

ISO/IEC 15938-17:2024-01 (E)

Information technology - Multimedia content description interface - Part 17: Compression of neural networks for multimedia content description and analysis

Contents

| | Page |
|---|-----------|
| Foreword | v |
| Introduction | vi |
| 1 Scope | 1 |
| 2 Normative references | 1 |
| 3 Terms and definitions | 1 |
| 4 Abbreviated terms, conventions and symbols | 3 |
| 4.1 General..... | 3 |
| 4.2 Abbreviated terms..... | 3 |
| 4.3 List of symbols..... | 3 |
| 4.4 Number formats and computation conventions..... | 6 |
| 4.5 Arithmetic operators..... | 6 |
| 4.6 Logical operators..... | 7 |
| 4.7 Relational operators..... | 7 |
| 4.8 Bit-wise operators..... | 7 |
| 4.9 Assignment operators..... | 8 |
| 4.10 Range notation..... | 8 |
| 4.11 Mathematical functions..... | 8 |
| 4.12 Array functions..... | 9 |
| 4.13 Order of operation precedence..... | 11 |
| 4.14 Variables, syntax elements and tables..... | 11 |
| 5 Overview | 13 |
| 5.1 General..... | 13 |
| 5.2 Compression tools..... | 13 |
| 5.3 Creating encoding pipelines..... | 14 |
| 6 Syntax and semantics | 15 |
| 6.1 Specification of syntax and semantics..... | 15 |
| 6.1.1 Method of specifying syntax in tabular form..... | 15 |
| 6.1.2 Bit ordering..... | 16 |
| 6.1.3 Specification of syntax functions and data types..... | 16 |
| 6.1.4 Semantics..... | 17 |
| 6.2 General bitstream syntax elements..... | 18 |
| 6.2.1 NNR unit..... | 18 |
| 6.2.2 Aggregate NNR unit..... | 18 |
| 6.2.3 Composition of NNR bitstream..... | 19 |
| 6.3 NNR bitstream syntax..... | 20 |
| 6.3.1 NNR unit syntax..... | 20 |
| 6.3.2 NNR unit size syntax..... | 20 |
| 6.3.3 NNR unit header syntax..... | 20 |
| 6.3.4 NNR unit payload syntax..... | 25 |
| 6.3.5 Byte alignment syntax..... | 31 |
| 6.4 Semantics..... | 31 |
| 6.4.1 General..... | 31 |
| 6.4.2 NNR unit size semantics..... | 31 |
| 6.4.3 NNR unit header semantics..... | 31 |
| 6.4.4 NNR unit payload semantics..... | 39 |

| | | |
|-----------|---|-----------|
| 7 | Decoding process | 45 |
| 7.1 | General | 45 |
| 7.2 | NNR decompressed data formats | 46 |
| 7.3 | Decoding methods | 47 |
| 7.3.1 | General | 47 |
| 7.3.2 | Decoding method for NNR compressed payloads of type NNR_PT_INT | 47 |
| 7.3.3 | Decoding method for NNR compressed payloads of type NNR_PT_FLOAT | 48 |
| 7.3.4 | Decoding method for NNR compressed payloads of type NNR_PT_RAW_FLOAT | 48 |
| 7.3.5 | Decoding method for NNR compressed payloads of type NNR_PT_BLOCK | 49 |
| 7.3.6 | Decoding process for an integer weight tensor | 50 |
| 8 | Parameter reduction | 51 |
| 8.1 | General | 51 |
| 8.2 | Methods | 51 |
| 8.2.1 | Batchnorm folding | 51 |
| 8.3 | Syntax and semantics | 52 |
| 8.3.1 | Sparsification using compressibility loss | 52 |
| 8.3.2 | Sparsification using micro-structured pruning | 52 |
| 8.3.3 | Combined pruning and sparsification | 52 |
| 8.3.4 | Unstructured statistics-adaptive sparsification | 53 |
| 8.3.5 | Structured sparsification (global and local approach) | 53 |
| 8.3.6 | Weight unification | 53 |
| 8.3.7 | Low rank/low displacement rank for convolutional and fully connected layers | 54 |
| 8.3.8 | Batchnorm folding | 54 |
| 8.3.9 | Local scaling adaptation (LSA) | 54 |
| 9 | Parameter quantization | 55 |
| 9.1 | General | 55 |
| 9.2 | Methods | 55 |
| 9.2.1 | Uniform quantization method | 55 |
| 9.2.2 | Codebook-based method | 55 |
| 9.2.3 | Dependent scalar quantization method | 55 |
| 9.2.4 | Predictive residual encoding (PRE) | 55 |
| 9.3 | Syntax and semantics | 55 |
| 9.3.1 | Uniform quantization method | 55 |
| 9.3.2 | Codebook-based method | 56 |
| 9.3.3 | Dependent scalar quantization method | 56 |
| 10 | Entropy coding | 56 |
| 10.1 | Methods | 56 |
| 10.1.1 | DeepCABAC | 56 |
| 10.2 | Syntax and semantics | 58 |
| 10.2.1 | DeepCABAC syntax | 58 |
| 10.3 | Entropy decoding process | 64 |
| 10.3.1 | General | 64 |
| 10.3.2 | Initialization process | 64 |
| 10.3.3 | Binarization process | 65 |
| 10.3.4 | Decoding process flow | 66 |
| | Annex A (normative) Implementation for NNEF | 73 |
| | Annex B (informative) Implementation for ONNX® | 75 |
| | Annex C (informative) Implementation for PyTorch® | 77 |
| | Annex D (informative) Implementation for TensorFlow® | 79 |
| | Annex E (informative) Recommendation for carriage of NNR bitstreams in other containers | 81 |
| | Annex F (informative) Recommendation for naming method regarding performance metric type | 83 |
| | Annex G (informative) Encoding side information for selected compression tools | 84 |
| | Bibliography | 95 |