Algorithmic enablers for compact Neural Network topology Hardware design: review and trends

William Guicquero, Arnaud Verdant

Univ. Grenoble Alpes, CEA-Leti

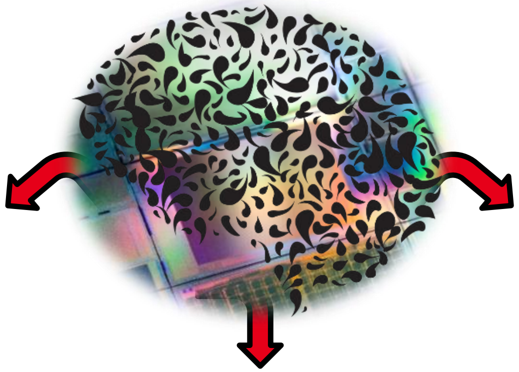
2020 IEEE International Symposium on Circuits and Systems
Virtual, October 10-21,2020

Introduction

AI device enablers

AI-related applications are growing, **dedicated devices** also…

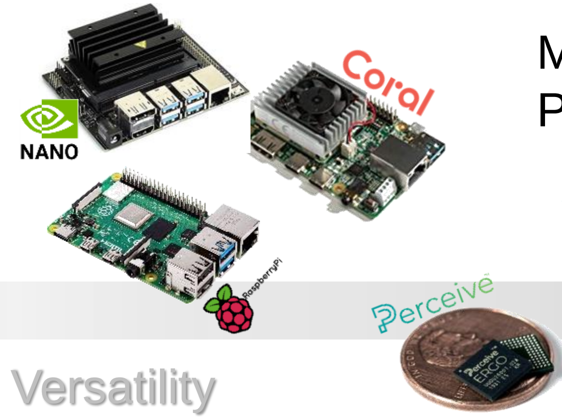
→ what are we actually talking about ?

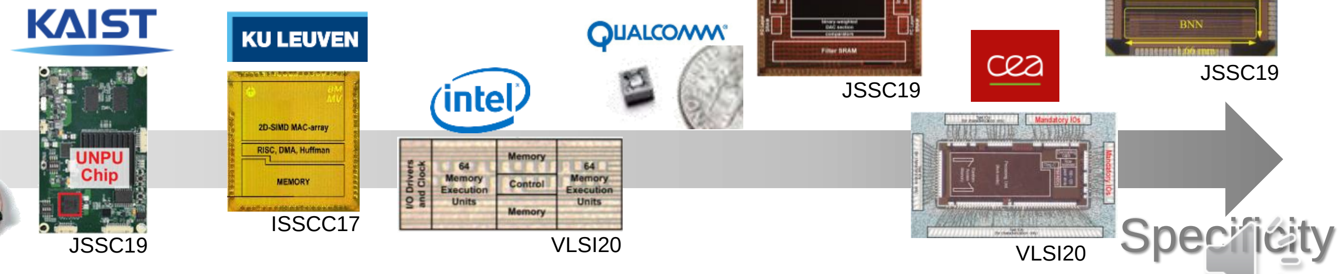


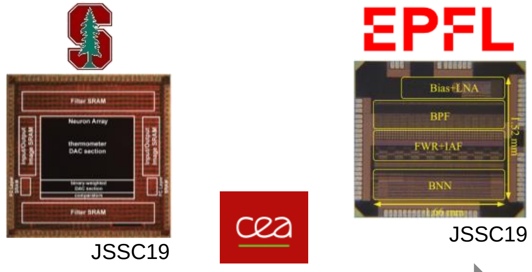
AI @ nano-workstation
Model size: ~100Mb
Power consumption: ~10W

AI @ the edge
Model size: ~1Mb
Power consumption: ~10mW

AI @ sensor node
Model size: ~10kb
Power consumption: ~10µW







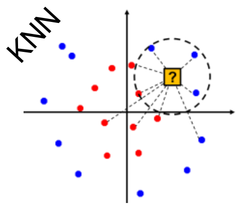
Introduction

AI algorithm enablers

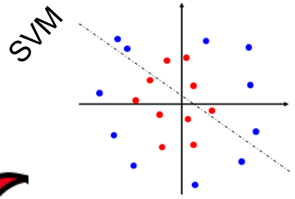
AI-related applications are growing, **dedicated algorithms** also…

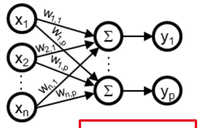
→ what are we actually talking about ?

Distance-based



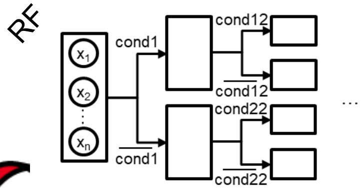
Linear projection



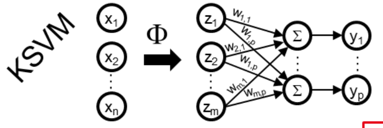


$\boxed{\mathbf{y} = W\mathbf{x}}$

Decision trees



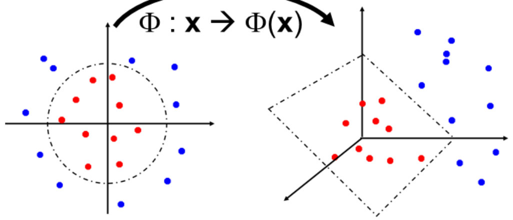
Kernel methods



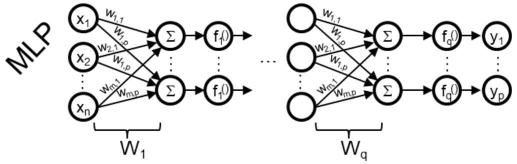
$\boxed{\mathbf{y} = W\Phi(\mathbf{x})}$

Higher Dimensional space

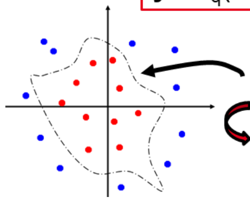
$\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x})$



Neural Networks



$\boxed{\mathbf{y} = f_q(W_q \ldots f_2(W_2 f_1(W_1 \mathbf{x})))}$



Properly describe learned data

⚠ overfitting

Powerful

Generic

Scalable

Introduction

Main Motivation

Report the main State-Of-The-Art algorithmic enablers for **compact** Neural Network topology design and **categorize** it

Dimensionality Reduction

→ acts on: **layers width**
- Reduce the number of MACs
- Reduce numbers of computational nodes
- Limit model size memory

Quantization with Normalization

→ acts on: **data dynamic range**
- Simplify HW core components
- Reduce local memory needs
- Limit model size memory

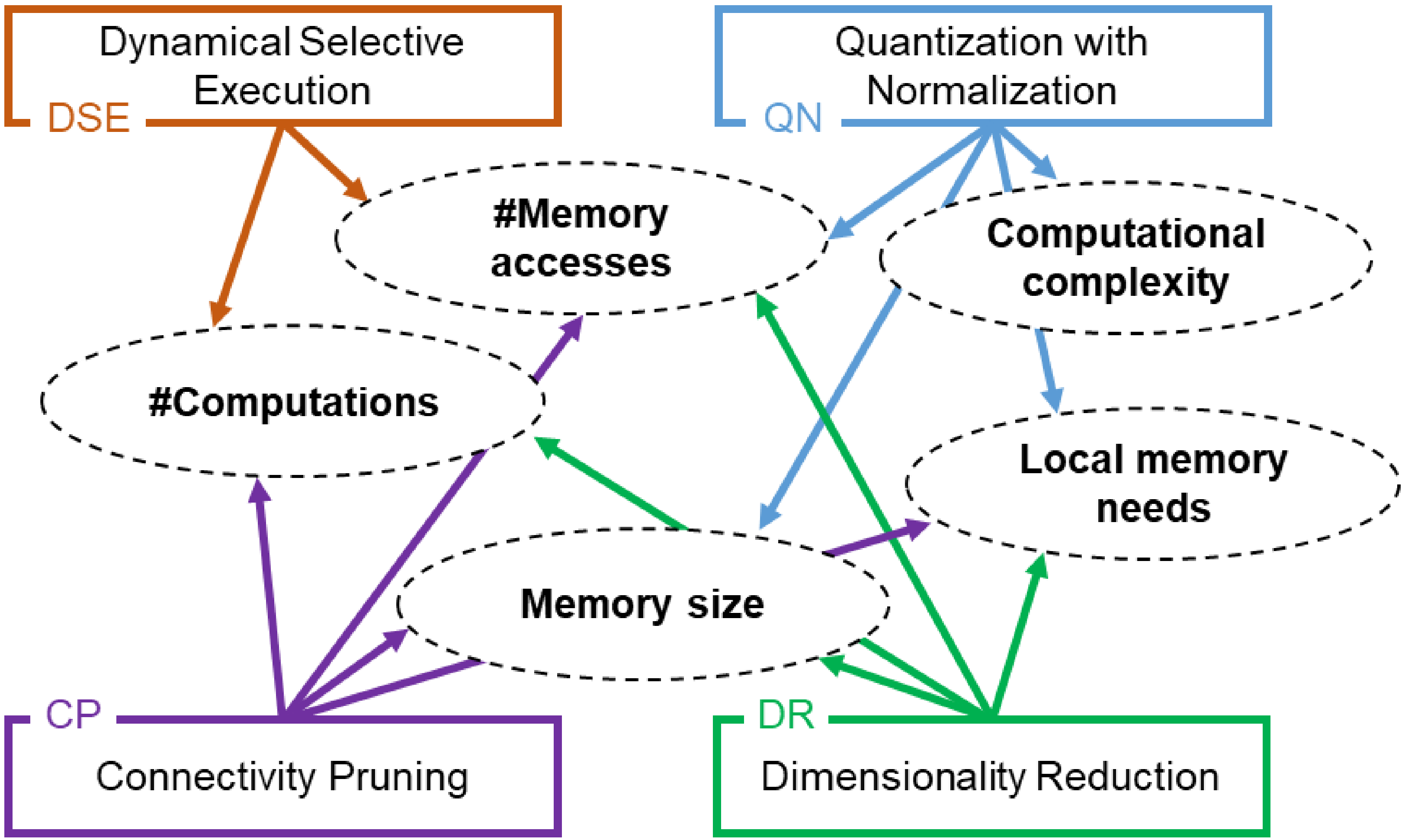
Connectivity Pruning

→ acts on: **coef. matrix sparsity**
- Limit needless computations
- Robustify models against overfitting

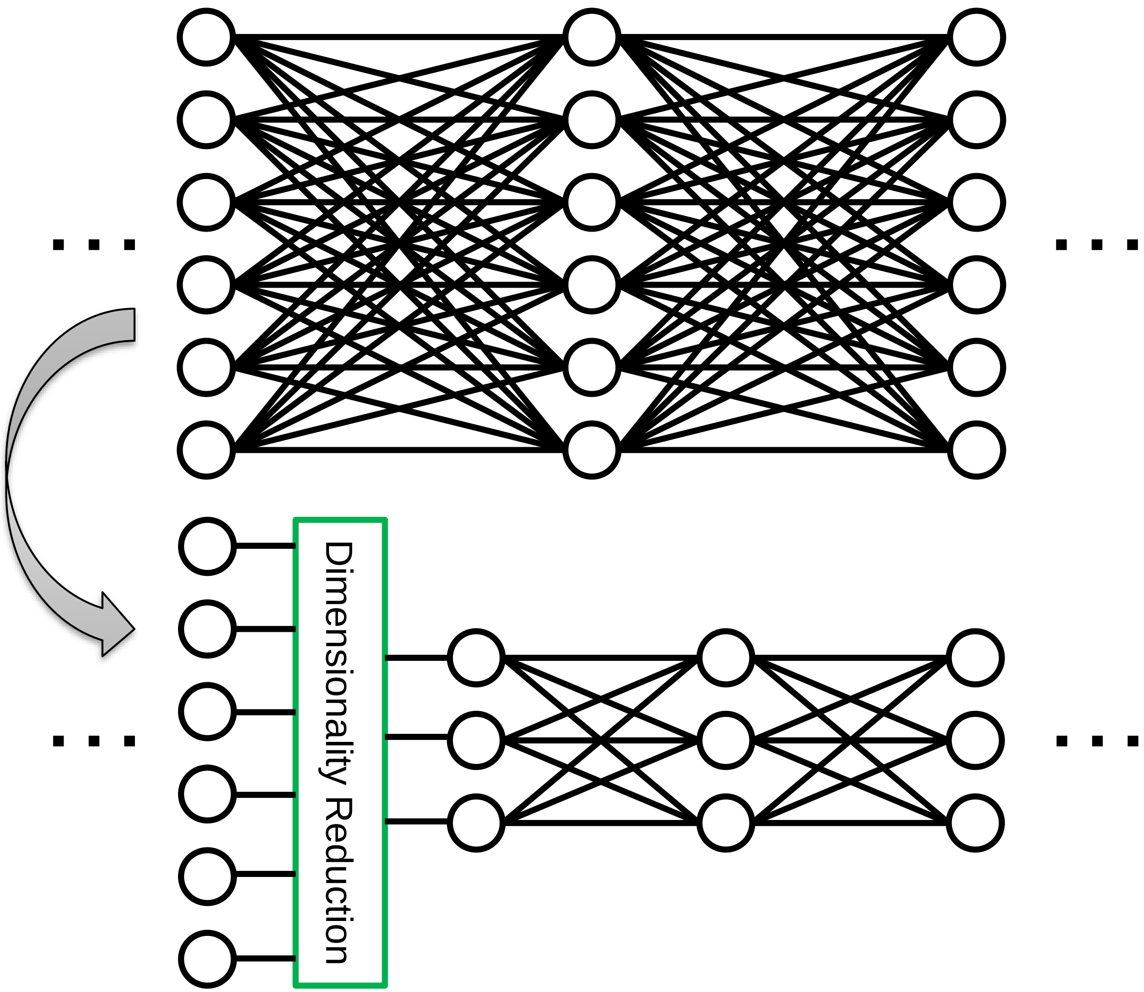
Dynamical Selective Execution

→ acts on: **topology activation**
- Limit the activation of HW components
- Cap average power consumption

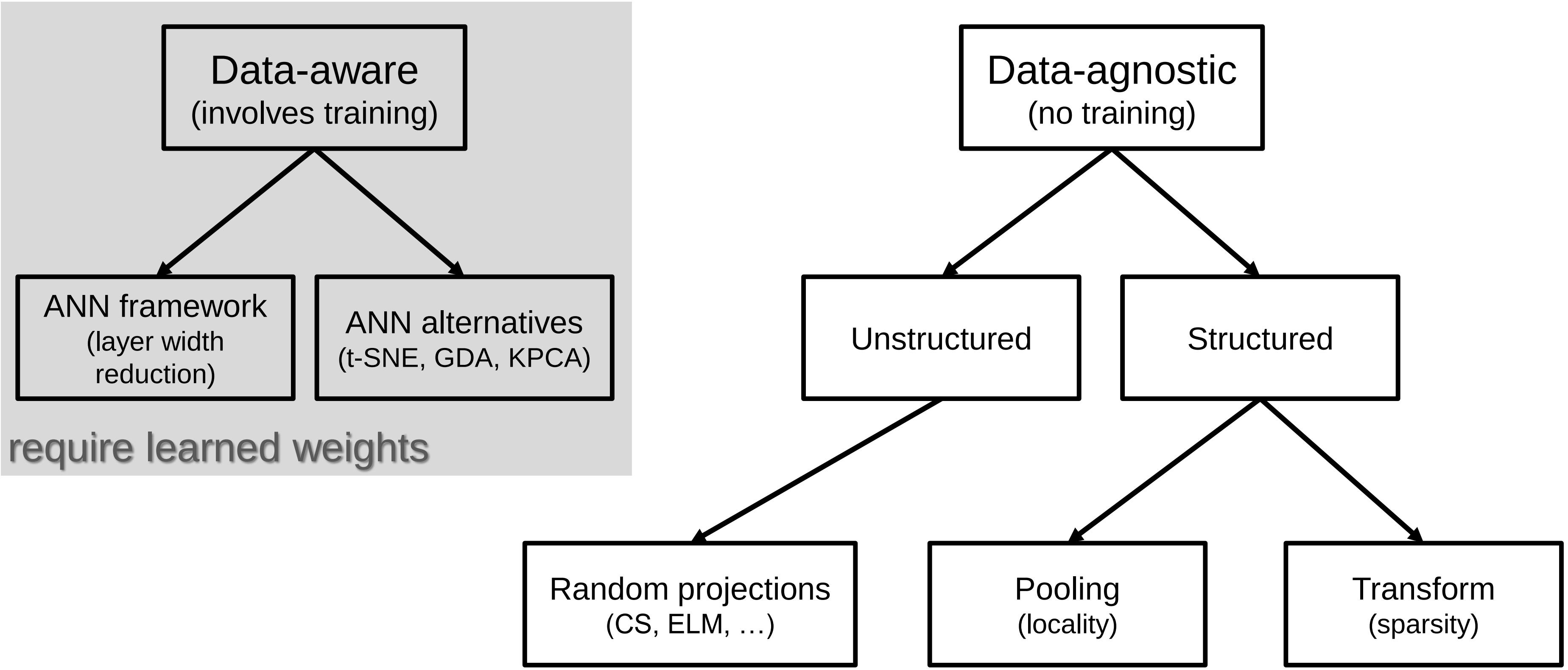
Actions of algorithmic enablers on HW limitations



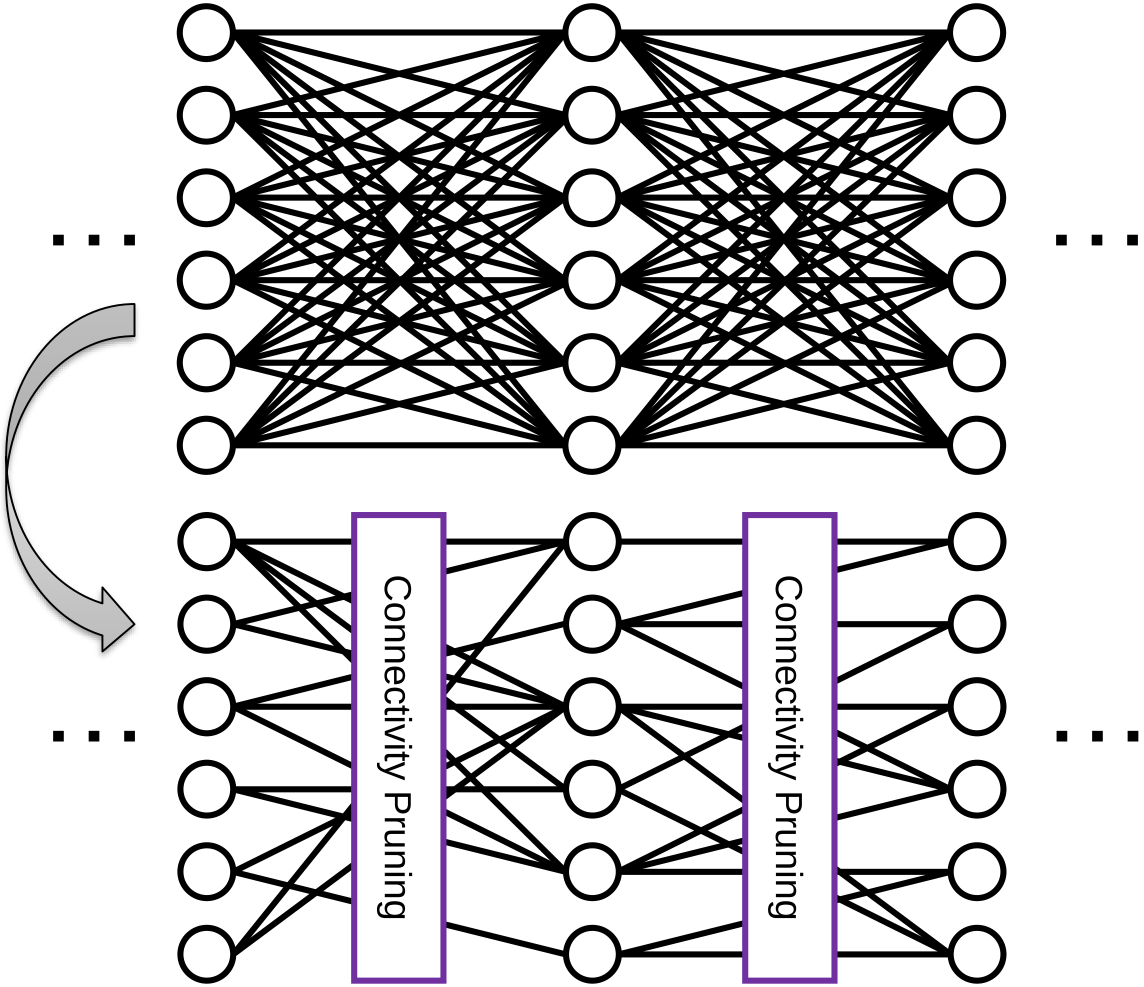
Dimensionality Reduction



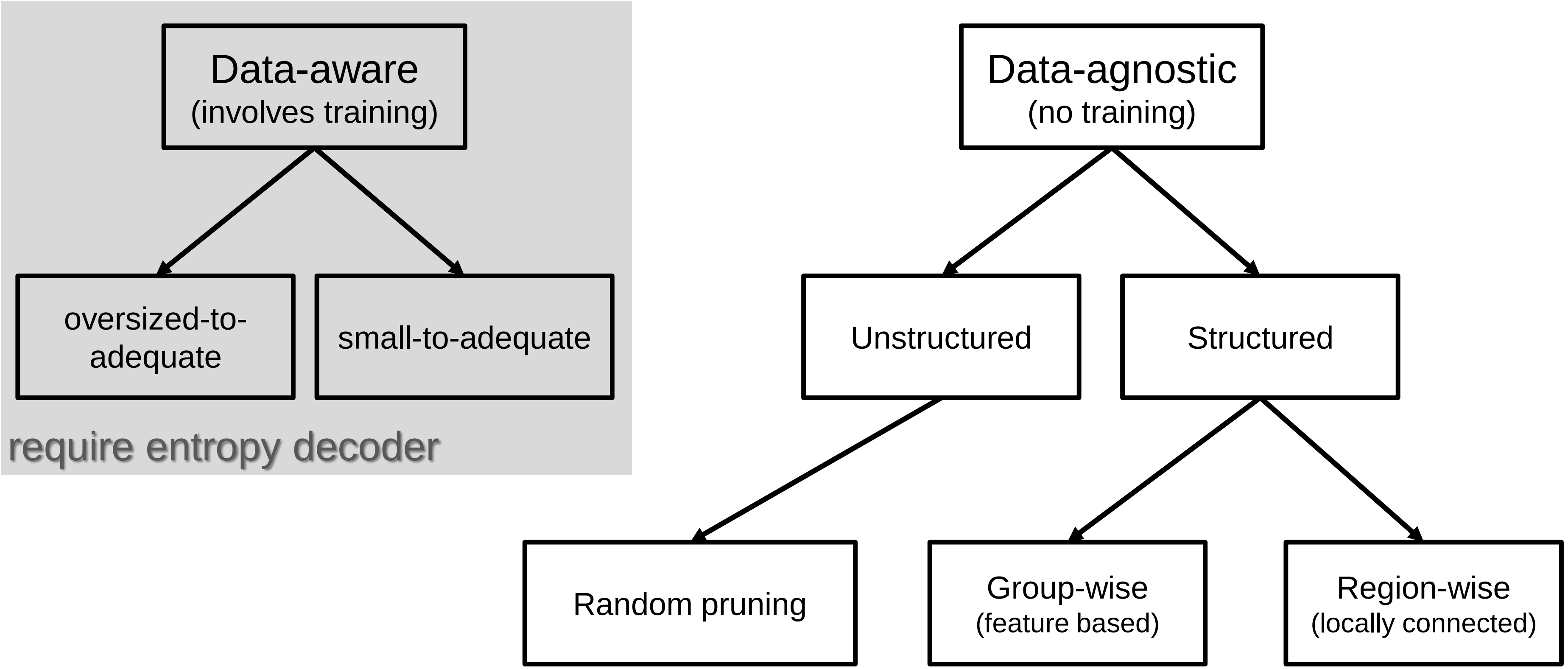
Dimensionality Reduction



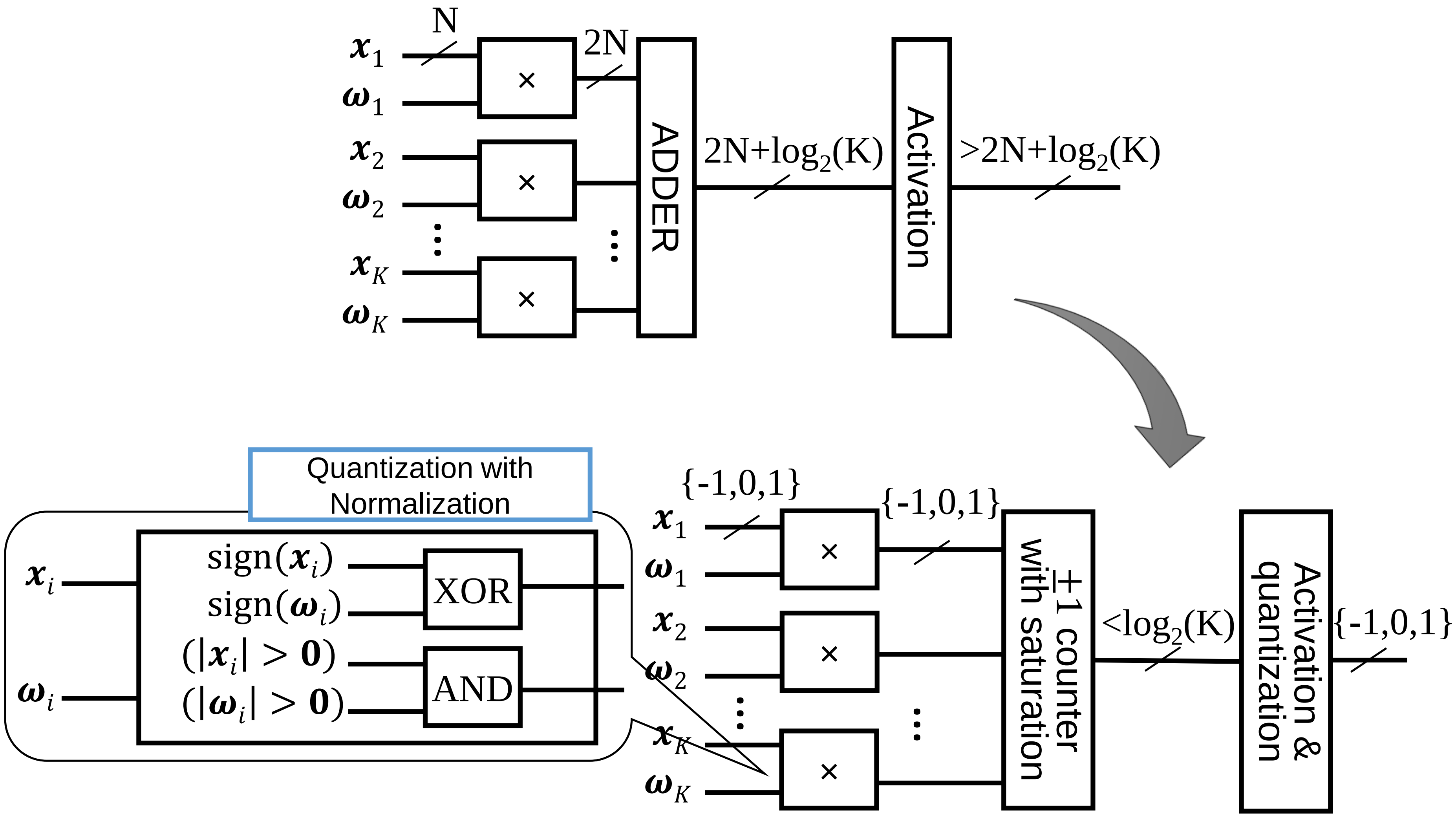
Connectivity Pruning



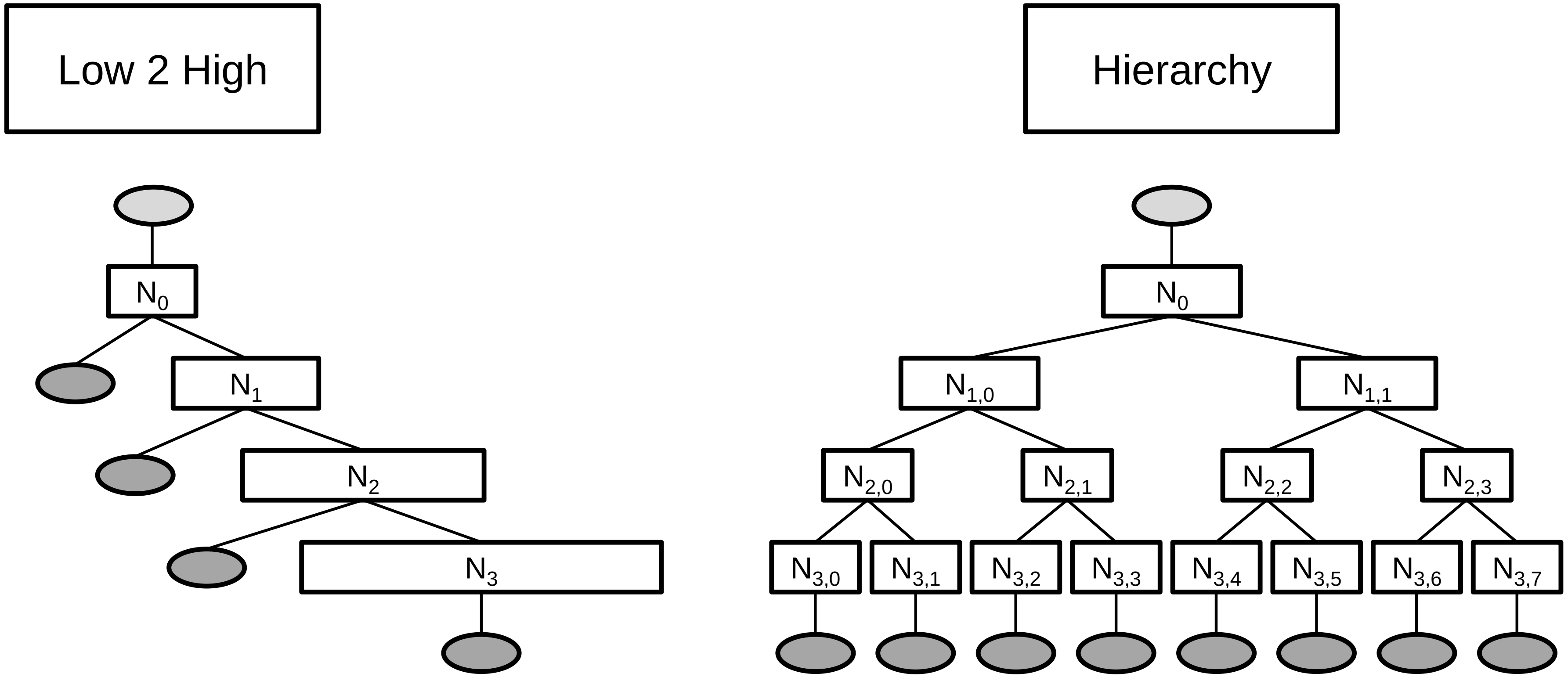
Connectivity Pruning



Quantization & Normalization



Dynamical Selective Execution



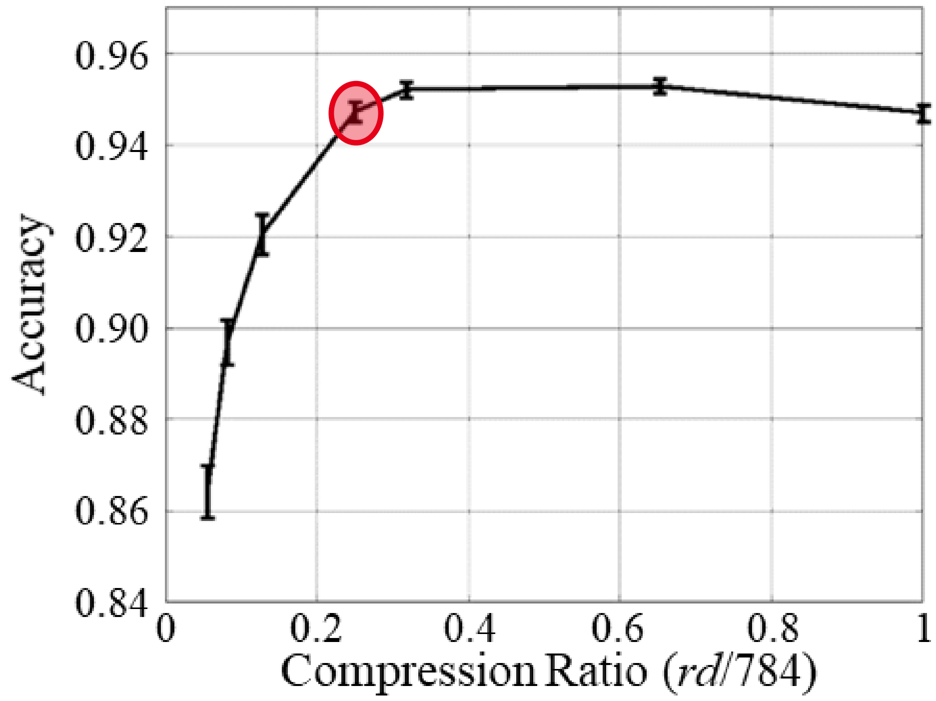
➡ Other topologies depending on event occurrence duty cycles and target applications

Numerical experiments (w/ a very basic topology)

MNIST database classification problem:

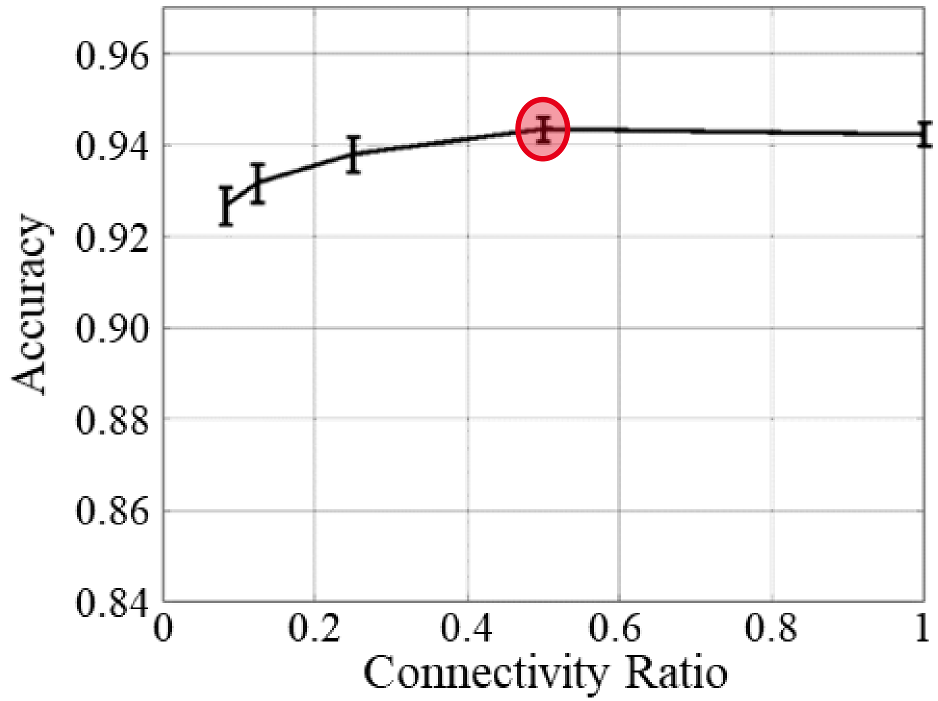
Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	rd	CS
Connected layer	rd	rd	CP, QN
ReLu	rd	rd	QN
Connected Layer	rd	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (DR)



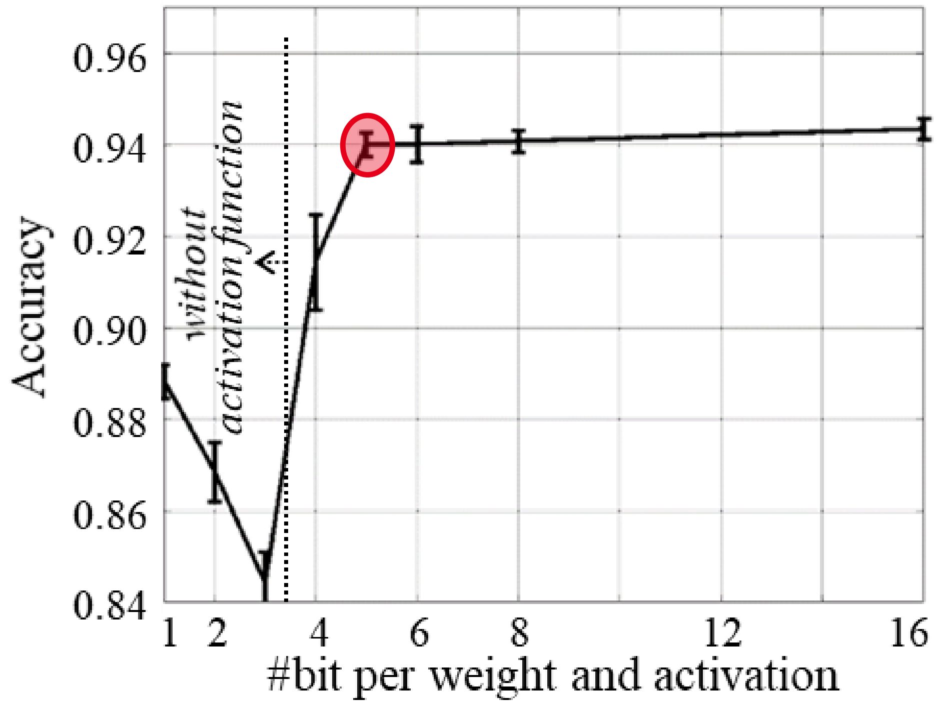
Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	rd	CS
Connected layer	rd	rd	CP, QN
ReLu	rd	rd	QN
Connected Layer	rd	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (CP)



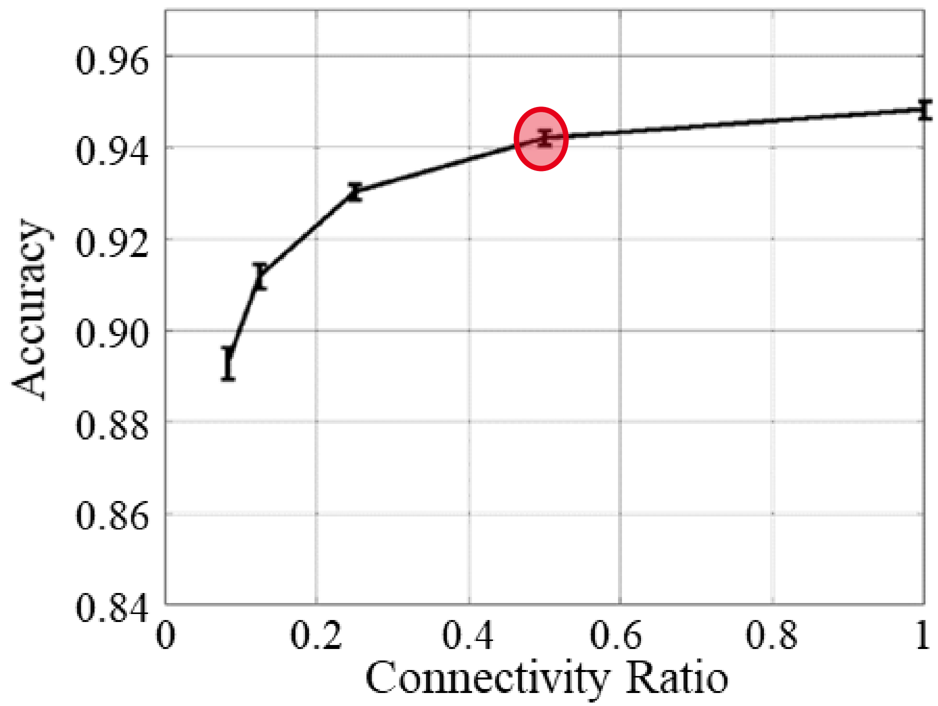
Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	784	CS
Connected layer	784	784	CP, QN
ReLu	784	784	QN
Connected Layer	784	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (QN)



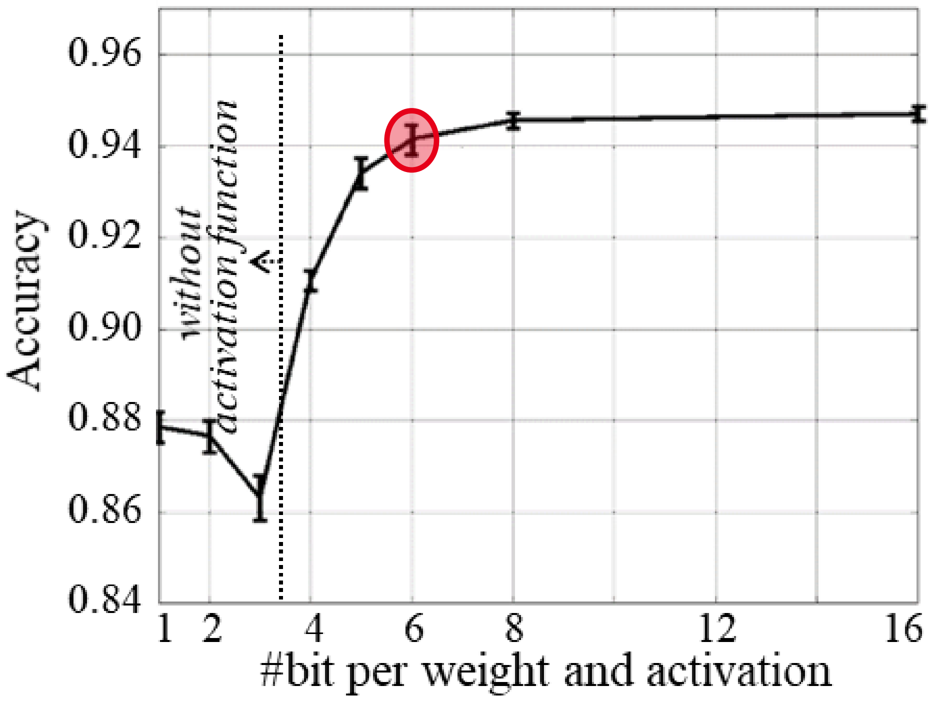
Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	784	CS
Connected layer	784	784	CP, QN
ReLu	784	784	QN
Connected Layer	784	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (CS+CP)



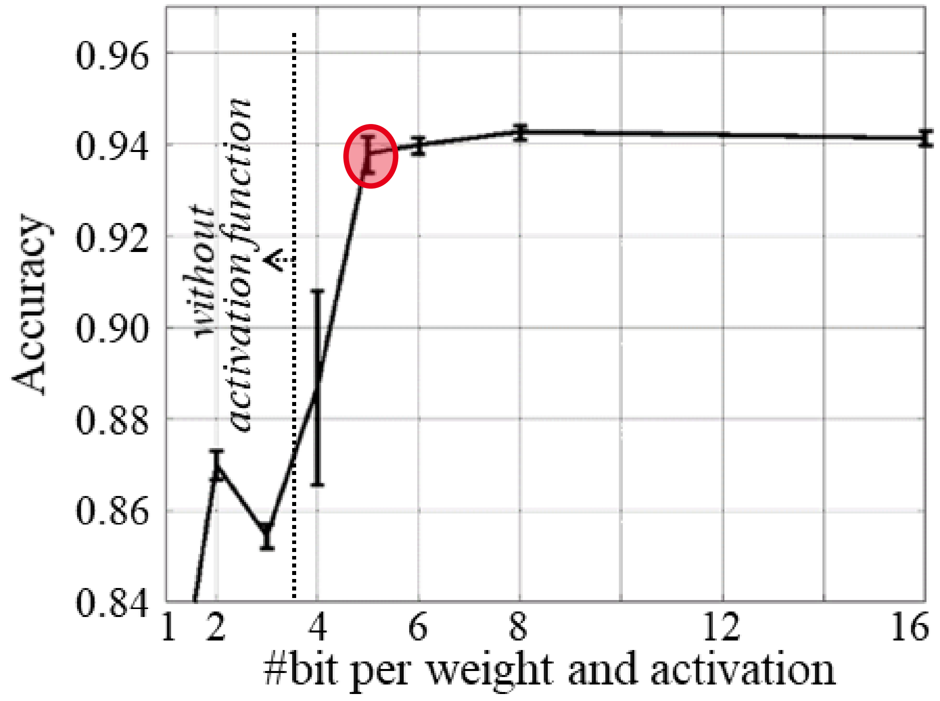
Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	196	CS
Connected layer	196	196	CP, QN
ReLu	196	196	QN
Connected Layer	196	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (CS+QN)



Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	196	CS
Connected layer	196	196	CP, QN
ReLu	196	196	QN
Connected Layer	196	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Numerical experiments (CS+CP+QN)



Layer type	input	output	enabler
Flatenned sample	28×28	784	-
Dimensionality Reduction	784	196	CS
Connected layer	196	196	CP, QN
ReLu	196	196	QN
Connected Layer	196	32	CP, QN
SoftSign	32	32	-
Connected Layer	32	10	-
Softmax	10	10	-

Conclusion & perspectives

Generic categorization of algorithmic enablers for Neural Network model compression

→ Algorithmic enablers to lower HW constraints
→ Data-agnostic DR and CP – dedicated HW
→ Quantized Network training algorithms pave the way to mixed-quantization
→ Complex Finite State Machine for Dynamical Selective Execution

Basic numerical experiments

→ Multiple enablers can be advantageously combined

➡ Neural Architecture Search (NAS) under HW constraints

HW design levers

→ In-line processing vs. single-core iterative processing
→ Digital design: data-centric vs. memory-centric vs. hybrid approaches
→ Technological opportunities: Near or in-memory computing

Thanks for your attention!

Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 rue des Martyrs | 38054 Grenoble Cedex | France
www.leti.fr