ICS 35.040
CCS L 71

中华人民共和国国家标准

GB/T 42382.2—2026

信息技术　神经网络表示与模型压缩 第2部分：大规模预训练模型

Information technology—Neural network representation and model compression—Part 2：Large scale pre-training model

2026-04-30 发布　　2026-11-01 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2部分。GB/T 42382已经发布了以下部分：

——第1部分：卷积神经网络；

——第2部分：大规模预训练模型；

——第3部分：图神经网络。

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、华为技术有限公司、北京百度网讯科技有限公司、厦门大学、杭州海康威视数字技术股份有限公司、中国电子技术标准化研究院、中国科学院自动化研究所、中科南京人工智能创新研究院、铁塔智联技术有限公司、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、陈光耀、郑侠武、彭军、纪荣嵘、韩凯、胡晓光、燕肇一、张一帆、沈岗、曹刘娟、周奕毅、张玉鑫、马跃萧、吴宇航、谢展豪、倪铭坚、张翀、彭佩玺、马艳军、于佃海、陈秋良、陈泽裕、陈醒濠、唐业辉、王云鹤、蓝朝祥、杨绮明、郑传杨、张凯、彭博、李哲暘、谭文明、任烨、叶挺群、任文奇、冯仁光、周智强、王培松、程健、麻文军、杨雨泽、鲍薇、郑若琳、沈芷月、张伟民、赵海英、黄铁军、高文。

引　言

人工智能领域正在发生迅速的范式转变，深度影响了计算机视觉、自然语言处理、机器人、自动驾驶、智慧医疗等领域的发展。大规模预训练模型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。该标准的目标在于提供大规模预训练模型可能涉及的表示和压缩技术方法参考，提升用户对模型的复用效果。使用时，对于大规模预训练模型的表示方法、传输方法需进行必要的支持，对于压缩技术可根据实际应用场景及技术构成做可选支持，具体的支持方法由后续标准进行补充。对于该标准规定的表示方法不要求平台原生支持，可以通过转换、工具包等形式进行支持，同时相关的定义可转化为与特定计算设备、框架匹配的形式和实现。GB/T 42382 旨在确立适用于不同种类神经网络的表示方法与模型压缩的规范，拟由三个部分组成。

——第 1 部分：卷积神经网络。目的在于确立适用千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。目的在于确立适应多种推理平台和计算要求的大规模预训练模型的基本表示方法与加速压缩过程。

——第 3 部分：图神经网络。目的在于确立适应多种计算要求的高效图神经网络模型的基本表示方法与压缩加速过程。

本文件的发布机构提请注意，声明符合本文件时，可能涉及到第 6 章、第 7 章和第 8 章与神经网络表示框架相关的专利的使用；第 7 章与基于层级特征的网络稀疏化技术、基于可微量化训练的视觉模型压缩技术、基于模型量化的任务处理技术、神经网络模型裁剪技术、基于注意力模型的特征提取技术相关的专利的使用；第 8 章与模型封装与模型解封装技术相关的专利的使用。

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

地址：北京市海淀区颐和园路 5 号；

专利持有人：华为技术有限公司

地址：深圳市龙岗区坂田华为总部办公楼；

专利持有人：厦门大学

地址：福建省厦门市思明区思明南路 422 号；

专利持有人：杭州海康威视数字技术股份有限公司

地址：浙江省杭州市滨江区阡陌路 555 号；

专利持有人：中科南京人工智能创新研究院

地址：江苏省南京市创研路 266 号麒麟人工智能产业园 3 号楼 3 楼；

专利持有人：中国科学院自动化研究所

地址：北京市海淀区中关村东路 95 号；

专利持有人：百度在线网络技术(北京)有限公司

地址：北京市海淀区上地十街 10 号百度大厦。

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2部分:大规模预训练模型

1　范围

本文件规定了适应多种计算机要求的大规模预训练的表示、压缩表示和封装表示,以及其对应的压缩流程、适配流程、封装流程和模型传输与分发。

本文件适用于大规模预训练模型的研制、开发过程,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,均通过转换、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

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

3　术语和定义

GB/T 5271.34—2006界定的以及下列术语和定义适用于本文件。

3.1

预训练模型　pre-trained model

通过自监督或者无监督技术,在大量的训练数据上训练得到初始模型,能被迁移到目标相近的任务中进行使用的一种深度学习模型。

3.2

大规模预训练模型　large scale pre-trained model

大模型　large scale model

大规模深度学习模型　large-scale deep learning model

基于大量数据训练得到,具有复杂计算架构,能够处理复杂任务,且具备一定泛化性的深度学习模型。

注:大模型的参数量由其功能和模态决定,一般不低于1亿。大模型训练使用的数据总量受参数量的影响,达到收敛的大模型的参数量的对数与其训练数据总量的对数成正比。

[来源:GB/T 45288.1—2025,3.1]

3.3

转换器　transformer

基于多头注意力机制,包含残差连接、层归一化和全连接的、能并行处理序列数据的、序列到序列架构(Encoder-Decoder架构)的网络。