电 子 科 技 大 学

专业学位研究生学位论文开题报告表

攻读学位级别： □博士 ☑硕士

培养方式： ☑全日制 □非全日制

专业学位类别及领域： 电子信息

学 院： 计算机科学与工程学院

学 号： 202022060814

姓 名： 张创

论文题目： 基于深度学习的轻量化

目标检测算法研究

校内指导教师： 曾金全

校外指导教师：

填表日期： 2021 年 12 月 25 日

电子科技大学研究生院

1. 学位论文研究内容

|  |  |  |
| --- | --- | --- |
| 课题类型 | | □应用基础研究 ☑应用研究 |
| 课题来源 | | □纵向 □横向 ☑自拟 |
| 学  位  论  文  研  究  内  容 | 学位论文的研究目标、研究内容及拟解决的关键性问题（可续页）   1. **研究目标**   目标检测是计算机视觉中的三大传统任务之一，广泛地应用在军事，安全，交通，医疗和生活领域。随着训练可用数据量的增长与计算平台处理能力的增强，基于深度学习的智能模型能够完成越来越复杂的任务，其在目标检测领域已经取得重大的突破。然而，这些深度模型具有庞大的参数规模，与此相伴的可畏的计算开销与内存需求使其在计算能力受限平台(例如移动嵌入式设备)的部署中遇到了巨大的困难与挑战。本文着手有设计一个可以部署在这些移动端设备上的轻量化目标检测模型，通过参数剪切、参数量化与参数共享等主流的模型压缩与加速方法降低模型大小，减少模型参数，提高检测的实时性，再结合最新提出的注意力机制算法，旨在增加少量模型参数量和计算复杂度的条件下，提高模型的检测准确率。   1. **研究内容**   针对轻量化目标检测算法，主要从以下几个方面进行研究：   1. 模型压缩与加速   模型压缩与加速方法，主要可以分为压缩参数和压缩结构两种方法。压缩参数主要包括参数剪枝，参数量化，低秩分解和参数分享等方法。压缩结构的方法主要包括使用紧凑网络和知识蒸馏的方法。本文中将采用结合参数剪枝，参数量化和紧凑网络的混合型模型压缩与加速方法：通过结合实际的轻量化要求以及对于模型准确度的要求，设计参数重要性评价准则，并基于此评价准则判断网络模型中参数的重要程度，删除冗余参数，并将网络参数从32位全精度浮点整形量化到更低位数。并从卷积核、特殊层和网络结构入手设计和优化网络模型。   1. 注意力机制   注意力机制本质上是一种资源的重分配机制，通过加入少量的参数，引导神经网络模型向重要的部分分配更多的计算资源，已达到在增加少量模型参数量和计算复杂度的条件下，提高模型的检测准确率。本文将采用一种空间注意力机制和通道注意力机制混合的注意力机制模型，用于轻量化模型后准确度的补充。   1. **拟解决的关键性问题**   对于轻量化网络模型的一个关键性问题在于，轻量化的网络模型意味着更浅的模型深度，更少的卷积核，以及更少的参数，这样的结果使得模型更加有利于布置到移动端的嵌入式设备中，但是也造成了模型对于目标特征的提取不够充分，检测准确度较低，于是如何保证在对于网络进行轻量化优化的同时，还能保证一个可接受的模型准确度，所以主要着手从模型压缩的过程和注意力机制两个方面保证轻量化模型的准确度。  （1）模型压缩与加速方法  在模型压缩与加速的算法中，采用结构化剪切的方法，剪切的颗粒度较粗，剪切的最小单位为filter内参数的组合，并且应为结构化剪切后的模型比较规整，便于后续的在现有软/硬件上的有效加速，而且结合设计参数重要性评价准则，尽量保留对准确度贡献大的参数结点  （2）注意力机制  采用一种空间注意力机制和通道注意力机制混合的注意力机制模型，用于在引入少量参数的情况下，增加轻量化模型的准确度。 | |

1. 学位论文研究依据

|  |
| --- |
| 学位论文的选题依据和研究意义，国内外研究现状和发展态势；选题在理论研究或实际应用方面的意义和价值；主要参考文献，以及已有的工作积累和研究成果。（2000字）   1. **选题依据和研究意义**   目标检测是计算机视觉的基础问题之一，目的就是在一副图像中以边界框的形式定位出相关物体的位置并识别出他们的类型。作为计算机视觉和图像理解的基石，目标检测是解决实例分割、目标追踪、事件检测、活动识别和场景理解等更复杂更深层次视觉任务的基础。  近几年随着深度学习和物联网的结合，人们开始逐渐体验到人工智能带来的便利，对于智能设备的需求日益增长。物联网为人工智能落地应用提供了一个良好的平台，其为人工智能技术提供大量的实景数据支撑，促进人工智能技术发展与改进。而人工智能技术在物联网中的应用推动了物联网的智慧化，使得物联网变得更有价值。人工智能技术的主要技术代表是深度学习，深度学习部署于物联网主要通过两种方式，第一种是让物联网中的边缘设备将数据传输到服务计算中心，服务器通过计算将结果返回给边缘设备，这种方式需要一个庞大的服务集群来处理各种服务请求，当使用量到达一定程度后，服务计算中心将承受巨大的计算负载和带宽负载。第二种方式是将深度学习模型部署到那些内存和计算资源有限的边缘设备，这种方式相对于第一种方法的优点在于不需要传输大量的数据，大大缓解带宽压力，同时因为在本地计算，不用考虑其网络延迟，加速了响应的速度。但部署需要庞大的计算和内存消耗的深度学习模型对于边缘设备来说是十分困难和昂贵的，尤其在近几年深度学习模型都在往更大的模型方向发展，深度学习模型将会产生更大的计算消耗和内存消耗，这样会造较差的用户体验。因此，在保持一定的精准度下研究一种轻量级的目标检测网络结构,对于未来在移动端嵌入式设备上部署目标检测模型有着重大意义。   1. **国内外研究现状和发展态势**   得益于计算机硬件的持续进步和反向传播算法的不断完善，神经网络在众多深度学习算法中脱颖而出，并且神经网络技术在多个计算机视觉任务中表现出色，包括分类，检测，分割等，其中以卷积神经网络最为突出，受到了广泛关注和研究。现今各种新型的神经网络算法或多或少的受到了基于 CNN 思想的启发和影响，甚至其中的多数都被认为是基于 CNN 的一个变体。2012 年 Hinton 成功用 CNN 构建了一个图像分类网络 AlexNet[1]，并获得了当年的ImageNet图像识别大赛第一名，AlexNet首次把Relu激活函数应用到图像分类网络中，并采用 GPU 进行图像加速计算，在网络尾部使用Dropout层降低过拟合，这些开创性的工作影响了后续的研究，给之后的发展指出了一条道路。从那之后，神经网络和新的深度学习算法就开始进入一个爆发性的发展周期，涌现了许多优秀的研究。  牛津大学在2004年提出了VGGNet[2]，该网络骨架部分借鉴了 AlexNet 的网络设计，夺得了该年度的ILSVRC大赛定位项目第一名和分类项目第二名，该网络充分证明了小卷积以及快速扩大网络深度对于特征提取和识别的重要性。  得益于神经网络的发展，目标检测领域在近些年也涌现了许多优秀的研究。目标检测起初都使用传统方法，通过相关关键点特征进行目标检测，出现了SVM[3]，DPM[4]等方法。从R-CNN[5]网络的提出开始，神经网络开始大规模的应用在目标检测中，R-CNN网络使用深层神经网络进行目标类别特征的提取和学习，通过Anchor锚点的选择，确定候选框对目标区域进行选择和判定，使目标检测进入深度学习时代。R-CNN系列网络经过几次改进已经是比较成熟的目标检测算法，具有良好的通用性和高精度，但是由于其两阶段算法结构的影响，检测速度较慢，后续出现了YOLO[6-9]系列网络，将目标类别判定和位置信息放在一个网络中进行回归训练，极大的加快了网络的检测速度，同时也保证了检测精度。  深度学习和人工智能技术的最新突破已实现了在众多移动设备上的部署。前沿的计算方法依赖于移动感测和云计算，而在移动设备上实施的深度学习则提供了许多优势，诸如低通信带宽，较小的云计算资源成本，快速的响应时间以及改进的数据保密性。最近在移动和嵌入式设备[10-11]上进行深度学习的研究与开发引起了广泛地关注。  网络结构优化指的是重新设计一种更少网络连接模型结构，从而实现计算量和参数量的减少，对于网络结构优化，一般采用的方法有：分组卷积、瓶颈结构、小尺寸卷积核。比如 F. N. Iandola等人提出的SqueezeNet[12]，其在SqueezeNet中提出的firemodule模块使用了瓶颈结构和使用 1x1 大小的卷积核替换部分 3x3 的卷积核。Google 公司的研究员们在 Mobilenet 系列[13-14]和 Xception[15]中提出了用深度可分离卷积(Depthwise Separable Convolution, DSC)替代常规卷积，DSC 将卷积分解为逐深度卷积(Depthwise, DW)和逐点卷积(Pointwise, PW)，在大量减少参数量和运行时间时保持模型的精度。旷世的研究员们提出了 ShuffelNet [16][17]系列轻型模型，采用分组卷积减少计算量，同时考虑分组会增加内存访问次数增加整体运行时间，去确定合适的分组组数；同时使用通道混洗加强分组之间的信息交流来加强特征的有效性，获得更好的效果。  模型剪枝则是去掉模型网络结构中不重要的连接，使得网络可以以更少的计算和内存消耗获得与原模型相当的水平。早在1990年LeCun就提出了最优脑损失[18](Optimal Brain Damage, OBD)去剪枝参数获取更高的准确率。Babak Hassibi[19]在OBD基础上，提出了对剪枝过后的模型参数进行恢复性更新，提高了剪枝后模型的泛化能力，也形成了当前模型剪枝的主要流程：训练、剪枝、恢复训练。一些研究人员认为可以根据参数的绝对值值去决定参数的重要性，Han S[20]将低于某一阀值的连接裁剪；Hu[21]等采用神经元激活为 0 的平均比例(Average Percentage of Zors, APoZ)作为裁剪标准；Luo[22]等人提出一种基于激活响应熵值的方法，裁减掉熵值晓得通道；Song Han在Deep Compression[23]中结合剪枝，量化和哈夫曼编码将 AlexNet 压缩了35倍。以上方法都是针对单个神经元或者卷积核层级进行的非结构化裁剪，经过裁剪之后的模型需要专门的计算库才能体现加速的效果。He[24]等人针对通道层级(Channel Pruning)提出了一种结构化剪枝方式，通过Lasso 回归的方法对卷积通道进行选择并删除。Hao Li[25]根据通道内所有卷积核参数的绝对值之和作为判断标准，删除数值较小的通道。Zhuang Liu 在 Network Slimming[26]中根据 BN 层前后映射的关系，以 Gamma 系数绝对值为标准，裁剪掉绝对值较小的低响应通道。  Gong 等人[27]的研究工作表明，网络剪枝和量化可以有效降低网络复杂性并解决过度拟合问题。在发现剪枝可以将正则化带入神经网络并因此提高泛化性之后，人们对压缩DNN[28]进行了广泛地研究。模型剪枝分为细粒剪枝，卷积核级别剪枝，向量级和内核级剪枝，群组级别剪枝。  细粒剪枝方面，Guo[29]提出了一个动态网络手术框架，该框架由两个操作组成：剪枝和剪接。剪枝操作旨在剪枝那些不重要的参数，而拼接操作旨在恢复错误剪枝的连接。该算法能够在更少的训练时间中获得更好的压缩率，但是这属于一种非结构化的压缩行为，因此压缩效率不高，且在硬件上的运行效率也不高。  卷积核级别剪枝，拟合器连接到通道号，He 等人[30]为每个卷积核引入了一个选择权重 β，然后在β上添加了稀疏约束。在 MobileNet-v1[31]中也有类似的行为，将通道数目缩减一些，从而进行网络瘦身。  向量级和内核级剪枝的相关的工作相对较少。Anwar[32]提出跨度剪枝子向量的方法。Mao 等人[33]研究了剪枝中不同的粒度级别，发现矢量级别的剪枝比细粒度的剪枝占用更少的存储空间，因为矢量级别的剪枝需要较少的索引来指示剪枝的参数。与细粒度的剪枝相比，它更像是结构性的剪枝，并且对内存访问更友好，因此在硬件实现中效率更高。这种算法更像是有组织的，成批型的剪枝算法，简单有效，在硬件上更有效。   1. **主要参考文献** 2. Technicolor T, Related S, Technicolor T, et al. ImageNet Classification with Deep Convolutional Neural Networks . 3. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014. 4. Burges C . A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998, 2(2):121-167. 5. Felzenszwalb, Pedro, F, et al. Object Detection with Discriminatively Trained Part-Based Models.[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 32(9):1627-1645. 6. Girshick R , Donahue J , Darrell T , et al. rich feature hierarchies for accurate object detection and semantic segmentation tech report[J]. 2017. 7. Redmon J , Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016. 8. Redmon J , Farhadi A . YOLO9000: Better, Faster, Stronger[C] IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525. 9. Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018. 10. Bochkovskiy A,Wang C Y,Liao H. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020. 11. 冒睿瑞,江波.面向嵌入式的高实时微小目标跟踪检测方法[J/OL].计算机工程:1-9[2021-02-25], 2021. 12. 崔家华, 张云洲, 王争, 刘及惟. 面向嵌入式平台的轻量级目标检测网络[J].光学学报,2019,39(04):307-313. 13. Howard A, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional NeuralNetworks for Mobile VisionApplications[J]. arXiv:1704.04861, 2017. 14. Sandler M, Howard A G, Zhu M, et al. MobileNetsV2: Inverted Residuals andLinear Bottlenecks[C]. Computer Vision and Pattern Recognition (CVPR), 2018,4510-4520. 15. Franc, ois Chollet, Google. Xception: Deep Learning with Depthwise SeparableConvolutions[C]. Computer Science. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 1063-6919. 16. Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun. ShuffleNet: An ExtremelyEfficient Convolutional Neural Network for Mobile Devices[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,6848-6856. 17. Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, Jian Sun. ShuffleNet V2:Practical Guidelines for Efficient CNN Architecture Design[C]. European Conference on Computor Vision (ECCV), 2018,122-138. 18. Lecun Yann, John Denker, Sara A Solla. Optimal Brain Damage[C]. Advances in Neural Information Processing Systems (NIPS) , 1989, 2:598-605. 19. Babak Hassibi, David G.Stork. Optimal Brain Surgeon and general network pruning[C]. Neural Information Processing Systems (NIPS), 1993, 0-7803-0999-5. 20. Han S, Pool J, Tran J, et al. Learning both weights and connections for efficient neural network[C]. Advances in Neural Information Processing Systems. 2015,1135-1143. 21. Hu H, Peng R, Tai Y W, et al. Network trimming: a data-driven neuron pruning approach towards efficient deep architectures[J]. 2017, arXiv:1607.03250[cs.NE]. 22. Luo J H, Wu J. An entropy-based pruning method for CNN compression [J].arXiv:1706.05791, 2017. 23. Song Han, Huizi Mao, William J. Dally. Deep Compression: Compression Deep Neural Networks with Pruning. Trained Quantization and Huffman Coding[C].International Conference on Learning Representations (ICLR), 2016. 24. Y. He, X. Zhang and J. Sun, "Channel Pruning for Accelerating Very Deep Neural Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice,2017, 1398-1406. 25. Hao Li, Asim Kadav, Igor Durdanovic, et al. Pruning Filters for Efficient ConvNets[J]. International Conference on Learning Representations(ICLR), 2017. 26. Zhuang Liu, Jianguo Li, Zhiqiang Shen, et al. Learning Efficient ConvolutionalNetworks through Network Slimming[C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017,978-1-5386-1032-9. 27. Gong Y, Liu L, Yang M, et al. Compressing deep convolutional networks using vector quantization[J]. Computer Science, 2014. 28. N. Najva,K. Edet Bijoy. SIFT and Tensor Based Object Detection and Classification in Videos Using Deep Neural Networks[J]. Procedia Computer Science,2016,93: 29. Han S, Mao H , Dally W J . Deep Compression: compressing deep neural networks with pruning, Trained Quantization and Huffman Coding[J]. Fiber, 2015, 56(4):3--7. 30. Yiwen Guo et.al. Dynamic network surgery for efficient dnns[J]. Neural Information Processing Systems. 2016. 31. He Y , Zhang X , Sun J . Channel Pruning for Accelerating Very Deep Neural Networks[C]. International Conference on Computer Vision, Venice, Italy, 2017. 32. Andrew G, Howard K, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861, 2017. 33. Anwar S , Hwang K , Sung W . Structured pruning of deep convolutional neural networks[J]. ACM Journal on Emerging Technologies in Computing Systems, 2015, 13(3).   [33] Mao H , Han S , Pool J , et al. Exploring the regularity of sparse structure in convolutional neural networks[J]. Tensor Methods in Computer Vision, 2017.   1. **已有的工作积累** 2. 参与课程实践《基于CNN网络的人脸戴口罩识别》，对于深度学习模型的搭建，优化以及调参都有一定的实践经验。 3. 参与修改和完善一篇专利《基于yolo和小目标识别的管线阀门渗漏检测方法》，并且阅读了Yolo系列论文，对于以yolo为代表的目标检测算法有理论上的积累。 |

1. 学位论文研究计划及预期目标

|  |
| --- |
| 1.拟采取的主要理论、研究方法、技术路线和实施方案（可续页）   1. **拟采取的主要理论**    1. **卷积算子**   卷积神经网络是一类包含卷积算子的深度前馈神经网络，其主要由卷积层、池化层以及全连接层等构成，使用反向传播算法对神经网络进行优化。经过多年的研究，卷积算子在标准卷积的基础上推陈出新，出现了一系列新的卷积运算。以下是对一些卷积算子的介绍  **1.1.1 标准卷积**  在数学分析领域，卷积是一种重要的运算，其通过两个函数I和K生成第三个函数。假设函数I和K为二维离散函数，那么卷积的定义如公式(1)：  (1)  其中函数I也被称为输入，函数K被称为核函数。卷积运算具有稀疏连接、权值共享和稀疏不变性的特征。相对于密集连接，它能够以更少的参数表征图像的特征。  **1.1.2空洞卷积**  空洞卷积(Dilated Convolution)是由 Holschneider 等人提出的，最早被用于信号分析当中。在二维离散空间，空洞卷积的定义如公式（2）所示：  （2）  其中，r为空洞率。当r = 1时，公式（2）可退化成标准卷积公式。当r大于1时，空洞卷积相当于在卷积核的相邻元素间插入了r − 1个空洞，即填充了r − 1个0。如果卷积核的大小为k×k，则空洞卷积的卷积核等效于一个稀疏的卷积核大小为× 的标准卷积。其中与k的关系如公式(3)所示：  （3）  从这种等效关系可以看出，空洞卷积相对于步长为 1 的标准卷积增大了输出的特征图的感受野(Receptive field, RF)。因此，空洞卷积常用于代替池化层或者步长大于1的卷积，在不降低特征图分辨率的情况下，增大其感受野。  **1.1.3转置卷积**  转置卷积(Transposed Convolution)是一个上采样过程，它可以视为卷积操作的一个反向的过程。卷积操作可以用矩阵乘法来表示，以单输入通道单输出通道的卷积为例，将4×4的输入区域展开成列向量 ，卷积核的权重依照与输入的对应关系重新排列成一个矩阵 ，将2×2的输出区域同样展开成列向量 ，则它们之间的关系如公式(4)所示：  （4）  反之，可以通过 X 得到 Y，如公式(1-5)所示：  （5）  其中，。根据卷积层的参数设置，可以快速地找到与之对应的转置卷积操作。图 1展示了步长为2、卷积核为3且填充为1的卷积对应的转置卷积。  图1 步长为2，卷积核为3且填充为1的卷积对应的转置卷积  **1.1.4深度可分离卷积**  深度可分离卷积(Depthwise Separable Convolution)是标准卷积分解的一种形式，它将标准卷积分解成逐层卷积(Depthwise Convolution)和逐点卷积(Pointwise Convolution)的组合。对于标准卷积来说，若其输入特征图的维度为，输出特征图的维度为，那么标准卷积的卷积核的维度为，而逐层卷积的卷积核的维度为 ，逐点卷积的卷积核的维度为。标准卷积利用卷积核生成特征并组合特征以产生新的特征，对于每一个输入通道，需要与个卷积核进行卷积，其结果分别作为组成个输出通道的一部分。标准卷积可以被解耦成使用卷积核生成特征和组合特征这两个部分，前者即为逐层卷积，后者为逐点卷积。逐层卷积对于每一个输入通道使用一个卷积核生成特征，随后逐点卷积通过1×1的卷积将生成的特征线性组合起来得到新的特征。通过将标准卷积进行分解，深度可分离卷积减少了运算量和时间复杂度，使其更加适用于嵌入式设备。   * 1. **SGD算法**   深度神经网络的训练可以看作是对其损失函数的优化问题，于是SGD (Stochastic Gradient Descent)算法及其变种，就是针对如何使得神经网络又好又快收敛的应用最多的算法。SGD 算法从小批量的样本中计算损失函数的梯度，通过计算它们的梯度均值，可以得到梯度的无偏估计。将损失函数记作J(θ)，其中θ为深度神经网络的参数，则 SGD 算法如公式(6)所示：  （6）  其中，为学习率。合适的学习率的设置对于SGD算法尤为重要。如果学习率太大，损失函数的值通常会增加。反之，如果学习率设置的过小，将会使优化过程非常缓慢，需要训练更长的时间。由于使用小批量的样本来估计梯度，SGD算法的学习曲线通常是振荡的。当损失函数收敛到局部最优解附近时，SGD算法往往会由于振荡无法收敛到局部最优解，因此，在神经网络的训练过程中，学习率通常是变化的，在训练后期通常会降低学习率以使神经网络收敛的更好。   * 1. **Momentum算法**   Momentum算法是在SGD算法基础上的优化和改进，核心思想是一种加速收敛的方法，尤其是处理带噪声以及高曲率、小且一致的梯度时。Momentum 算法的思想来源自物理学，将单位质量的物体的速度作为其动量，将梯度类比为物体所受的力，物体下一时刻的速度将由物体当前的速度及所受的力共同决定。Momentum 算法的更新规则如公式(7)和公式(8)所示：  （7）  （8）  其中，为衰减项。在当前梯度的基础上，动量积累了之前梯度的指数衰减平均，使得之前的梯度也能影响到当前参数的更新。衰减项决定了之前梯度衰减的程度，越大，之前梯度的影响也越大。在实际使用中，的值通常被设为0.9。相对于SGD算法，Momentum 算法能够减少由于噪声或者随机采样带来的梯度的振荡，使得网络能够更快且更加稳定的收敛。   * 1. **DAC算法**   DAC算法是Li等人提出的一种参数分解的方法，与其他的参数分解方法不同，它将一层卷积核维度为的标准卷积分解为类似于MobileNet的形式。而且，该算法在进行参数分解的过程中不需要任何的训练或者数据，它直接从已经训练好的标准卷积的参数中得到分解后的层的参数。DAC算法的整体流程如图2所示。  图2 DAC算法  DAC算法将一个卷积核大小为的标准卷积分解成两个部分。第一个部分为卷积核大小为的逐层卷积，其中𝑟𝐶 = 𝑟∗𝑐。第二个部分为卷积核大小为𝑛×1 × 1×𝑟𝐶的逐点卷积。另外，第一部分的逐层卷积不使用偏置，而标准卷积的偏置不做改变地作为第二部分的逐点卷积的偏置。这种形式与MobileNet的深度可分离卷积尤为相似，不过 MobileNet的深度可分离卷积在每一层都使用了激活函数，而由DAC算法分解得到的逐层卷积并不使用激活函数。如果将的卷积核调整为的形式，其中𝑛𝐶 =𝑛∗𝑐，那么将得到一个输出通道为𝑛𝐶的逐层卷积，再加上一个稀疏的权重由0或者1组成的逐点卷积，即可无损地得到与标准卷积的一致的输出。这在一定程度上表明了深度可分离卷积与标准卷积的等效性。  与其他的参数分解算法的出发点一致，DAC算法基于的是标准卷积中存在的冗余性。在标准卷积中，对于输入特征图的每一个通道 ，标准卷积都使用𝑛个的卷积为生成𝑛个通道的特征，分别对应地用于输出特征图的每一个通道。如果将𝑛个的卷积核表示成一个行为𝑛列为的矩阵，将输入特征图的每一个块表示成相应地列向量的形式，那么上述过程可以用矩阵乘法来表示。然而，的值一般远小于𝑛，因此这个矩阵的秩将不超过，这表明为每一个输入通道生成𝑛个通道的特征是冗余的，可以用更少的通道来表示，比如𝑟个。假设为每一个输入通道生成𝑟个通道的特征，然后以这𝑟个通道为基，通过线性组合的形式可以近似地还原𝑛个通道的特征。是否能无损的还原则取决于𝑟的个数。将标准卷积、DAC算法分解后的逐层卷积和逐点卷积分别记作、和，对应地将这些层对每一个通道的操作记作、和，则DAC算法分解标准卷积的目标函数如公式(9)所示  （9）  其中表示矩阵范数。针对这一目标函数，DAC算法提出了基于SVD的求解方法，通过将、和转换成、和，使用SVD分解得到和。  算法的具体实现如下所示：  输入：标准卷积的权重 ，分解使用的秩𝑟  输出：标准卷积的权重，分解使用的秩𝑟  过程：  1. 初始化list\_d和list\_s为空集  2. for i  c do  3.  4.  5.  6.  7.  8.  9.  10. end for  11.  12.   1. **拟采取的主要研究方法**    1. **参数分解**   此次研究将采用结合参数剪.枝，参数量化和紧凑网络的混合型模型压缩与加速方法：通过结合实际的轻量化要求以及对于模型准确度的要求，设计参数重要性评价准则，并基于此评价准则判断网络模型中参数的重要程度，删除冗余参数，并将网络参数从32位全精度浮点整形量化到更低位数。并从卷积核、特殊层和网络结构入手设计和优化网络模型。   * 1. **知识蒸馏**   知识蒸馏是一种常见的模型轻量化方法,通过学生网络学习教师网络的特征表达,将教师网络的特征表达迁移至学生网络,以提升学生网络的模型检测性能。本论文使用轻量化的网络模型作为学生网络，应用基于输出响应图(Response-Based Knowledge)的知识蒸馏方法，在降低模型复杂度的同时提高模型的检测性能。在知识蒸馏过程中，学生网络会同时受到教师网络和图像真实框的监督，如图3所示,在训练过程中,学生网络会不断地学习教师网络的特征表达,以提高模型的泛化能力。    图3 CenterNet蒸馏示意图  CenterNet的头结构(head)输出包含三个部分,表征响应值的热力图、目标尺寸预测和目标位置偏移。   * 1. **注意力机制**   注意力机制本质上是一种资源的重分配机制，通过加入少量的参数，引导神经网络模型向重要的部分分配更多的计算资源，已达到在增加少量模型参数量和计算复杂度的条件下，提高模型的检测准确率。本文将采用一种空间注意力机制和通道注意力机制混合的注意力机制模型，用于轻量化模型后准确度的补充。   1. 通道注意力机制   图4 通道注意力机制原理  通道注意力机制首先将 spatial维度进行压缩，例如将7×7 🡪 1×1，然后通过两个全连接网络来学习通道注意力，再用Sigmoid归一化，然后再将这个学习到的矩阵与原来的矩阵相乘，就可以得到空间维度加权之后的特征。   1. 空间注意力机制   不是图像中所有的区域对任务的贡献都是同样重要的，只有任务相关的区域才是需要关心的，比如分类任务的主体，空间注意力模型就是寻找网络中最重要的部位进行处理。因此空间注意力的本质就是定位目标并进行一些变换或者获取权重。  在处理上，空间注意力机制和通道注意力机制相类似，先在通道维度进行两次池化操作后，将两个特征进行拼接，然后用7x7的卷积来提取空间注意力（之所以用7x7是因为提取的是空间注意力，所以用的卷积核必须足够大）。然后做一次归一化，就得到了空间的注意力矩阵。将这个学习到的矩阵与原来的矩阵相乘，就可以得到空间维度加权之后的特征。  图5 空间注意力机制原理   1. **拟采取的技术路线**   **轻量化目标检测模型的技术路线：**   1. 以在目标检测训练中使用最多的COCO和VOC数据集，作为轻量化目标检测模型研究的数据集。 2. 结合结构化剪切、参数分解和知识蒸馏等多种模型压缩与加速算法，对网络模型进行轻量化处理。 3. 加入Channel Attention 和 Spatial Attention 增强轻量化后目标检测模型的精准度。 4. **拟采用的实施方案** 5. 目标检测数据集处理   在目标检测领域，PASCAL VOC(PASCAL Visual Object Classes)数据集和MS COCO(Microsoft COCO: Common Objects in Context)数据集是较为主流的数据集。早期的目标检测模型如 Fast R-CNN、SSD 和 YOLO 等都是基于 PASCAL VOC 数据集提出的。图6为PASCAL VOC 的类别及层次结构。  图6 PASCAL VOC 的类别及层次结构  在 PASCAL VOC 数据集上的研究基本都是在 VOC 2007 和 VOC 2012 上面进行的。VOC 2012 的检测和分类数据集包含了从 2008 年到 2011 年的所有数据，与 VOC 2007 的检测和分类数据集互不包含，为互斥的。VOC 2007 和 VOC 2012的数据集均划分为三部分，即train、val和test。  MS COCO 数据集是由微软构建，并于 2014 年公开的一个数据集。MS COCO数据集共包含 91 个类别，328 千个图像和 250 万个实例标注。相较于 PASCAL VOC 数据集，MS COCO 数据集具有更多的类别，每种类别的实例也更多。MS COCO 的目标检测任务共包含 80 个类别，为总共的 91 个类别的一个子集。  在本次的实验中将首先在COCO数据集上进行预训练，再使用VOC 2007的trainval+test 以及VOC 2012的trainval进行训练，使用 VOC 2012 的test进行测试。   1. 模型压缩与加速   在模型压缩与加速的算法中，采用结构化剪切的方法，剪切的颗粒度较粗，剪切的最小单位为filter内参数的组合，并且应为结构化剪切后的模型比较规整，便于后续的在现有软/硬件上的有效加速，而且结合设计参数重要性评价准则，尽量保留对准确度贡献大的参数结点。   1. 注意力机制   采用一种空间注意力机制和通道注意力机制混合的注意力机制模型，用于在引入少量参数的情况下，增加轻量化模型的准确度。 |
| 2.研究计划可行性，研究条件落实情况，可能存在的问题及解决办法（可续页）  **1. 研究计划可行性**  通过查阅相关论文，依据现有国内外相关研究成果，对已有技术已有全面的了解，对于基于不同类型的模型压缩与加速算法以及注意力机制算法已有一定程度了解，在国内外相关实验平台也有一定研究成果。  **2. 研究条件落实情况**  实验室拥有用于模型训练的计算服务器，摄像头等实验必须材料，实验室有着在目标检测领域充足的理论以及试验积累。  **3. 可能存在问题以及解决方法**  问题一：论文阅读较少，解决思路有限  解决方法：大量调研相关领域文献，多读顶会顶刊论文，多做论文总结。  问题二：不同的模型压缩与加速算法需要进行大量试验，计算不足  解决方法：导师额外提供了一块30系列显卡，增加了试验进行的速度。  问题三：注意力机制是最近两年提出的新的思想，相关的论文和实现代码较少  解决方法：花时间着重理解最关键的几篇经典注意力机制相关论文，然后在已有的网络模型上做对比试验，摸透原理。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2021.11-2022.01 | 查阅资料，对于研究内容有一定认识或实验，完成开题 |
| 2022.01-2022.02 | 数据集处理及实验平台搭建 |
| 2022.02-2022.04 | 模型压缩与加速各方法对比研究 |
| 2022.04-2022.06 | 注意力机制算法研究设计 |
| 2022.07-2022.10 | 完成各个研究目标 |
| 2022.11-2023.03 | 撰写论文、准备答辩 |
| 预  期  创  新  点  及  成  果  形  式 | 预期创新点：  1.面向边缘计算场景下，轻量级目标检测网络的设计。  2.对比不同模型压缩与加速算法效果，得到面向边缘计算场景的科学性评价；  3.对比不同注意力机制算法效果，得到面向边缘计算场景的科学性评价  4.将注意力机制与参数剪切等模型压缩与加速算法相结合，在保证轻量化网络模型的同时，提高模型准确度。  预期成果：  1.轻量级目标检测算法一套。  2.硕士学位论文一篇。 | |

1. 开题报告审查意见

|  |  |  |  |
| --- | --- | --- | --- |
| 1.导师对学位论文选题和论文计划可行性意见，是否同意开题：  校内导师（组）签字： 年 月 日  校外导师签字： 年 月 日 | | | |
| **2.开题报告考评组意见** | | | |
| 开题日期 |  | 开题地点 |  |
| 考评专家 |  | | |
| 考评成绩 | 合格 票 基本合格 票 不合格 票 | | |
| 结 论 | □通过 □原则通过 □不通过  **通过：**表决票均为合格  **原则通过：**表决票中有1票为基本合格或不合格，其余为合格和基本合格  **不通过：**表决票中有2票及以上为不合格 | | |
| 考评组对学位论文的选题、研究计划及方案实施的可行性的意见和建议： | | | |
| 考评组签名：  年 月 日 | | | |
| **3.学院意见：** | | | |
| 负责人签名： 年 月 日 | | | |