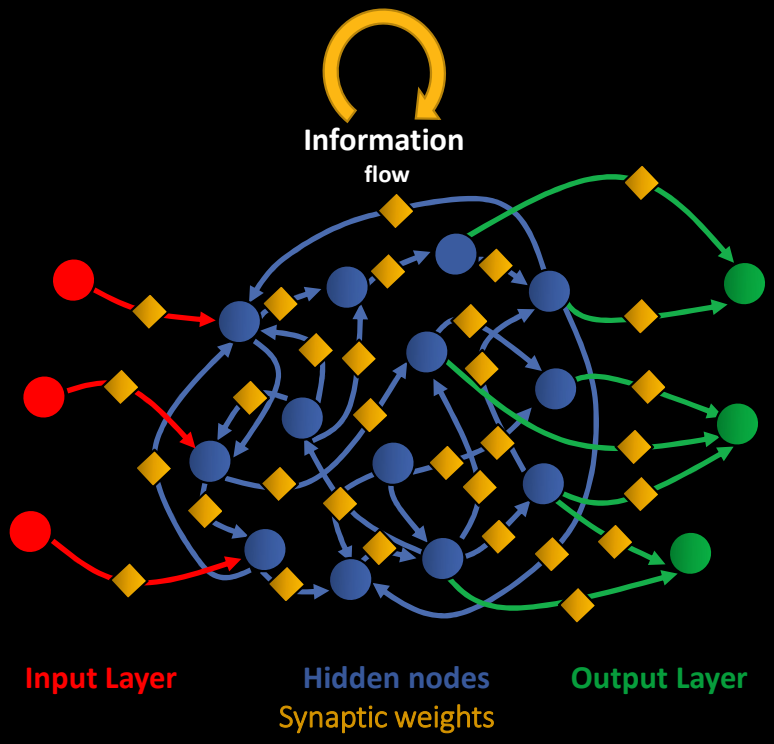
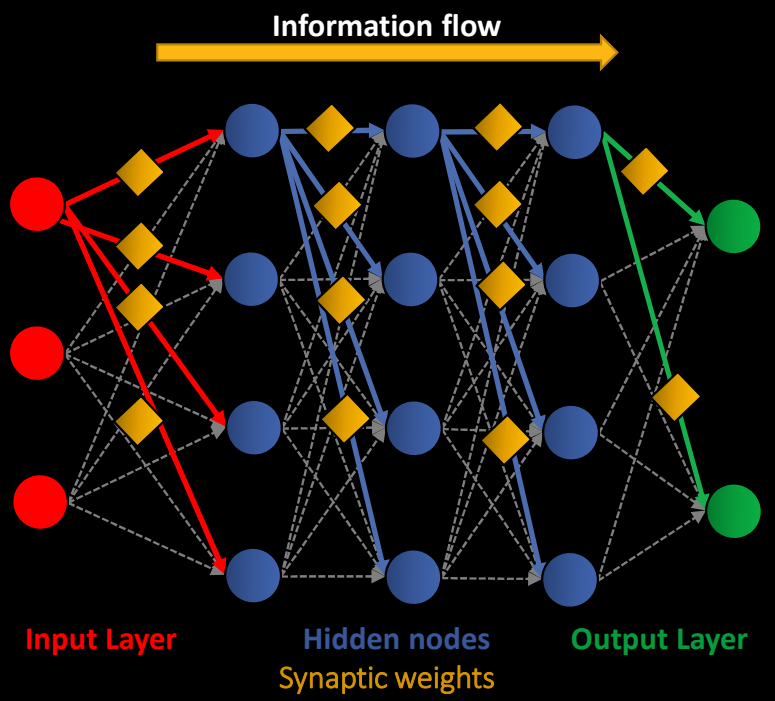
Cascaded Mach-Zehnder Interferometers for Efficient AI Acceleration

by Felix Hermann, Pascal Stark, Mustafa Yildirim, Folkert Horst, Jonas Weiss, Bert Jan Offrein

Neuromorphic Systems and Devices





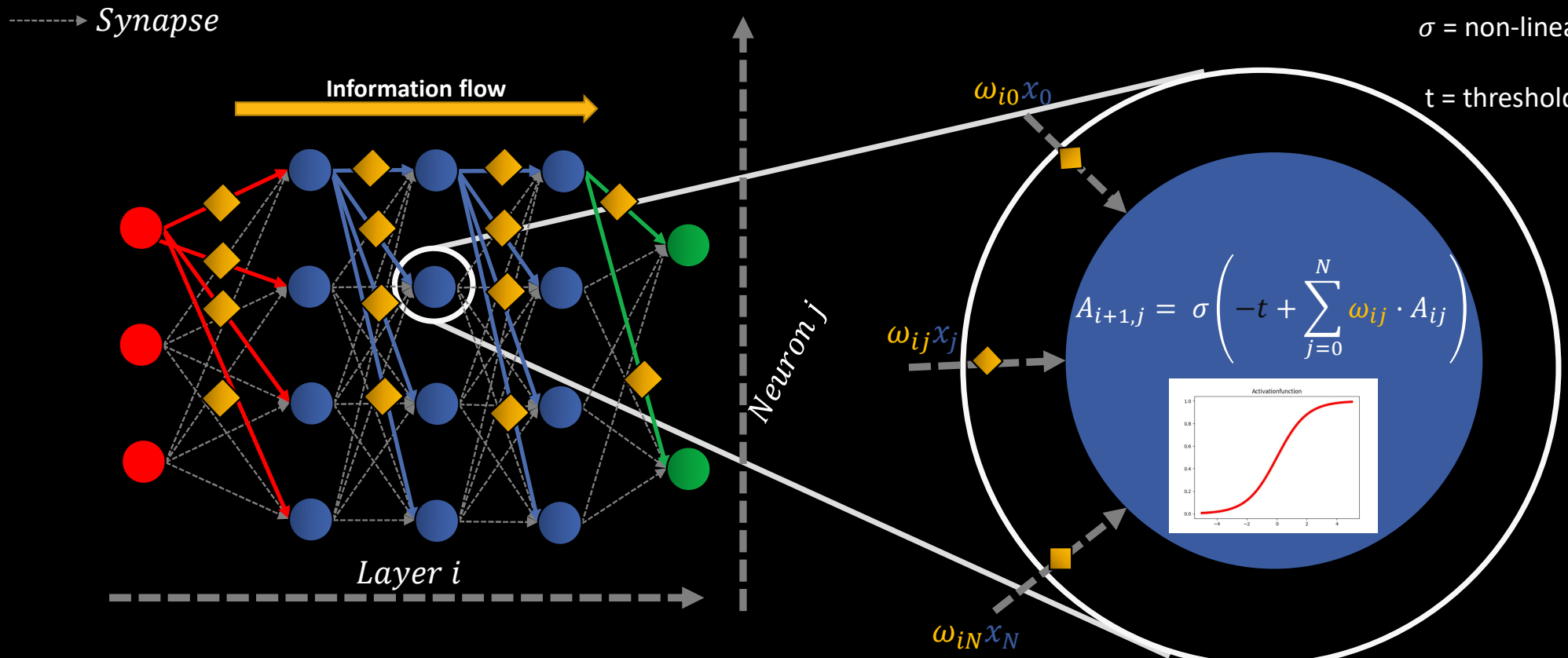
human brains

- recurrent connections
- information is stored in network
- training by creating new and tuning existing synapses

neural network

- layer-by-layer signal processing
- training by tuning synapses

Felix Hermann eli@zurich.ibm.com



Neural network

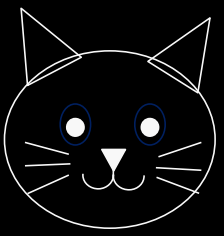
- layer-by-layer signal processing
- training by tuning synapses

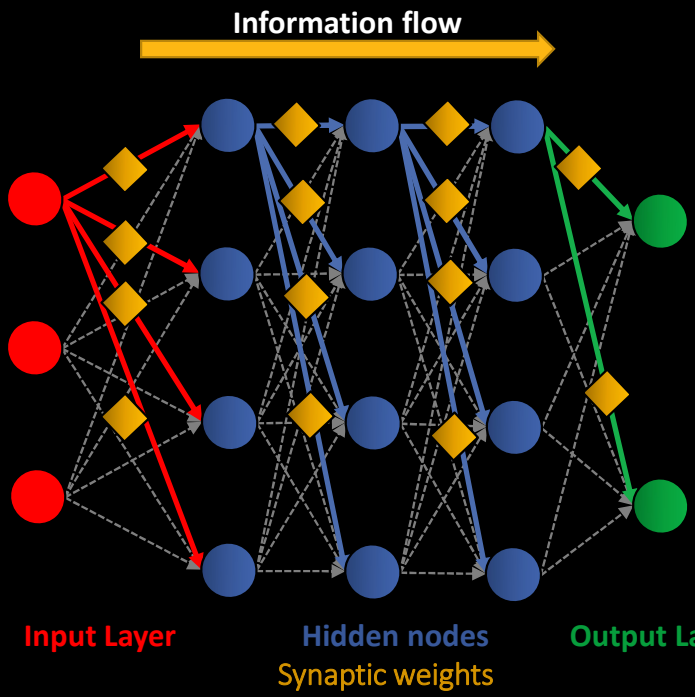
Neurons

- connected by weighted connections → Synapses
- activation is derived from evaluating activation function for a weighted sum of inputs

Felix Hermann eli@zurich.ibm.com

Tasks for Neural Networks





Dog

Cat

recognize:

- images from a camera,
- language from microphone or
- patterns in data set

- detect features by e.g., convolution

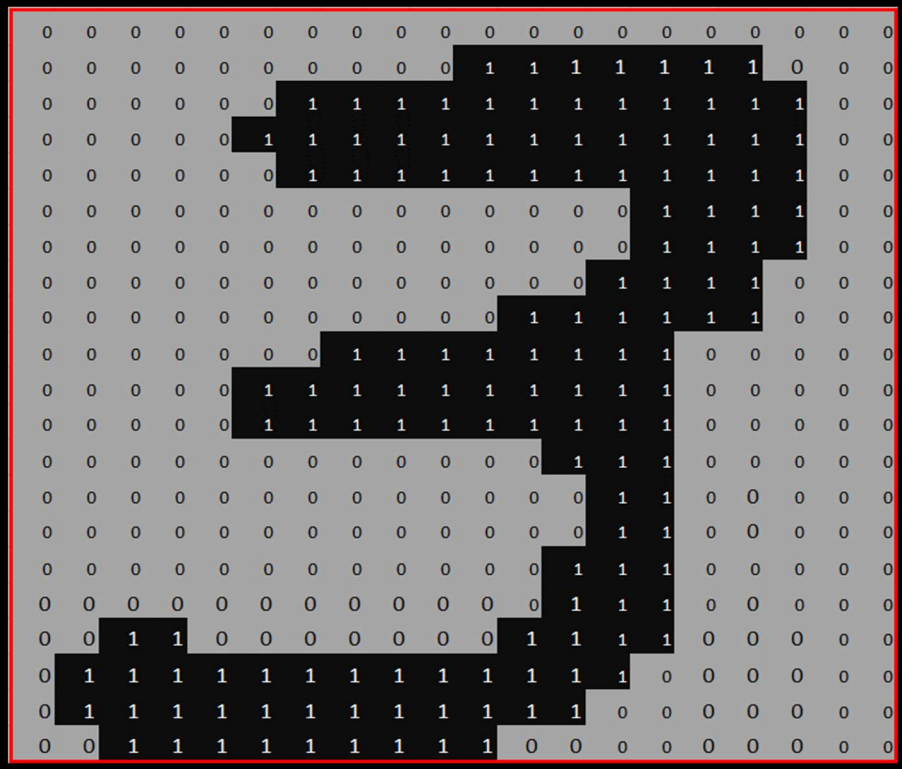
→ requires massive computing power

What is a Convolution in Image Processing?

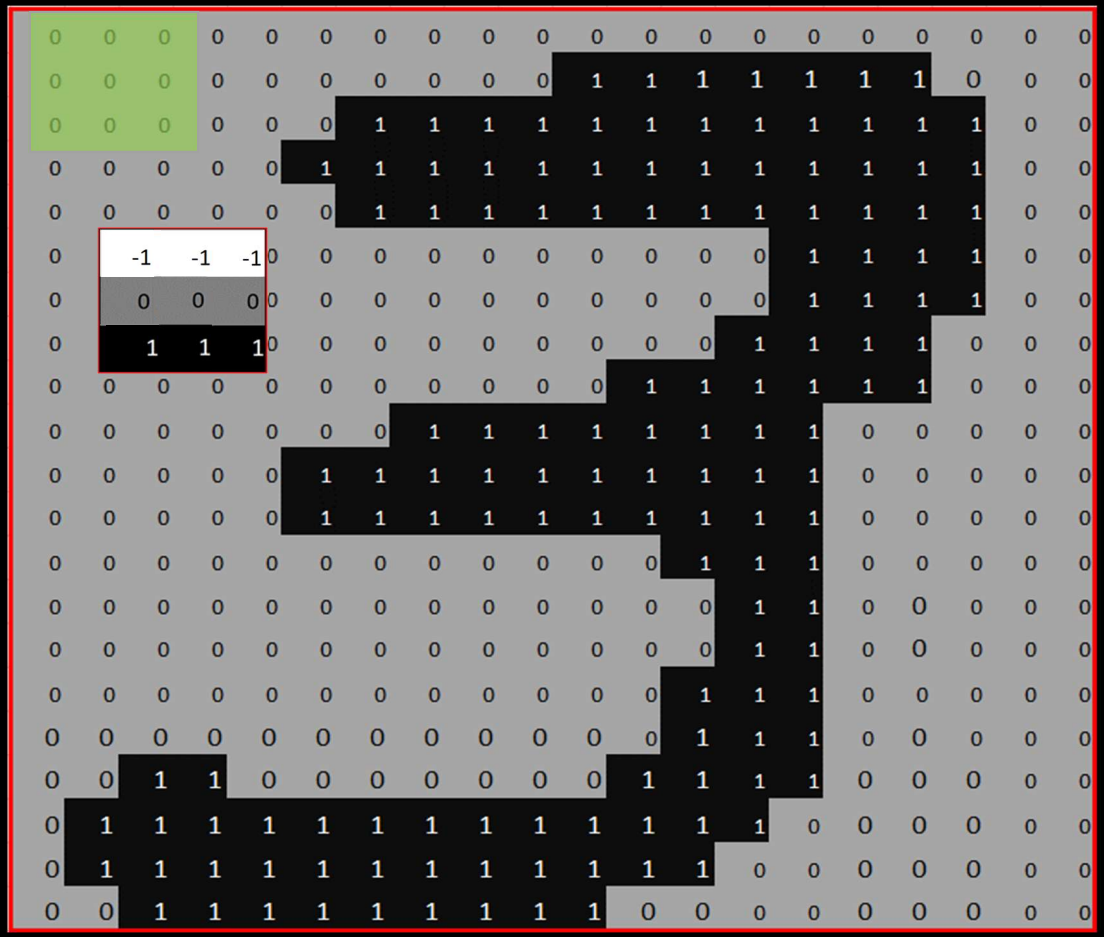
defined filter-kernel

-1	-1	-1
0	0	0
1	1	1

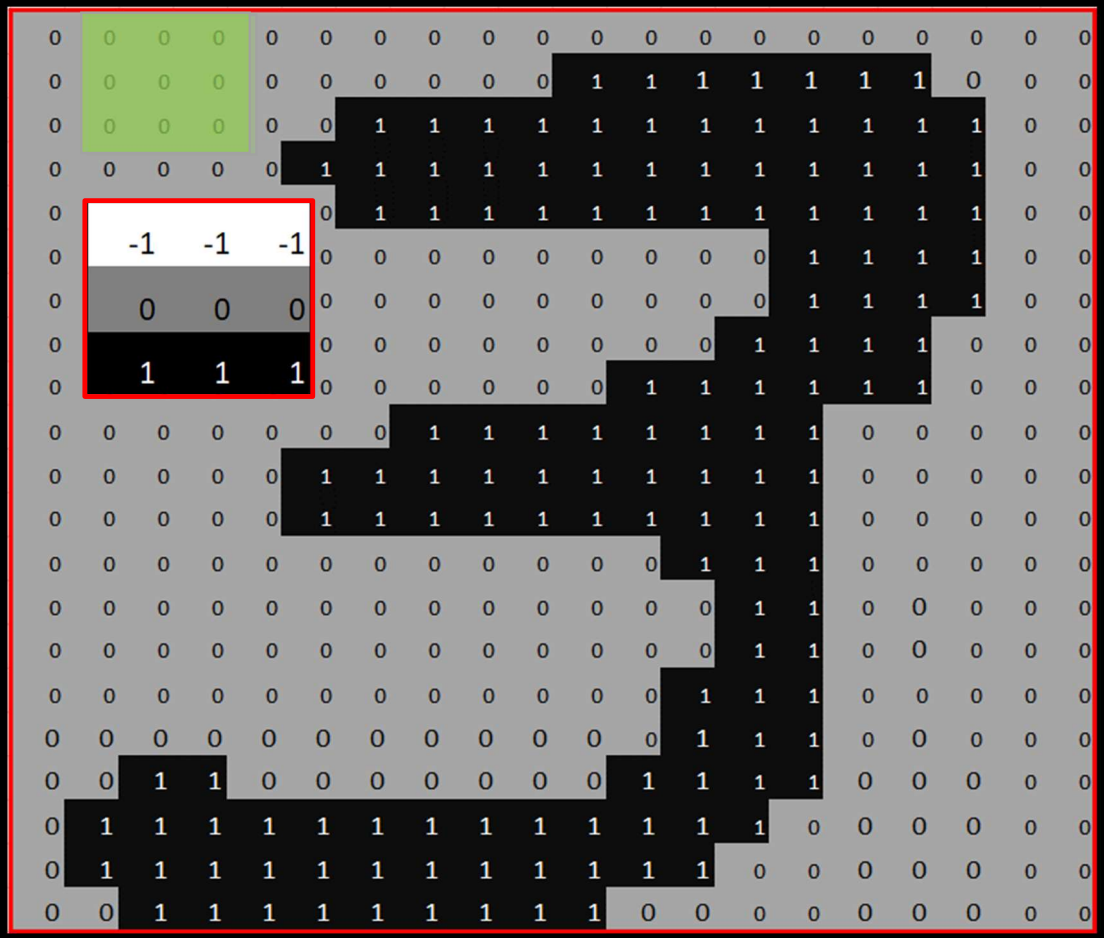
unknown image

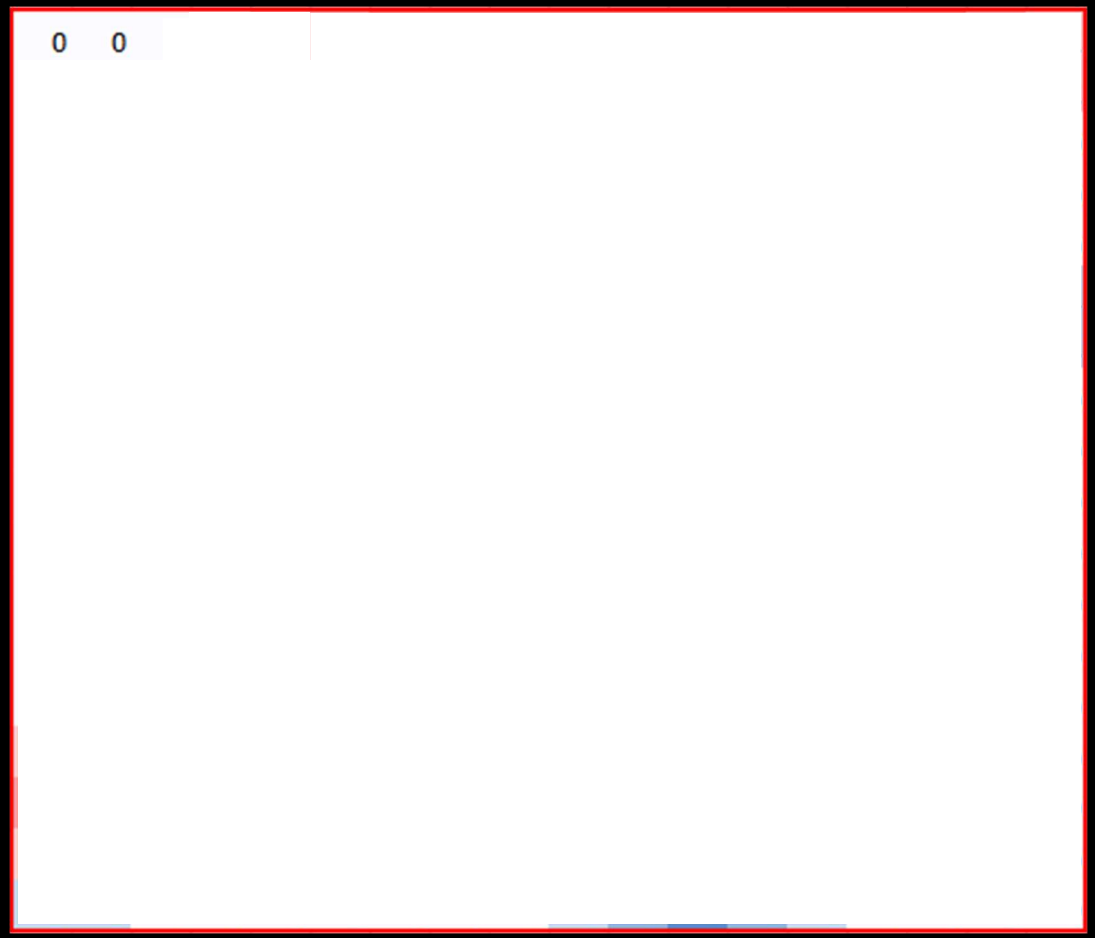


What is a Convolution in Image Processing?

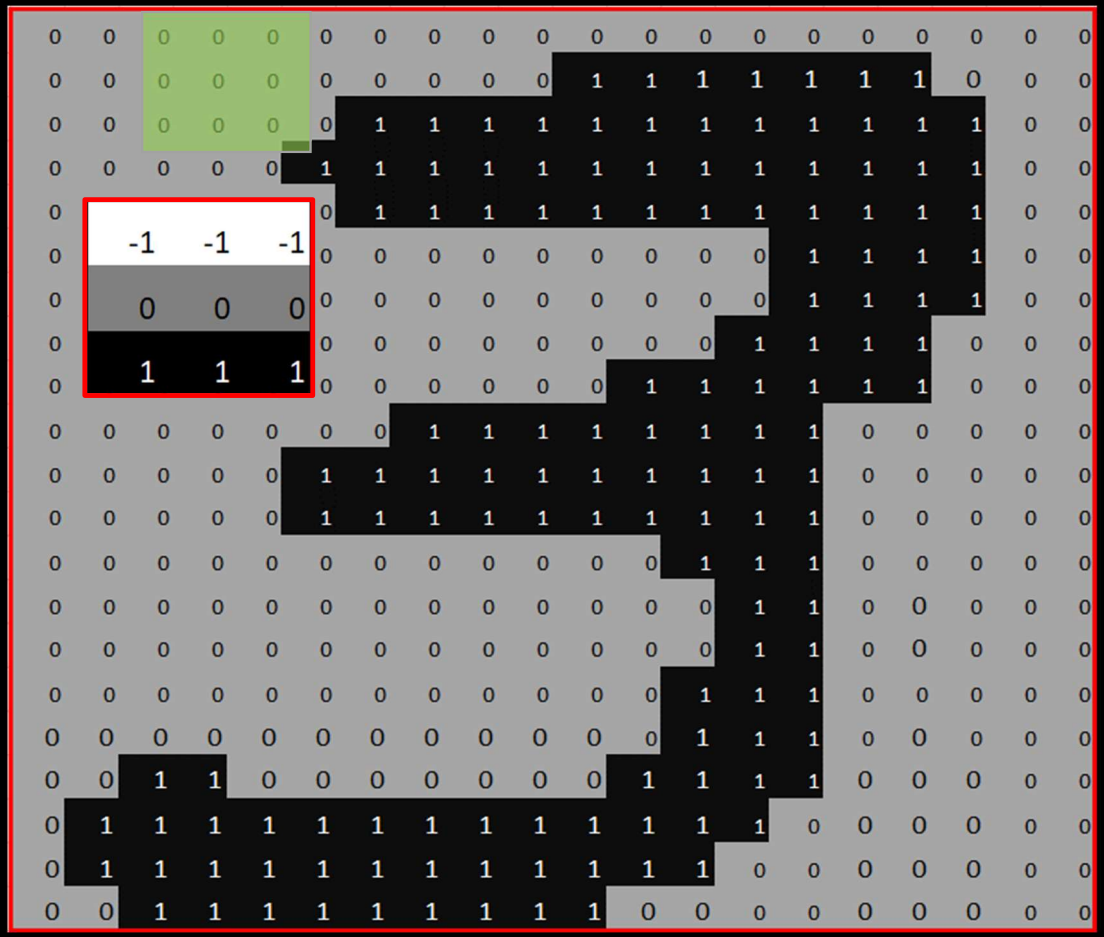


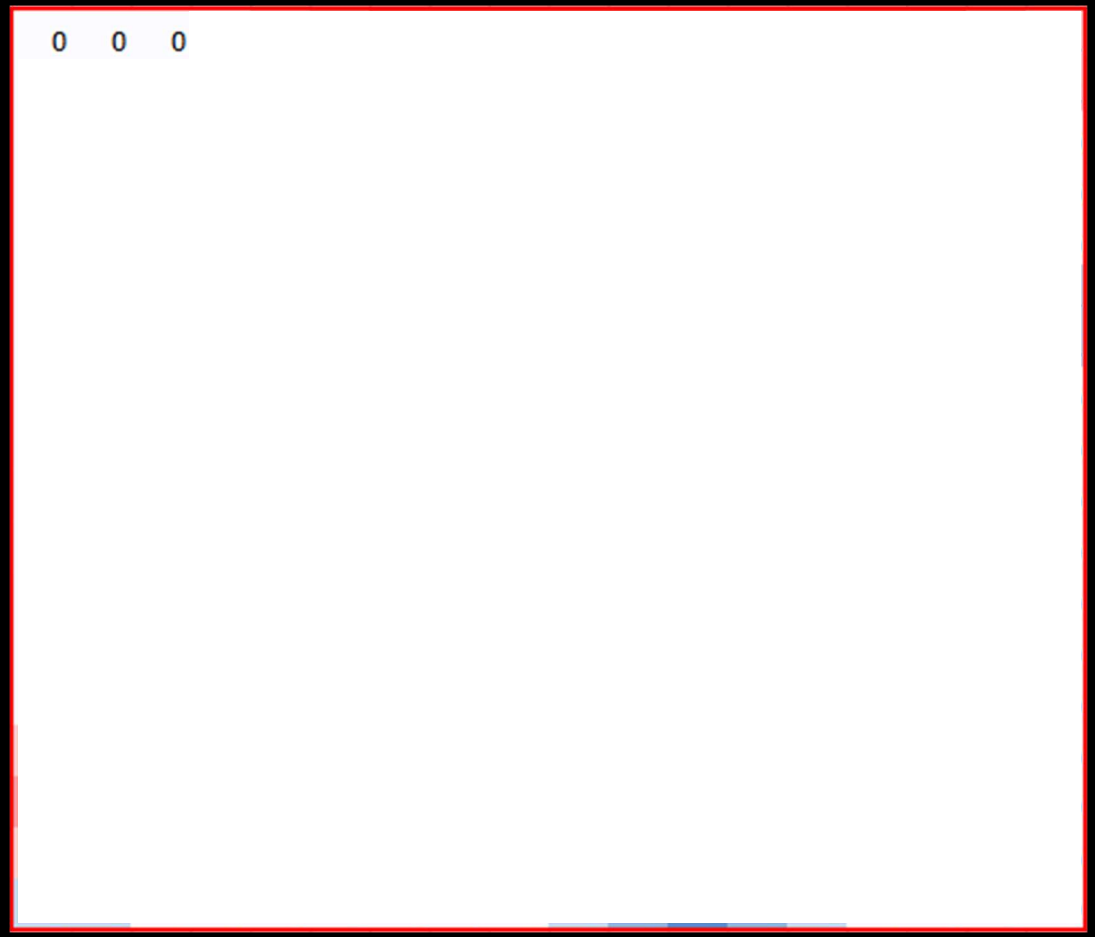
What is a Convolution in Image Processing?



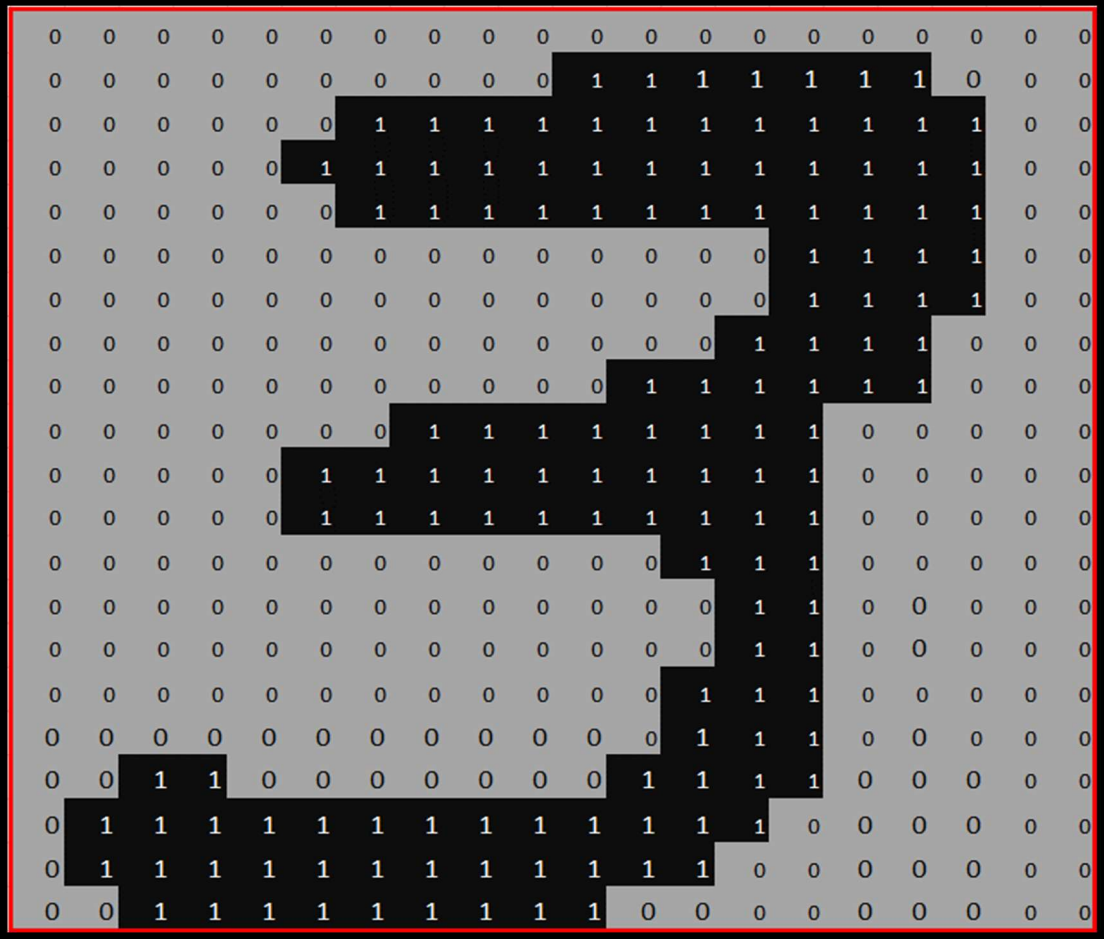


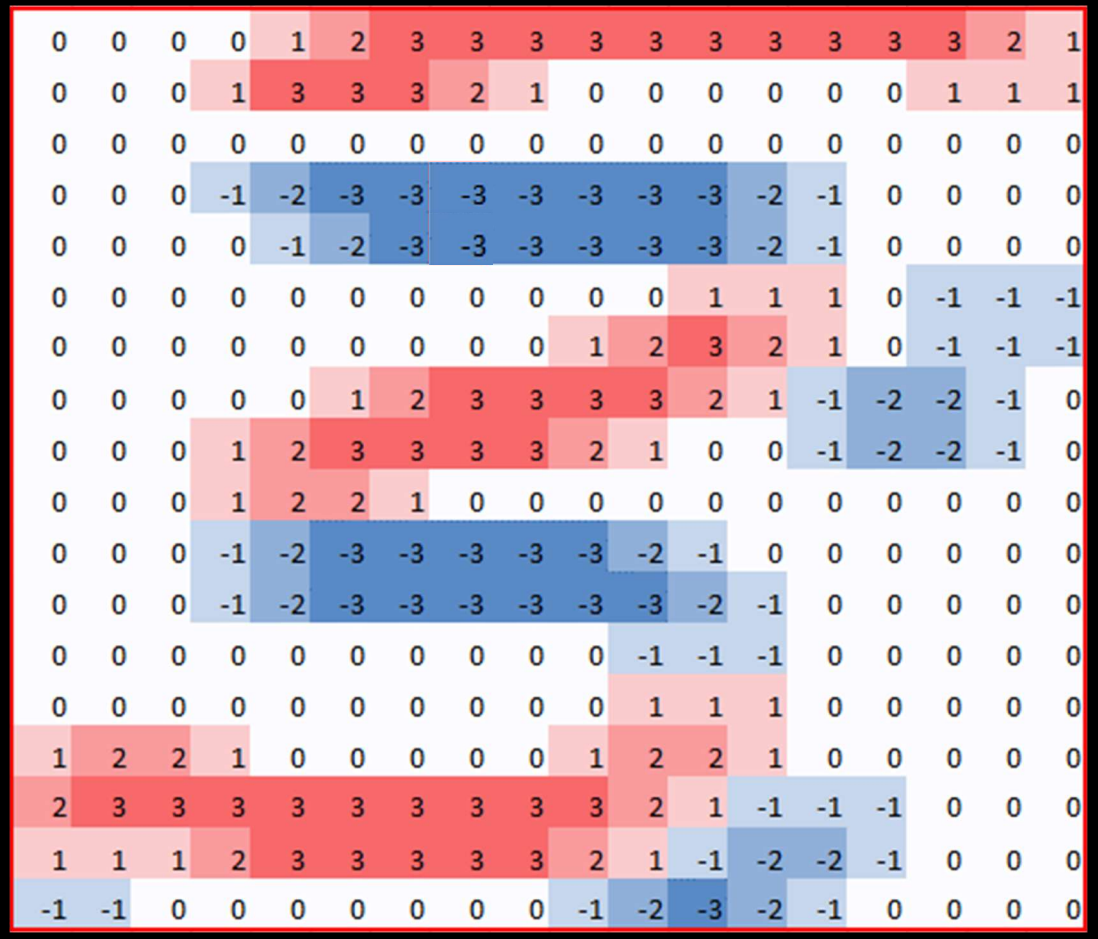
What is a Convolution in Image Processing?



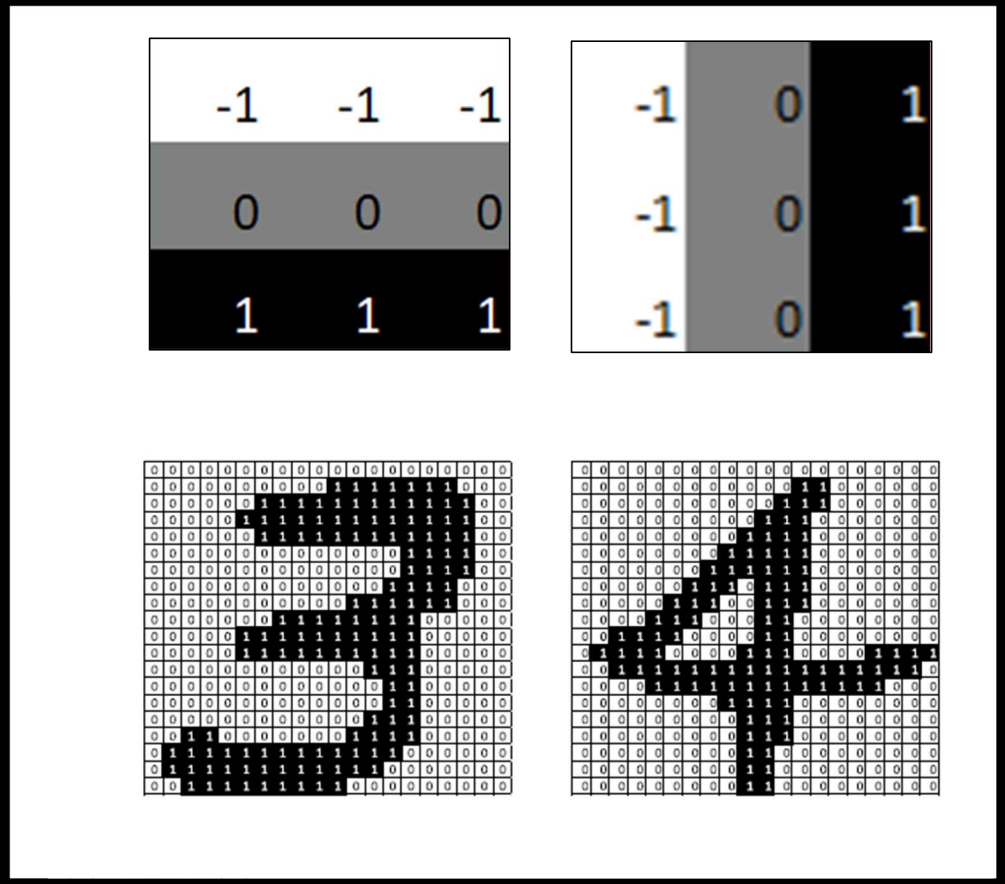


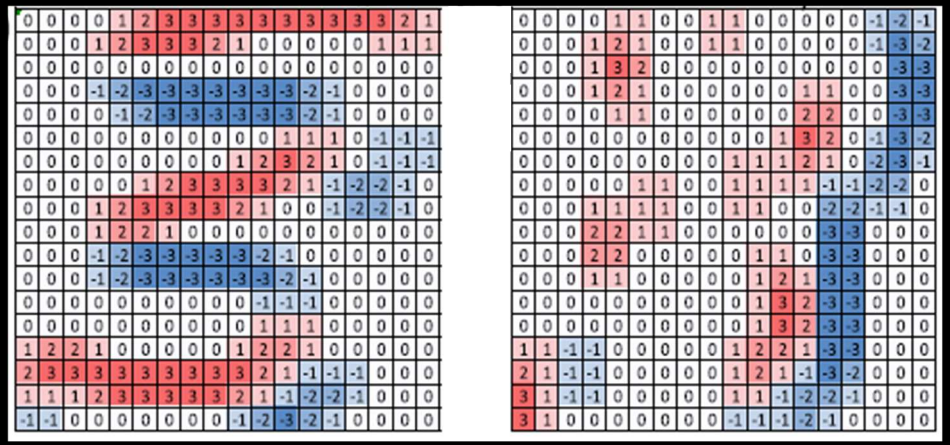
What is a Convolution in Image Processing?





What is a Convolution in Image Processing?



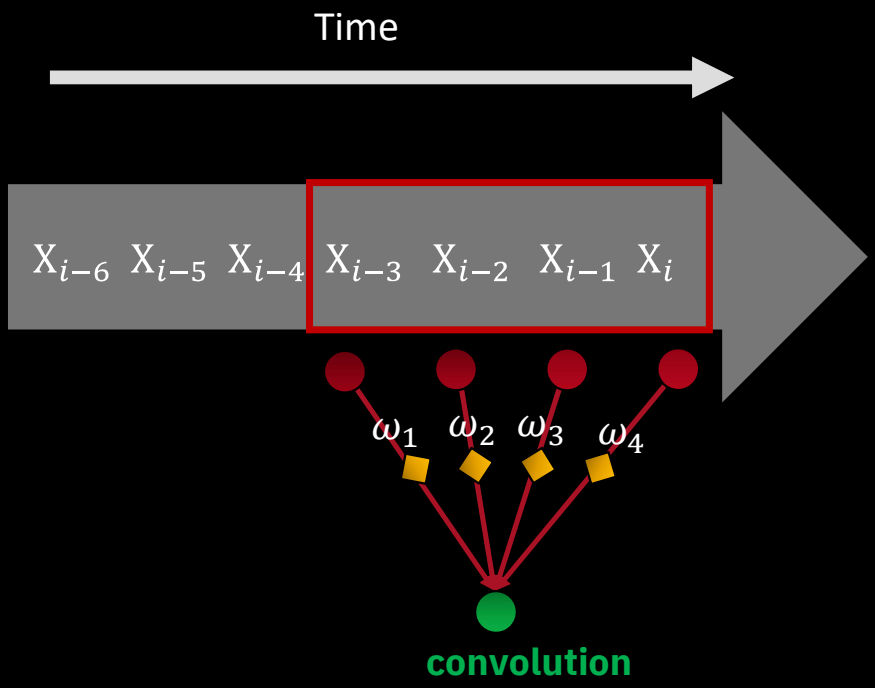


- convolution detects features, which allows for interpretation further processing

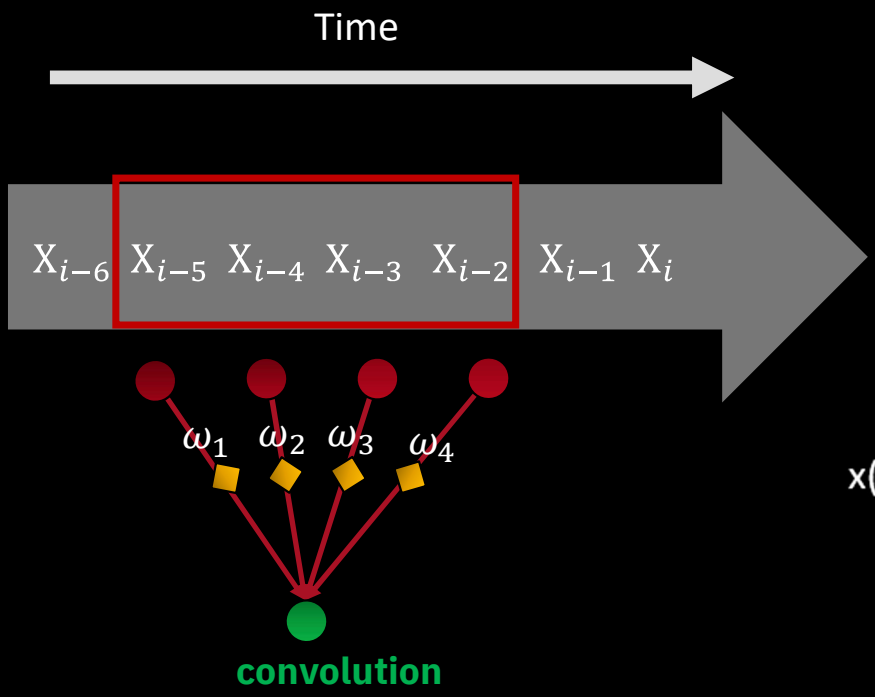
Gao et al. IEEE ELECTRON DEVICE LETTERS, VOL. 37, NO. 7, JULY 2016

Felix Hermann eli@zurich.ibm.com

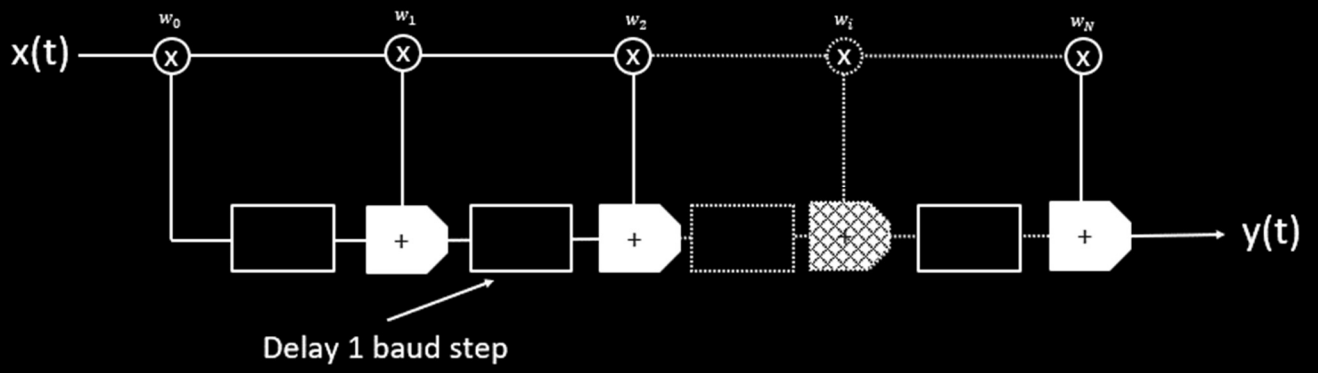
Convolution in Neural Networks



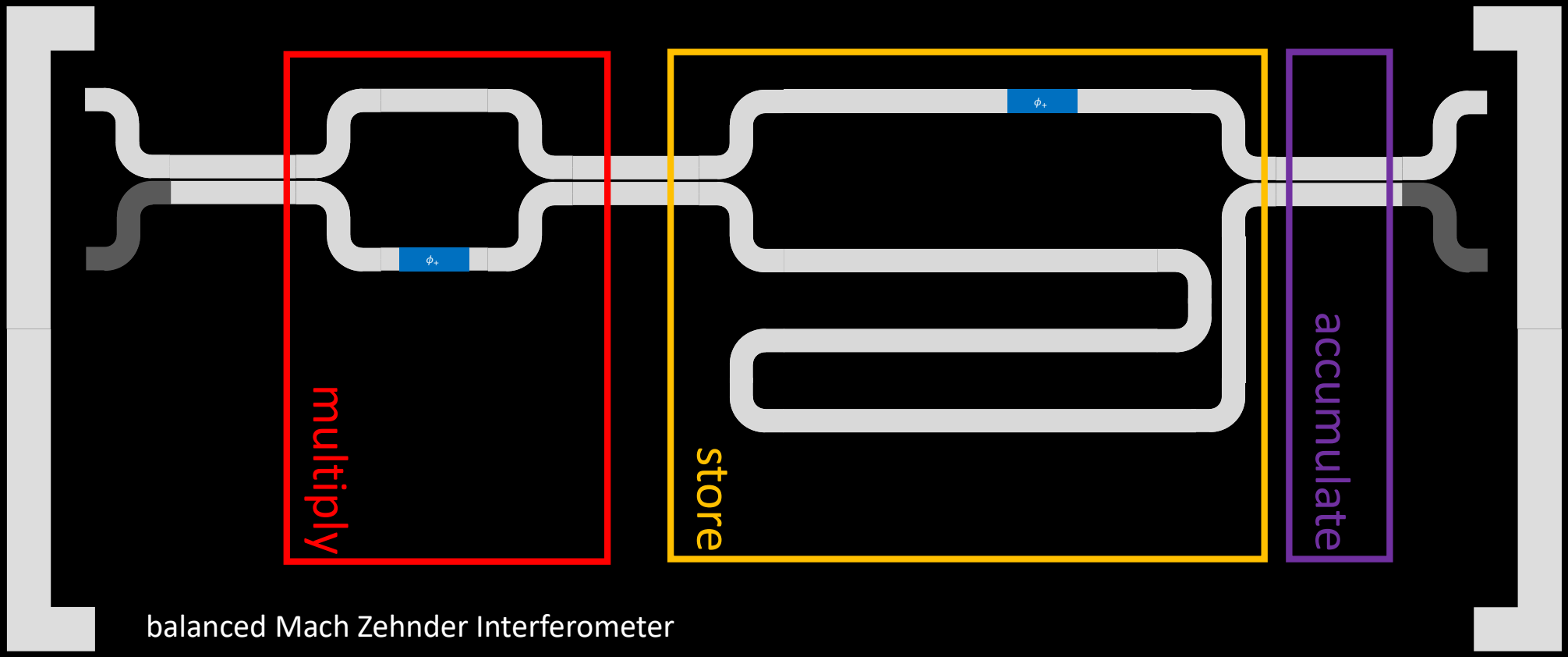
Convolution in Neural Networks



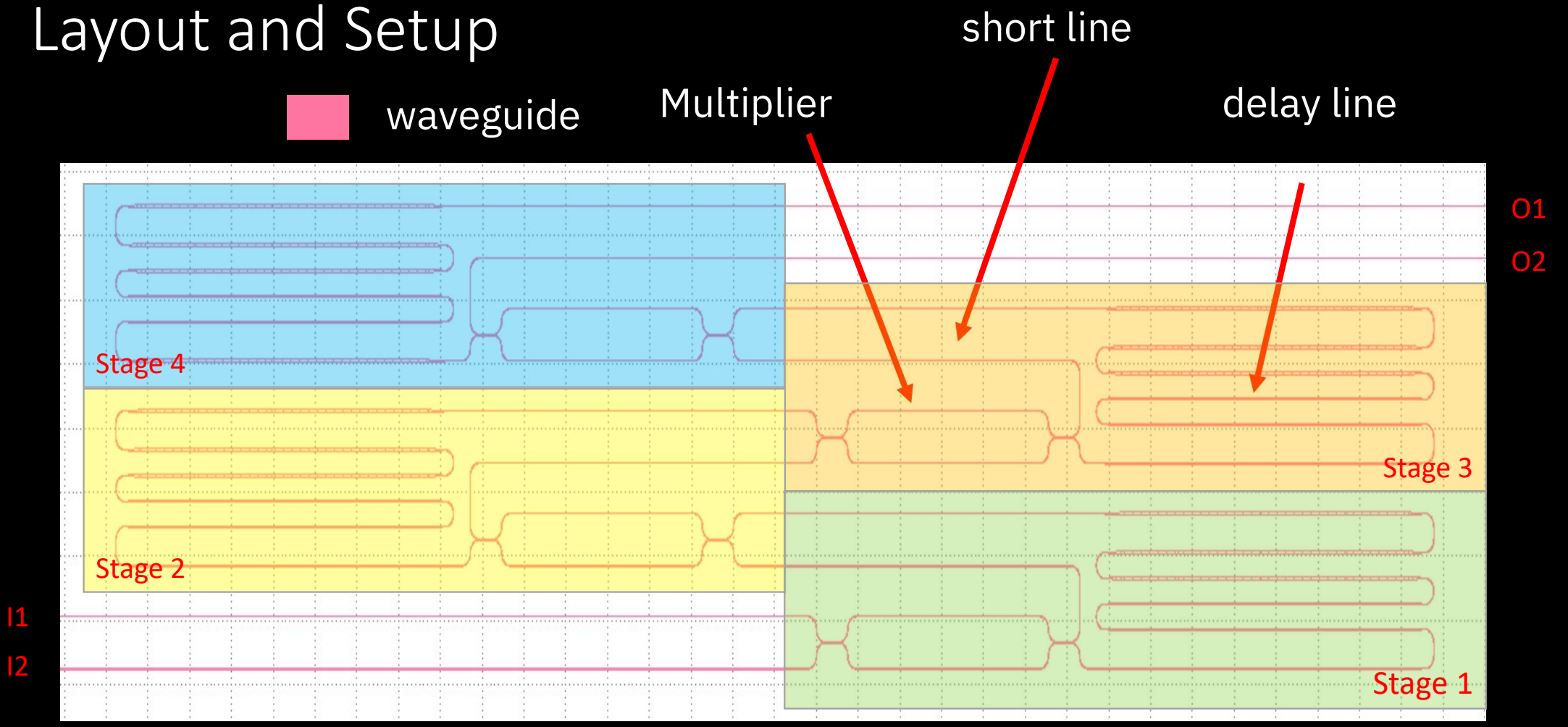
$$y[n] = \sum_{i=0}^{N} \omega_i \cdot x[n-i]$$



Convolution Optical Processor

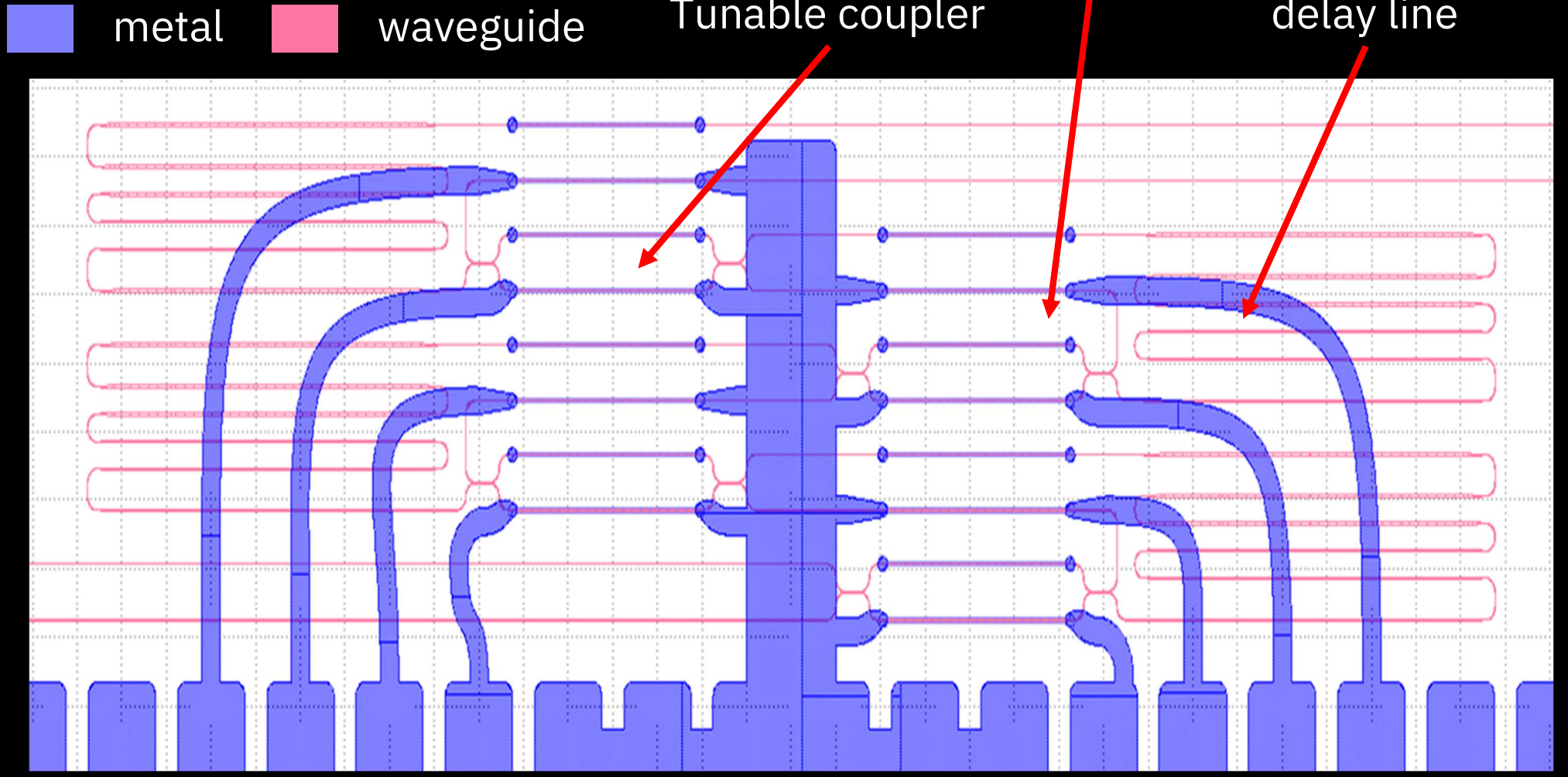


Layout and Setup



Felix Hermann eli@zurich.ibm.com

Layout and setup

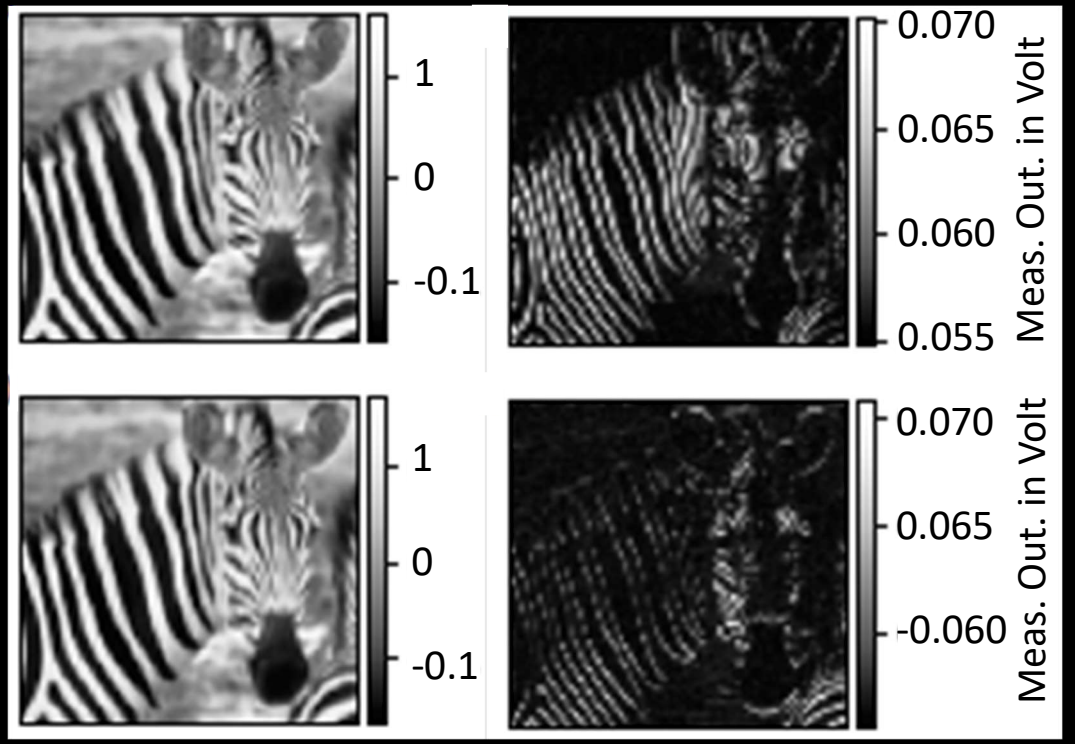


proof of principle results: Sobel filter

- 2-element kernel:

$\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ vertical

$(1 \quad -1)$ horizontal

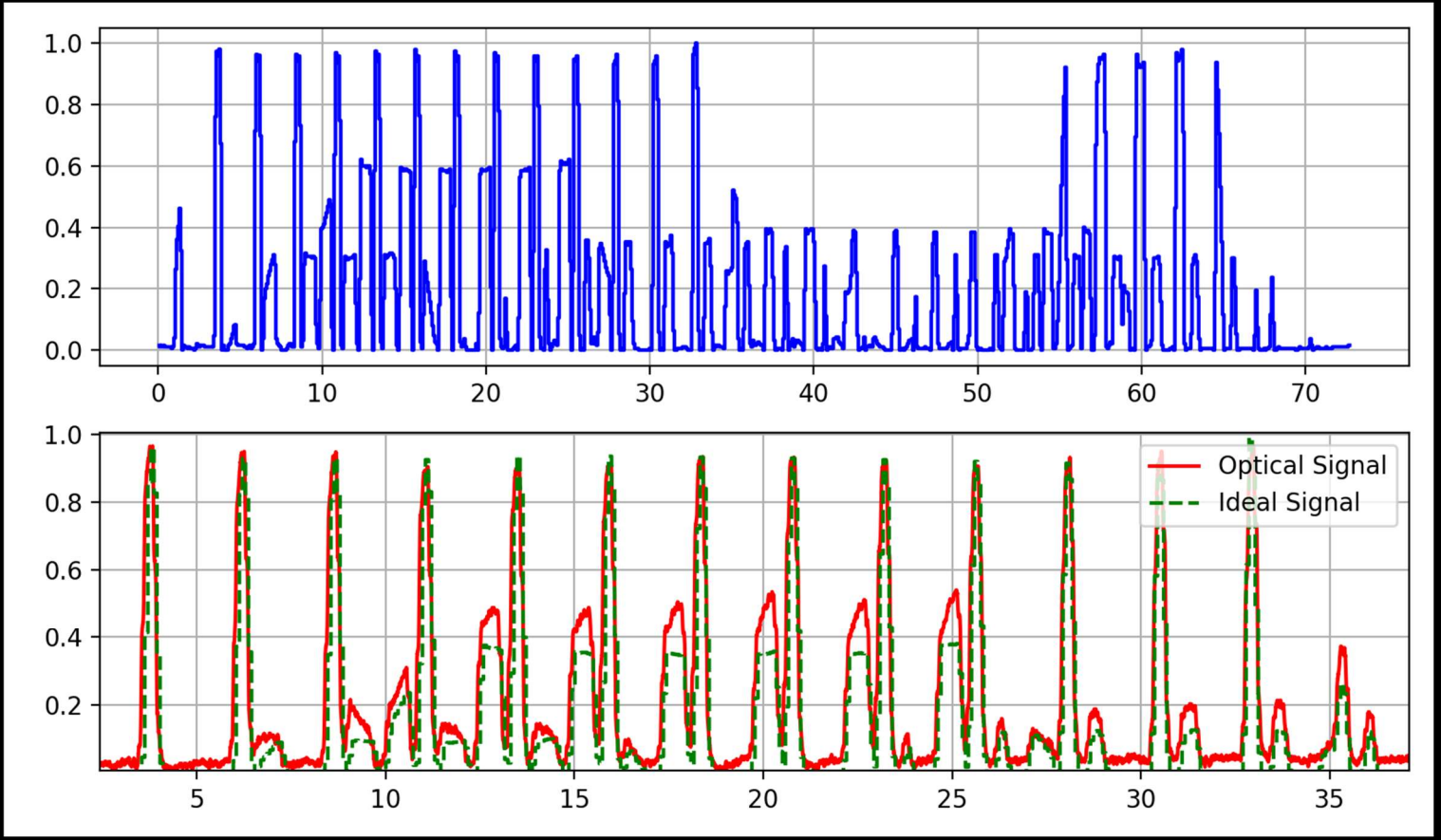


reprinted from Doctoral Thesis by Pascal Stark

proof of principle results: Sobel filter

- 3-element kernel:

$$(1 \quad 2 \quad 1)$$

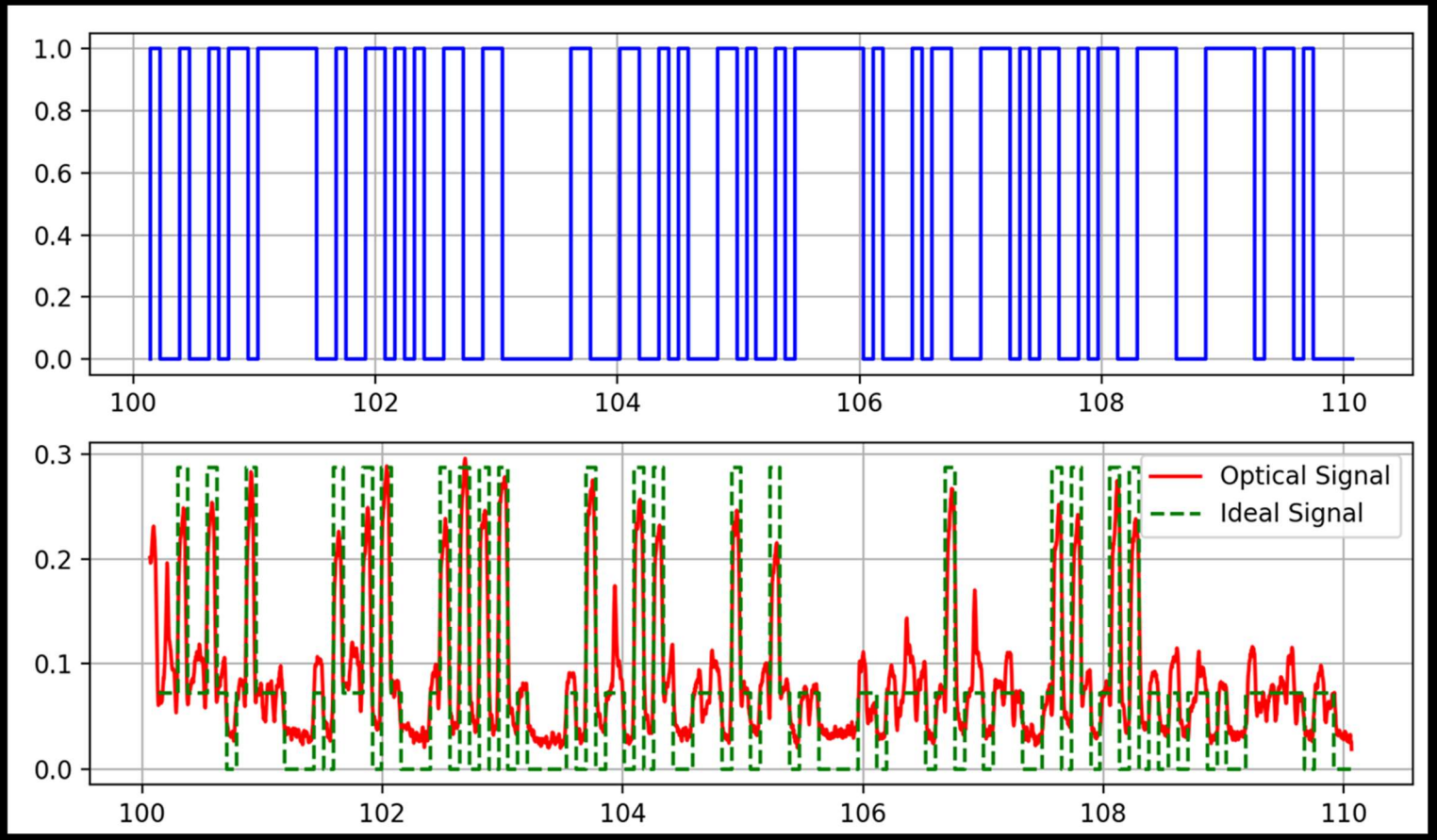


Felix Hermann eli@zurich.ibm.com

proof of principle results: Sobel filter

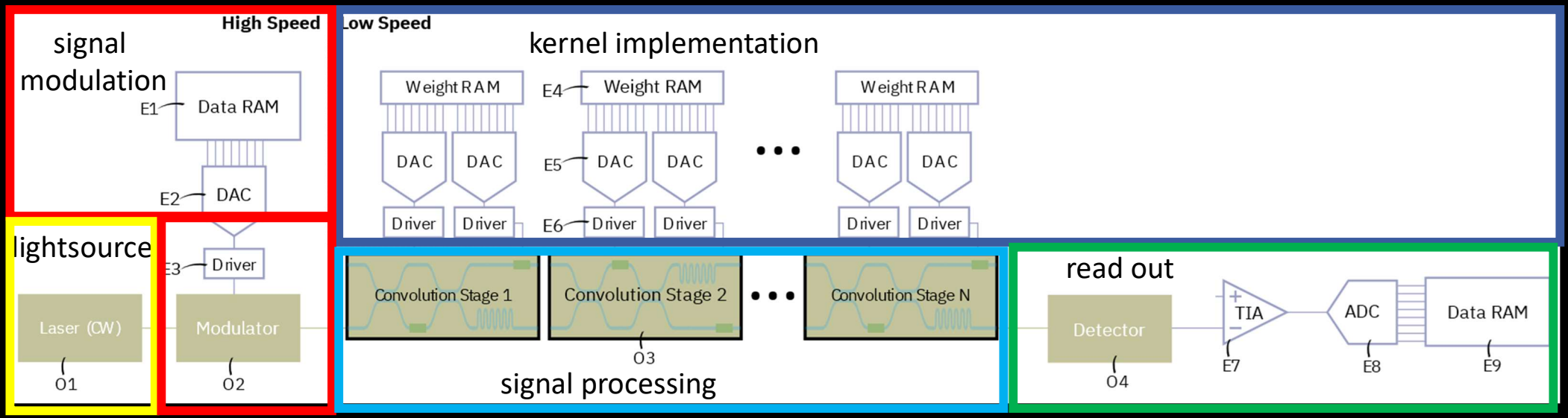
- 4-element kernel:

$$(1 \quad -1 \quad -1 \quad 1)$$



Felix Hermann eli@zurich.ibm.com

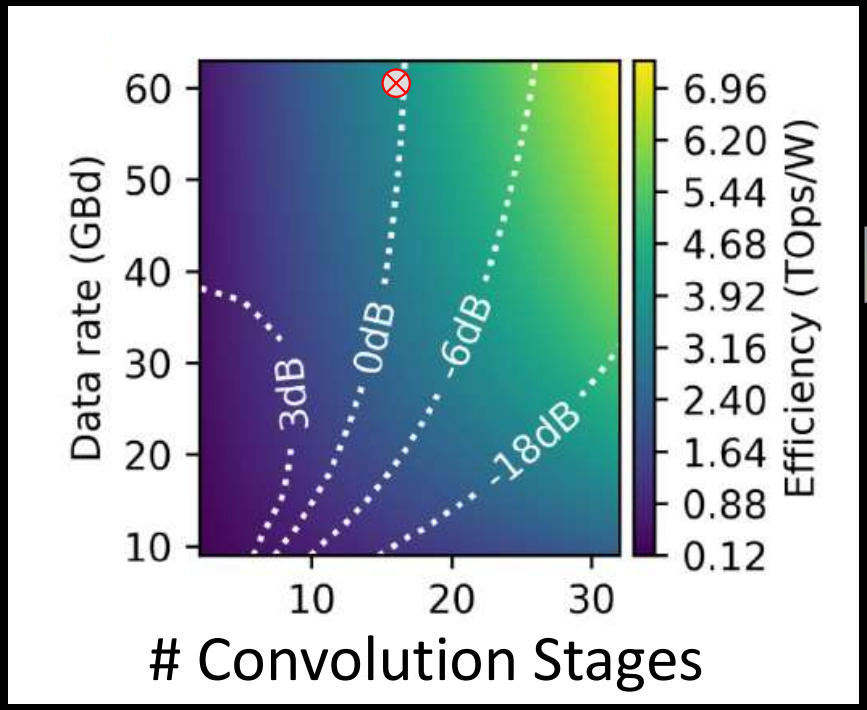
Benchmarking



Felix Hermann eli@zurich.ibm.com

Benchmarking

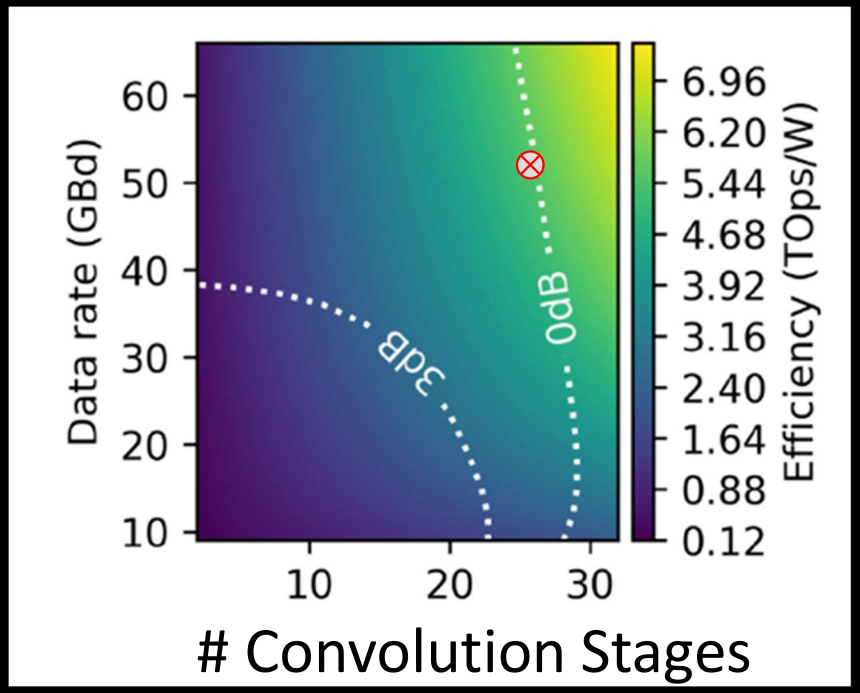
in-house Si platform + BTO



⊗ = most efficient device

··0dB·· = power margin at receiver

commercial Si platform + BTO



low latency, high speed, fast kernel implementation

25 stage - devices with up to 6TOps/W are possible with today's available technology!

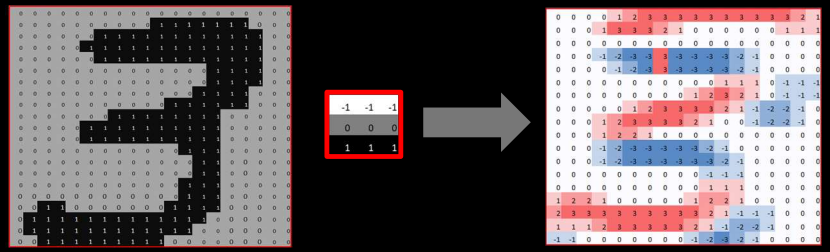
Acknowledgment

- my team at IBM
- Swiss National Science Foundation grant no. 175801 (Napreco)

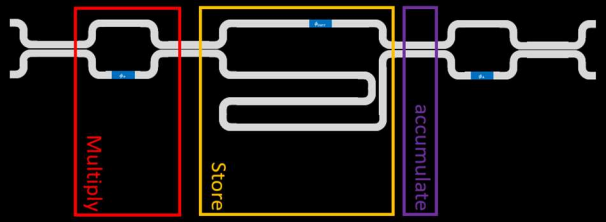
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Thank you and let's discuss!

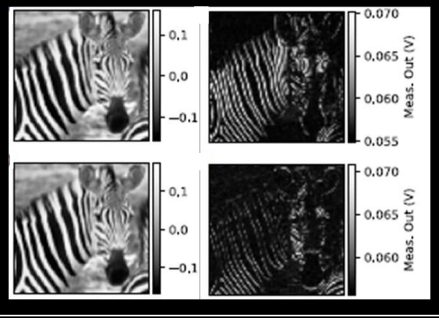
- convolutions for NNs

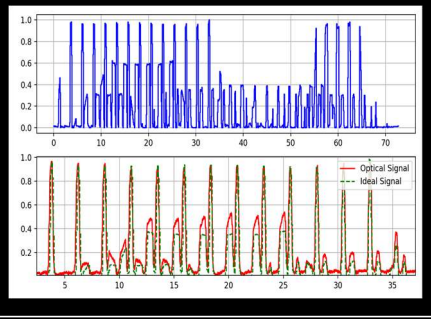


- device concept



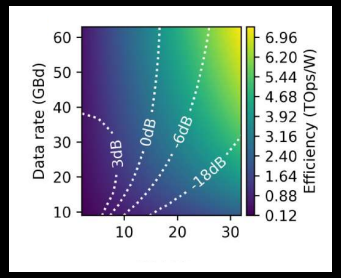
- proof of principle

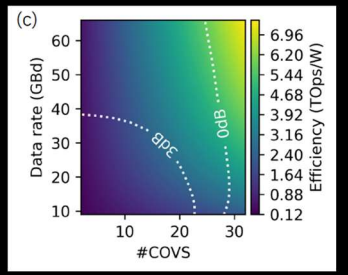




- benchmark

6TOps/s





Felix Hermann eli@zurich.ibm.com