Verifiable and Energy Efficient Medical Image Analysis with Quantised Self-attentive Deep Neural Networks

Rakshith Sathish(✉), Swanand Khare, and Debdoot Sheet

Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India
rakshith.sathish@kgpian.iitkgp.ac.in

Abstract. Convolutional Neural Networks have played a significant role in various medical imaging tasks like classification and segmentation. They provide state-of-the-art performance compared to classical image processing algorithms. However, the major downside of these methods is the high computational complexity, reliance on high-performance hardware like GPUs and the inherent black-box nature of the model. In this paper, we propose quantised stand-alone self-attention based models as an alternative to traditional CNNs. In the proposed class of networks, convolutional layers are replaced with stand-alone self-attention layers, and the network parameters are quantised after training. We experimentally validate the performance of our method on classification and segmentation tasks. We observe 50–80% reduction in model size, 60–80% lesser number of parameters, 40–85% fewer FLOPs and 65–80% more energy efficiency during inference on CPUs. The code will be available at https://github.com/Rakshith2597/Quantised-Self-Attentive-Deep-Neural-Network.

Keywords: Self-attention · Quantisation · Medical image analysis

1 Introduction

Deep neural networks have played a significant role in medical image analysis. Since the advent of UNet [18] to UNetr [5], the performance of neural networks on various tasks like classification, segmentation, and restoration has improved considerably. Deeper and broader convolutional neural networks generally show an improvement in performance at the cost of an increase in the number of learnable parameters, model size and total floating-point operations performed during a single forward pass of the data through the network. Moreover, these models require specialised high-performance hardware even during inference. This reliance on larger models and high-performance hardware hinders the last-mile delivery of AI solutions to improve the existing healthcare system, especially in resource constrained developing and under-developed countries.

S. Albarqouni et al. (Eds.): DeCaF 2022/FAIR 2022, LNCS 13573, pp. 178–189, 2022.
https://doi.org/10.1007/978-3-031-18523-6_17

Challenges: The performance and trustability of deep neural network-based methods are of utmost importance, especially in the medical domain. The performance of these methods decreases as we try to reduce the number of learnable parameters in the model. As an example, in the case of image classification, deeper networks have been shown to be superior to shallow networks with fewer parameters [6,8]. Despite the good performance measured in terms of quantitative evaluation metrics, deep neural network (DNN) are known to make the right decision for the wrong reasons [4]. This limits the trustability of DNN-based frameworks in practical application. Additionally, the black box nature of the convolutional neural networks makes them unreliable for clinical applications. Developing a method that relies on fewer parameters and is clinically verifiable is a challenging task. Also, an efficient model is expected to replicate the performance during inference at a reasonable execution speed even in the absence of GPUs.

Attention-based networks were proposed to augment DNNs with explainability in the case of natural images. However, due to the inherent differences in the nature of images, we cannot assume an equivalent performance in the medical images. As an example, in detecting objects in natural images, the objects of interest often have a well-defined shape and structure, which are absent in the case of medical images. In the case of medical image classification, the biomarkers are usually unstructured pathologies with variable appearance. In this work, we try to verify the effectiveness of replacing convolutions with attention in neural networks for medical images.

Related Works: Transformers [21], based solely on attention mechanisms has revolutionised the way models are designed for natural language tasks. Motivated by their success, [17,26,27] and [25] explored the possibility of using self-attention to solve various vision tasks. Among these, the stand-alone self-attention proposed by [17] established that self-attention could potentially replace convolutional layers altogether. Even though it is efficient compared to other DNNs, such models can be further improved by quantising the weights and activations of the networks [15]. The quantisation of deep neural networks has shown significant progress in recent years [1,24]. The ability to quantise the neural network trained in high precision without substantial loss in performance during inference simplifies the process.

Our Approach: Inspired by the success of [17] in natural image classification tasks, we propose the design of a new class of networks for medical image classification and segmentation, in which we replace the convolution layers with self-attention layers. Furthermore, we optimise the networks for inference by quantising the parameters thereby decreasing energy consumption. To the best of our knowledge, a quantised fully self-attentive network for classification and segmentation of medical images and comparison with its convolutional counterparts has not been attempted so far. Schematic overview of the proposed method is illustrated in Fig. 1.

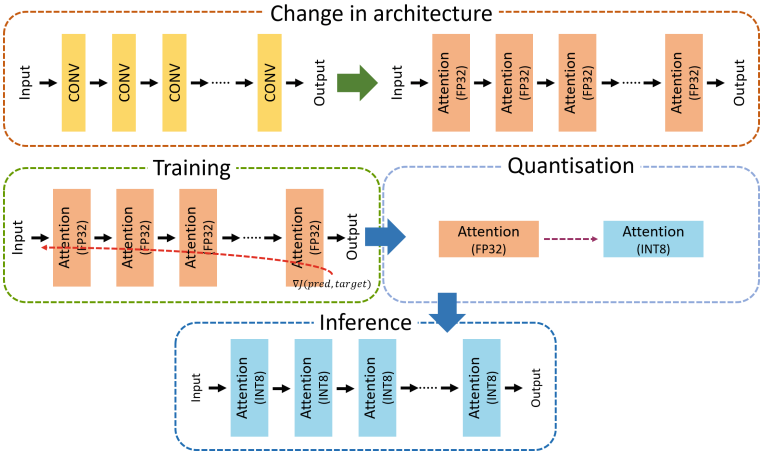


Fig. 1. Overview of the proposed method. Convolutional layers in deep neural network architectures are replaced with self-attention layers and networks with parameters at FP32 precision are trained till convergence. To optimise the model for storage and faster inference, the network parameters are quantised without loss in performance.

2 Method

2.1 Stand-Alone Self-attention

Attention was introduced by [3] for a neural machine translation model. Attention modules can learn to focus on essential regions within a context, making it an important component of neural networks. Self-attention [21] is defined as attention applied to a single context instead of across multiple contexts; that is, *Key*, *Query* and *Values* are derived from the same context. [17] introduced the stand-alone self-attention layer, which can replace convolutions to construct a fully attentional model. Motivated by the initial success of [17] in natural images, we explore the feasibility of using such modules in the proposed class of networks for medical image analysis.

To compute attention for each pixel $\mathbf{x}_{i,j} \in \mathbb{R}^{C_{in} \times 1 \times 1}$ in an image or an activation map, local regions with spatial extent $h \times w$ around $\mathbf{x}_{i,j}$ are used to derive the *keys* and *values*. Learned linear transformations are performed on $\mathbf{x}_{i,j}$ and its local regions to obtain *query* ($\mathbf{Q}$), *keys* ($\mathbf{K}$) and *values* ($\mathbf{V}$) as

$$\mathbf{Q} = \mathbf{W_Q} \mathbf{x}_{i,j} \tag{1}$$

$$\mathbf{K} = \mathbf{W_K} \mathbf{x}_{h,w} \tag{2}$$

$$\mathbf{V} = \mathbf{W_V} \mathbf{x}_{h,w} \tag{3}$$

where $\mathbf{W_Q} \in \mathbb{R}^{C_{out} \times C_{in}}$, $\mathbf{W_K} \in \mathbb{R}^{C_{out} \times C_{in}}$ and $\mathbf{W_V} \in \mathbb{R}^{C_{out} \times C_{in}}$ are learnable transformation matrices and $\mathbf{x}_{h,w} \in \mathbb{R}^{C_{in} \times h \times w}$ is the local region centered at $\mathbf{x}_{i,j}$.

Self-attention on its own does not encode any positional information, which makes it permutation equivariant. Relative positional embedding [19] as used in [17] are incorporated into the attention module. The keys $\mathbf{K} \in \mathbb{R}^{\mathrm{C_{out}} \times h \times w}$ are split into $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{\mathrm{C_{out}/2} \times \mathrm{h} \times \mathrm{w}}$ each and column offset $\mathbf{E}_{col}$ and row offset $\mathbf{E}_{row}$ of the positional embedding are added to these separately. After this, we concatenate $\mathbf{K}_1, \mathbf{K}_2$ to obtain a new key ($\mathbf{K}' \in \mathbb{R}^{\mathrm{C_{out}} \times \mathrm{h} \times \mathrm{w}}$) which contains the relative spatial information of pixels in the local region of size $h \times w$. Thus, the relative spatial attention for a pixel x_{ij} is mathematically defined as in Eq. 4 and is graphically illustrated in Fig. 2.

$$\mathbf{y_{i,j}} = \sum_{\{u,v\} \in N_{h,w}(i,j)} softmax_{u,v}(\mathbf{Q}_{i,j}^{\top} \mathbf{K}_{u,v}) \mathbf{V}_{u,v} \tag{4}$$

where $N_{h,w}(i,j)$ is the neighbourhood of size $h \times w$ centered at (i,j).

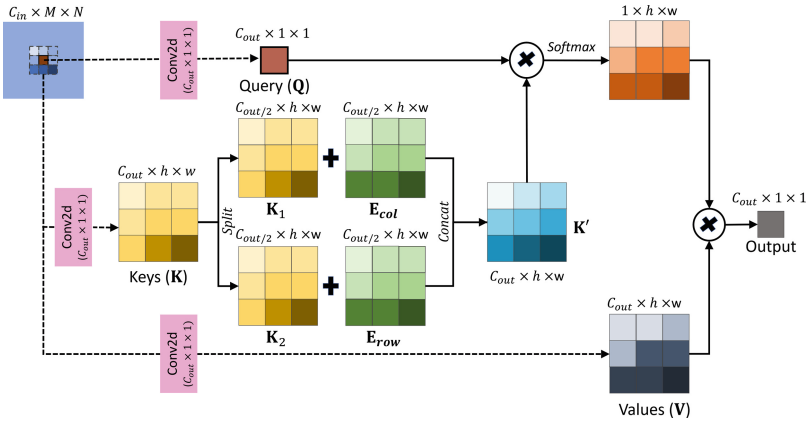


Fig. 2. Self-attention mechanism with local context. Operations are performed on a per-pixel basis to compute attention as shown in the figure. Linear transformations for obtaining query, keys and values are implemented using 2D convolution (*Conv2d*) operation. The learnt relative positional embedding are added to the keys to incorporate the inter-pixel relationships within the local context.

We use these attention blocks instead of 2D convolutional blocks in our networks. During training, all the weights and activations are represented and stored with a precision of FP32. The parameters are quantised to INT8 precision for inference.

2.2 Quantisation of Network Parameters

We perform quantisation using the FBGEMM (FaceBook GEneral Matrix Multiplication) [10] backend of PyTorch for x86 CPUs, which is based on the quantisation scheme proposed by [9]. In order to be able to perform all the arithmetic

operations using integer arithmetic operations on quantised values, we require the quantisation scheme to be an affine mapping of integers q to real numbers r as

$$r = S(q - Z) \tag{5}$$

where S and Z are quantisation parameters. We have employed a post-training 8-bit quantisation of all the weights and operations for our proposed model.

2.3 Network Architecture

Classification: The architecture of the proposed classification network is illustrated in Fig. 3(a) with the details of the constituent modules in Fig. 3(c). The network consists of a series of alternating attention blocks and attention down blocks followed by fully-connected linear layers. The feature maps are downsampled using the max-pooling operation. The size of the output linear layer is equal to the number of target classes. The network is trained to perform multi-label classification using a binary cross-entropy loss.

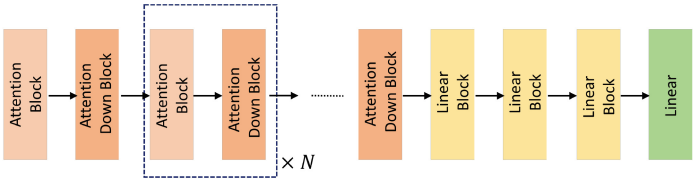
Segmentation: The proposed segmentation network has a fully attention-based encoder-decoder architecture as shown in Fig. 3(b). The encoder unit consists of stand-alone self-attention blocks with ReLU activation and max-pooling operations with the number of feature maps increasing progressively with each attention block. The decoder consists of attention blocks and max-unpooling operations. The size of activation maps of the decoder matches with the corresponding layer in the encoder. The unpooling operations are performed using the indices transferred from the pooling layers in the encoder. To prevent the loss of subtle information, we employ activation concatenation in the decoder, similar to UNet [18]. The network is trained using soft dice loss [12].

3 Experiments

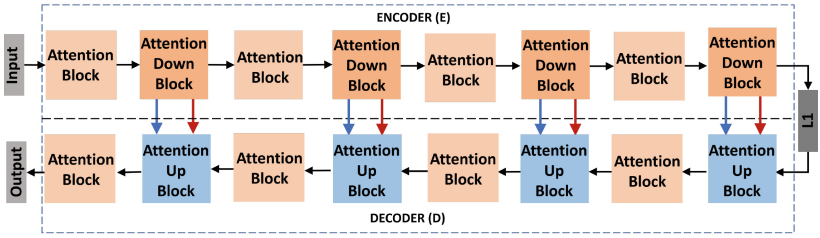
3.1 Datasets

Classification: To evaluate the performance of the fully self-attentive network (SaDNN-cls) on classification tasks, we have used the NIH Chest X-ray dataset of 14 Common Thorax Disease [22]. The dataset comprises 112,120 frontal-view X-ray images of 30,805 patients with fourteen disease labels. These disease classes can co-occur in an image; therefore, the classification problem is formulated as multi-label classification. The train, validation and test split provided in the dataset was used for the experiments.

Segmentation: A subset of the medical segmentation decathlon dataset [2] is used to evaluate the performance of the proposed fully-attentive network (SaDNN-seg) for liver segmentation. Out of the 131 ground truth paired 3D CT volumes-Ground truth pairs available in the dataset, 80% were randomly chosen for training, and the remaining 20% were used for testing.



(a) Classification network (SaDNN-cls)



(b) Segmentation network (SaDNN-seg)

(c) Network components

Fig. 3. Architecture of the proposed Self-attentive Deep Neural Networks (SaDNN). Detailed architecture of the networks for classification and segmentation are shown in (a) and (b) respectively. Components of the various blocks in these networks are detailed in (c).

3.2 Implementation Details

Training: The proposed models were trained using an Adam Optimiser [11] with a learning rate of 1×10^{-4}. The models for classification task were trained for 15 epochs and the models for segmentation were trained for 25 epochs.

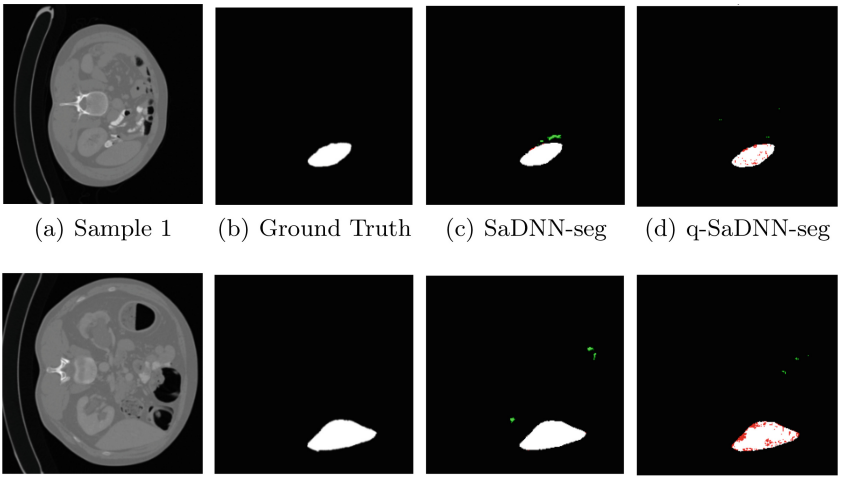
Baselines: Performance of the proposed quantised self-attention network for the classification task is compared with ResNet-18, ResNet-50 and their 8–bit quantised versions q-ResNet-18, and q-ResNet-50. To assess the performance of the segmentation network, we chose a modified UNet [18] (UNet-small) and SUMNet [13] architecture trained on the same dataset split and their quantised versions q-UNet-small and q-SUMNet as baselines.

System Specifications: All networks were trained on a high-performance server with a NVIDIA $V100$ GPU, $x86_64$ Intel(R) Xeon(R) Silver 4110 CPU @ 2.10 GHz, 96 GB RAM and 1 TB HDD running on Ubuntu 18.01.1 LTS OS. The inference of quantised models was also performed on the same class of CPUs.

4 Results and Discussions

4.1 Qualitative Analysis

visualisation of predictions of the proposed q-SaDNN-seg network and its unquantised version SaDNN-seg are presented in Fig. 4. Over-segmented regions in the predicted segmentation maps are marked in green, under-segmented regions are marked in red and correctly segmented region is shown in white. We observe that the tendency of the original unquantised network SaDNN-seg to over-segment is significantly reduced post quantisation. However, the quantisation of network parameters causes the q-SaDNN-seg to under-segment the target organ. This is reflected in the slightly lower Dice coefficient (DSC) of the proposed model as seen in Table 2.



(a) Sample 1 (b) Ground Truth (c) SaDNN-seg (d) q-SaDNN-seg

(e) Sample 2 (f) Ground Truth (g) SaDNN-seg (h) q-SaDNN-seg

Fig. 4. Comparison of segmentation predictions. Figure shows sample input CT images in (a) and (e) with the corresponding ground truths of liver in (b) and (f) respectively. Segmentation map as predicted by SaDNN-seg, with the over-segmented region marked in green and under-segmented region marked in red are presented in (c) and (g) for the two sample images. Similar visualisation of segmentation by the proposed q-SaDNN-seg are presented in (d) and (h). (Color figure online)

4.2 Quantitative Analysis

The performance of the proposed quantised fully self-attentive network and baselines for multi-label classification task is reported in terms of accuracy in Table 1. It can be observed that the proposed network can achieve performance slightly better than the existing deep residual convolutional neural networks. Table 2 shows the comparison of the proposed segmentation network with the baselines in terms of DSC. The proposed quantised network performs almost as good as the quantised versions of the baseline convolutional neural networks.

Table 1. Evaluation of classification

Model	Accuracy
ResNet-18	0.89
q-ResNet-18	0.88
ResNet-50	0.84
q-ResNet-50	0.83
SaDNN-cls (ours)	**0.90**
q-SaDNN-cls (ours)	0.89

Table 2. Evaluation of segmentation

Model	DSC
UNet-small	0.88
q-UNet-small	0.88
SUMNet	0.89
q-SUMNet	0.89
SaDNN-seg (ours)	**0.88**
q-SaDNN-seg (ours)	0.85

4.3 Computational Analysis

The DNNs used for the experiments exhibited superior classification and segmentation performance in terms of quantitative metrics, but they require a considerable amount of computations and memory access operations to be performed. Deploying a framework which needs excessive computations to be performed results in large energy consumption, which is not feasible in diverse resource-constrained scenarios. Therefore, it is key to have an energy-efficient model without degradation in performance. A rough estimate of energy cost per operation in $45nm$ $0.9V$ IC design can be calculated using Table 3 presented in [7,14,23].

Table 3. Approximate energy cost in 45 nm 0.9 V for different multiplication and addition operations

Operation	Energy (pJ)	
	MUL	ADD
8-bit INT	0.2 pJ	0.03 pJ
16-bit FP	1.1 pJ	0.40 pJ
32-bit FP	3.7 pJ	0.90 pJ

The number of multiplication and addition operations in a standalone self-attention layer [20] can be calculated as

$$Ops_{mul} = Ops_{add} = 2b^2c \tag{6}$$

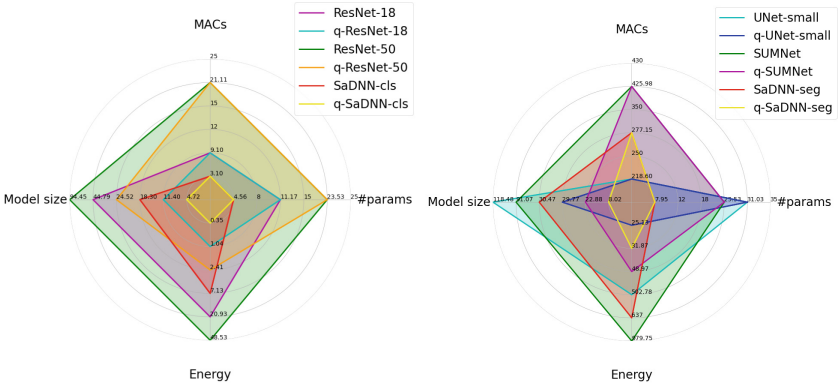
where b is the block (local region) size and c is the number of channels.

The total number of parameters, MACs, energy consumed during forward pass and model size of the proposed q-SaDNN-cls and q-SaDNN-seg networks are reported in Table 4 and Table 5 with graphical comparisons in Fig. 5. Models with the least area in the radar charts are more efficient. The proposed q-SaDNN-cls network is 58.59% smaller than quantised ResNet-18 and 80.75% smaller than quantised ResNet-50 in terms of model size. In terms of total MAC units, the propsed networks have 65.93% fewer MACs than ResNet-18, 85.32% fewer than

ResNet-50. Similarly, in terms of the total trainable parameters, the proposed networks have 59.17% lesser parameters than ResNet-18 and 80.62% lesser than ResNet-50.

Table 4. Comparison of classification networks

Model	#Params	MACs	Model size	Energy
ResNet-18	11.17 M	9.10 G	44.79 MB	20.93 J
q-ResNet-18	11.17 M	9.10 G	11.40 MB	1.04 J
ResNet-50	23.53 M	21.11 G	94.45 MB	48.53 J
q-ResNet-50	23.53 M	21.11 G	24.52 MB	2.41 J
SaDNN-cls	**4.56 M**	3.10 G	18.30 MB	7.13 J
q-SaDNN-cls	**4.56 M**	3.10 G	**4.72 MB**	0.35 J



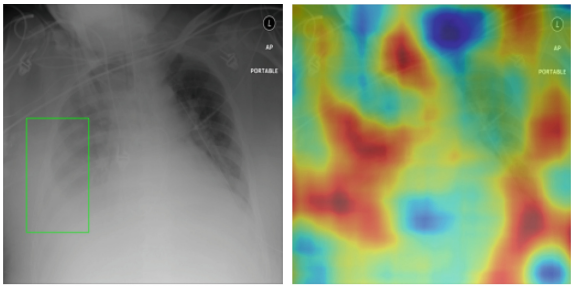
(a) Classification (b) Segmentation

Fig. 5. Graphical comparison of proposed networks. Figure shows radar chart based comparison of proposed (a) classification network and (b) segmentation network in terms of number of parameters, MACs, model size and energy. The model with the least area within the plot is the best one.

Similar improvement in efficiency of computing can be observed in the case of segmentation as well. The segmentation network q-SaDNN-seg is 73.06% smaller than q-UNet-small and 64.94% smaller than q-SUMNet in terms of model size. In terms of total MAC units, the q-SaDNN-seg has 34.94% fewer than SUMNet. In terms of the trainable parameters, q-SaDNN-seg has 74.37% lesser parameters than UNet-small and 66.21% lesser than SUMNet. It is to be noted that the proposed models are superior in terms energy consumption as well.

Table 5. Comparison of segmentation networks

Model	#Params	MACs	Model size	Energy
UNet-small	31.03 M	218.60 G	118.48 MB	502.78 J
q-UNet-small	31.03 M	218.60 G	29.77 MB	25.13 J
SUMNet	23.53 M	425.98 G	91.07 MB	979.75 J
q-SUMNet	23.53 M	425.98 G	22.88 MB	48.97 J
SaDNN-seg	**7.95 M**	277.15 G	30.47 MB	637 J
q-SaDNN-seg	**7.95 M**	277.15 G	**8.02 MB**	31.87 J



(a) Test image (b) Saliency map

Fig. 6. Figure shows (a) a sample image from the test set used in our experiments with the clinically relevant region as provided in the dataset marked in green and (b) saliency map of q-SaDNN-cls. Regions shown in red in the saliency map are perceived as most important and those in blue to be least important by the network during prediction. (Color figure online)

4.4 Analysis of Clinical Relevance

Validating the results of the model with respect to clinically relevant information to provide some explanations for the decision made by the model is an important factor that determines trustability. The clinically relevant region provided in the NIH Chest X-ray dataset as marked by a radiologist and the saliency map based explanation generated using RISE [16] for the proposed quantised self-attention deep neural network for classification are shown in Fig. 6. It can be observed that the proposed model focuses on the clinically relevant region while making the decision.

5 Conclusion

We proposed a class of quantised self-attentive neural networks which can be used for medical image classification and segmentation. In these networks, convolutional layers are replaced with attention layers which have fewer learnable parameters. Computation of attention while considering a small local region

surrounding a pixel prevents degradation of performance despite the absence of local feature extraction which is typically performed in a CNN. We show that our energy efficient method achieves performance at par with the commonly used CNNs with fewer number of parameters and model size. These attributes make our proposed models affordable and easy to adopt in resource constrained settings.

References

1. Aji, A.F., Heafield, K.: Compressing neural machine translation models with 4-bit precision. In: Workshop on Neural Generation and Translation, pp. 35–42 (2020)
2. Antonelli, M., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Chakravarty, A., Ghosh, N., Sheet, D., Sarkar, T., Sethuraman, R.: Radiologist validated systematic search over deep neural networks for screening musculoskeletal radiographs (2019)
5. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Horowitz, M.: 1.1 computing's energy problem (and what we can do about it). In: IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 10–14 (2014)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
9. Jacob, B., et al.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2704–2713 (2018)
10. Khudia, D., et al.: FBGEMM: enabling high-performance low-precision deep learning inference. arXiv preprint arXiv:2101.05615 (2021)
11. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571 (2016). https://doi.org/10.1109/3DV.2016.79
13. Nandamuri, S., China, D., Mitra, P., Sheet, D.: SUMNet: fully convolutional model for fast segmentation of anatomical structures in ultrasound volumes. In: IEEE International Symposium on Biomedical Imaging, pp. 1729–1732 (2019)
14. Park, S.S., Chung, K.S.: CENNA: cost-effective neural network accelerator. Electronics **9**(1), 134 (2020)
15. Paupamah, K., James, S., Klein, R.: Quantisation and pruning for neural network compression and regularisation. In: International SAUPEC/RobMech/PRASA Conference, pp. 1–6. IEEE (2020)

16. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: British Machine Vision Conference (2018)
17. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909 (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
19. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
20. Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J.: Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12894–12904 (2021)
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)
23. Wu, S., Li, G., Chen, F., Shi, L.: Training and inference with integers in deep neural networks. arXiv preprint arXiv:1802.04680 (2018)
24. Xu, J., Yu, J., Hu, S., Liu, X., Meng, H.M.: Mixed precision low-bit quantization of neural network language models for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 3679–3693 (2021)
25. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5791–5800 (2020)
26. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10502–10511 (2019)
27. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)