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： 航空航天学院

学 号： 202152100327

姓 名： 彭京伟

论文题目： 基于云-边协同的图像目标

检测方法研究

校内指导教师： 赖俊宇

校外指导教师： 黄硕

填表日期： 2023 年 1 月 7 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）  **（一）研究目标**  本论文旨在提出一种轻量级图像目标检测算法，并将在云-边协同架构中实现快速更新,以解决工业检测场景基于传统云计算范式中存在的问题。  近年来，工业物联网迅速发展，物联网技术通过传感器进行图像数据采集并传输到云端进行数据分析和模型训练，能够为工业场景提供服务。但是传统模式中深度学习模型被固化在边缘设备当中，无法满足柔性制造中一条产品线使用不同模型检测不同产品的需求，也很难利用边缘端采集的数据快速完成模型的更新。另一方面，随着深度学习不断进步，对于模型的性能要求更高，云服务器训练出一个庞大的目标检测模型，在普通服务器上推理时需要的资源和算力很难满足要求。  因此本论文的研究目标就是以实现图像目标检测为任务，针对目前传模式中存在的算法更新更换不及时、边缘设备运行大模型性能不高的问题，提出一种能够部署在边缘节点的轻量级目标检测算法，设计一种实现深度学习模型快速更新更换的目标检测系统，使其能够在实际生产应用中提供敏捷反应、实时响应的模型训练与推理服务。  **（二）研究内容**  在明确了本学位论文的研究目标之后，得出目前需要研究的主要研究内容如下：   * 分析基于云-边协同的图像目标检测方法应用需求与建模 * 研究以模型快速更新更换算法为目标的云-边协同方法 * 设计实现模型边缘部署的轻量级的目标检测算法 * 构建云-边协同的图像目标检测验证系统，评估本方法性能  1. **应用需求分析与建模**   在当前工业生产中，柔性制造作为一种新型生产模式，非常考验生产线的反应速度，同时也考验部署在产品线上的检测模型算法的能力。一个工厂通常会有条产品线，每条产品线上用于检测的摄像头有枚。在传统模式中，当生产线更换产品类型，就需要更换部署了专门用于检测特定产品算法的摄像头，每一枚更换时间为,那么将需要的总时间，将严重影响产品线的运行效率。通过云-边协同，可以将云中心存储或训练好的新模型，通过网络在需要时推送到边缘检测设备，实现快速更换。当模型大小为，网络速度为,模型在设备上部署时间为时，完成新算法部署的时间为。  另一方面，边缘设备性能一般比较低，大模型虽然精度较高，但是模型大小过于庞大，当网络速度为时，模型传输时间也很久，且边缘设备可能无法提供模型能运行的硬件环境。使用轻量化的目标检测算法模型，能够在提供与大模型精度接近的情况下压缩模型大小，使传输时间减小，并减小资源消耗能够在边缘设备正常运行。   1. **研究以模型快速更新更换为目标的云-边协同方法**   传统依靠云计算的图像目标检测的实现方式，使将边缘设备会将所有数据上传至云端，随后在云端进行数据处理、模型训练和数据计算等一系列工作，之后再将检测结果返回至边缘，不仅增加了许多传输等待时间，边缘端也不能离线推理和灵活更新更换算法。  本论文提出了基于云-边缘协同的方法。中心云服务器收到控制端的命令后，任务信息通过消息队列传递到任务执行单元，并按照任务信息执行。对于训练任务，前端将模型训练的相关参数传递至云中心训练服务器的后端，训练服务器根据指定的参数开始训练，并实时记录和显示训练进度、模型相关信息，训练任务完成后将模型保存在云中心服务器，并在需要时推送至边缘设备实现离线推理。当对于模型更换任务，前端将部署的相关参数传递至云中心服务器后端，并通过云-边通信模块与边缘端设备完成连接，从而能够推送新模型至边缘端设备，边缘端设备开始根据相关参数加载模型并完成推理任务，并向云中心服务器返回结果，实现模型的快速更换。对于模型更新任务，需要从边缘端上传新采集的数据集至云中心节点，云中心训练服务器根据相关参数重新训练模型并推送至边缘设备，完成模型更新任务，并对比新旧模型的精度，选择准确率更高的模型。   1. **设计轻量级的目标检测算法**   本论文拟通过借助剪枝、知识蒸馏等方式，改进其主干网络，使得其无需依靠昂贵的高性能设备就能完成深度学习的训练与计算。目前常见的目标检测算法的主干特征提取网络结构复杂，网络的训练与推算对硬件要求高，难以部署到计算能力有限的边缘嵌入式设备。一般地，大模型往往是单个复杂网络或者是若干网络的集合，拥有良好的性能和泛化能力，而小模型因为网络规模较小，表达能力有限。因此，本论文通过剪枝修改删除非关键的权重，利用大模型学习到的知识去指导小模型训练，即为知识蒸馏，使得轻量的小模型具有与大模型相当的性能，但是参数数量大幅降低，从而实现模型压缩与加速，训练出的模型足够轻量化，能够部署在边缘设备。   1. **构建平台，评估性能**   为了能够验证本论文提出的基于云-边协同的图像目标检测方法，并评估模型快速更换更新与边缘设备运行轻量级算法的性能，需要搭建一套云-边协同的平台。拟搭建的平台由一个云中心服务器和若干边缘端设备组成的实验平台。本论文需要研究其实现方式，并通过云端组件实现云-边节点对的应用和配置状态同步、管理边缘设备设备、完成与边缘设备的通信及任务下发，通过边缘端组件间接收云端资源、发送边缘端资源至云端、在边缘端部署容器及管理容器的生命周期，通过接口管理组件中的应用程序编程接口完成对各模块的管理通信与控制，并确保整个系统可靠、安全、稳定地运行，使其能够验证本论文提出的方法。另一方面，本论文提出针对改进的性能评价指标体系，包括对比传统方法更新更换模型时间与本方法的时间的部署时间减少量,更换轻量级模型后模型平均精度损失量,模型传输时间减少量，以及模型在边缘设备的运行情况。  **（三）拟解决的关键性问题**  以上为本论文的研究目标以及研究的主要内容，下面将解释本论文拟解决的关键性问题：  **1. 实现模型轻量化，能够部署在边缘设备**  随着深度学习技术的发展，越来越多研究人员发现使用卷积神经网络对研究物体进行目标检测，准确度可获得较大程度提升[1]。卷积神经网络具有提取高层特征的能力，可提高特征表达能力，并将特征提取、特征选择与特征分类集成到同一模型中。然而，深度学习网络模型具有更深层次的结构与更高的计算复杂性，且存储网络模型所需的空间非常大，导致其无法在资源有限的设备上有效地工作。如果一个参数较少、重建效果较差的轻量网络尽最大程度地去学习一个参数较多、重建效果好的深度神经网络，则既节省了相应算力及空间，又有效提高了轻量网络目标检测的准确性与精度。因此，在不影响深度网络模型效果的前提下，现有的各种模型压缩与加速技术受到人们欢迎。压缩网络模型[2]是一个重要的研究问题。本文将使用剪枝、只是蒸馏、量化等方法，研究模型网络轻量化的问题，是其在轻量的前提下，依然能够保持较好的性能。  **2. 通过云-边协同，实现实时更新、离线推理**  在目前需要部署在边缘设备、终端设备的部分场景中，新采集到的数据集会在使用过程中源源不断地产生，但是现有的方法一般是提前将网络训练好并直接部署在终端设备，而不考虑后续对网络的优化。本论文通过云-边协同的方案，利用这些在检测过程中得到的新样本对网络模型进行更新，能够使网络的性能得到动态的更新，持续提高网络的准确率。同时将新训练的网络模型推送到边缘端设备进行推理，实现边缘端设备的网络模型实时更新，将训练和部署分割在云-边两端，实现边缘端离线推理。 | |

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）  **（一）学位论文的选题依据和研究意义**  **1.** **模型轻量化成为边缘设备算法部署的重要方法**  目标检测是计算机视觉中的重要分支，主要任务是根据用户感兴趣的目标信息找出输入图像中所有符合要求的目标,并对目标所属类别和所处位置进行分析，其广泛应用于图像分割、物体追踪、关键点检测等需求。当前目标检测算法主要分为Two stage和One stage两类算法。相较于Two Stage目标检测算法复杂的网络结构，One Stage目标检测算法具有检测速度快、模型简单轻量的特点。其中，YOLO算法在经过数次迭代之后，逐渐弥补了许多缺陷，同时在保持速度优势的前提下，兼顾了检测准确度和实时性能。 尽管YOLO系列是目标检测领域中速度-精度均衡的佼佼者，但其主要工作是面向电脑端，同时模型也逐渐复杂。然而，庞大的网络参数意味计算量大、存储成本高、模型复杂等特性，而增长的浮点型训练运算次数意味着训练成本和计算时间的增长，这极大地限制了在资源受限设备上的部署。其中卷积层和全连接层含有大量的参数，网络经过训练之后，参数存在大量冗余，这些冗余的参数是不重要的、可以删除的，去除这些参数并不影响网络的精 度。由于参数减少，网络的计算得以简化并且速度大幅提高，从而能提升网络的整体性能。当前，边缘计算以及云-边协同的架构已成为人工智能发展的重要趋势之一，面向性能较弱的AI计算设备，通过模型压缩方法，例如剪枝、知识蒸馏、量化，使目标加测算法更轻更快已成为当前研究的热点。本论文研究聚焦于在模型训练时使用模型压缩手段，以实现模型轻量化。  **2.** **图像目标检测算法对实时性与更新便捷性要求更高**  近些年，工业物联网迅速发展，物联网技术通过传感器进行数据采集并传输到云端进行数据分析和逻辑控制，能够很好地存储分析生产过程中的关键数据。但在传感器经历长时间运行，所产生的数据量也海量爆炸，传统云计算模式很难满足实时目标检测的要求。基于上述问题，以边缘计算模型为核心的面向网络边缘设备所产生数据计算的边缘大数据处理应运而生，其与现有以云计算模型为核心的集中式大数据处理相结合，形成云-边协同合力，二者相辅相成，实现实时更新、离线推理与扩展优化，很好地解决了工业物联网目前暴露的问题。wang等[3]提出了一种基于云计算与边缘计算协同的智能表面检测系统，将表面检测计算任务部署到边缘端，避免了数据泄露的风险，同时保证了计算的实时性。尹子会等[4]提出了基于云计算与边缘计算协同的变电站设备典型视觉缺陷检测系统，与传统模式直接上传到云端计算相比，传输量减少90％以上，同时检测速率也有较大提高。上述研究将云计算与边缘计算融合的模式和检测行业相结合，取得不错的成绩，但所采用的深度学习模型都需要采用多块GPU进行训练推理，对硬件成本要求极高。本文提出一种基于云-边协同的轻量级图像目标检测算法，能够通过在部署了深度学习框架的边缘端设备进行实时的目标检测；同时对边缘端和云端进行调度和管理，搭建边缘端和云端的数据通道实现边缘端和云端的信息交互，使更加轻量敏捷、更加准确稳定的模型能够方便地在边缘端实现目标检测算法。  **（二）国内外研究现状**  **1. 云-边协同**  云计算是一种计算范式,它可以根据用户的需求随时随地为最终用户提供无限的计算资源,用户只需为使用的服务付费。云中可以提供各种类型的服务,如资源池、弹性和灵活性、可扩展性(水平和垂直)、性能高可用性、托管服务等[5]。正是因为具有强大的服务能力,云计算成为了所有业务之首,为全世界提供就业机会,近十年来被学术界和工业界广泛研究。然而,根据数据机构IDC的预测,2020年底将有超过500亿的终端与设备联网[6],从而产生海量的异构数据。此时,传统的云计算已经不能满足一些对实时性比较敏感的应用,并且将全部数据都上传到云数据中心也会给网络带宽带来很大的压力。因此,以解决数据传输延迟、降低网络带宽为目标的边缘计算正迅速兴起。边缘计算是指在网络边缘执行计算的一种新型计算模式。边缘计算中的下行数据表示云服务,上行数据表示万物互联服务,边缘计算的边缘是指从数据源到云计算中心的路径之间的任意计算和网络资源[7]。边缘计算架构中,用户数据不再需要全部上传到云数据中心,而是通过部署在网络边缘的边缘节点快速处理部分数据,从而大大减轻了网络带宽的压力,大幅降低了网络边缘端智能设备的能耗。为此,针对边缘计算的探索性研究已经广泛展开。随着其市场规模的逐渐扩大,边缘计算成为了与云计算同台竞技的解决方案。为了更好地结合云计算与边缘计算的优势,云-边协同作为一种新型计算模式成为了新的研究趋势。  云计算在近些年已经获得了巨大的发展和应用，同时也已有部分边缘计算产品被逐步推出，但云-边协同的发展仍处于探索阶段。随着数据密集型应用与计算密集型应用的增加,需要利用云计算强大的计算能力以及通信资源与边缘计算短时传输的响应特性来实现并完成相应的应用请求。通过两者协同工作、各展所长,将边缘计算和云计算协作的价值最大化[8],从而有效地提高应用程序的性能。目前,针对云-边协同的研究大多数集中在物联网、工业互联网、智能交通、安全监控等诸多领域的应用场景上,主要目的是减少时延、降低能耗以及提高用户体验质量等。Ren等[9]提出的云-边缘协作方法能够有效地提高延迟性能。Ding 等[10]提出了一种云-边缘协作框架,通过浅层卷积神经网络模型提供持续时间长、响应速度快的认知服务,给用户带来了良好的体验。Zhang等[11]在工业互联网中提出了一个 Cloud-Edge 协作的工业设备管理服务系统,在一定程度上提高了工业现场系统的响应速度,减轻了数据传输带来的网络带宽负载压力,推动了工业物联网向智能化发展。  **2. 目标检测算法**  目标检测任务要求不仅判断输入数据包含物体种类,还要定位物体位置并用矩形框框出。现如今目标检测算法可分成两种:(1)传统目标检测算法,(2)基于深度卷积神经网络的算法。传统目标检测法由区域选择、特征信息提取[12]以及分类器构成。区域选择较多使用基于滑动窗口 [13] 的方法,特征信息主要涉及颜色、边缘、尺度不变特征(SIFT)[14]以及方向梯度直方图(HOG)特征[15] ,分类器则有支持向量机[16][17]和AdaBoost[18]。传统算法在区域选择以及特征提取上时间复杂度较高,最终分类准确率也较低。随着深度卷积神经网络不断开拓创新,将其融合到目标检测算法成为研究的热点。现如今,基于深度卷积神经网络目标检测算法[19]可分成两种:一种是基于候选区域的Two-Stage 检测算法,另一种为基于回归的 One-Stage 检测算法。  **2.1 基于候选区域的 Two-Stage 检测算法**  Grishick R 等人[20]于 2014 年提出 R-CNN,第一次将深度卷积网络融入到目标检测领域。该算法框架首先采用选择搜索算法(Selective Search,SS)[21]生成 2000 个候选框,并将不同尺寸候选框缩放至 227×227;然后算法使用 AlexNet 卷积网络对候选框进行特征信息提取;最后使用 SVM 算法确定目标的类别,使用回归器对候选框进行位置矫正。He 等人[22]提出 SSP-Net 目标检测算法,通过特征金字塔将由 SS 算法得到的候选框统一尺度,避免 R-CNN 中候选框卷积计算量。2015 年 Grishick R [23]基于 SSP-Net 思想改进 R-CNN 提出 Fast-RCNN,该算法创新提出 ROIPooling 层,避免了对候选框缩放操作,有效提高网络计算速度。算法还替换特征提取网络为 VGG16,SVM 分类层替换为 Softmax 函数,采用多任务学习即同时实现目标类别预测与位置回归计算,有效降低模型参数量与计算量,在 Pascal VOC 2007 取得 70%的 mAP (mean Average Precision),但由于 Fast R-CNN 仍采用 SS 方法选择候选框,导致算法无法实现低延时检测。Ren S等人[24]提出了 Faster-RCNN 目标检测算法,该算法提出候选框生成网络(RPN),降低候选区域选择带来的复杂计算量。Faster-RCNN 将候选框生成、特征信息提取、目标分类和位置回归融合进单个算法框架,实现目标检测算法端到端训练,网络检测精度以及速度得到较高提升,在 VOC2007 测试集 mAP 达到 73.2%。;2016 年,Dai J 等人[25]提出了 R-FCN 目标检测算法,该算法采用残差网络模块作为特征提取网络,使用卷积神经网络代替 ROIPooling 层之后的全连接层,有效减少模型参数量。R-FCN 还加入位置感受得分图解决卷积平移不变性,将目标位置信息添加进 ROIPooling 层中,在 PascalVOC2007 测试集检测精度提升到 80.5%。但是该模型计算复杂,检测实时性较低。  **2.2 基于回归的 One-Stage 检测算法**  2016 年,Redmon J 等人[26]首次将目标检测看待为一种回归问题,提出单阶段目标检测算法 YOLO V1,输入图像通过单次卷积就能得到图像中物体分类与所在位置结果。YOLO V1 相较于两阶段目标检测器在目标检测速度实现大幅提升,但算法对于小目标、密集分布物体预测精度较低,在 VOC2007 上 mAP 仅为 66.4%。2017 年,Redmon J 等人[27]在 YOLO V1 基础上提出 YOLO 9000,该网络对卷积层输出都采用批量归一化处理,同时采用了锚框机制并设计 Darknet-19 作为 YOLO 9000 的特征提取网络,有效提升了目标检测算法准确度,在 VOC2007 数据集上的 mAP 达到 78.6%;2018 年, RedmonJ 等人[28]基于 YOLO 9000 提出 YOLO V3,它使用了特征提取效果更好的 Darknet-53,并且还采用多尺度分类预测方法以及锚框机制,使得网络在保证检测速度同时,预测精度也有效提高。2020 年,Bochkovskiy A 等人[29]提出了 YOLO V4 算法,该算法是在原有 YOLO 系列目标检测算法基础上,采用了近些年 CNN 领域中最优秀的优化策略,从数据预处理、主干网络设计、网络训练方法、激活函数、损失函数等各个方面着手对算法进行优化处理,在 COCO 数据集上,可达 43.5%AP,速度达到 65FPS。Liu W 等人[30]提出 SSD 目标检测框架,该方法融合 YOLO 以及 Faster-RCNN 中锚框机制,采用 VGG16部分卷积层用于特征提取,并新增 6 个卷积层获得更多特征信息,使用多尺度检测实现对不同大小物体精确检测,在 VOC2007 上测试得到 75.1%的 mAP,检测速度达到 58FPS。  **2.3模型轻量化与云-边架构部署**  受限于边缘侧资源限制，深度学习在边缘侧网络上有以下两个主要研究优化方向：首先在算法模型规模上尝试新的模型部署技巧，这类方式可以通过减少网络中的模型参数实现（如进行模型剪枝与压缩、知识蒸馏和使用更高效的算法层）；另一个方向根据边缘硬件结构重新设计网络模型，实现高效的网络推理。  比较著名的边缘侧网络举例如下。SqueezeNet[31]参考 AlexNet 进行改进，减少卷积核数量，使用shortcut结构，将参数减少到1/50。ShuffleNet[32]提出将图像不同的部位使用不同的卷积进行运算，以减少卷积的重复计算进而减少整体计算量。CondenseNet[33]使用分组卷积，经过训练过程学习分组策略并进行参数修剪，去除次要特征图，实现与 ShuffleNet 相似的精度，而参数却减少 50%。NASNet[34]提出使用未定义块结构，通过强化学习确定块属性。相似地， PNASNet[35]使用序列模型优化算法确定块属性。MNASNet[36]根据特定应用场景自动生成网络模型，并通过强化学习获取平衡硬件性能和算法时间复杂度的最佳网络结构。MobileNet[37]是一个面向边缘环境的目标检测网络，它以较小的精度降低为代价，有效减少权重和计算量。通过减少输入图大小和逐点卷积的单核计算实现性能上的优化。MobileNetv2[35]通过使用新的残差块有效减少通道数量，减少30%的参数数量和50%的算力消耗，并提高准确率。  同时，云-边协同架构在今年的工业互联网领域也得到了广泛的应用与研究。文献[38]提出了基于机器学习的边缘云框架，在数据边缘网关采用随机切割森林和隔离森林算法检测异常变化的数据，并将预处理的时间序列数据传输到云，结合原始数据采用长短期记忆循环神经网络进行数据趋势预测和补全，大大提高了时间序列数据分析的效率和准确性。文献[37]构建了边云协同模式下基于深度学习算法的大型光伏电站线性缺陷检测系统，通过边缘设备，边缘服务器和云服务器之间的任务卸载，实现了对视觉数据的智能化治理，减少了通信开销，实现了对光伏电站的高效快速检测。  由于边缘侧资源有限，模型在具备准确性外，还要考虑模型的算力消耗、内存占用和能耗优化。与云侧算法模型以准确率为导向相比，边缘侧模型需要考虑性能、成本、功耗等多方面的因素[40]，需要更加轻量的网络模型来匹配边缘侧的算力。  **（三）发展态势**  新型人工智能算法服务平台融合了云计算、边缘计算、 5G等新兴技术实现云-边协同，在云端部署业务系统、研发工具。边缘云部署在接入网关的边缘侧，将云端的部分计算、存储能力和服务扩展到边缘节点上。中心云与边缘云协同对全网算力进行调度和统一管控，发掘工业互联网的内在能力，提高产线效率和用户体验。图像目标检测算法作为应用最为广泛的深度学习算法，在新型的平台中，将传统的应用部署到云端，并将应用中需要部署在边缘端的模块剥离后做成镜像的方式。通过云化镜像，云端具备即时开通应用服务的能力。另外通过云应用市场和云管理平台的协同共同实现云端一键开通的能力。算法实时推理计算边缘化利用边缘云集群承载实时性业务并赋能给终端设备，云端智能化利用人工智能将图像目标检测任务的结果进行分析并做出预测。结合应用上云、云端一键开通、算法实时推理边缘化和云端智能化构成云-边协同的全流程。在基于云-边协同的图像目标检测架构中，边端通过算法部署边缘为网络卸载，云端通过云计算能力对边端反馈的目标检测算法推理结果进行智能分析并即时反馈给用户，用户能够实时根据分析结果进行后续的操作。  **（四）选题在理论研究/实际应用的意义和价值**  选题在工业领域有实际应用价值与意义。  **1.实现资源的高效可靠利用**  深度学习模型的训练往往要进行大量数据的处理工作，对硬件的内存和计算能力要求较高，因此需要在资源充足的数据中心完成。在视觉检测系统中，检测器本身从硬件成本和检测实时性考虑，都不适合充当数据中心。而在工厂内各车间、产线部署服务器单独进行模型训练，既效率低，也无法实现模型的统一管理和部署。本论文提出的方法能够实现边缘计算与云计算互为补充，对接入集群的设备统一管理与编排任务，实现硬件资源的最大化利用，同时避免了给核心网络带来沉重负担。另一方面，通过云-边协同的方式训练和部署深度学习模型，可以实现模型更加便捷地发布到检测设备，确保生产车间数据不出厂，降低了网络负载和数据泄露风险。  **2.快速部署实时性应用**  由于现代工业应用场景的多样性和复杂性，工厂站点具有不同形式的计算任务请求。例如，实时记录设备信息和状态，智能识别工人是否戴着安全帽，传感器收集的多数据信息以及其他计算任务要求针对这些计算任务的深度学习。本论文提出的方法，通过容器化管理应用的云-边协同平台，可以简化开发流程，将复杂的机器学习、图像识别、事件处理等高级应用部署到边缘设备并保持更新，其有利于推动深度学习在智能制造、工业质检领域的商业化应用，在工业中具有广泛的应用前景。  **（五）主要参考文献**   1. LAWRENCE S，GILES C L，BACK A D，et al. Face recognition：a convolutional neural-network approach [J]. IEEE Transactions on Neural Networks，1997，8（1）：98-113 2. AI Y J，HAO S F，HUANG D J. Deep neural network model compression methods and progress[J]. Journal of East China Normal University （Natural Science），2020（5）：68-82. 3. WANG Y B，LIU M G，ZHENG P，et a1．A smart surface inspection system using faster R—CNN in cloud—edge computjng engjronment[J]．Advanced engjneering informatics，2020，43：101037．1—101037．9 4. 尹子会，孟荣，范晓丹，李冰，赵振兵．融合边缘计算和改进Faster R—CNN的变电站设备典型视觉缺陷检测系统[J]．中国科技论文，2021，16(3):343—348． 5. KUMAR M，SHARMA S C,GOEL A,et al.A comprehensive survey for scheduling techniques in cloud computing[J].Journal of Network and Computer Applications,2019,143:1-33. 6. SHI W S，SUN H,CAO J,et al.Edge computing:a new computing model for the Internet era [J].Journal of Computer Research and Development,2017,54(5):907-924. 7. BOUSSELHAM M，BENAMAR N，ADDAIM A.A new Security Mechanism for Vehicular Cloud Computing Using Fog Computing System[C]//2019 International Conference on Wireless Technologies,Embedded and Intelligent Systems (WITS ).IEEE,2019:1-4 8. REN J，HE Y，YU Get al.Joint communication and computation resource allocation for cloud-edge collaborative system[C]∥2019 IEEE Wireless Communications and Networking Conference (WCNC).IEEE,2019:1-6. 9. DING C，ZHOU A,LIU Y,et al.A Cloud-Edge Collaboration Framework for Cognitive Service[J/OL].IEEE Transactions on Cloud Computing,2020.https://ieeexplore.ieee.org/abstract/document/8895891 10. ZHANG H，CHEN S,ZOU P,et al.Research and Application of Industrial Equipment Management Service System Based on Cloud-Edge Collaboration[C]∥2019 Chinese Automation Congress (CAC).IEEE,2019:5451-5456. 11. NIXON M S. Feature Extraction and Image Processing[M]. Publishing House of Electronics Industry, 2013 12. XU Y, D XU, LIN S, et al. Sliding Window and Regression Based Cup Detection In Digital Fundus Images for Glaucoma Diagnosis[J]. Springer-Verlag, 2011. 13. LOWE D G.Distinctive Image Features from Scale-Invariant Keypoints[J].International Journal of Computer Vision, 2004, 60(2):91-110. 14. 刘方园,王水花,张煜东.方向梯度直方图综述[J].计算机工程与应用,2017,53(19). 15. LIN, ZHU, YANG, et al. Large-scale image classification: Fast feature extraction and SVM training. IEEE Computer Society, 2011. 16. 乔风娟,郭红利,李伟等.基于SVM的深度学习分类研究综述[J].齐鲁工业大学学报,2018,32(5):39-44. 17. 张溪樾.基于 Adaboost 的行人检测综述[J].电子制作,2019,(1):59-61. 18. 黄健,张钢.深度卷积神经网络的目标检测算法综述[J].计算机工程与应用,2020,56(17):12-23.DOI:10.3778/j.issn.1002-8331.2005-0021. 19. GIRSHICK R,DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. IEEE Computer Society, 2013. 20. UIJLINGS, J.R.R., Van De Sande, et al. Selective search for object recognition[J].International Journal of Computer Vision,2013,104(2):154-171. 21. HE K, ZHANG X, REN S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(9):1904-16. 22. GIRSHICK R. Fast R-CNN[J]. Computer Science, 2015. 23. REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149. 24. DAI J, LI Y, HE K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. Advances in Neural Information Processing Systems, 2016. 25. REDMON J,DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection[C]// Computer Vision & Pattern Recognition. IEEE, 2016. 26. REDMON J,FARHADI A. YOLO9000: Better, Faster, Stronger[J]. IEEE Conference on Computer Vision & Pattern Recognition, 2017:6517-6525. 27. REDMON J,FARHADI A. YOLOv3: An Incremental Improvement[J]. Computer Science,2018,4(1):1-6. 28. BOCHKOVSKIY A, WAND C Y, LIAO H. YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020. 29. LIU W,ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[J]. Springer Cham, 2016. 30. IANDOLA F N ，HAN S , MOSKEWICZ M W ,et al.SqueezeNet:AlexNet-level accuracy with 50x fewer parameters and<0.5 MB model size[J]. arXiv preprint arXiv:1602.07360, 2016. 31. ZHANG X ，ZHOU X , LIN M ,et al.Shufflenet: an extremely efficient convolutional neural network for mobile devices[C]// Proceedings of the2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2018: 6848-6856. 32. HUANG G ，LIU S , VAN D M L ,et al.Condensenet:an efficient densenet using learned group convolutions[C]// Proceedings of the2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2018: 2752-2761. 33. ZOPH B，VASUDEVAN V , SHLENS J ,et al.Learning transferable architectures for scalable image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8697-8710. 34. LIU C,ZOPH B , NEUMANN M ,et al.Progressive neural architecture search[C]// Proceedings of the European Conference on Computer Vision (ECCV). Piscataway:IEEE Press, 2018: 19-34. 35. TAN M ，CHEN B , PANG R ,et al.Mnasnet: platform-aware neural architecture search for mobile[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2820-2828. 36. HOWARD A G , ZHU M , CHEN B ,et al.Mobilenets:efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017. 37. SANDLER M , HOWARD A , ZHU M ,et al.Mobilenetv2:inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4510-4520. 38. HI F，YAN L P，ZHAO X，et a1．Machine learning-based time-series data analysis in edgecloud—assisted oil industrial loT system[J]．Mobile Information Systems，2022，2022． 39. TANG W Q，YANG Q，HU X C，et a1．Deep learning-based linear defects detection system for large-scale photovoltaic plants based on an edge-cloud computing infrastructure[J].Solar Energy,2022，23 1：527-535． 40. YANG T J , HOWARD A , CHEN B ,et al.Netadapt: platform-aware neural network adaptation for mobile applications[C]// Proceedings of the European Conference on Computer Vision (ECCV). Piscataway: IEEE Press, 2018: 285-300. |

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）   **（一）主要技术理论**  完成本次毕业设计预计要掌握的技术理论有：KubeEdge架构、YOLO系列目标检测网络相关知识  **1. KubeEdge**  本论文云-边协同的实现主要依靠KubeEdge架构。KubeEdge 是一个开源的边缘计算平台，其基 于 Kubernetes（K8s）原生的容器编排和调度能力实现了云-边协同、计算下沉、海量边缘设备管理、边缘自治等功能，具有完全开放、可扩展、易开发、易维护、支持离线模式和跨平台等特点。相较于K8S，此框架在边 缘端的部署上有显著优势，同时支持 MQTT 协议，允许开发者在边缘端自定义接入边缘设备。KubeEdge架构图如图1所示。    图1 KubeEdge架构图  KubeEdge 以 Kubernetes 为载体构建，通过提供核心基础架构支持，在边缘终端进行重写的 edged 实现 kubelet 的功能，使得边缘有可以独立处理数据的能力。云端与边端的通信由传输控制协议完成，而通信安全和网络构建均基于Kubernets Master 节点生成的电子认证服务相关证书来实现，切实从源头上保证了指令传输的便捷性及通讯的安全可靠性。边缘终端上内置有 meta manager 和devicetwin 数据模块，可快速实现基础数据本地化，由此达到边缘自治的目的。evenbus 和 servicebus 这两个模块的主要功能是存储数据、事件管理、数据分析等，待大规模数据汇聚、处理后反馈给云端，由此可见，它们为边缘端开发提供了强大技术支持。  **2. YOLO目标检测网络**  本论文通过YOLO网络实现图像目标检测算法。YOLO 网络具有检测精度高、速度快的特点，同时具 有一定的鲁棒性，比较适合用于实时目标检测。YOLO 网络将目标检测过程当成回归问题 去解决，直接通过整张图片得到 bounding box 的坐标、box 包含目标的置信度、class probabilities。如此，只需要看一眼图片就可以知道图片包含哪些目标和目标的具体坐标。 符合 YOLO 的只看一次的初衷。初代YOLO 检测过程如图 2 所示。  IMG_256  图2 YOLO检测过程示意图  首先将图片填充缩放到大小为 448\*448 并输入到 YOLO 网络中，从而得到 bounding box 的坐标、box的置信度和 class probabilities，最后通过 nms，对 bounding box 进行筛选得到正确的目标检测结果。随着YOLO网络的升级，其使用的网络结构也在不断更新，并引入了新的机制提升算法能力。  **（二）研究方案**  为实现本论文的研究，总体方案如图3所示。  演示文稿2图3 基于云-边协同的图目标检测方法研究总体方案  其中云端服务使用搭载高性能GPU的设备，其需要完成图像目标检测模型训练任务，并搭载大数据存储与数据通信模块，接受边缘设备传输的图像，进行模型的更新，并将模型推送至边缘端。边缘端设备具有一定AI运算能力的服务器，部署轻量的图像目标检测模型，能够满足检测的实时性要求，并将推理结果实时显示。边缘端设备配备数据存储模块和数据通信模块，接收到终端所采集的图像；部署边缘端软件平台对应用进行管理和调度，同时完成与云端设备的通信。控制节点面向用户实现对于系统的任务下达。  **（三）技术路线**  为实现基于云-边协同的目标检测方法的研究，本论文的技术路线主要分为平台搭建、算法研究与平台应用算法三个方面。  **1. 平台搭建方面**  首先需要熟悉并掌握containerd、k8s技术以及相关通信协议，并确定系统的实现架构与本论文需要使用的接口与组件，在本地实现KubeEdge的部署，借助KubeEdge平台的各API，能够完成基本流程的实现，包括对于集群的管理，对于边缘节点、设备、应用和任务的管理，服务暴露，云-边消息通信等。  基于云-边缘协同计算的图像目标检测算法平台采用基于Kubernetes技术的KubeEdge平台架构。KubeEdge平台架构分为云端、边缘端2部分，本论文暂不考虑设备端部分。其通过将Kubernetes的功能扩展部署到边缘端完成云端与边缘端设备之间的协同、调度和管理工作。 搭建的平台具有以下两部分组件：  （1）云端组件。云端设备部署Kubernetes控制平面，包含云端通信中心cloudHub组件、边缘管理控制器和设备管理控制器。其中，cloudHub实质为套接字服务器，负责完成与边缘端设备EdgeHub的数据通信；边缘管理控制器负责管理边缘端设备，同时负责边缘元数据在边缘端和云端之间的同步；设备管理控制器负责管理设备，同时负责设备元数据的同步工作。  （2）边缘端组件。边缘端设备主要包含边缘端通信中心EdgeHub组件、Edged边缘节点管理容器等组件。EdgeHub组件与cloudHub相对应，是一个套接字客户端，用于负责接收云端的资源，同时将设备信息和缺陷图像等数据传输至云端设备分析存储；元数据管理器MetaManager组件用于通信中心和容器组件之心的信息处理，用于管理元数据；设备孪生DeviceTwin组件通过事件总线和边缘端通信中心同步设备端所采集到的数据，同时管理和处理设备端的元数据；边缘容器管理模块相当于轻量化的Kubelet组件，主要用于在边缘端部署容器，管理容器的生命周期。  **2. 算法研究方面**  本论文拟采用YOLOv3网络结构，基于PyTorch框架，通过剪枝和知识蒸馏方法来实现轻量化图像目标检测算法。  IMG_256YOLOv3的基础网络为Darknet-53，如图4所示，它借鉴了ResNet的残差结构，加深网络结构的同时，又防止了网络梯度爆炸引发的网络难以收敛的问题。前向传播过程中，移除  图4 DarkNet-53  池化层和全连接层，通过改变卷积核的步长来改变张量的尺寸。与v2类似，Darknet-53会将输出特征缩小到输入的1/32，因此通常要求输入图片分辨率是32的倍数。同时YOLOv3采用张量拼接扩充张量的维度，以提取更多信息，具体操作是将Darknet-53中间层和后面的某一层经过上采样之后进行拼接。Darknet-53从第0～74层，共有53个卷积层， 其余为残差层。第75～105层为YOLOv3的特征融合层，其中YOLOv3增加了多尺度检测(相当于颈部)，使用了3种尺度，其输出分别是52×52, 26×26,13×13用于检测小、中、大目标，每种尺度预测3个锚框。  为了使训练出来的目标检测算法能够部署在边缘端设备，采用剪枝和知识蒸馏技术。  **2.1模型剪枝**  /home/yijiahe/桌面/icon/_cgi-bin_mmwebwx-bin_webwxgetmsgimg__&MsgID=552237690124841592&skey=@crypt_b8c3caf8_87e103e61c0d962b9d4ff55257c9e077&mmweb_appid=wx_webfilehelper.jpeg_cgi-bin_mmwebwx-bin_webwxgetmsgimg__&MsgID=552237690124841592&skey=@crypt_b8c3caf8_87e103e61c0d962b9d4ff55257c9e077&mmweb_appid=wx_webfilehelper神经网络模型中有大量的神经元与权重冗余。真正对模型产生影响的只有5％～10％。将模型中不重要的参数去掉来减小模型大小的方法越来越多，最常见的模型压缩方法是模型剪枝。剪枝不仅可以降低模型复杂度、防止过拟合，还可以压缩模型大小、提高模型的速度。剪枝方式可分为通道剪枝、权值剪枝、层剪枝。层剪枝效果最粗糙，但无软硬件要求、灵活性好；权值剪枝灵活性最好，但是对软硬件有要求，需要专门的系数矩阵运算库或者硬件来做加速；通道剪枝集成了层剪枝和权值剪枝的优点，精细且灵活性好．本论文拟采用通道与层相结合方式来对模型进行剪枝。在通道剪枝结束后，将模型再汇总进行层剪枝，最后输出剪枝完的模型。模型剪枝示意图如图5所示。  图5 模型剪枝示意图  **2.2 知识蒸馏**  目前，深度学习需要对海量的数据进行训练来提取特征，这些数据数量较大且冗余，需要消耗大量资源，在没有实时性要求下训练得到的模型精度较高，但是模型结构笨重．在实际应用中，考虑各方面成本，特别是在计算资源方面有所限制．使用一个较大的神经网络模型有较多局限性．因此，本文采用知识蒸馏模型，让大模型为小模型进行知识迁移。  知识蒸馏的原理是将复杂的教师网络中的 dark knowledge迁移到简单的学生网络中，知识蒸馏模型示意图如图7所示。  /home/yijiahe/桌面/icon/222.jpeg222图7 知识蒸馏模型示意图  综上，本论文的目标检测算法采用YOL0 v3模型对数据集进行训练，接着对网络进行模型剪枝，最后采用知识蒸馏技术对剪枝后的模型进行蒸馏，补偿剪枝后模型的检测精度损失，最终预期能够得到轻量化但是仍有较高精度的模型。  **3．平台应用算法方面**  基于图3的总体方案，借助KubeEdge的APIServer来实现如下功能。  当用户下发训练任务时，任务管理模块创建训练任务，并调用云端任务控制模块，加载用户上传的参数配置，开始执行模型训练任务。当模型训练完成后，通过云端数据管理模块将模型与日志存入云端数据库。  当用户下发推理任务时，任务管理模块创建推理任务，并通过云-边通信模块通知边缘任务管理模块进行推理准备，云端数据管理模块从云端数据库读取模型文件，通过云-边通信模块推送至边缘端任务控制模块，边缘端任务控制模块启动推理任务，并将推理结果予以显示。推理任务结束后将模型与日志通过边端数据管理模块保存至数据库，并将日志通过云-边通信模块上传至云端备份。  当用户下发模型更新任务时，任务管理模块通知云端任务控制模块并通过云-边通信模块通知边端任务管理模块准备进行模型更新，边端任务控制模块启动模型更新任务，通过边端数据管理模块从数据库中读取平常接受到的新采集的数据，并将其通过云-边通信模块传送至云端数据管理模块，云端数据管理模块将数据集传输至云端任务控制模块，并开始重新训练模型，当训练模型结束后再进行部署，将模型推送至边端。边端任务控制模块进行模型精度验证并与之前保存的模型精度进行对比，选择效果更好的模型保存在边端数据库中。 |
| 1. 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）   **（一）计划可行性**  计划可行性是对当下技术和硬件设施能否满足系统开发的分析。  算法上，通过YOLO网络训练目标检测模型方案已经十分成熟，同时对于模型压缩所采用的剪枝和知识蒸馏近些年也有许多研究成果，所以通过此方案可以训练出在边缘节点部署的轻量化图像目标检测模型；  硬件上，联合培养基地和教研室均拥有足够性能的服务器，同时可以通过虚拟机搭建仿真平台，因此在硬件配置上具有可行性。  软件上，KubeEdge作为一款开源的提供云-边协同支持的框架，有官方的文档和代码提供学习和使用；算法实现上，MMDetection框架把数据集构建、模型搭建、训练策略等过程都封装成了一个个模块，可以以很少的代码量实现算法；后端框架包括 Django等框架的技术都非常成熟；数据库包括MySQL、SQLLite等各有各的特点，这些技术已被广泛使用；开发环境Python、 cuda、和开发工具 PyCharm、VSCode 也相当成熟。  综上，基于云-边协同的目标检测算法研究具有可行性。  **（二）条件落实情况**  联合培养单位提供一台超高性能的服务器与若干台性能较强的服务器。性能超高的服务器可以作为整套云-边系统中的云中心节点，其他若干台服务器可以作为边缘节点，设备运行情况与网络状况良好，方便搭建实验系统。 同时也可以借助虚拟机搭建边缘计算仿真平台  **（三）可能存在的问题及解决办法**  可能存在问题：搭建KubeEdge本地集群的方式繁琐且容易出现错误，K8S与容器运行时版本不匹配可能导致集群运行失败。  解决办法：前期调研搭建集群所需要组件的版本匹配情况，查阅官方文档与相关教程，完成集群的部署。同时研究计算机网络相关知识，通过CNI（Container Network Interface，容器网络接口）配合合理网络环境使节点网络通信正常。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2022.11-2023.1 | 调研相关文献，查看相关资料，完成论文开题报告 |
| 2023.01-2023.03 | 本地实现基于KubeEdge的集群完备功能部署 |
| 2023.04-2023.05 | 实现接入设备的管理，实现模型在云中心节点的训练，在边缘节点的推理部署 |
| 2023.06-2023.08 | 实现模型轻量化，对目标检测算法进行优化 |
| 2023.09-2023.11 | 在集群中完成新算法模型的训练与推理全流程的实现 |
| 2023.12-2024.03 | 完成论文的撰写，并准备答辩 |
| 预  期  创  新  点  及  成  果  形  式 | **（一）预期创新点**  1. 目前业界使用的传统云计算模式暴露了实时性不够、带宽不足、能耗较大及数据不安全等不足。本文将创新性地提出将云-边协同架构应用于图像目标检测任务中，边缘计算可以实现数据在边缘侧的预处理或完全处理,从而减少了数据传输的带宽压力,满足实时应用对时延的要求；云计算继续发挥出强大计算能力与存储能力的优势。云端设备与边缘设备各司其职又联合工作，实现资源、数据、智能全面协同，解决传统模式存在的问题，提升算法在实际应用场景中的性能。  2. 传统网络结构复杂，训练消耗资源大，很难部署到低性能设备。本文提出一种轻量化网络，能够在精简网络参数的同时也能保持良好的性能。同时依靠云-边协同架构加持，能够实时灵敏地对网络模型进行迭代，使模型效果得到不断提升。  **（二）预期成果形式**  1. 搭建云-边协同图像目标检测仿真系统；  2. 研究生毕业论文1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 | | | |