同行专家业内评价意见书编号：20240854151

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名： 姜博丰

学号： 22160068

申报工程师职称专业类别（领域）： 电子信息

浙江工程师学院（浙江大学工程师学院）制

2024年03月27日

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.专业基础理论知识和专业技术知识掌握情况

本人基本掌握电子信息专业控制工程方向所需的理论知识，包含微积分、线性代数、概率论与统计、常微分方程、偏微分方程等高等数学知识；包含自动控制理论、模拟电路与数字电路技术、嵌入式系统、现代控制理论等控制工程相关知识；本人英语CET-6分数为561，托福分数为85，英语的听说读写能力良好。本人还熟悉计算机基础知识，包括计算机组成原理、数据结构、机器学习、深度学习、模型压缩等计算机领域的相关知识，掌握C++、Python、JAVA等编程语言的使用，掌握VScode、Pycharm等编辑器的应用。

2.工程实践经历

本人在2022年6月1日-2023年7月20日在杭州大数云智科技有限公司进行专业工程实践，项目名称是基于模型压缩的网络加速推理研究。

3.实际工作中综合运用所学知识解决复杂工程问题的案例

这个项目中，我承担了研究动态阈值量化方法的工作，并成功实现了8位和16位量化神经网络的前向计。通过这个项目，我获得了以下方面的知识和能力提升：

数据集准备：为了进行模型训练和评估，我首先收集了适用于图像分类任务的数据集，如ImageNet。然后，我对数据集进行预处理，包括图像的缩放、裁剪和标准化等操作，以确保数据的质量和一致性

模型选择和优化：在开始实验之前，我研究了不同的神经网络架构，并选择了适合图像分类任务的基准模型，如ResNet或MobileNet。后，我使用常规的浮点数权重初始化模型，并使用传统的反向传播算法进行训练。

理解量化神经网络：通过研究动态阈值量化方法，我深入了解了量化神经网络的原理和技术。我学会了将浮点数参数和活值映射为低精度整数类型，以减少计算和存储销。

动态阈值调整：在该项目中，我学会了如何根据模参数和激活值的分布动态调整量化的阈值范围。通过不断采样和更新阈值范围，我可以改变映射系数，从而提高量化后模型的精度。

激活函数优化：我设计了带有可学习截断参数的激活函数，使其能够在训练过程中动态调节输出阈值。这种方法使得激活函数能够自我学习，进一步提高了量化模型的性能。

批归一化的处理：我研究了批归一化层的均值和方差固定方法，并将其折叠进卷积层，实现了模型的低度计算。这种处理方案有效地减少了浮点计算的需求，提高了量神经网络的计算效率。

通过个企业社会实项目，我获得了许多必要的识和能力，这些是在课堂上无法获的：

实际应用经：通过实现化神经网络并在ImageNet数据集进行验证，我获得了实际应用的经验。我学会了如何处理真实世界中的大规模数据集，并评估量化模型的性能损失情况。

解决问题的能力：在项目中，我面临了许多挑战，例如确定映射系数优化激活函数和处理批归一化层等。通过解决这些问题，我培养了决复杂问题的能力和创新思维。

团队合作与沟通：虽然这个项目是个人小项目，但我也需要与导师和其他团队成员进行交流和讨论这锻炼了我的队合作和沟通能力，使我能够更好地与他人合作，共同完成项目目标。

总之，通过参与个企业社实践项目，我不仅在知识和能方面得到了提升，还获得了实际用经验和解问题的能力。通过这些工作，我获得了丰富的实践经验和技能提升。我学会了处理实际数据集、选择合适的模型结构、实现化方法以及估和优化量化模型的性能。这些经验将对我未来在深度学习人工智能领域的研究和工作中非常有价值。这些是通过课堂学习无法获得的，对我的职业发展和未来的学习都具有要意义。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】

1.
公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
一种适配硬件的深度学习压缩转换框架	发明专利申请	2022年02月28日	申请号：202210186122.5	2/3	
Single-shot pruning and quantization for hardware-friendly neural network acceleration	TOP期刊	2023年08月23日	Engineering Applications of Artificial Intelligence	1/3	

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 85 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1.1 年（要求1年及以上） 考核成绩： 86 分（要求80分及以上）
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ 申报人签名： 姜博丰	

2216068

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀　☐良好　☐合格　☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：　　年　月　日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过　☐不通过（具体原因：　　） 工程师学院教学管理办**公室**审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22160068	姓名：姜博丰	性别：男	学院：工程师学院	专业：电子信息	学制：2.5年
毕业时最低应获：24.0学分	已获得：27.0学分			入学年月：2021-09	毕业年月：2024-03
学位证书号：1033532024602150			毕业证书号：103351202402600376		授予学位：电子信息硕士

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2021-2022学年秋季学期	研究生英语基础技能		1.0	免修	公共学位课	2021-2022学年春季学期	数学建模		2.0	90	专业选修课
2021-2022学年冬季学期	传感器前沿技术及应用		2.0	90	专业选修课	2021-2022学年夏季学期	机器人智能控制		3.0	82	专业学位课
2021-2022学年秋冬季学期	中国特色社会主义理论与实践研究		2.0	89	公共学位课	2021-2022学年夏季学期	自然辩证法概论		1.0	84	公共学位课
2021-2022学年冬季学期	标准与知识产权		2.0	96	专业选修课	2021-2022学年夏季学期	工程伦理		2.0	79	公共学位课
2021-2022学年秋冬季学期	研究生论文写作指导		1.0	92	专业学位课	2021-2022学年春夏季学期	工程技术发展前沿		2.0	95	专业学位课
2021-2022学年冬季学期	智能工业机器人		2.0	81	专业学位课	2023-2024学年秋季学期	游泳		1.0	优	公共素质课
2021-2022学年冬季学期	研究生英语		2.0	免修	公共学位课	2023-2024学年秋季学期	歌唱艺术		1.0	90	公共素质课
2021-2022学年春季学期	人工智能制造技术		2.0	93	专业学位课	2023-2024学年秋季学期	攀岩		1.0	70	公共素质课

说明：1.研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2.备注中“*”表示重修课程。

学院成绩校核章：

成绩校核人：张梦依

打印日期：2024-04-02

(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114595817 A
(43)申请公布日 2022.06.07

(21)申请号 202210186112.5

(22)申请日 2022.02.28

(71)申请人 浙江大学
地址 310000 浙江省杭州市西湖区余杭塘路866号

(72)发明人 刘勇 姜博丰 陈军

(74)专利代理机构 杭州泓呈祥专利代理事务所(普通合伙) 33350
专利代理师 张婵婵

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)

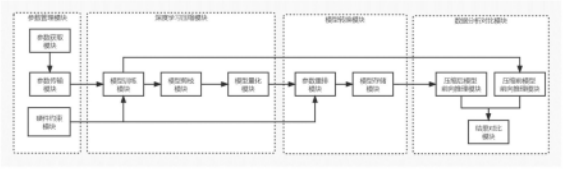
权利要求书2页 说明书6页 附图1页

(54)发明名称

一种适配硬件的深度学习压缩转换框架

(57)摘要

本发明公开了一种适配硬件的深度学习压缩转换框架，包括参数管理模块，与深度学习压缩模块、模型转换模块分别连接，用于获取用户自定义的参数并传输给深度学习压缩模块，将硬件资源信息传输给深度学习压缩模块，将硬件的接口信息传输给模型转换模块；深度学习压缩模块，与模型转换模块、数据分析对比模块分别连接，利用用户自定义的参数结合硬件资源信息，训练生成深度学习目标检测模型后进行压缩处理；模型转换模块，与数据分析对比模块连接，按照用户需要的格式，将压缩处理后的深度学习目标检测模型进行格式转换并存储；所述数据分析对比模块，将压缩前的深度学习目标检测模型、压缩并格式转换后的深度学习目标检测模型进行比较并输出结果。



CN 114595817 A

Engineering Applications of Artificial Intelligence 126 (2023) 106816

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Single-shot pruning and quantization for hardware-friendly neural network acceleration

Bofeng Jiang[1], Jun Chen[1,*], Yong Liu[*]

Institute of Cyber-Systems and Control, Zhejiang University, China

ARTICLE INFO

Keywords:
Pruning
Quantization
Hardware-friendly
CNN acceleration

ABSTRACT

Applying CNN on embedded systems is challenging due to model size limitations. Pruning and quantization can help, but are time-consuming to apply separately. Our Single-Shot Pruning and Quantization strategy addresses these issues by quantizing and pruning in a single process. We evaluated our method on CIFAR-10 and CIFAR-100 datasets for image classification. Our model is 69.4% smaller with little accuracy loss, and runs 6–8 times faster on NVIDIA Xavier NX hardware.

1. Introduction

Convolutional Neural Networks (CNNs) are currently widely used in computer vision. The widespread use of CNNs is due to the outstanding computational performance and the growth of large-scale datasets. The performance of CNNs is increasing, and the recognition error rate is decreasing, but at the same time, their spatial and temporal complexity is increasing; the number of parameters and the number of computational operations during network training are also increasing. For example, VGG-16 has up to 138 million parameters, and its overall model size is over 500 M (Simonyan and Zisserman, 2014). It requires 15.5 billion floating-point operations to classify a single image. Moreover, the introduction of ResNet (He et al., 2016) solved the problem of degradation that occurs when the model depth is increased, thus raising the parameter and computation levels of the model to unprecedented heights. Furthermore, as the number of parameters increases, the cost of both CNN training and inference is rising. For high-performance inference devices such as GPUs, this is not a difficult task, but for inference platforms with limited resources, high computational costs and high performance requirements make performing visual tasks difficult. To address this issue, one possible solution is to reduce the cost of training and inference. By implementing techniques that can lower the cost of these processes, we can improve the efficiency and effectiveness of CNNs, which is good to energy structure on carbon emissions and energy storage (Zhang et al., 2023; Licheng and Wang, 2022; Yu et al., 2023).

At the same time, the Internet of Things (IoT) development allows small models to extend deep learning to a more extensive application space. The need for image classification, target detection and OCR (optical character recognition) text recognition algorithms in robots, drones, and other mobile embedded devices is constantly on the rise, necessitating algorithms that are highly accurate and have minimal latency. Researchers strive to meet these demands by continuously delving into cutting-edge technologies and exploring novel methodologies. Their aim is to develop more sophisticated algorithms that allow mobile devices to operate more intelligently and serve people better. Moreover, the compression of neural networks holds a crucial significance in enhancing the cognitive (Li et al., 2022) as well as perceptual facets (Li et al., 2023) of computer vision. How to compress and employ CNN models on embedded devices without compromising accuracy has become a frontier hotspot for network structure optimization.

Quantization and pruning are commonly used to reduce the number of model parameters and computation operations. Quantization is to convert parameters from 32-bit floating point numbers to 16-bit floating point numbers or 8-bit integer to reduce memory occupied by parameters and operation run-time (Jacob et al., 2018). Pruning is adding a judging mechanism to the network training process, eliminating unimportant connections, filters, and layers, to achieve the purpose of streamlining the network structure (Han et al., 2015a)

Current pruning methods are divided into weight pruning, channel pruning, and inter-layer pruning according to granularity (Guo et al., 2016). Weight pruning is a kind of sparse pruning. The convolutional kernel obtained from pruning has unstructured characteristics. However, this unstructured convolutional kernel requires a particular hardware configuration to hit the acceleration effect. Inter-layer pruning reduces the network's depth, effectively reducing the number of network parameters, but the performance degradation is very problematic (Anwar and Sung, 2016). The performance of the model and

* Corresponding authors.
E-mail addresses: 22160068@zju.edu.cn (B. Jiang), junc@zju.edu.cn (J. Chen), yongliu@iipc.zju.edu.cn (Y. Liu).
[1] Bofeng Jiang and Jun Chen contributed equally to this work.

https://doi.org/10.1016/j.engappai.2023.106816
Received 4 January 2023; Received in revised form 3 June 2023; Accepted 13 July 2023
Available online 27 August 2023

经检索《Web of Science》、《Journal Citation Reports（JCR）》及《中国科学院文献情报中心期刊分区表》数据库，《Science Citation Index Expanded (SCI-EXPANDED)》收录论文及其期刊影响因子、分区情况如下。（检索时间：2024 年 1 月 2 日）

第 1 条，共 1 条

标题: Single-shot pruning and quantization for hardware-friendly neural network acceleration

作者: Jiang, BF(Jiang, Bofeng);Chen, J(Chen, Jun);Liu, Y(Liu, Yong);

来源出版物:ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE **卷:**126 **文献号:**106816 **提前访问日期:**AUG 2023 **DOI:**10.1016/j.engappai.2023.106816 **出版年:**NOV 2023

入藏号:WOS:001066913400001

文献类型: Article

地址:

[Jiang, Bofeng; Chen, Jun; Liu, Yong] Zhejiang Univ, Inst Cyber Syst & Control, Hangzhou, Peoples R China.

通讯作者地址:

Chen, J; Liu, Y (corresponding author), Zhejiang Univ, Inst Cyber Syst & Control, Hangzhou, Peoples R China.

电子邮件地址:22160068@zju.edu.cn; junc@zju.edu.cn; yongliu@iipc.zju.edu.cn

IDS 号:R8PE5

ISSN:0952-1976

eISSN:1873-6769

期刊《ENG APPL ARTIF INTEL》2022 年的影响因子为 8.0，五年影响因子为 7.4。

期刊《ENG APPL ARTIF INTEL》2022 年的 JCR 分区情况为：

Edition	*JCR® 类别*	*类别中的排序*	*JCR 分区*
SCI	AUTOMATION & CONTROL SYSTEMS	7/65	Q1
SCI	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE	25/145	Q1
SCI	ENGINEERING, ELECTRICAL & ELECTRONIC	30/275	Q1
SCI	ENGINEERING, MULTIDISCIPLINARY	5/90	Q1
SCI	ENGINEERING	N/A	N/A
SCI	ROBOTICS & AUTOMATIC CONTROL	N/A	N/A

期刊《ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE》2023 年升级版的中科院期刊分区情况为：

刊名	ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE
年份	2023
ISSN	0952-1976

	学科	分区	Top 期刊
大类	计算机科学	2	是
小类	ENGINEERING, MULTIDISCIPLINARY 工程：综合	1	-
小类	AUTOMATION & CONTROL SYSTEMS 自动化与控制系统	2	-
小类	COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE 计算机：人工智能	2	-
小类	ENGINEERING, ELECTRICAL & ELECTRONIC 工程：电子与电气	2	-

注：

1. 期刊影响因子及分区情况最新数据以 JCR 数据库、《中国科学院文献情报中心期刊分区表》最新数据为准。

2. 以上检索结果来自 CALIS 查收查引系统。

3. 以上检索结果均得到委托人及被检索作者的确认。

教育部科技查新工作站（Z09）

检索人（签章）：杨柳

2024 年 1 月 2 日