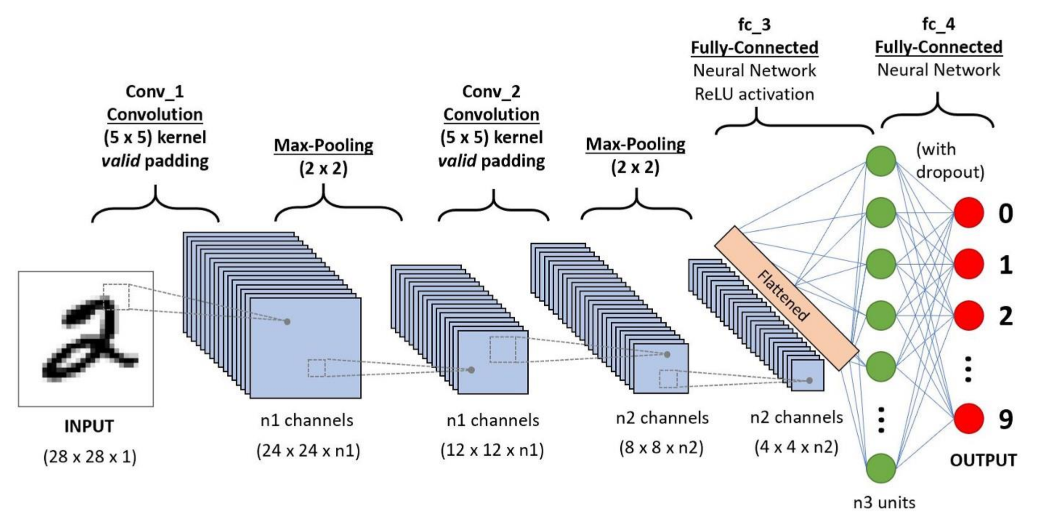
Mixed Intra Layer CNN Quantization for CIM Architectures

A. Vardar, S. Hu, S. Jain, S. Mojumder, N. Laleni, A. Shrivastava, S. De, T. Kämpfe
Fraunhofer IPMS, Center Nanoelectronic Technologies, An d. Bartlake 5 01109 Dresden/Germany
alptekin.vardar@ipms.fraunhofer.de / thomas.kaempfe@ipms.fraunhofer.de

Convolutional Neural Networks (CNN)

- In computer vision tasks, such as image classification, object detection or segmentation, CNNs have achieved state-of-the-art performance due to their shift-invariant ability to capture representative patterns.
- The convolution operation is the result of sliding the convolution kernel across the input matrix of the layer to produce a feature map which is the input of the next layer.
- As convolution and pooling takes into account spatial relations between features, convolutional neural networks are ideal for data with a grid-like structure, such as images.

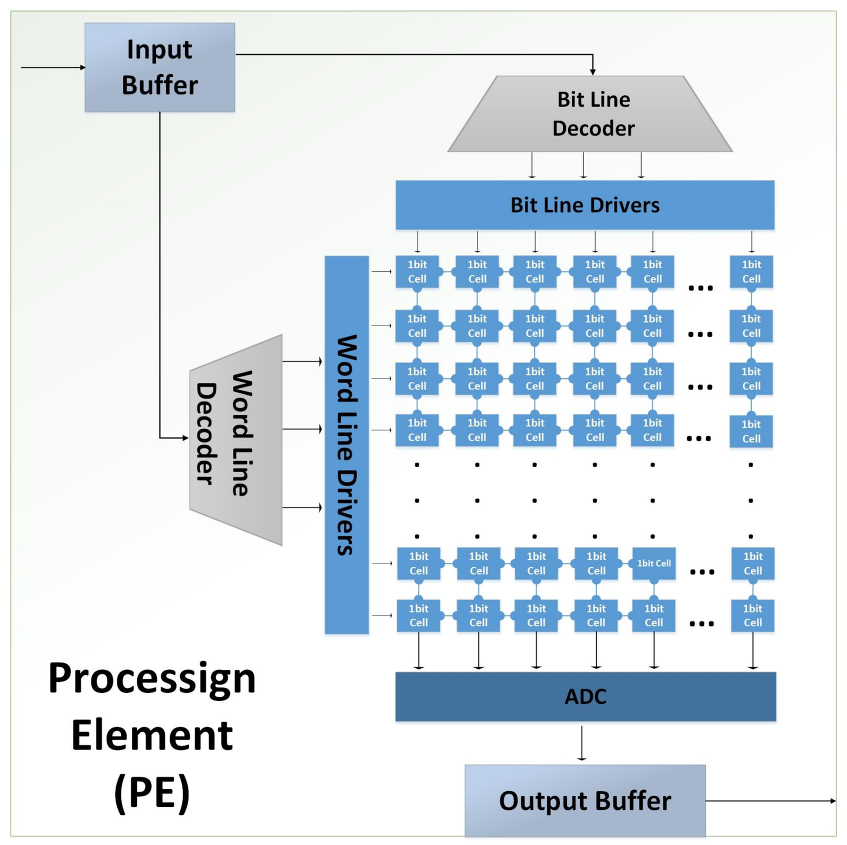
Example CNN for hand written digit recognition (MNIST) [1]



Edge Computing

- Data at the edge is produced continuously. This makes it more efficiency to process there, which can integrate the decision making by NNs with the sensors and actuators for fast automated responses [2].
- Operations like dot-products can be accelerated using special purposed memory arrays, called Processing Elements(PE). This helps to achieve high parallelism while decreasing data traffic.

Crossbar structure of an IMC Processing Element



Acknowledgement

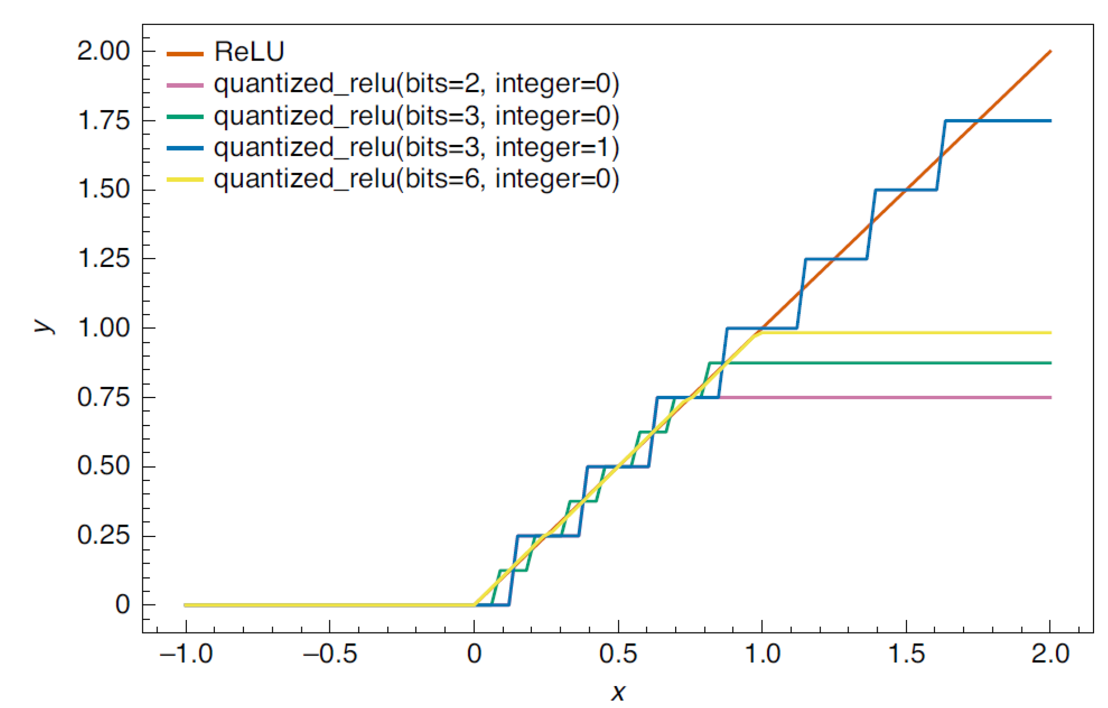
This work received funding within the ECSEL Joint Undertaking project TEMPO in collaboration with the European Union's H2020 Framework Program (H2020/2014-2020) and National Authorities, under grant agreement number 826655.

Quantized Neural Networks (QNN)

- Quantization is a common way to reduce the demand on hardware.
- When the activations are quantized, the number of MAC operations vastly reduces, resulting in with a better latency and energy consumption.
- On the other hand, weight quantization decreases both memory footprint and the number of MAC operations, also helping with area reduction.
- To obtain independent quantization of trainable parameters, QKeras library is used. Mathematically, the mantissa quantization for a give input x is: [3]

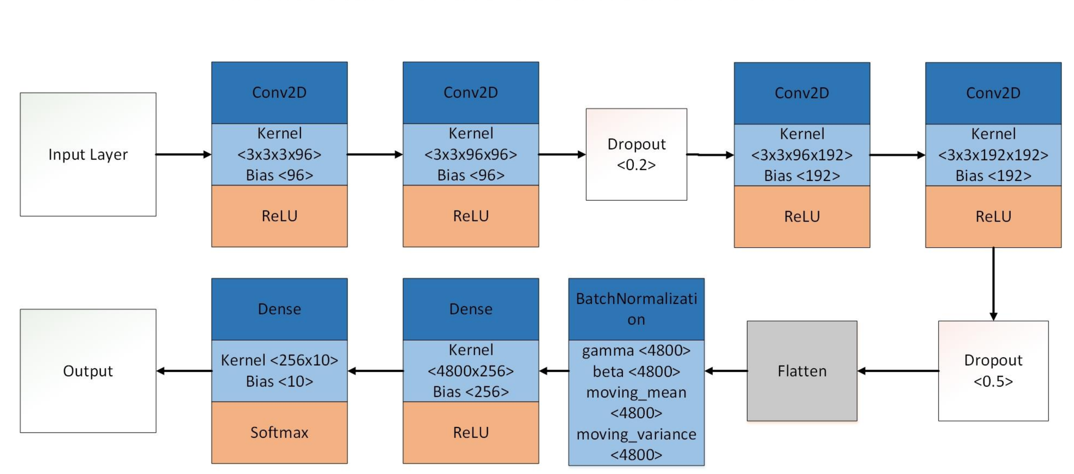
$$2^{int.-bits+1} clip(round(x * 2^{bits-int.-1}), -2^{bits-1}, 2^{bits-1}-1)$$

Quantized ReLU options available in Qkeras



- Previous studies have been done on 8-bit quantization schemes and other fixed lower precision levels. [4]
- Experiments have been conducted using a light-weight network on the CIFAR10 dataset [5].

Architecture of the used neural network



- Adapting an intra-layer mixed quantization training technique for both weights and activations, with respect to layer sensitivities, a memory reduction of 2/8 times and a number of MAC operation reduction of 2/30 times can be achieved compared to their 8bit/FP32 counterparts while sacrificing virtually no accuracy against 8bit and around 2% against the FP32 model.

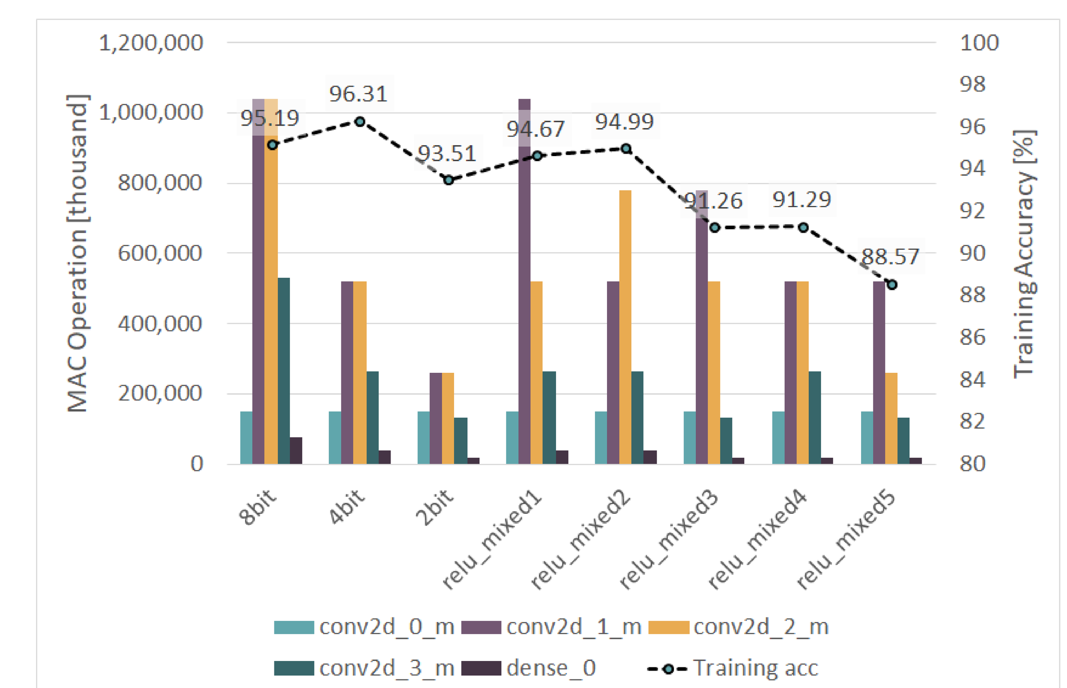
Weight and Activation Quantization

- The experiments showed that the tradeoff between accuracy and quantization differs in each layer. We call this layer sensitivity.
- For example, middle layers are much more robust to quantization whilst also housing majority of the weights

	mixed1	mixed2	mixed3	mixed4	mixed5
Activations	8|4|4|4|2	4|6|4|4|4	6|4|2|2|2	4|4|4|2|2	4|2|2|2|2
Weights	8|6|6|6|4|8	8|6|4|4|4|8	8|4|4|2|2|8	4|4|4|2|2|8	4|2|2|2|2|8

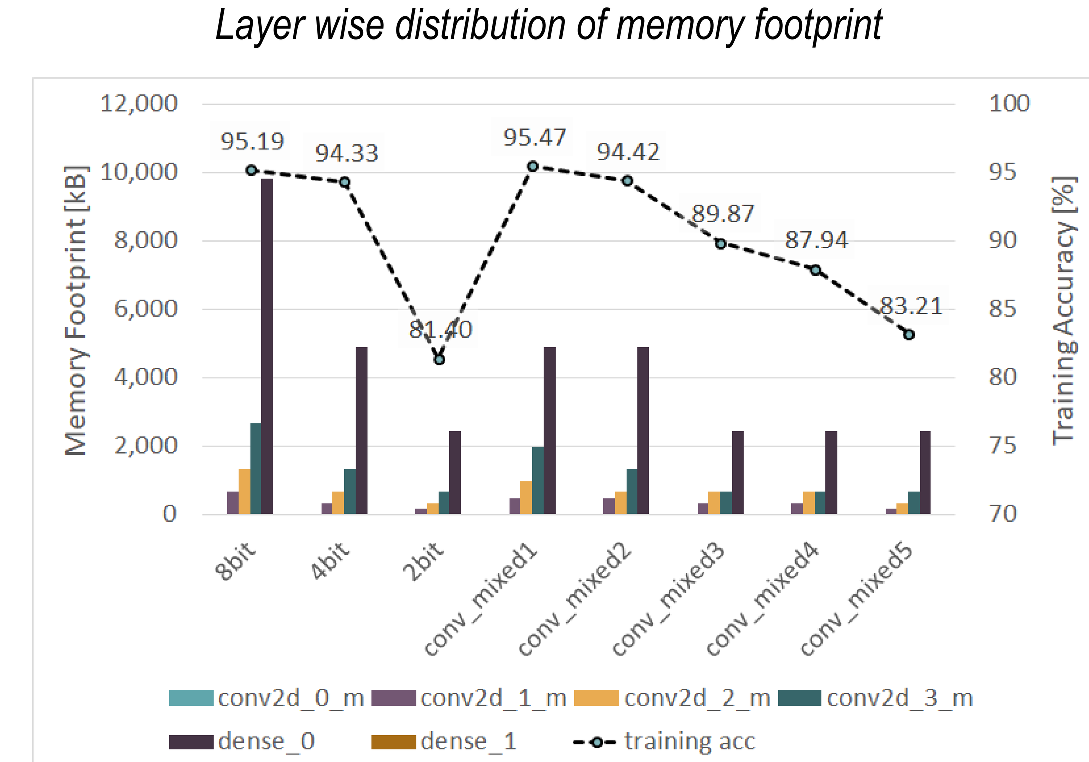
- As shown in Figure, majority of the MAC operations are conducted in the second, third and fourth convolutional layers.
- Using mixed2 scheme, more than 3 times saves is achieved at the MAC operations. Hence, the energy consumption and latency numbers can also be significantly reduced with only sacrificing 0.2% accuracy.

Layer wise distribution of multiply-accumulate (MAC) operations



- In the next Figure , the distribution of weights per layer is shown.
- It is clear that the majority of weights reside in the first dense layer, followed by the fourth convolutional layer.
- These are also some of the least sensitive layers. Meaning, by deeply quantizing these layers, majority of the memory space can be saved.

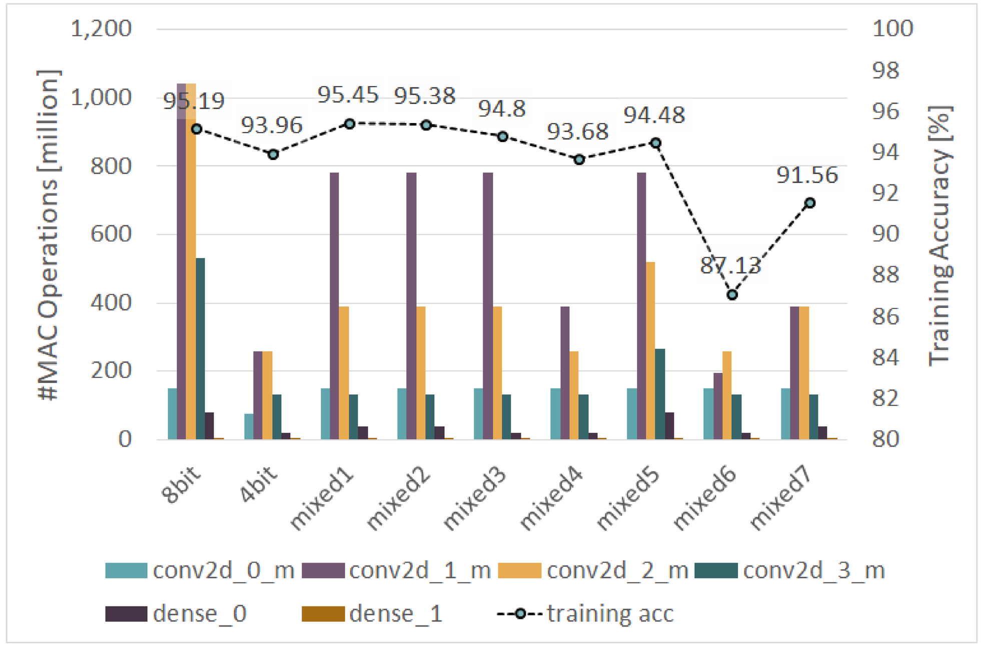
Layer wise distribution of memory footprint



Mixed Quantization

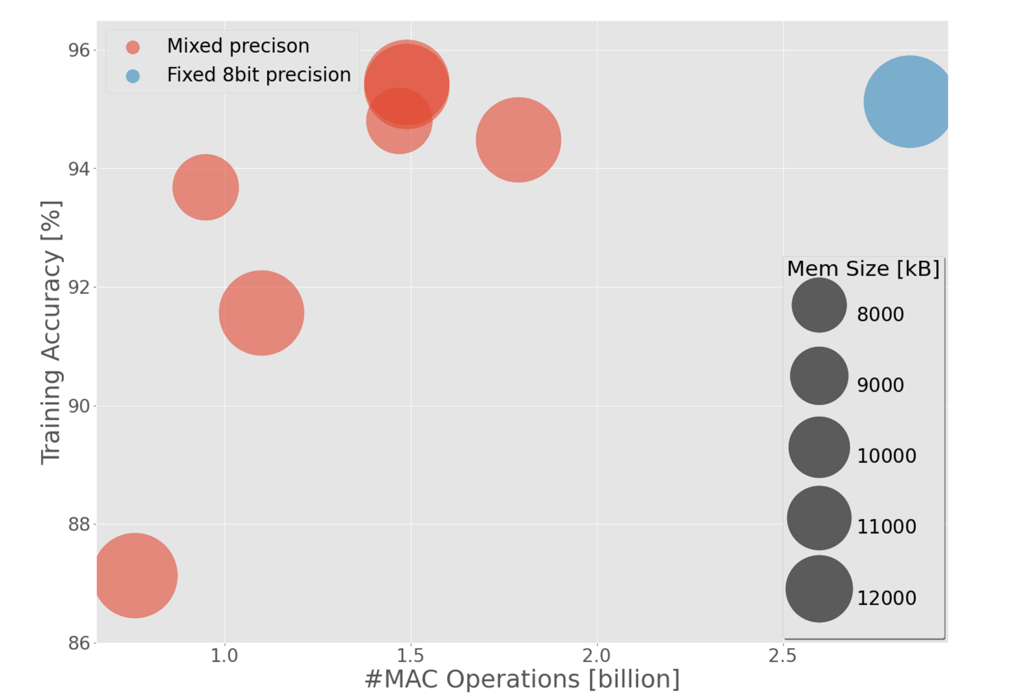
- In order to achieve the best results, a smart mixture of both techniques should be utilized.

Effect of mixed quantization on number of MAC operations



- Depending on the accuracy and resource requirements, ideal configuration can be selected from the given bubble graph.

Bubble graph of accuracy vs memory vs quantization



- For a more detailed analysis of the effect of each quantization combination, the table shows the change in both the training and testing accuracy with respect to saved MAC operations and memory space.

Conf. Name	Weight Quant.	Activation Quant.	Training acc. [%]	Testing acc. [%]	#MAC ops [GOPs]	Memory [Mb]
Full	All 32	All 32	98.12	80.58	45.44	58.09
8bit	All 8	All 8	95.19	78.21	2.84	14.52
4bit	All 4	All 4	93.96	78.38	0.75	7.26
mixed1	8|6|4|4|8|8	8|6|4|4|8	95.45	78.35	1.49	12.36
mixed2	8|6|4|4|8|4	8|6|4|4|8	95.38	77.97	1.49	12.35
mixed3	8|6|4|4|4|8	8|6|4|4|4	94.80	78.36	1.47	7.45
mixed4	8|6|4|4|4|8	All 4	93.68	77.53	0.95	7.45
mixed5	8|6|4|4|8|8	All 8	94.48	77.14	1.79	12.36
mixed6	8|6|4|4|8|8	All 2	87.13	76.14	0.76	12.26
mixed 7	8|6|4|4|8|8	4|6|4|4|4	91.56	77.10	1.10	12.36

References

[1] Saha, S., 2018. A Comprehensive Guide to Convolutional Neural Networks—the ELI5 way. [online] Medium. Available at: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
[2] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, "Edge Computing: Vision and Challenges," in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.
[3] Coelho, Claudionor N., et al. arXiv: Ultra Low-latency, Low-area Inference Accelerators using Heterogeneous Deep Quantization with QKeras and hls4ml. No. arXiv: 2006.10159. 2020.
[4] Z. Li et al., "Laius: An 8-Bit Fixed-Point CNN Hardware Inference Engine," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), 2017, pp. 143-150, doi: 10.1109/ISPA/IUCC.2017.00030.
[5] A. Krizhevsky et al. Cifar-10. 2009.