

**论文答辩情况和学位授予决议书**

**答 辩 人**  阮晓峰

**学 号**  201718014628053

**论文题目** 深度卷积神经网络压缩与加速方法研究

**指导教师** 胡卫明 研究员

**学位类别** 工学博士

**学科(专业)** 模式识别与智能系统

**研究方向** 计算机视觉

**培养单位** 中国科学院自动化研究所

**中国科学院大学制**

**填 表 说 明**

1. 本表内容须真实、完整、准确。
2. “学位类别”名称填写：哲学博士、教育学博士、理学博士、工学博士、农学博士、医学博士、管理学博士，哲学硕士、经济学硕士、法学硕士、教育学硕士、文学硕士、理学硕士、工学硕士、农学硕士、医学硕士、管理学硕士，工程硕士、工商管理硕士等。
3. “学科(专业)”名称：学术型学位填写“二级学科”全称，专业学位填写“培养领域”全称。
4. 本表“学位授予决议书”中：学术型学位填写“同意（或不同意）授予博士学位”、“同意（或不同意）授予硕士学位”。

**学位论文答辩记录**

|  |  |  |  |
| --- | --- | --- | --- |
| 学位论文题目 | 深度卷积神经网络压缩与加速方法研究 | | |
| 答辩日期 | 2021.05.26 | 答辩地点 | 中国科学院自动化研究所  智能化大厦17层第七会议室 |
| 用简明语言记录答辩过程，答辩委员会提出问题和研究生回答情况：  (彭宇新教授)  问题1、论文中提出的三种深度卷积神经网络压缩方法可否联系起来？  回答：在本论文提出的三种算法中，我们都采用了松弛的稀疏优化方式，自适应地获得轻量化的网络结构。在算法1中，我们通过在原始网络中构造一个可压缩的模块，采用通用的稀疏正则方式，可以将分解与剪枝方法有效统一；在算法2中，我们改进了通用的稀疏正则方式，提出了一种动态渐进式稀疏正则方式。在算法3中，我们充分利用模型本身的特征表征信息，将提出的动态稀疏正则方式进行了扩展，应用到了无类别标签数据场景下的模型压缩与加速任务中。算法3扩展了算法2，算法1可以使用算法2提出的动态稀疏正则。  问题2、表格目录页标题太长，建议缩短标题？  回答：已在论文中修改。  (陆耀教授)  问题3、如何评价模型的压缩结果？  回答：在论文中，我们使用参数减少率、FLOPs减少率以及实际加速倍率来反映模型压缩的力度，同时需要考虑模型的精度，即需要平衡压缩力度与模型性能。实验部分中，在保持baseline精度前提下，我们会给出模型参数、FLOPs减少率以及实际加速倍率；在较大的压缩率下（比如40%以上），我们也会给出模型的精度。  (王文成研究员)  问题4、网络层数越深，性能越强，压缩后的网络层数是否可以减少？  回答：在结构化剪枝中，可以对网络整个层进行剪枝，比如，在我们所提出的算法中，如果某一个层所有通道对应的权值很小，这个时候就可以将整个层移除。  问题5、分类任务压缩后的模型迁移到其他任务，效果如何？  回答：深度神经网络压缩算法通常是在分类数据集下获得轻量化的模型，然后将其作为backbone应用到下游任务。在提出的算法中，我们将压缩后的模型应用到检测任务，在性能下降不到1%的前提下，速度有了一定的提升，验证了所提出的算法具有良好的泛化性能。目前也有一些直接针对特定压缩任务的模型压缩方法，可能会获得更优秀的性能，是未来值得探究的方向。  (孙立峰教授)  问题6、在第二个工作中，动态渐进式稀疏正则什么时候停止？  回答：在第二个工作中，施加的动态渐进式稀疏正则发生在整个训练过程中。为了满足预设剪枝率，我们只在预期稀疏的参数上施加结构化稀疏正则，当网络稀疏率满足预设剪枝率时，稀疏的参数保持稳定。  问题7、压缩后的模型在终端设备上需要考虑什么问题？  回答：在终端设备上，通常会使用一些推理加速工具，比如Nvidia推出的TensorRT，同时还会和一些量化技术相结合，以便获得更好的加速效果。  (张文生研究员)  问题8、卷积核的核心是提取特征，剪枝后卷积核哪些部位比较重要？  回答：本论文提出的剪枝算法是一种粗粒度的裁剪方法，直接将整个通道对应的所有卷积核移除。在确定冗余通道时，会考虑滤波器之间的相关性。一些细粒度的裁剪方法会考虑卷积核裁剪后的模式，通常靠近卷积核中心的权值比较重要。  (李兵研究员)  问题9、工作3中的基准模型是否需要预训练？  回答：在工作3中，压缩子模型需要学习基准模型的特征表征信息，因此基准模型需要预训练。  答辩委员会记录人（签字）： | | | |

注：可另附页

**答辩委员会决议（限1200字）**

|  |
| --- |
| 答辩委员会对论文的学术评语（论文选题意义，论文创新性成果及学术水平；论文写作规范化和逻辑性；论文存在的主要不足之处，答辩情况。）:  深度神经网络压缩与加速是近年来计算机视觉与机器学习领域的一个热门研究课题，论文选题具有重要的理论意义和应用价值。  论文取得如下创新性研究成果：  1. 提出了一种基于有效统一分解与剪枝的深度卷积神经网络压缩方法，利用稀疏正则化实现了低秩分解和通道稀疏化两个任务的联合优化，并利用剪枝合并操作，进一步减少了网络冗余，实验表明了该方法的有效性。  2. 提出了一种基于渐进式稀疏正则的深度卷积神经网络剪枝方法，通过渐进式稀疏正则和动态稀疏的迭代阈值收缩算法，实现了预设剪枝率的结构化稀疏网络高效生成，简化了整个结构化剪枝过程。  3. 提出了一种基于模型特征学习增强的动态剪枝方法，利用基准模型的特征表征信息以及压缩子模型之间的互补信息，提高了基于无类别标签数据的端到端模型压缩效果。  论文工作表明，作者在本学科领域掌握了坚实宽广的基础理论和系统深入的专门知识，独立从事科研工作的能力强。论文写作规范、条理清晰、逻辑性强，是一篇优秀的博士学位论文。答辩过程中讲述清楚，回答问题正确。答辩委员会经无记名投票，一致同意通过阮晓峰的博士论文答辩，并建议授予阮晓峰工学博士学位。  答辩委员会主席（签字）：  年 月 日 |

**答辩委员会组成**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 答辩委员会成员 | 姓 名 | 专业技术职 务 | 是否博导 | 研究领域 | 工作单位 | 签 字 |
| 主 席 | 彭宇新 | 教授 | 是 | 跨媒体分析 | 北京大学 |  |
| 委 员 | 陆耀 | 教授 | 是 | 计算机视觉、模式识别与跟踪 | 北京理工大学 |  |
| 王文成 | 研究员 | 是 | 计算机图形学 | 中国科学院软件研究所 |  |
| 孙立峰 | 教授 | 是 | 网络多媒体 | 清华大学 |  |
| 张文生 | 研究员 | 是 | 人工智能理论与方法、社会计算与大数据解析 | 中国科学院自动化研究所 |  |
| 李兵 | 研究员 | 是 | 图像与视频处理 | 中国科学院自动化研究所 |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 表决结果 | 记录 | 投票 6 人， 其中同意 6 票，不同意 0 票, 弃权 0 票  综合评价意见：优秀 6 票，良好 0 票，中 0 票，差 0 票 | | | | |
| 结论 | ☑ 通过 □修改论文重新答辩 □ 不通过 | | | | |
| 答辩委员会秘书（签字）： 答辩委员会主席（签字）：  年 月 日 | | | | | |