ISO/IEC 15938-17:2022-08 (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 Termsanddefinitions 1

4 Abbreviated terms, conventions and symbols 2
4.1 General 2
4.2 Abbreviated terms 2
4.3 List of symbols 3
4.4 Number formats and computation conventions 5
4.5 Arithmetic operators 5
4.6 Logical operators 6
4.7 Relational operators 6
4.8 Bit-wise operators 6
4.9 Assignment operators 7
4.10 Range notation 7
4.11 Mathematical functions 7
4.12 Array functions 8
4.13 Order of operation precedence 9
4.14 Variables, syntax elements and tables 10

5 Overview 11
5.1 General 11
5.2 Compression tools 11
5.3 Creating encoding pipelines 12

6 Syntax and semantics 13
6.1 Specification of syntax and semantics 13
6.1.1 Method of specifying syntax in tabular form 13
6.1.2 Bit ordering 14
6.1.3 Specification of syntax functions and data types 14
6.1.4 Semantics 16
6.2 General bitstream syntax elements 17
6.2.1 NNR unit 17
6.2.2 Aggregate NNR unit 17
6.2.3 Composition of NNR bitstream 18
6.3 NNR bitstream syntax 18
6.3.1 NNR unit syntax 18
6.3.2 NNR unit size syntax 19
6.3.3 NNR unit header syntax 19
6.3.4 NNR unit payload syntax 24
6.3.5 Byte alignment syntax 29
6.4 Semantics 29
6.4.1 General 29
6.4.2 NNR unit size semantics 29
6.4.3 NNR unit header semantics 29

6.4.4 NNR unit payload semantics 36

7 Decoding process 41
7.1 General 41
7.2 NNR decompressed data formats 42
7.3 Decoding methods 42
7.3.1 General 42
7.3.2 Decoding method for NNR compressed payloads of type NNR_PT_INT 43
7.3.3 Decoding method for NNR compressed payloads of type NNR_PT_FLOAT 43
7.3.4 Decoding method for NNR compressed payloads of type NNR_PT_RAW_ FLOAT 43
7.3.5 Decoding method for NNR compressed payloads of type NNR_PT_BLOCK 43
7.3.6 Decoding process for an integer weight tensor 45

8 Parameter reduction 46
8.1 General 46
8.2 Methods 46
8.2.1 Sparsification using compressibility loss 46
8.2.2 Sparsification using micro-structured pruning 46
8.2.3 Combined pruning and sparsification 47
8.2.4 Parameter unification 49
8.2.5 Low rank/low displacement rank for convolutional and fully connected layers 50
8.2.6 Batchnorm folding 50
8.2.7 Local scaling adaptation 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 Weight unification 53
8.3.5 Low rank/low displacement rank for convolutional and fully connected layers 53
8.3.6 Batchnorm folding 53
8.3.7 Local scaling 54

9 Parameter quantization 54
9.1 Methods 54
9.1.1 Uniform quantization method 54
9.1.2 Codebook-based method 54
9.1.3 Dependent scalar quantization method 54
9.2 Syntax and semantics 54
9.2.1 Uniform quantization method 54
9.2.2 Codebook-based method 55
9.2.3 Dependent scalar quantization method 55

10 Entropy coding 55
10.1 Methods 55
10.1.1 DeepCABAC 55
10.2 Syntax and semantics 56
10.2.1 DeepCABAC syntax 56
10.3 Entropy decoding process 59
10.3.1 General 59
10.3.2 Initialization process 60
10.3.3 Binarization process 61
10.3.4 Decoding process flow 61

Annex A (normative) Implementation for NNEF 67

Annex B (informative) Implementation for ONNX® 69

Annex C (informative) Implementation for PyTorch® 71

Annex D (informative) Implementation for TensorFlow® 73

Annex E (informative) Recommendation for carriage of NNR bitstreams in other containers 75

Bibliography 77