INTERNATIONAL STANDARD

ISO/IEC 15938-18

First edition
2023-07

Information technology — Multimedia content description interface —

Part 18: Conformance and reference software for compression of neural networks

Technologies de l'information — Interface de description du contenu multimédia —

Partie 18: Conformité et logiciel de référence pour la compression des réseaux neuronaux

Reference number
ISO/IEC 15938-18:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see https://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

A list of all parts in the ISO/IEC 15938 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

This document describes conformance testing and the reference software for ISO/IEC 15938-17 Compression of neural networks for multimedia content description and analysis. The reference software includes both encoder and decoder functionality.

The reference software is useful in aiding users of a standard for coding neural networks to establish and test conformance and interoperability, and to educate users and demonstrate the capabilities of the standard. For these purposes, the accompanying software is provided as an aid for the study and implementation of 15938-17 compression of neural networks for multimedia content description and analysis.

The purpose of this document is to provide the following:

— A set of reference bitstreams conforming to ISO/IEC 15938-17.

— Description of procedures to test conformance of bitstreams and decoders to ISO/IEC 15938-17.

— Reference decoder software capable of decoding bitstreams that conform to ISO/IEC 15938-17 in a manner that conforms to the decoding process specified in ISO/IEC 15938-17.

— Reference encoder software capable of producing bitstreams that conform to ISO/IEC 15938-17.

Information technology — Multimedia content description interface —

Part 18: Conformance and reference software for compression of neural networks

1 Scope

This document specifies conformance testing procedures for implementations of ISO/IEC 15938-17 and provides conformance bitstreams. It also provides the reference software for ISO/IEC 15938-17 which is an integral part of this document.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 15938-17:2022, *Information technology — Multimedia content description interface — Part 17: Compression of neural networks for multimedia content description and analysis*

ISO/IEC 21778, *Information technology — The JSON data interchange syntax*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 15938-17 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
array
ordered list of elements where all elements are of the same type

3.2
dictionary
ordered list of key/value pairs where each key/value pair is an list of two elements with the first element being denoted ‘key’ and with the second element being denoted ‘value’

4 Conformance testing

4.1 General

Subclauses 4.2 through 4.6 specify tests for verifying the conformance of bitstreams as well as decoders. These tests make use of test data (bitstreams and related models) provided at https://standards.iso

.org/iso-iec/15938/-18/ed-1/en (described in detail in 4.4) and follow the procedure described in subclause 4.5.

4.2 Conformance testing for decoder

The decoder conformance is specified in ISO/IEC 15938-17:2022, Clause 7.

4.3 Conformance testing for bitstreams

The bitstream conformance is specified in ISO/IEC 15938-17:2022, Clause 7.

4.4 Models and reference bitstreams

A set of bitstreams and related neural network models is provided for conformance testing. When neural network models are provided, they are the source models used to generate one or more compressed bitstreams. The code defining these models is provided as part of the reference software distribution at https://standards.iso.org/iso-iec/15938/-18/ed-1/en. A bitstream can be generated:

— from a commonly used neural network model,

— from synthetic data by encoding it with the reference encoder,

— by creating a conformant synthetic bitstream directly, without using an encoder.

Where applicable, the dataset used to train the model is provided for information only. However, the dataset is not needed for conformance testing and thus not provided.

Table 1 summarizes the provided bitstreams. It also lists the features tested with the bitstreams, and the reference encoder configuration used to generate the bitstream (where applicable). If decoding the bitstream requires out-of-band parameters (e.g. information that is derived from the network topology description), those parameters are also provided. Source models and datasets are referred to using the following names:

— ImageNet for the dataset described in Reference [5],

— CIFAR-100 for the dataset described in Reference [6],

— DCase for the dataset described in Reference [7] and the model described in Reference [8],

— MobileNet V2 for model described in Reference [9],

— UC12B for the model described in Reference [10] and

— VGG16 for the model described in Reference [11].

Table 1 — Bitstreams and related models for conformance testing

Bitstream id	Source model	Data-set	Relevant technology ISO/IEC 15938-17:2022	Features tested	Reference encoder configuration
perf_map_sparse_MobileNetV2.nctm	MobileNetV2	Image Net	6.3.4.3	sparsification performance map	qp_density = 2 scan_order = 1 approx_method = "codebook" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
perf_map_prune_DCase.nctm	DCase	DCase	6.3.4.3	pruning performance map	qp_density = 2 scan_order = 1 approx_method = "codebook" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
perf_map_sparse_prune_UC12B.nctm	UC12B		6.3.4.3	Sparsification and Pruning Performance map	qp_density = 2 scan_order = 1 approx_method = "codebook" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
perf_map_sparse_VGG16.nctm	VGG16	Image Net	6.3.4.3	Sparsification Performance Map (pruned model)	qp_density = 2 scan_order = 1 approx_method = "codebook" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
prune_tpl_cont_sparse_bm_DCase.nctm	DCase	DCase	6.3.4.5	Prune Topology - sparse bitmask	encode_tpl_only = True partial_data_counter = 0

Table 1 *(continued)*

Bitstream id	Source model	Data-set	Relevant technology ISO/IEC 15938-17:2022	Features tested	Reference encoder configuration
prune_tpl_cont_prune_bm_VGG16.nctm	VGG16	Image Net	6.3.4.5	Prune Topology - prune bitmask	encode_tpl_only = True partial_data_counter = 0
prune_tpl_cont_comb_bm_VGG16.nctm	VGG16	Image Net	6.3.4.5	Prune Topology - combined bitmask	encode_tpl_only = True partial_data_counter = 0
prune_tpl_cont_prune_dictionary_DCase.nctm	DCase	DCase	6.3.4.5	Prune Topology - prune dictionary	encode_tpl_only = True topology_indexed_reference_flag = False partial_data_counter = 0
prune_tpl_cont_prune_dictionary_idx_ResNet50.nctm	ResNet50	Image Net	6.3.4.5	Prune Topology - prune dictionary (indexed elem id)	encode_tpl_only = True topology_indexed_reference_flag = True partial_data_counter = 0
tpl_reflist_DCase.nctm	DCase	DCase	6.3.4.5, 6.3.3.7	Topology Reflist	encode_tpl_only = True partial_data_counter = 0
partial_data_counter_VGG16_ndu_size_65536.nctm	VGG16	Image Net	6.3.3.1	Partial data counter	max_ndu_nnr_unit_size = 65536
partial_data_counter_VGG16_ndu_size_32768.nctm	VGG16	Image Net	6.3.3.1	Partial data counter	max_ndu_nnr_unit_size = 32768
partial_data_counter_VGG16_ndu_size_16384.nctm	VGG16	Image Net	6.3.3.1	Partial data counter	max_ndu_nnr_unit_size = 16384
partial_data_counter_DCase_ndu_size_2048.nctm	DCase	DCase	6.3.3.1	Partial data counter	max_ndu_nnr_unit_size = 2048
partial_data_counter_DCase_ndu_size_1024.nctm	DCase	DCase	6.3.3.1	Partial data counter	max_ndu_nnr_unit_size = 1024
deepCABAC_ResNet50_1_qp-38_qp_density2.nctm	ResNet50	Image Net	10 9.1.1 / 9.2.1	DeepCABAC entropy coding, uniform quantization	see verify_all.sh
dependent_quantization_ResNet50_2_qp-38_qp_density2.nctm	ResNet50	Image Net	9.1.3 / 9.2.3 6.3.3.7	Dependent scalar quantization	see verify_all.sh
deepCABAC_qp_density_MobileNetV2_3_qp-38_qp_density2.nctm	MobileNetV2	Image Net	9.2	QpDensity	see verify_all.sh

Table 1 *(continued)*

Bitstream id	Source model	Data-set	Relevant technology ISO/IEC 15938-17:2022	Features tested	Reference encoder configuration
deepCABAC_qp_density_MobileNetV2_4_qp-76_qp_density3.nctm	MobileNetV2	Image Net	9.2	QpDensity	see verify_all.sh
block_scan_order_8x8_cabac_entry_points_ResNet50_5_qp-38_qp_density2.nctm	ResNet50	Image Net	4.12 / 6.4.3.7 / 6.4.3.8 / 7.3.6	Block scan order / cabac entry points	see verify_all.sh
block_scan_order_16x16_cabac_entry_points_ResNet50_6_qp-38_qp_density2.nctm	ResNet50	Image Net	4.12 / 6.4.3.7 / 6.4.3.8 / 7.3.6	Block scan order / cabac entry points	see verify_all.sh
codebook_signaling_MobileNetV2_7_qMse0.00001.nctm	MobileNetV2	Image Net	9.1.3 / 9.2.2	Codebook-based quantization	see verify_all.sh
local_scaling_DCase_8_qp-38_qp_density2.nctm	DCase	DCase	8.2.7 / 8.3.7	Local scaling	see verify_all.sh
batchnorm_folding_Mobile NetV2_9_qp-38_qp_density2.nctm	MobileNetV2	Image Net	8.2.6 / 8.3.6	BatchNorm Folding	see verify_all.sh
out_of_band_signaling_Res Net50_10_qp-38_qp_density2.nctm	ResNet50	Image Net	6.3.3.7 / 6.4.3.7	Out-of-band signaling [a]	see verify_all.sh
deepCABAC_8bit_ResNet50_PYTzoo_11_qp0_qp_density4.pt.nctm	ResNet50	Image Net	9.1.1 / 9.2.1	Uniform quantization with limited precision (8bit)	see verify_all.sh
deepCABAC_8bit_MobileNetV2_PYTzoo_12_qp0_qp_density4.pt.nctm	MobileNet V2	Image Net	9.1.1 / 9.2.1	Uniform quantization with limited precision (8bit)	see verify_all.sh
deepCABAC_4bit_VGG16_PYTzoo_13_qp0_qp_density4.pt.nctm	VGG16	Image Net	9.1.1 / 9.2.1	Uniform quantization with limited precision (4bit)	see verify_all.sh
deepCABAC_8bit_UC12B_14_qp0_qp_density4.nctm	UC12B	CIFAR-100	9.1.1 / 9.2.1	Uniform quantization with limited precision (8bit)	see verify_all.sh

Table 1 *(continued)*

Bitstream id	Source model	Data-set	Relevant technology ISO/IEC 15938-17:2022	Features tested	Reference encoder configuration
deepCABAC_4bit_DCase_15_qp0_qp_density4.nctm	DCase	DCase	9.1.1 / 9.2.1	Uniform quantization with limited precision (4bit)	see verify_all.sh
perf_map_sparse_MobileNetV2_bw8.nctm	MobileNetV2	Image Net	6.3.4.3	sparsification performance map (8bit)	qp_density = 2 scan_order = 1 approx_method = "uniform" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
perf_map_prune_DCase_bw8.nctm	DCase	DCase	6.3.4.3	pruning performance map (8bit)	qp_density = 2 scan_order = 1 approx_method = "uniform" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0
perf_map_sparse_prune_UC12B_bw8.nctm	UC12B		6.3.4.3	Sparsification and Pruning Performance map (8bit)	qp_density = 2 scan_order = 1 approx_method = "uniform" qp = 35 qp_density = 2 opt_qp = False disable_dq = True lambda_scale = 0.0 cb_size_ratio = 5000 q_mse = 0.00001 param_opt_flag = False cabac_unary_length_minus1 = 9 partial_data_counter = 0