ICS 35.040
CCS L 71

中华人民共和国国家标准

GB/T 42382.1—2023

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

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

2023-03-17 发布　　2023-10-01 实施

国家市场监督管理总局
国家标准化管理委员会　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件是GB/T 42382《信息技术　神经网络表示与模型压缩》的第1部分。GB/T 42382已经发布了以下部分：

——第1部分：卷积神经网络。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：北京大学、鹏城实验室、深圳市海思半导体有限公司、赛灵思电子科技(北京)有限公司、杭州海康威视数字技术股份有限公司、北京百度网讯科技有限公司、深圳市腾讯计算机系统有限公司、华为技术有限公司、厦门大学、中国电子技术标准化研究院、中国科学院自动化研究所、浙江大学、中国科学技术大学、上海交通大学、清华大学、中关村视听产业技术创新联盟。

本文件主要起草人：田永鸿、杨帆、纪荣嵘、单弈、陈光耀、燕肇一、郑侠武、浦世亮、谭文明、李哲暘、彭博、钟刚、赵恒锐、段文鸿、胡浩基、李翔、骆阳、王炜、许奕星、李慧霞、林绍辉、王培松、赵依、胡晓光、郑辉煌、蒋佳军、马金成、程健、江帆、朱文武、汪小娟、高文、黄铁军、赵海英、马珊珊。

引　言

神经网络的表示和模型压缩是人工智能技术体系的重要组成部分，是国民经济各行业应用人工智能的前提。然而，多源算法平台不能协同工作，模型不可以相互转换，制约了人工智能技术的传播和应用。为了保证人工智能技术的跨平台可操作性，提升模型复用效果，本标准将对神经网络的表示和模型压缩进行规范，带动人工智能产业的健康、快速发展。GB/T 42382 旨在确立适用于卷积神经网络、大规模预训练网络及图神经网络的神经网络表示与模型压缩的规范，拟由三个部分组成：

——第 1 部分：卷积神经网络。目的在于确立适用于卷积神经网络的表示与模型压缩标准。

——第 2 部分：大规模预训练模型。目的在于确立适用于大规模预训练网络的模型表示、模型压缩及模型传输标准。

——第 3 部分：图神经网络。目的在于确立图数据和图神经网络表示，确定图神经网络模型的编码格式标准。

本文件的发布机构提请注意，声明符合本文件时，可能涉及到 6、7、8、9、10、11 与《神经网络表示标准框架结构》(专利号：201810575097.7)；7.1.4、8.1.3、9.3.2 与《基于神经网络差分的量化方法及系统》(专利号：201910478617.7)；7.2、8.2、9.4 与《一种基于参数范数的神经网络量化方法》(专利号：201810387893.8)；7.4.2、8.4.2、9.6.2 与《一种神经网络计算方法和装置》(专利号：PCT/CN2018/101598)；7.1.3、8.1.2、9.3.1 与《一种神经网络模型、数据处理方法及处理装置》(专利号：PCT/CN2019/085885)、《一种神经网络模型、数据处理方法及处理装置》(专利号：201810464380.2)；7.2.3.3、7.2.4.3、8.2.3.3、8.2.4.3、9.4.1.1、9.4.2.1 与《Neural Network Quantization Method using Multiple Refined Quantized Kernels for Constrained Hardware Deployment》(专利号：PCT/EP2019/053161)；9.2.2 与《图像生成方法、神经网络的压缩方法及相关装置、设备》(专利号：201910254752.3)；7.2.4.1、8.2.4.1、9.4.2.2 与《模型训练方法、装置、存储介质和程序产品》(专利号：PCT/CN2019/129265)；7.2、8.2 与《神经网络模型和训练方法和装置》(专利号：CN202010144315.9)、《神经网络模型的量化方法和装置》(专利号：CN202010143782.X)、《神经网络模型的量化方法和装置》(专利号：CN202010144339.4)、《神经网络模型压缩方法以及装置》(专利号：CN201610943049.X)；6.2、6.3 与《DEEP LEARNING PROCESSING APPARATUS AND METHOD, DEVICE AND STORAGE MEDIUM》(专利号：US17/017,600)；7.4、8.4 与《NEURAL NETWORK DATA PROCESSING APPARATUS, METHOD AND ELECTRONIC DEVICE》(专利号：US16/893,044)；7.2、8.2、9.4 与《基于多比特神经网络非线性量化的深度神经网络压缩方法》(专利号：201910722230.1)；7.3、8.3、9.5 与《一种基于结构化剪枝的高效图像分类方法》(专利号：201910701012.X)；7.3、8.3、9.5 与《一种基于增量正则化的神经网络结构化稀疏方法》(专利号：201910448309.X)；11 与《一种模型数据的处理方法、装置及设备》(专利号：201911230340.2)；7.2.3.4、7.2.4.4、7.2.4.5 与《一种应用有界线性整流单元的全整数神经网络系统》(专利号：2019104537988)；7.4、8.4、9.6 与《基于张量分解的深度卷积神经网络的加速与压缩方法》(专利号：201610387878.4) 相关专利的使用。

本文件的发布机构对于该专利的真实性、有效性和范围无任何立场。

该专利持有人已向本文件的发布机构承诺，他愿意同任何申请人在合理且无歧视的条款和条件下，就专利授权许可进行谈判。该专利持有人的声明已在本文件的发布机构备案，相关信息可以通过以下联系方式获得：

专利持有人：北京大学、华为技术有限公司、北京百度网讯科技有限公司、厦门大学、浙江大学、赵恒锐、中科院自动化研究所

地址：北京市海淀区颐和园路 5 号理科 2 号楼 2604 室 邮编：100871；北京市上地信息路 3 号华为大厦 邮编：100085；北京市海淀区上地十街 10 号百度大厦 邮编：100085；福建省厦门市思明区厦门大学信息学院 邮编：361005；浙江省杭州市西湖区浙大路 38 号浙江大学玉泉校区 邮编：310000；安徽省合肥市蜀山区中国科学技术大学西区 邮编：230027；北京市海淀区中关村东路 95 号 邮编：100190

联 系 人：黄铁军

通讯地址：北京大学理科 2 号楼 2641 室

电子邮件：tjhuang@pku.edu.cn

电话：+8610-62756172

请注意除上述专利外，本文件的某些内容仍可能涉及专利。本文件的发布机构不承担识别专利的责任。

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

1　范围

本文件规定了卷积神经网络离线模型的表示与压缩过程。

本文件适用于各种卷积神经网络模型的研制、开发、测试评估过程,以及在端云领域的高效应用。

注:对于本文件规定的表示与模型压缩方法不要求机器学习框架原生支持,可以通过转换、工具包等形式支持。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;不注日期的引用文件,其最新版本(包括所有的修改单)适用于本文件。

GB/T 5271.34—2006　信息技术　词汇　第34部分:人工智能　神经网络

3　术语和定义

下列术语和定义适用于本文件。

3.1

编解码表示　codec representation

使用压缩技术减少模型的规模。

注:具体定义参考编解码表示,见第10章。

3.2

层　layer

神经网络中的分级结构。

注:每个网络层包含多个个算子,例如输入层,卷积层,全连接层。

3.3

参考随机向量　reference random vector

整个网络共存的基础符号向量。

3.4

多重INT4量化　multiple INT4 quantization

一种量化的方式,将一个张量量化为多个INT4张量的组合的量化方式。

3.5

封装表示　encapsulation representation

支出安全信息和身份验证等接口。

注:具体定义参考模型保护,见第11章。

3.6

分块结构化矩阵　block structured matrix

可以分为多个块,且每个分块均按照某种规律排列的矩阵。