The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

ANDANTE

AI
ECSEL JU

AI4DI

TEMPO

AI

Vienna, Austria

25-26 July 2022

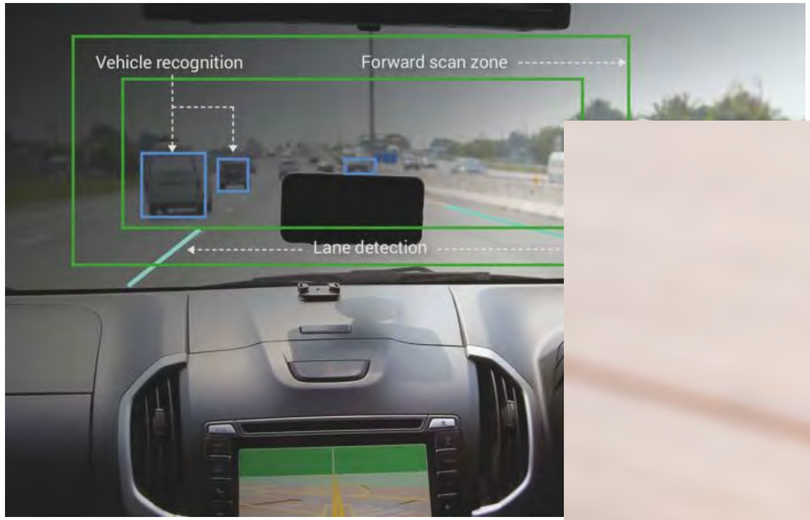
The International Workshop on Edge Artificial Intelligence for Industrial Applications (EAI4IA)

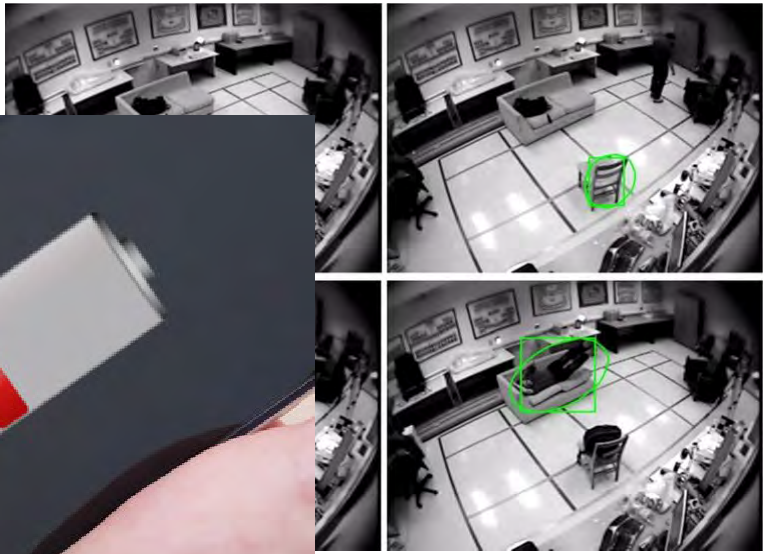
Temporal Delta Layer: Exploiting Temporal Sparsity in Deep Neural Networks for Time-Series Data

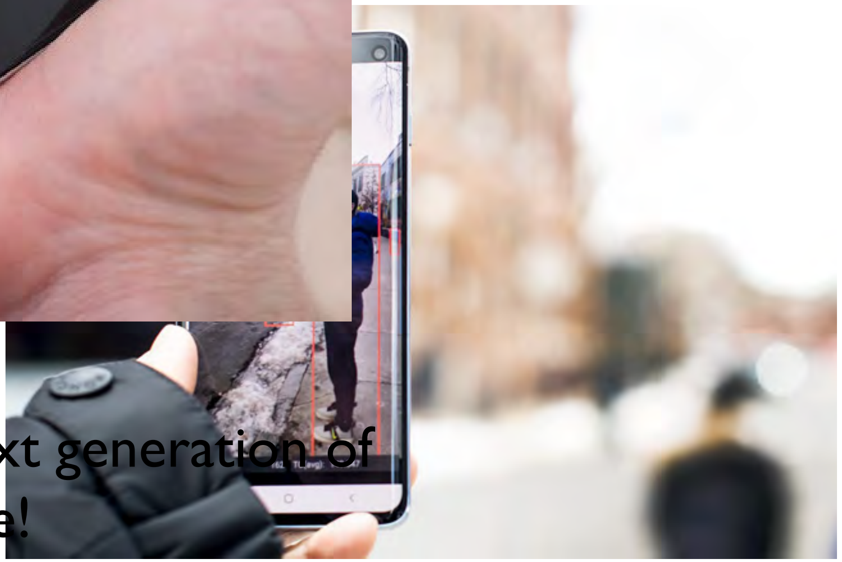
Preetha Vijayan
TU Delft / IMEC
Netherlands

Vienna, Austria 25-26 July 2022

Introduction







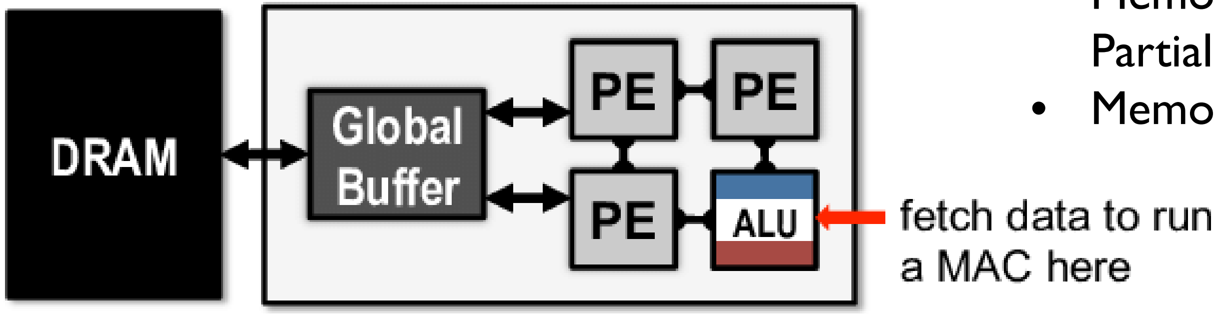
Energy Efficient DNNs are the next generation of Artificial Intelligence!

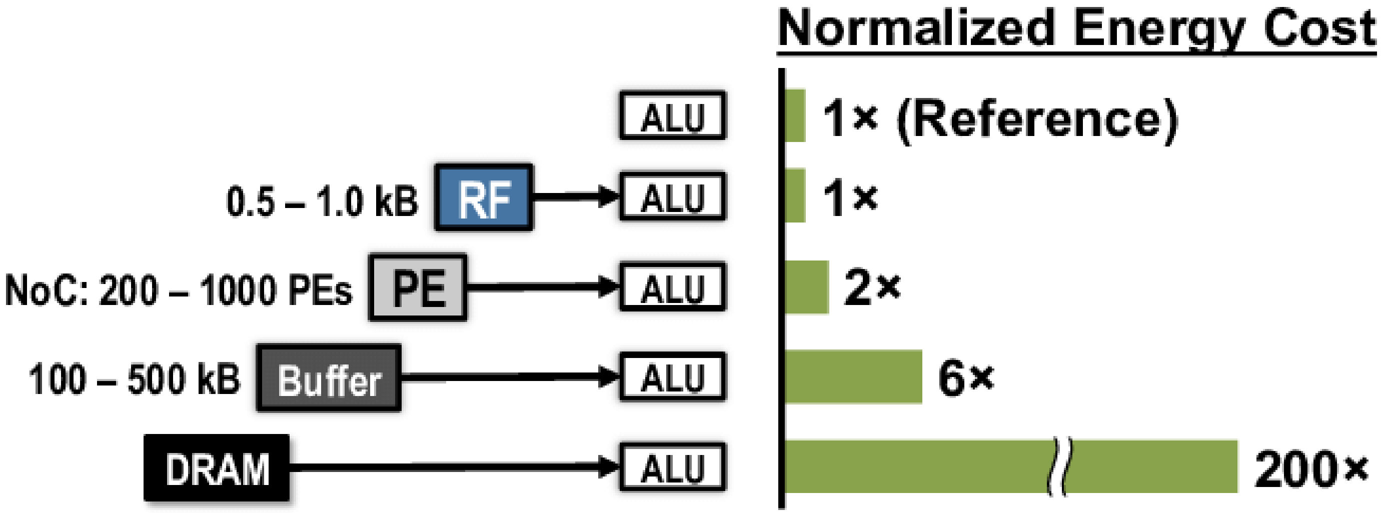
Problem Description

Each computation requires,

- Memory Read – Weights, Activations and Partial Sums
- Memory Write – Updated Partial Sums

Basic memory hierarchy





Memory access is the bottleneck.

Image source - [Eyeriss]

Brain-Inspired Solution

- Human brain achieves impressive accuracy and speed with very little power consumption.
- Brain uses spatio-temporal redundancy or sparsity available in the natural input to accomplish this.
- Brain relies on change based processing rather than frame-based processing.

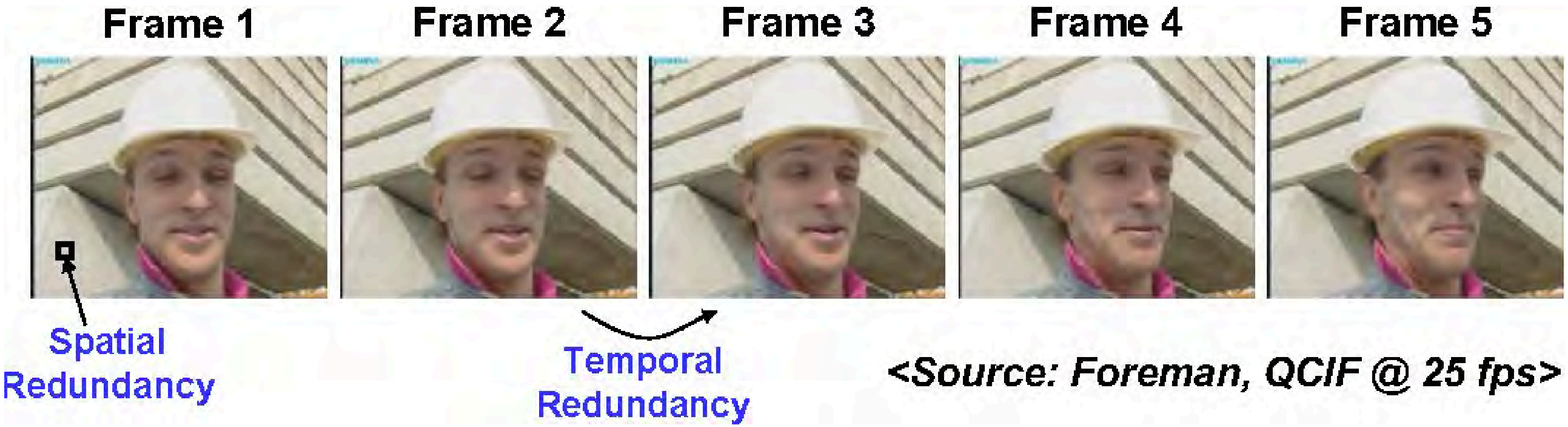


Image source: [Parallelizing-H.264-Motion-Estimation-Algorithm]

Exploiting Sparsity for Energy Efficient DNN

- DNN inference – dominated by multiplication between weight matrix and activation vector.

- Sparse data can be compressed.

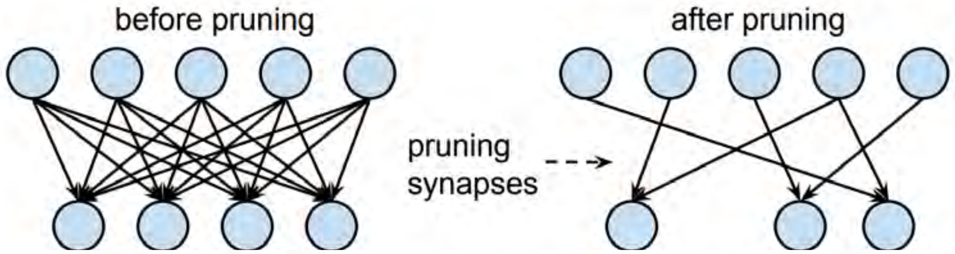
Can save space and energy by avoiding manipulation of zero values

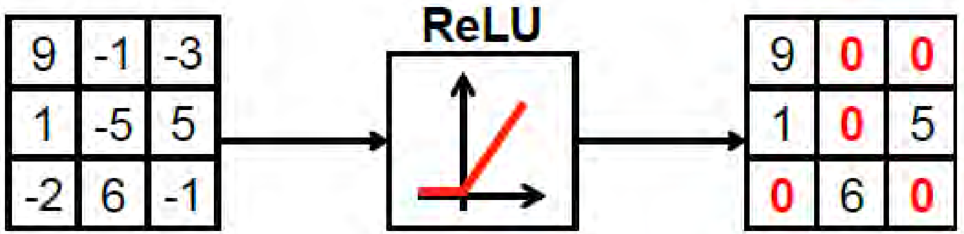
- Y x 0 = 0
- Y + 0 = Y

Can save time and energy by avoiding fetching unnecessary operands and avoiding computations

Sparsity in DNN

- Structural sparsity
  - Lottery-ticket hypothesis
- Spatial sparsity
  - Most pixels in a frame have no relevant feature
  - Results in zero-valued activations
- Temporal sparsity
  - Little change going from frame to frame
  - Wasteful to re-process the whole frame





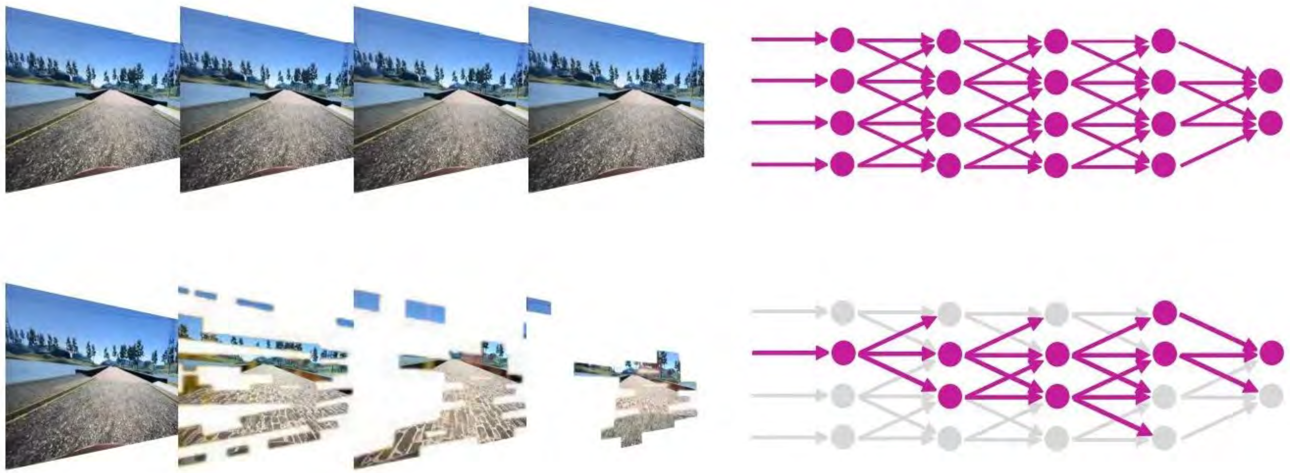


Image source : Pruning, ReLU and GrAI matter

Temporal Sparsity – Related Work

"CBInfer: Change-Based Inference for Convolutional Neural Networks on Video Data"

- Change-based inference of CNNs for video exploiting the spatio-temporal sparsity of pixel changes.
- Only change maps are propagated forward instead of entire frames.
- Trained and tested on static camera inputs.
- ! Change detection is based on thresholds, which are fixed offline.

Research Goals

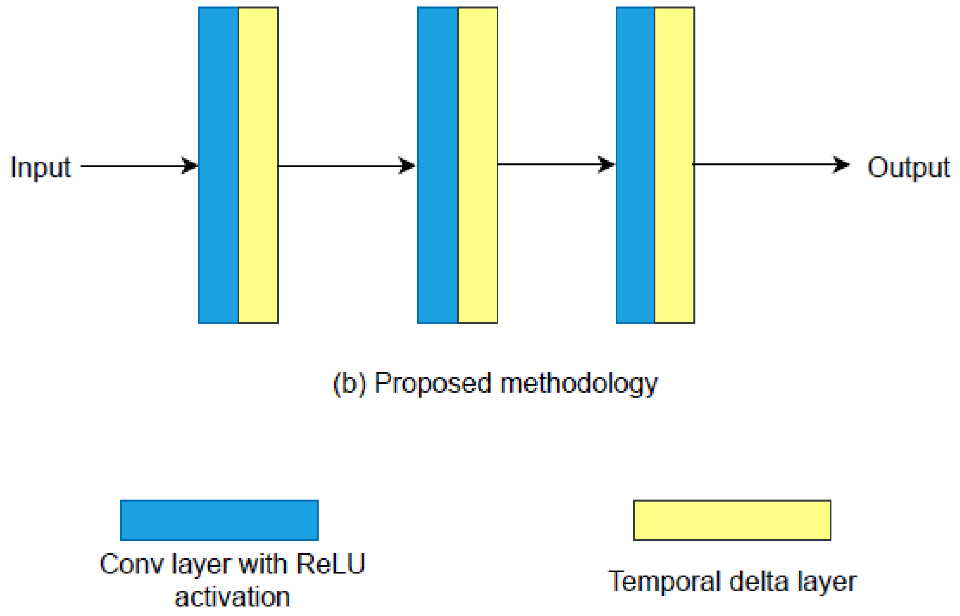
- To induce sufficiently high temporal activation sparsity without suffering too much accuracy loss.

- To make the method flexible enough to be integrated with existing architectures.

- To study the potential of spatial sparsification methods in facilitating the induction of temporal sparsity.

Proposed Approach : Temporal Delta Layer

- The layer consists of 3 main components,
  - Delta Inference
  - Activation Quantization
  - Sparsity Penalty



(b) Proposed methodology

Delta Inference

Standard DNN layer

$$Y_t = W X_t + B$$

$$Z_t = \sigma(Y_t)$$

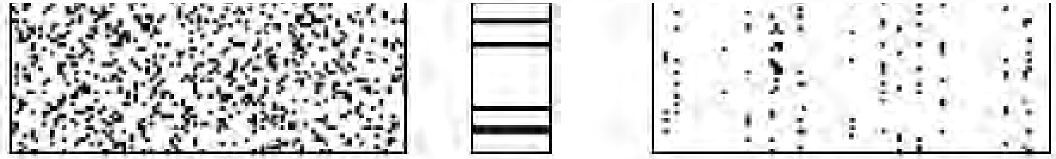
W – Weights
X_t – Input activation at time *t*
B – Bias
Y_t – Transitional state at time *t*
σ - Non-linear activation
Z_t – Output activation at time *t*

Proposed layer

$$\Delta Y_t = W \Delta X_t = W(X_t - X_{t-1})$$

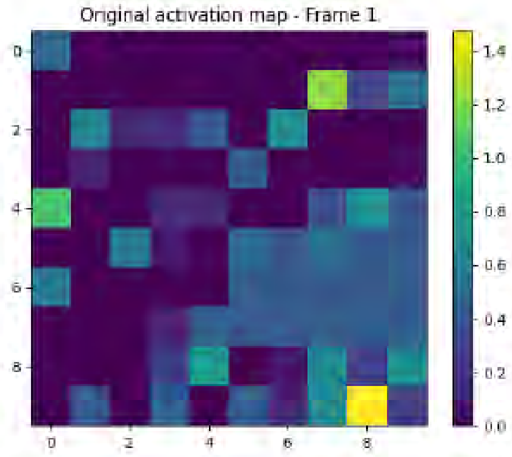
$$\begin{aligned} Y_t &= \Delta Y_t + Y_{t-1} \\ &= W(X_t - X_{t-1}) + W(X_{t-1} - X_{t-2}) + \ldots + Y_0, \quad where\ Y_0 = B \\ &= W X_t + B, \end{aligned}$$

$$\Delta Z_t = Z_t - Z_{t-1} = \sigma(Y_t) - \sigma(Y_{t-1}), \quad where\ \sigma(Y_0) = 0$$

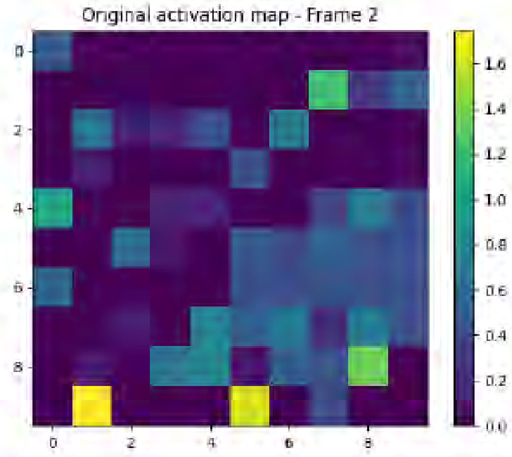


- As input is temporally redundant, ΔX_t is temporally sparse, and by association, so is ΔZ_t
- Temporal sparsity between feature maps is cast onto the spatial sparsity of delta map which is propagated forward.

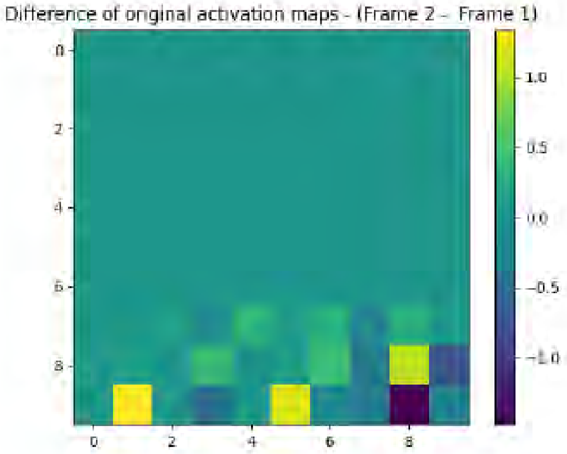
Activation Quantization – Why?



(a)



(b)

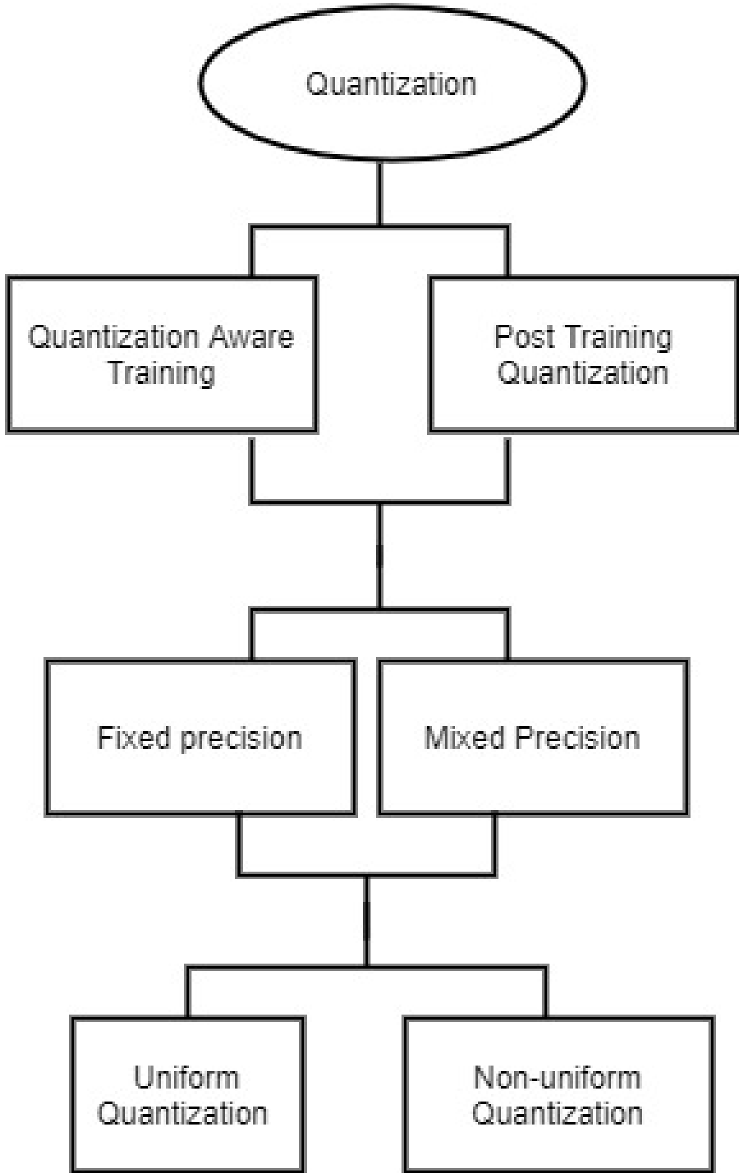


(c)

- A lot of near zero values in the delta map!
- Solution: Reduce precision!

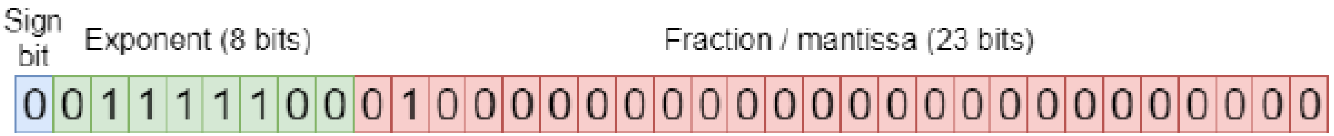
Activation Quantization

- Two methods are considered:
  - Fixed point quantization
  - Learnable step size quantization



Fixed Point Quantization

- Floating point – IEEE 754



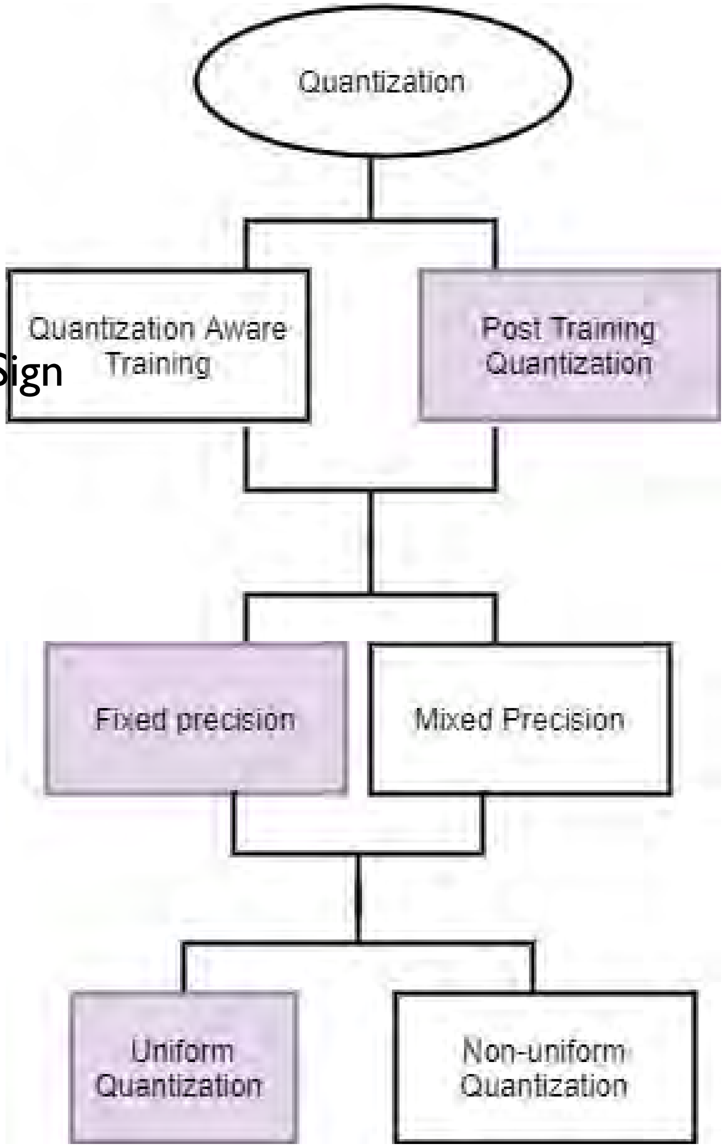
- For fixed point representation, Given BW – Bitwidth, x – Input Activation and S – Sign bit = 1, Integer bits is,

$$I = 1 + \lfloor log_2 \left(\max_{1<i<N} |x| \right) \rfloor$$

- Fractional bits = BW – Sign bit – Integer bits
- Uniform Quantization:

$$Q(\mathbf{x})_{(BW,F)} = \frac{C(R(\mathbf{x} \cdot 2^F), -t, t)}{2^F}$$

- R (.) – Round function, (-t, t) is the range for the given bitwidth and C(.) – Clipping function



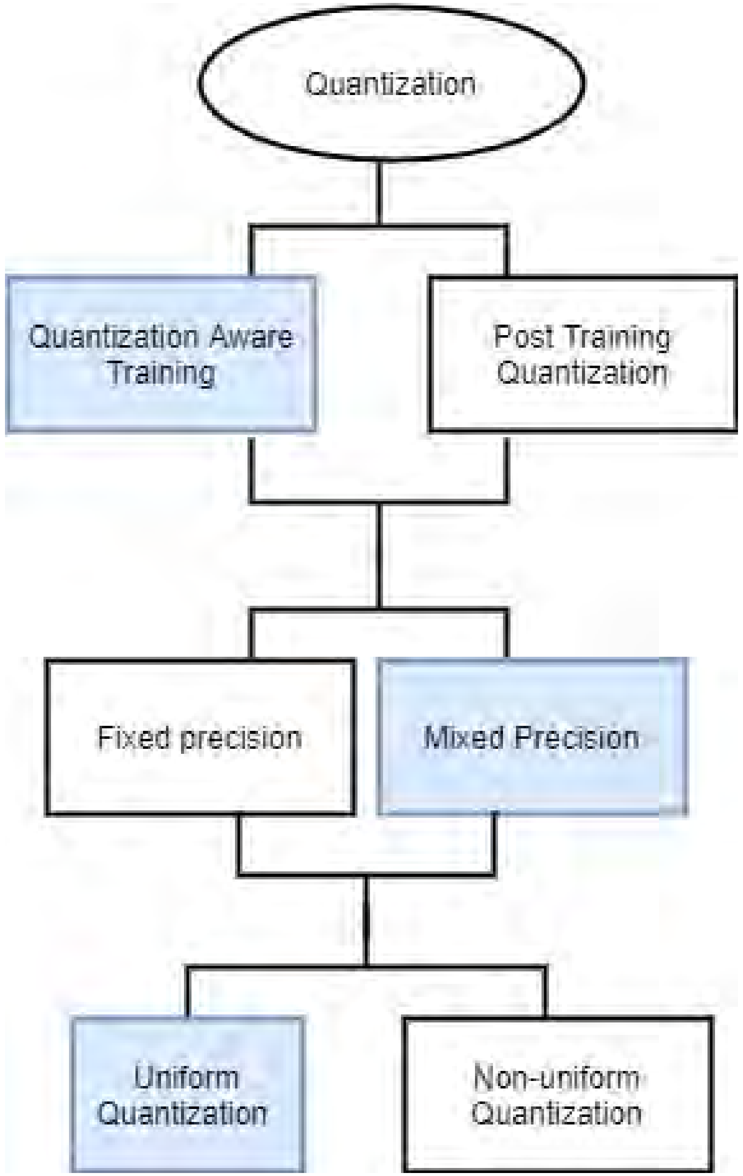
Learnable Step Size Quantization

x – input activation to be quantized
s – step size, that is learnable

- Smaller s, more number of quantization levels → Larger bitwidth
- Larger s, less number of quantization levels → Smaller bitwidth

Forward $q(x;s) = \lfloor \frac{x}{s} \rceil . s$

Backward $\nabla_s q(x;s) = \lfloor \frac{x}{s} \rceil - \frac{x}{s}$



Sparsity Penalty

- As the number of trials or learning increases, the number of neurons required for inference decreases.

- *Optimizing the new layer to decrease the activation density as a part of the overall objective.*

Minimizes the prediction error

Minimizes the activation density within delta map

Proposed Algorithms

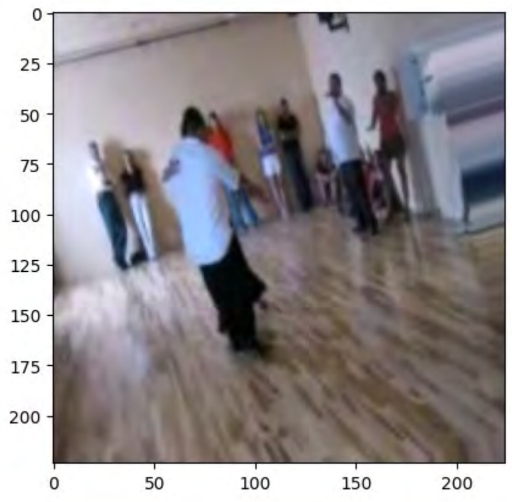
1. Temporal delta layer + sparsity penalty + fixed point quantization

2. Temporal delta layer + sparsity penalty + learnable step size quantization

Experimental Setup

- Application : Human action recognition
- Dataset used : UCF101
- Model architecture : 2 stream network
  - Spatial stream – RGB frames
  - Temporal stream – Absolute difference of RGB frames
- Both streams uses ResNet50

Baseline – Two Stream Network



Spatial Stream

Single RGB frame
(224, 224, 3)
Val accuracy – 75%

Temporal Stream

Stack of Absolute Diff of RGB frames
(224, 224, 10x3)
Val accuracy – 70%

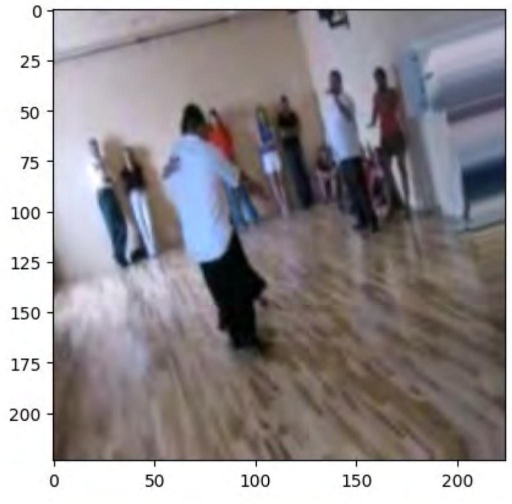
Average Fusion
Val accuracy – 82%

Accuracy v/s Activation Sparsity

Model setup (Spatial stream)	Accuracy	Activation sparsity
Baseline	75%	48%
Temporal delta layer with fixed point quantization	73%	74%
Temporal delta layer with learned step-size quantization	**69%**	**86%**

Model setup (Temporal stream)	Accuracy	Activation sparsity
Baseline	70%	47%
Temporal delta layer with fixed point quantization	68%	67%
Temporal delta layer with learned step-size quantization	**65%**	**89%**

Two Stream Network with Temporal Delta Layer



Spatial Stream

Single RGB frame
(224, 224, 3)
Val accuracy – 69%
Activation sparsity – 86%

Temporal Stream

Stack of Absolute Diff of RGB frames
(224, 224, 30)
Val accuracy – 65%
Activation sparsity – 89%

Average Fusion
Val accuracy – 77%
Activation sparsity– 88%

Conclusion

- The proposed method (temporal delta layer with LSQ) resulted in 88% activation sparsity with an accuracy drop of 5% on UCF-101 dataset for human action recognition.
- The proposed layer can be deployed after any activation layer, and its incorporation does not require any adjustment to the preceding or following layer.
- As the quantization step-size is learnable in LSQ, similar to weights, the initialization of step-size is important and is found heuristically in this work which can be an “annoyance".
- The drawback of using temporal delta layer derives from its requirement to keep track of the previous activations to perform delta operations, so there is memory overhead.

Event Organisers

The Key Digital Technologies Joint Undertaking - the Public-Private Partnership for research, development and innovation – funds projects for assuring world-class expertise in these key enabling technologies, essential for Europe's competitive leadership in the era of the digital economy. KDT JU is the successor to the ECSEL JU programme. www.kdt-ju.europa.eu

The AI4DI project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826060. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the national authorities. www.ai4di.eu

The TEMPO 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, The Netherlands, Switzerland. www.tempo-ecsel.eu

The ANDANTE project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, The Netherlands, Portugal, Spain, Switzerland. www.andante-ai.eu

Thank You

For your attention

@ preethai35@gmail.com