different implementations is integrated over the energy axis at various for a range of DNPU operating frequencies.