ICS 35.040
CCS L 71

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 42382.1-2023

Information technology - Neural network representation and model compression - Part 1: Convolutional neural network

信息技术 神经网络表示与模型压缩 第1部分：卷积神经网络

Issued on: March 17, 2023 **Implemented on: October 1, 2023**

Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC

Table of Contents

1 Scope

The document lays down the representation and the compression process of the offline model of a convolutional neural network.

It applies to the development, the design, the testing and the evaluation of convolutional neural network models of every kind, and to their efficient application in the device and cloud domains.

A note adds that the representation and model compression methods laid down in the document do not require native support from a machine learning framework, and may be supported by means of conversion, of a toolkit or of similar forms.

2 Normative references

One document is cited: GB/T 5271.34-2006, Information technology, Vocabulary, Part 34: Artificial intelligence, neural networks. For dated references only the edition cited applies; for undated references the latest edition, including any amendments, applies.

3 Terms and definitions

3.1 codec representation: use of compression technology to reduce the size of a model. A

note refers to Clause 10 for the detailed definition.

3.2 layer: hierarchical structure within a neural network. A note adds that every network layer contains several operators, for instance the input layer, the convolution layer and the fully connected layer.

3.3 reference random vector: basic symbol vector shared by the whole network.

3.4 multiple INT4 quantization: form of quantization in which one tensor is quantized into a combination of several INT4 tensors.

3.5 encapsulation representation: as printed, the definition reads that it puts out interfaces such as security information and identity verification. A note refers to model protection in Clause 11 for the detailed definition.

3.6 block structured matrix: matrix that can be divided into several blocks, each block being arranged according to some rule.

3.7 block circulant matrix: matrix each of whose blocks is a circulant matrix.

3.8 shared weight, shared weight value, shared weight tensor: where one neural network handles several different tasks, the portion of the weights that those tasks share.

3.9 shared bias, shared bias vector: where one neural network handles several different tasks, the portion of the bias that those tasks share.

3.10 convolutional neural network: feedforward neural network that uses convolution in at least one of its layers. A note adds that it is a representative deep learning algorithm in artificial intelligence applications, and refers to ISO/IEC DIS 22989:2022 for the definition.

3.11 compact representation: serialized format output after compression by any of several neural network compact representation algorithms. The example given lists structured matrices, weight sharing between multiple models, the sparse matrix of sparse coding, the quantization table of weight quantization, the decomposed matrices of low-rank decomposition and the bit representation of a binary neural network. A note refers to the compression process, the decompression process and the data generation methods in Clause 7 to Clause 9.

3.12 basic symbol vector: comprises the basic symbol vector shared by the whole network and the basic symbol vector corresponding to each layer, the latter being obtained from the basic symbol vector of the network according to an agreed rule.

3.13 correct input vector: vector obtained by multiplying the elements of the input vector of a network layer by the elements at the corresponding positions in the disturbance vector.

3.14 structured matrix: particular class of matrix that can be built into a complete matrix from a smaller quantity of data together with a definite arrangement rule.

3.15 convolution kernel granularity quantization: form of quantization in which the weights of

a convolution layer are divided by output channel and each 3D convolution kernel is quantized one by one.

3.16 interoperable representation: definition of the syntax of a neural network, of the arithmetic operations it supports and of the weight formats, which are integer, floating point and fixed point. A note refers to the syntax and semantics of the neural network model in Clause 6.

3.17 quantitative scale factor: scaling ratio between the original tensor and the quantized tensor.

3.18 model compression: method that reduces the size of a neural network model and raises the efficiency with which the model runs and is transmitted.

3.19 model: neural network corresponding to the completion of a single task or of several tasks.

3.20 bias, offset, bias vector, offset vector: systematic deviation with respect to a reference value. A note refers to ISO/IEC 2382:2015 for the definition.

3.21 weight, weight tensor, connection weight, connection strength, synaptic weight: coefficient by which the input value of an artificial neuron is multiplied before it is combined with the other input values. A note refers to GB/T 5271.34-2006.

3.22 weight aggregation dimension: dimension along which the shared weight tensor and the unique weight vector are concatenated.

3.23 task: some function realized by means of a neural network. The examples given are image super-resolution and image denoising.

3.24 disturbance vector, disturbance symbol vector: symbol vector obtained by expanding the random vector so that its number of dimensions equals that of the input vector of the network layer.

3.25 neural network representation: method of defining the basic syntax, the semantics, the operations, the hyper-parameters and the bitstream of a neural network.

3.26 random vector, random symbol vector: basic symbol vector corresponding to a single network layer.

3.27 unique weight, unique weight value, unique weight tensor: where one neural network handles several different tasks, the portion of the weights that is not shared, for any one of those tasks.

3.28 unique weight list: where one neural network handles several different tasks, the portion other than the shared weight values, that is, the combination of the weight portions not shared among the several tasks.

3.29 specific offset vector: where one neural network handles several different tasks, the