AI for Performance-Optimized Quantization in Future SAR Systems

Nicola Gollin[a], Michele Martone[a], Ernesto Imbembo[b], Stefan Knoll[c], Gerhard Krieger[a], and Paola Rizzoli[a]
[a]Microwaves and Radar Institute, German Aerospace Center (DLR), 82234 Wessling, Germany
[b]Radio Frequency Payloads and Technology Division, European Space Agency (ESA), 2200 Noordwijk, The Netherlands
[c]Microwave Instruments, Airbus Defence and Space GmbH, 88039 Friedrichshafen, Germany

Abstract

Next-generation SAR systems will be capable of high-resolution wide-swath acquisitions, which will inevitably result in a significant increase of the onboard data volume to be acquired by the system. This, in turn, will lead to severe constraints in terms of onboard memory requirements and downlink capacity. In this context, the onboard quantization of SAR raw data represents an aspect of crucial importance, since it acts as a trade-off between achievable product quality and resulting on-board data volume. In this paper, we investigate the use of artificial intelligence (AI), and in particular of deep learning (DL), for flexible on-board SAR raw data quantization, with the aim of deriving an optimized and adaptive data rate allocation given a desired performance metric and requirements in the resulting focused SAR/InSAR products without relying on a priori information on the acquired scene. The derived bitrate maps (BRMs) can then be used for adapting a BAQ quantizer to the local characteristics of the input data and to the desired output performance. Different performance parameters can be used, such as the Signal-to-Quantization Noise Ratio (SQNR), the InSAR coherence loss or the resulting interferometric phase error, extending the capabilities of the architecture and, ideally, providing multiple bitrate estimations for a single input scene, depending on the specific application requirement. In view of a potential on-board implementation, a possible hardware architecture for the proposed compression scheme is presented as well.

1 Introduction

Future generation SAR systems will bring a huge improvement in performance by means of large bandwidths and digital beam forming (DBF) techniques in combination with multiple acquisition channels. This will overcome the limitations imposed by conventional SAR imaging for the acquisition of wide swaths at fine resolution. The remarkable improvements that can be achieved in terms of performance result in the generation of large volumes of data. This aspect sets stringent requirements for the onboard memory and downlink capacity of the SAR system. For instance, present global SAR mapping missions, such as Sentinel-1, or future ones, such as NISAR and especially ROSE-L and Sentinel-1 Next Generation (NG), will acquire data over selected areas with a temporal sampling down to one week, hence resulting in large data volumes to be stored onboard and downlinked to the ground.

In this scenario, an efficient quantization of SAR raw data is of critical importance, as it defines the amount of onboard memory and it directly affects the quality of the generated SAR products. These two aspects must be carefully considered due to the limited acquisition capacity and onboard resources of the system and, at the same time, to allow for the achievement of the specified product requirements and quality. At present, one of the most widely used methods for SAR raw data digitization is the Block-Adaptive Quantization (BAQ) [1]. In the last years, starting from the principle of BAQ, novel algorithms have been proposed, allowing for an improved performance and resource optimization. In particular, these are acquisition-dependent compression schemes, as for the case of the Flexible Dynamic BAQ (FDBAQ) [2], that may even be combined with the implementation of non-integer data rates [3]. However, the FDBAQ carries out the bitrate optimization based on the SAR raw data statistics only, i.e. it does not take into account the actual performance degradation in the SAR products. In the context of the Performance-Optimized BAQ (PO-BAQ), the basic concept of the BAQ is extended according to the approach proposed in [4], which represents the first attempt for an optimization of the resource allocation depending on the final performance requirement defined for the resulting higher-level SAR/InSAR product. As quantization errors are significantly influenced by the local distribution of the SAR intensity, such an optimization is achieved by exploiting the a priori knowledge of the SAR backscatter statistics of the imaged scene. Given the severe constraints imposed by the downlink capacity, a performance-optimized bitrate allocation contributes to tune the resulting data rate based on the target higher-level application performance. PO-BAQ requires the use of a-priori knowledge of the underlying SAR scene, which has to be uplinked on board, in the form of, e.g., look-up-tables (LUTs)[4] or backscatter maps. For this reason, this technique requires additional complexity and is not fully adaptive with respect to the acquired raw data, since the quantization settings are derived from prior considerations and do not account for the local conditions at the time of the survey.

In general, the quantization performance depends on the local characteristics of the illuminated scene on ground, which are linked to the local topography, radar backscat-

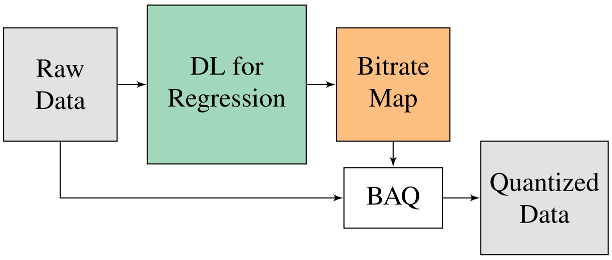


Figure 1 Flow chart of the proposed method: the raw data matrix is fed into the trained DL model which predicts the required two-dimensional bitrate map (BRM), needed to achieve the desired performance. An adaptive quantizer (i.e., BAQ) performs the raw data encoding with the estimated BRM.

ter characteristics (its absolute levels and degree of heterogeneity) and illumination geometry.

In this scenario, Artificial Intelligence (AI) is one of the most promising approaches in the remote sensing community, enabling scalable exploration of big data and bringing new insights on information retrieval solutions [5]. In this contribution we investigate the potential of an AI-based performance-optimized quantization to define a flexible approach for onboard SAR raw data quantization in future SAR missions, where the bitrate is derived depending on a desired target performance in the focused data domain without a priori information on the imaged scene. The description of the proposed method, named AI-BAQ, as well as of the DL architecture and of the used dataset is presented in Section 2. In Section 3 results are shown including the validation on the final SAR product, and a framework for a possible onboard hardware implementation is discussed in Section 4. Finally, conclusions and outlook are provided in Section 5.

2 Deep Learning for SAR Raw Data Quantization

Nowadays, Artificial Intelligence (AI) and, in particular, Deep Learning (DL) represent a very flexible and powerful tool to approach and solve different problems, which go well beyond the remote sensing image processing and interpretation. In our case, we have approached the onboard bitrate estimation for SAR raw data as a deep supervised regression task. In particular, the number of quantization bits to be allocated for a given portion of the raw data is estimated by a DL architecture within a continuous range of possible values, typically between 2 and 6 bits/sample. The principle of our method is depicted in Figure 1. First, the input raw data is fed to the DL architecture which estimates a two-dimensional bitrate map (BRM). A standard BAQ is then considered to compress the raw data by applying the estimated (variable) BRM, relying on azimuth-switched quantization to implement non-integer rates [3].

2.1 DL Architecture Description

The DL architecture that we have defined for performing the considered regression task is presented in Figure 2. It is composed of a sequence of three convolutional layers (with 64, 128 and 256 3×3 kernels, respectively) with rectified linear unit (ReLU) activation function, interleaved by max pooling layers which halves the dimensions of the input features at each layer. Afterwards the feature maps are "flattened" and given as input to a fully-connected dense layer with 128 units, followed by a final linear regression layer which returns an M vector of bitrate values (where M represents the number of optimization parameters considered during the training process). This means that, at inference, one single BAQ rate will be estimated and applied to blocks of 128×128 pixels within the input raw data. As loss function we utilized the mean squared error (MSE) between the network output and the reference bitrate map, estimated from the corresponding focused SAR data, as presented in Section 2.2. The architecture's hyperparameters (number of layers, number of kernels, size of the dense layer and size of the input patches) have been selected through empirical hyperparameter tuning, as a trade-off between achievable performance and onboard computational complexity, in a direct synergy with the hardware feasibility assessment presented later in Section 4. As an example, a raw data patch of size 128×128 samples (in range and azimuth dimensions, respectively) implies the storage in the onboard memory of 128 azimuth lines, which is still a manageable size with currently hardware components for spaceborne SAR. On the other hand, 128 range samples represent the standard range block size for the application of the BAQ quantizer in current spaceborne SAR missions. Additionally, the number and size of the convolutional kernels and of the dense layers directly impact the required onboard processing load.

2.2 Dataset Generation and Training Phase

For the generation of a descriptive and consistent dataset to train, validate and test the proposed architecture, we have exploited TanDEM-X data acquired in bypass configuration, i.e., raw data are quantized with a uniform 8-bit Analog-to-Digital Converter (ADC). The acquisitions are covering a variety of land cover types including desert, ice, forest, urban areas and variable topography. For the generation of reference bitrate maps to be used during the supervised training, we have re-quantized the acquisitions on ground using different BAQ rates (i.e., 2, 3, 4 and 6 bits/sample), and then performed the complete SAR processing, allowing for the derivation of SAR and InSAR products for each quantization rate. In order to achieve more granularity in the reference data, even if only integer (BAQ) bitrate values are available, we have performed an interpolation on the obtained performance, such that we were able to define a fractional bitrate which satisfies the requirement, as it is presented in [4]. Afterwards, a binary mask is derived for each re-quantized raw data, by setting a threshold on the specific target performance parameter. An overall reference bitrate map is then derived by selecting the minimum number of bits which ensures a certain

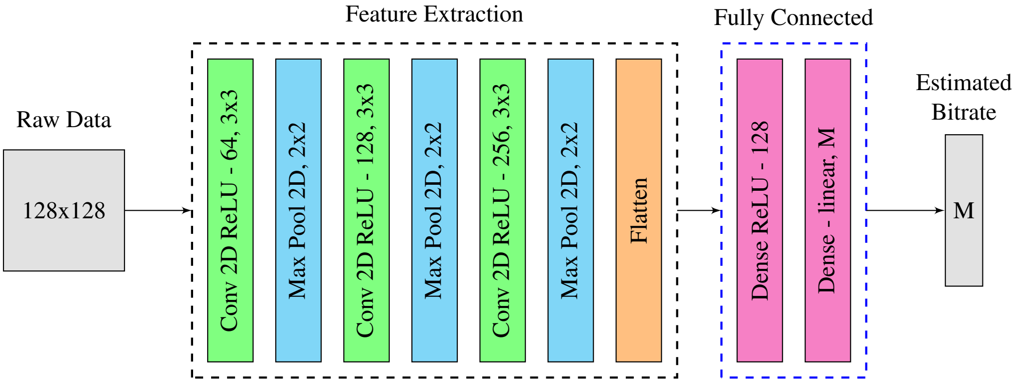


Figure 2 Block scheme of the proposed DL architecture. The initial feature extraction blocks consist in a sequence of two-dimensional convolutions with ReLU activation funaction and max pooling terminated by a flattening operation. The fully connected dense layer of 128 elements with ReLU activation is linked to the output regression element consisting of an M-elements dense layer with linear activation function, where M represents the number of target SAR optimization parameters.

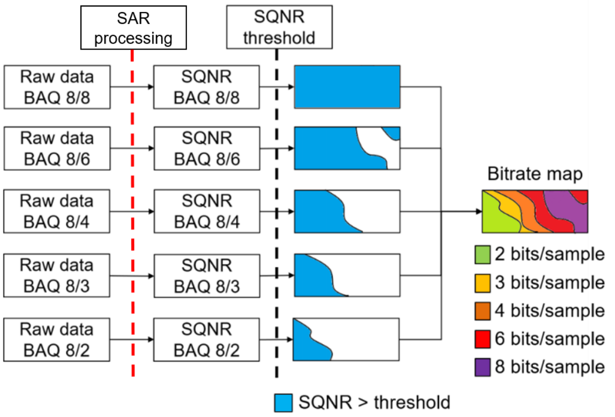


Figure 3 Approach used to derive the reference BRMs for training the DL architecture based on thresholding on a given performance requirement. In this case, the SQNR is selected as performance parameter, but the same method can be applied to other metrics as well (e.g., phase error, coherence loss).

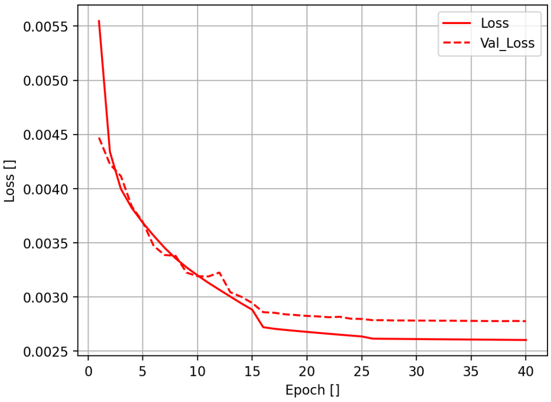


Figure 4 Training curve of the proposed DL architecture over 40 epochs. The considered loss function is the MSE.

performance within the focused SAR data. Since the quantization errors in SAR images are integrated within a large area on ground, given by the chirp length in range and by the synthetic aperture in azimuth, the derived BRM shows a smooth spatial variability (in the order of several hundreds of meters) [4]. An example is depicted in Figure 3 for the exemplary case of the signal-to-quantization noise ratio (SQNR) as target performance metric, which is defined as

$$\mathrm{SQNR} = \frac{\sigma_s^2}{\sigma_q^2}, \quad \text{with} \quad q = s - s_q. \tag{1}$$

In the above equation s and s_q represent the reference (non-quantized) signal and the quantized one, respectively.

During the training phase, the input to our DL architecture consists of 128×128 samples patches of uncompressed raw data amplitude. In order to link this information to the corresponding reference bitrate value, the derived reference BRM is averaged within a window of the same size of the corresponding raw data patch (128×128 samples), centered around the patch center sample. In this way, a single reference bitrate value is associated to the entire input raw data patch. The achieved granularity (1 bitrate value per patch) does not cause a loss of information, thanks to the previously mentioned smooth spatial variability of the original reference BRM.

In this contribution we optimize for specific values of SQNR, but it is worth noting that the SQNR is one possible optimization parameter, the same process could also be perform for deriving the required bitrate maps based on other performance metrics.

Overall, we have trained the network using a dataset of almost 11 million data patches, derived from 17 TanDEM-X SAR images, whose 80% (randomly selected) have been considered as training samples, while the remaining 20% have been used as validation samples.

Figure 4 shows the learning curve and verifies the convergence of the training process. As already mentioned, after training, the architecture has been evaluated on a set of 4 TanDEM-X test acquisitions, which were not part of the training/validation dataset.

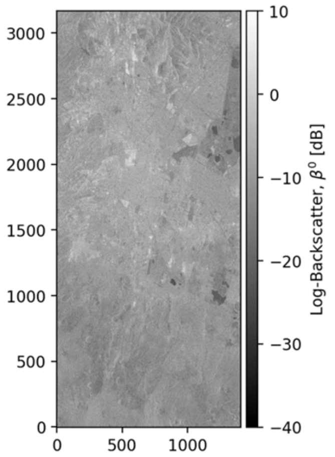


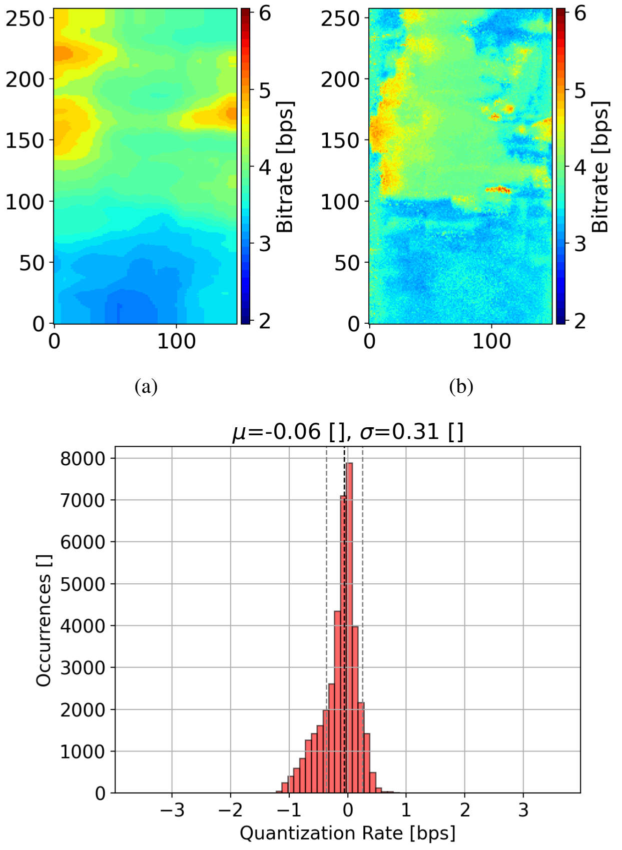
Figure 5 Log-Backscatter of the Mexico City (Mexico) area selected for testing the proposed method.

3 Results

As inference example, we consider a TanDEM-X acquisition over the urban area of Mexico City, whose Log-backscatter is depicted in Figure 5. This represents a highly heterogeneous scene characterized by the presence of urban structures, lakes and high-relief topography. Figure 6 depicts the reference BRM (Figure 6(a)), the estimated BRM (Figure 6(b)) and the histogram of the difference between the reference BRM and the estimated one (Figure 6(c)).

One can note that the difference between the estimate and the reference bitrate map is unbiased and well confined between ±1 bit/sample, being able to follow the local backscatter characteristics of the scene.

In order to properly assess the effectiveness of the proposed method, we need to evaluate the performance on the final quantized SAR product. To do so, we have applied the estimated BRM for variable quantization of the uncompressed raw data, and carried out the complete SAR processing. The results of this analysis, including the performance assessment on the SQNR, are depicted in Figure 7. It is worth noting that the estimation is consistent also in presence of the high degree of heterogeneity of the scene, which represents a worst-case scenario. The resulting SQNR calculated after SAR processing meets the input requirement of 10 dB and 15 dB, respectively. In Table 1 we report the SQNR calculated for all the four considered test areas: Greenland, Uyuni (Bolivia), Las Vegas (USA) and Mexico City (Mexico) and four different performance targets (SQNR=10, 15, 20 and 25 dB, respectively). The State-of-the-Art BAQ is also reported for 2, 3 and 4 bit/sample in Table 1 for comparison. These results highlight the capability of the architecture to meet the desired performance requirement with respect to the considered optimization parameter.



(a)

(b)

(c)

Figure 6 Inference results over the urban area of Mexico City for the target case of SQNR=15dB. (a) Reference bitrate map, (b) estimated (test) bitrate map and (c) distribution (histogram) of the estimation error. It is possible to see that the estimation error has zero mean (unbiased) and relatively narrow distribution with a standard deviation of only about 0.3 bits/sample.

4 Hardware Feasibility Assessment

In this section a possible hardware architecture for the proposed CNN-based data compression method is presented and discussed. For high-performance FPGA implementations, the SAR raw data should be available in Fixed Point number representation at the CNN input. A possible architecture to perform 2D/3D convolution with all feature inputs (channels) of the previous layer is shown in Figure 8.

In particular, the necessary steps to perform a convolution are listed in the following:

- Loading the image from external double-data-rate (DDR) SDRAM to the input buffer,
- Loading the weights from DDR to the input buffer,
- Perform the calculations,
- Storing the results from the output buffer to DDR.

In order to increase the performance, multiple blocks with different kernel weights need to work in interleaved mode (Interleaved Memory Reads / Writes and calculations).

Table 1 SAR Performance (in terms of mean and standard deviation of SQNR) on the final SAR products on the four test acquisitions. The proposed method (AI-BAQ) with four different performance targets (and its resulting average bitrate) and the State-of-the-Art BAQ at 2, 3 and 4 bps are reported below.

Method	Target	Greenland	Uyuni	Las Vegas	Mexico City
AI-BAQ	SQNR=10dB	10.7±0.1@2.2bps	10.2±0.5@2.2bps	9.7±1.3@2.5bps	9.6±0.9@2.7bps
	SQNR=15dB	15.6±0.2@3.2bps	15.3±0.6@3.1bps	14.7±1.3@3.5bps	14.5±0.9@3.7bps
	SQNR=20dB	18.7±0.6@4.2bps	20.5±0.5@4.4bps	20.0±1.3@5.0bps	19.7±1.0@5.1bps
	SQNR=25dB	22.6±1.1@5.1bps	25.0±0.6@5.4bps	23.8±1.3@5.8bps	24.0±1.1@5.8bps
BAQ@2bps	-	9.3±0.2	9.5±0.2	7.7±1.3	6.6±1.4
BAQ@3bps	-	15.1±0.2	15.0±0.4	12.9±1.5	11.6±1.8
BAQ@4bps	-	18.7±0.4	19.8±0.7	17.8±1.6	16.5±1.8

Xilinx Versal DPUs are well suited for CNN AI applications and can be considered for a hardware implementation of the proposed CNN-based data compression method. In the tool chain several AI functions (e.g., 2D/3D convolution, Rectified Linear Unit, Max Pooling, Flattening, Fully Connected Layer) and their architectural interconnections are supported.

The number of Multiply and Accumulate operations (MAC OPs) in a convolution layer can be calculated as follows:

$$\text{MAC OPs}_{\text{Conv}} = K_{\text{h}} \cdot K_{\text{w}} \cdot F_{\text{IN}} \cdot F_{\text{OUT}} \cdot R_{\text{h}} \cdot R_{\text{w}}, \quad (2)$$

where K_{h} and K_{w} are the kernel height and width, F_{IN} and F_{OUT} are the input and output features and R_{h} and R_{w} are the resulting height and width, respectively. The resulting MAC OPs for all the convolutional layers are of about 613.5M for the considered architecture. The number of MAC OPs for a Fully Connected Dense layer are calculated as:

$$\text{MAC OPs}_{\text{Dense}} = F_{\text{IN}} \cdot F_{\text{OUT}}, \quad (3)$$

resulting in 8.4M operations for the considered architecture and leading to a total number of MAC OPs of 621.9 M.

In future SAR missions the expected data rates will be of about 3000 Mbits/s. By assuming 8 bit/sample and a patch size (frame) of 128x128 pixels, this leads to a resulting 131072 bits per frames. From here, it is possible to derive the real-time performance requirement in frames per second (fps_{r}) as follows:

$$\text{fps}_{\text{r}} = \frac{3000M\,\frac{bit}{s}}{131072\,\frac{bit}{frame}} = 22888. \quad (4)$$

For performance estimation the CNN benchmark of Xilinx was used with the VGG16 CNN (Vitis-AI Model Zoo Name: tf_vgg16_imagenet_224_224_30.96G) for comparison. As Hardware platform, we selected the VCK190 considering 1xDPUCVDX8G 192 AIEs (C32B6CU1L2S2) @1250MHz with fixed point calculations. The VGG16 CNN uses the same kernel size (3x3) and downsampling (stride) size (2x2) and shares a very similar basic structure with the proposed architecture.

In Table 2 we report the details by considering two more convolutional layers. The selected VGG16 model has 30.96G OPs, while the proposed CNN has 0.62G OPs. In order to compare the frames per second, the fps_{r} value is multiplied by the factor $30.96/0.62 \approx 50$ for a first qualitative analysis of the performance.

VCK190 1* DPUCVDX8G 192 AIEs (C32B6CU1L2S2) @ 1250MHz		
Parameter	VGG16	Proposed Arch.
E2E fps Single Thread	505.43 fps	25271 fps
E2E fps Multi Thread	621.19 fps	31060 fps

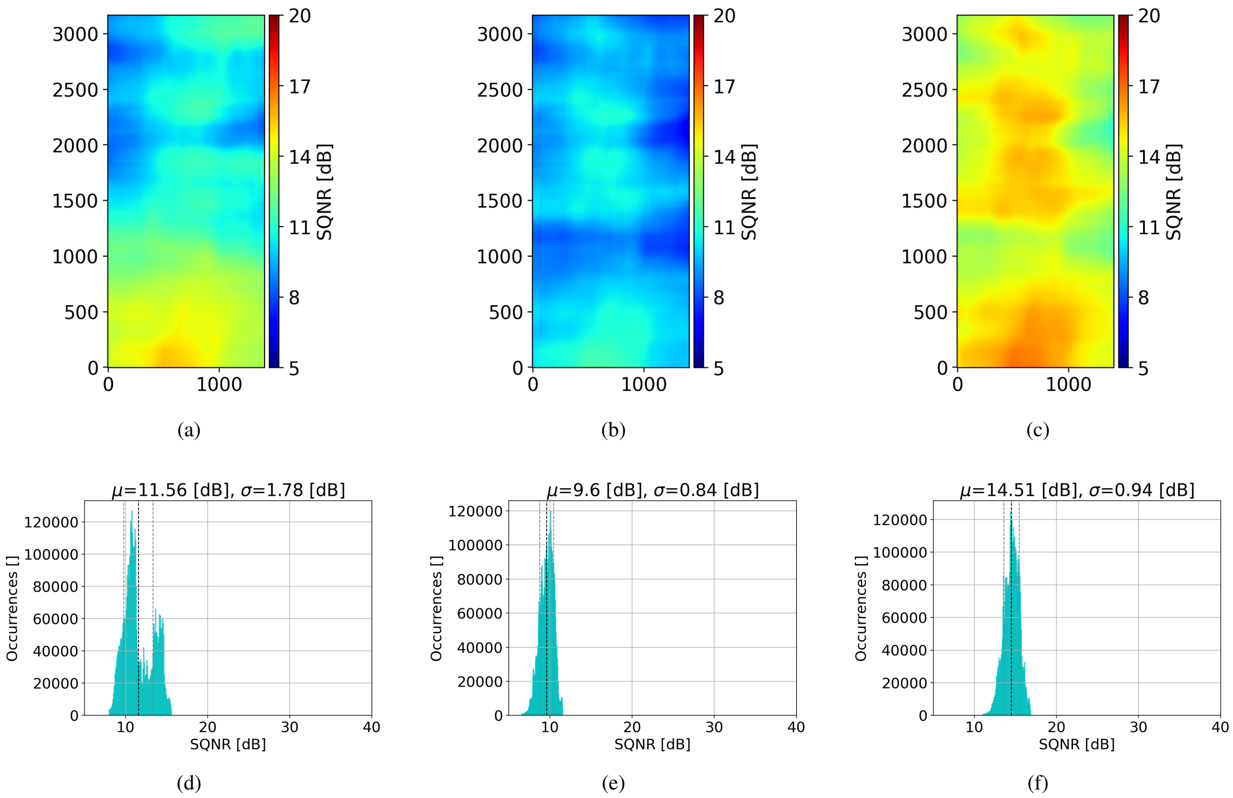
Table 2 Performance comparison VGG16 and CNN proposed by DLR.

By considering the above assumptions, the obtained fps_{r} of 22888 represents a feasible hardware requirement in a next-generation SAR missions.

5 Conclusions and Outlook

In this paper we investigate a novel approach to perform a performance-optimized bitrate allocation for SAR and InSAR systems by means of a Deep Learning-based regression architecture. The main advantage of the proposed method relies in the fact that no a priori information is required by the system for its implementation, hence allowing for different bitrate allocation depending on the considered performance parameter and target requirement.

We have presented relevant aspects and details of the network as well as the definition of the training, validation, and testing datasets and strategies, together with an assessment of the estimation performance on an independent test acquisition. The results are promising and show that an accurate bitrate estimation can be achieved by the proposed architecture, which is then consistently confirmed when the performance parameters are evaluated on the final SAR product. The comparison with the State-of-the-Art BAQ highlights the flexibility of the method to meet the desired performance on different scenes. Finally, we introduced and evaluated a possible hardware architecture for a next on-board implementation. As outlook to this work, a further optimization of the architecture is foreseen in order to further improve the performance, as well as the number of optimization parameters which can be handled by the architecture. The exploitation of a larger dataset would allow for the training of a more robust model and for a global-scale assessment of the data rate for future SAR missions.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7 Quantization performance results in terms of SQNR (dB) after SAR processing over the urban area of Mexico City. The State-of-the-Art BAQ at 3 bps is depicted in (a) and its distribution in (d). The performance metrics obtained with the proposed AI-BAQ method with a target case of SQNR=10 dB and SQNR=15 dB are depicted in (b) and (c), with the corresponding distributions in (e) and (f), respectively. After SAR processing the resulting SQNR map confirms the expected target performance with rather small deviation (about 1 dB).

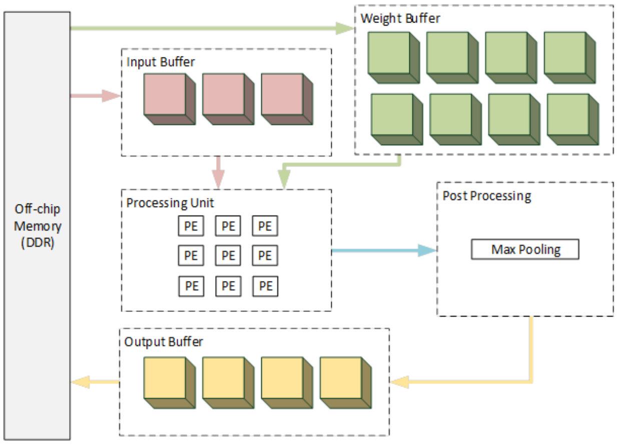


Figure 8 Block-diagram for convolution implementation.

Acknowledgements

This research was funded by the European Space Agency (ESA/ESTEC) project "Adaptive SAR Signal Compression Through Artificial Intelligence" Contract Nr. ESA AO/1-11419/22/NL/GLC/my. The method is patented under the German patent Nr. 10 2022 123 897 and European patent pending (Nr. 23 192 942.3).

6 Literature

[1] R. Kwok and W.T.K. Johnson, "Block adaptive quantization of magellan sar data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 4, pp. 375–383, 1989.

[2] P. Snoeij, E. Attema, A. Monti Guarnieri, and F. Rocca, "FDBAQ a novel encoding scheme for Sentinel-1," in *2009 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2009, vol. 1, pp. I–44–I–47.

[3] M. Martone, B. Bräutigam, and G. Krieger, "Azimuth-switched quantization for SAR systems and performance analysis on TanDEM-X data," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 181–185, 2014.

[4] M. Martone, N. Gollin, P. Rizzoli, and G. Krieger, "Performance-optimized quantization for SAR and InSAR applications," *IEEE Transactions on Geoscience and Remote Sensing*, 14 Jun. 2022.

[5] X. X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets SAR: Concepts, models, pitfalls, and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 4, pp. 143–172, 2021.