



专题：通信与 AI 融合

基于云边协同的计算机视觉推理机制

唐博恒, 柴鑫刚

(中国移动通信有限公司研究院, 北京 100053)

摘要: 深度学习和云计算的普及推动了计算机视觉在各行业中的广泛应用。但集中化的云端推理服务存在带宽资源消耗大、图像数据隐私泄露、时效性难以满足等问题, 难以充分满足计算机视觉在行业应用上的多样化应用需求。而通信网络的双吉比特升级将促进视觉算法云边算法深层次协同。对基于云边协同的计算机视觉推理机制开展研究。首先对近年主流的云侧和边缘侧计算机视觉推理模型的优劣势进行了分析和阐述, 然后在此基础上对云边协同计算机视觉推理模型框架、部署机制等开展研究, 详细讨论模型分布式推理模型分割策略, 云边协同网络部署优化策略。最后通过数据协同、网络分区协同、业务功能协同 3 方面对云边协同深度推理未来的发展挑战进行了展望。

关键词: 计算机视觉; 深度学习; 云边协同

中图分类号: TP393

文献标识码: A

doi:10.11959/j.issn.1000-0801.2021107

Cloud-edge collaboration based computer vision inference mechanism

TANG Boheng, CHAI Xingang

China Mobile Research Institute, Beijing 100053, China

Abstract: The popularity of deep learning and cloud computing has promoted the widespread application of computer vision in various industries. However, centralized cloud inference services have problems such as high bandwidth resource consumption, image data privacy leakage, and high latency. It is hard that satisfy demand which requires diversified computer vision application. The dual gigabit upgrade of the communication network will promote depth collaboration of computer vision cloud-edge algorithms. Aiming to study the computer vision inference mechanism based on cloud-edge collaboration. Firstly, the advantages and disadvantages of the mainstream cloud and edge computer vision inference models in recent years were analyzed and explained, and on this basis, research on the cloud-edge collaborative computer vision inference model framework and deployment mechanism was carried out, model distributed reasoning model segmentation strategy, cloud-side collaborative network deployment optimization strategy was discussed in detail. In the end, the challenge and prospect of deep learning cloud-edge collaboration inference in future was discussed through data collaboration, network partition collaboration, and business function collaboration.

Key words: computer vision, deep learning, cloud-edge collaboration

1 引言

过去十几年中,计算机视觉的研究在公共安全、交通管理、城市治理、政府决策、医疗健康等领域取得了巨大进步。人工智能(artificial intelligence, AI)图形图像计算也纳入了“十四五”规划,加快布局行业产业应用。算法推理基于计算机视觉算法,在非人为干预环境下提取图像有效信息,并对图像进行理解与分析。单独使用云侧或边缘侧服务器进行推理面临网络时延导致服务时效性差的问题,极大影响用户体验,数据爆发式增长带来算力短缺,导致使用成本提高,因此有必要对计算机视觉算法部署与推理方式进行新的尝试。

云边协同是融合通信、算力、数据存储、应用服务的分布式开放平台,相对于云侧的全局性、长周期、高时延、大数据的计算特点。边缘计算的短周期特性可以更好地支持本地业务。因此边缘侧与云侧并不是简单的替代关系,而是互补协同的合作关系。通过对资源协同、数据协同、应用协同、服务协同等协同领域构建统一高效的协同框架,实现云边互补,资源融合。

云边协同的推理机制不同于传统的云边协同,除业务、数据、服务等协同外增加了算法协同。算法协同具备算力资源分配、准确率验证、模型分区调用等机制。云边协同的推理机制弥补边缘侧算力不足导致的低准确率现象,避免数据短时增长带来的算力波峰,减少云侧带宽压力,提升服务速率,解决数据不出域问题,有效地提高了系统稳定性和资源的整体使用效率。

本文着重介绍基于 AI 技术的推理性能优化方法,从系统以及模型等层面阐述推理性能优化的一般方法,并从实际应用出发,给出了具体优化的实践案例和取得的部分成果以及对未来发展方向的一些展望和思考。

2 主流云侧以及边侧计算机视觉推理框架

在过去几年提出的深度学习模型中,卷积神

经网络(convolutional neural network, CNN)是应用最广泛的视觉算法模型之一。以往,科研领域往往将云计算作为人工智能深度学习训练推理的首选方案,得益于云侧强大的计算资源,算法被快速迭代验证,因此准确性更高、功能性更多样的算法层出不穷。由于更接近数据源,边缘侧可以实现更高的服务吞吐量和资源利用率,有效地减少了数据传输带宽,提高服务执行效率。

2.1 云侧推理的现状与挑战

深度学习在图像领域的应用飞速发展。1989年,LeNet^[1]中首次提出使用卷积解决图像分类问题;2012年,AlexNet^[2]和VGG^[3]通过增加网络卷积深度,并使用更大的卷积核得到更高精度。2015年,GoogleNet^[4]引入Inception结构,开创式地对同一特征图并行运算不同大小和规模的卷积层。ResNet^[5]引入残差块的概念,增加了网络深层与浅层特征连接的shortcut结构,从而有效地缓解深层网络的梯度消失问题。至此网络深度首次超过100层。DenseNet^[6]提出一种修改后的残差块连接方式,通过每层的相互连接实现特征重用,提升网络效率。SENet^[7]在特征图间添加挤压和激励块以实现比ResNet更高的特征提取效率。

现有的网络结构和算法性能显示更深的网络深度,可以得到更高的模型精度,见表1,随之而来的是计算资源的指数级增长,云侧计算中心数据处理能力也将迅速耗尽。虽然可以在中心云侧部署大规模GPU资源池,但是在高并发数据流场景中,实施完善的算力资源池的整合、负载均衡和容灾策略仍然是一项重大挑战。

在数据传输方面同样面临更高风险,由于远离数据产生源,需要传输大量未经处理的数据至云侧服务器,无形间消耗大量网络带宽,不能保证较短的通信时延,因此云侧推理模型无法适用于实时性要求较高的领域(如合作式自动驾驶等)。

云侧推理模式在数据安全与保护方面有着先天性的不足,对于某些数据不出局的业务,云侧

表 1 云侧模型算法资源与精度比较^[8]

网络名称	网络深度	参数数量/MB	浮点运算数量/ 10^9 次	Top1 错误率
AlexNet	8	61	1.45	36.7%
VGG16	16	138	31.0	25.6%
Inception-v1	22	5	2.86	30.2%
Inception-v2	42	11	4.0	25.2%
Inception-v3	48	24	11.4	21.2%
Inception-v4	76	35	24.5	19.9%
ResNet-152	152	60	22.6	19.4%
ResNeXt	50	68	8.4	19.1%
DenseNet	201	20	8.4	13.6%

推理则无法有效满足需求。同时有数据截获、数据丢包、网络病毒等外部因素影响系统稳定性。

2.2 边缘侧推理的现状与挑战

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^[9-11]、知识蒸馏^[12-13]和使用更高效的算法层^[14]）；另一个方向根据边缘硬件结构重新设计网络模型^[15-16]，实现高效的网络推理。

比较著名的边缘侧网络举例如下。SqueezeNet^[17]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。SqueezeNext^[18]使用可分离卷积模块替换原有卷积层，网络参数量减少到 AlexNet 的 1/112。ShuffleNet^[19]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^[20]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^[21]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^[22]使用序列模型优化算法确定块属性。MNASNet^[23]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^[24]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^[25]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。MobileNetv3^[26]发布了两个版本，分别为 MobileNetv3-large（简称 large 版）和 MobileNetv3-small 版（简称 small 版），large 版相对于 small 版具有更多的瓶颈模块，使用神经体系结构搜索 NAS（neural architecture search）算法^[27]和 NetAdapt 算法^[28]改进网络结构，使 large 版比 small 版运行速度提升 25%，而在相同运行速度下 small 版的精度比 MobileNetv2 高 6.6%。

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^[29]。更准确的模型并不一定最适用于边缘侧推理，因此边缘侧的挑战是如何实时响应来自多个终端设备的请求，同时均衡硬件占用和能耗等资源消耗。这也就导致单纯的边缘侧计算不足以完成高精度模型推理。

由于边缘侧注重本网关下的数据处理，不足以完成在大场景（如城市级的交通调度、多网关下的数据处理等）的任务。同时边缘节点的算法模型加载与更新不具备灵活性，在更换计算任务后需要手动进行算法替换，增加人力和时间成本。

现有边缘侧只是单纯执行任务复用，即一个边缘节点作为一个任务接口，没有完全发挥计算网络节点的复用价值，依然存在不同边缘节点的重复计算而导致算力浪费的情况。

2.3 云侧和边缘侧推理优化策略带来的启示

由于云侧和边缘侧存在以上的种种问题，一种新的推理系统架构开始出现，云侧、边缘侧和

端侧相互延伸、协同运作。云侧作为复杂运算和资源调度中心,边缘侧辅助云侧进行基础模型运算或扩展算力进行多任务并行运算。相对于单一的云侧推理和边缘侧推理,云边协同的分布式计算构架有以下几个优点:

- 云侧可以提供更复杂、更精细的预测功能,弥补端侧功能单一的特点;
- 缓解由于数据流集中在云侧处理,导致计算资源紧缺的问题,有效消除算力波峰;
- 将骨干网络部署在边缘侧,减少原始数据传输对带宽的压力;
- 有效提升服务速率与服务反馈速率;
- 满足对特殊数据的保密性需要。

3 云边协同推理

随着应用场景的多样性,简单而单一的边缘侧网络模型已经不能满足大规模系统部署的要求。如何体现边缘侧靠近数据收集端的优势,同时具备云侧强大的计算能力,是下一代深度学习推理系统的研究方向。云边协同深度学习的研究从以下几个切入点进行探讨。首先在原有的边缘侧模型基础上,将云侧模型作为补充提高整体模型精度的方案。然后深入探讨了根据网络层进行云边划分的分布式部署策略。最后是从性能优化方向,寻求模型分割的最优效率和数据传输的最

优策略。

3.1 云侧对边缘侧进行补充和验证

Teerapittayanon 等^[30]提出分布式深度神经网络(distributed deep neural network, DDNN),该分布式神经网络可以根据云侧、边缘侧和终端侧不同节点的地理层级进行扩展与缩放。同时参考文献[30]中提出了6种不同的分布式深度神经网络,如图1所示。

对云侧网络和边缘侧网络使用联合训练同时进行权重训练,分别得到各自权重并进行部署。边缘侧网络使用判定器判断特征信息是否足以进行准确的推理,当判定为错误推理时,边缘侧抛出数据至云侧,由云侧更复杂的网络进行推理,得到更准确的推理结果。边缘侧判定器使用正规化熵值确定,定义为:

$$\eta(x) = -\sum_{i=1}^{|C|} \frac{x_i \ln x_i}{\ln |C|} \quad (1)$$

其中, C 为标注值, x 为预测值,最终 η 值分布在 $[0,1]$ 。设定抛出阈值 T ,边缘抛出的分类结果越多,系统准确性越低,但边缘侧算力资源利用率也就越高。为了使系统推理性能和准确性达到平衡,设置 $T=0.8$,此时边缘抛出率为 60%,准确率为 97%,系统达到最优的效率和准确性。对于多个边缘侧情况(如图1(c)、图1(e)、图1(f)所示),还提出了3种本地聚合器以聚合不同来源

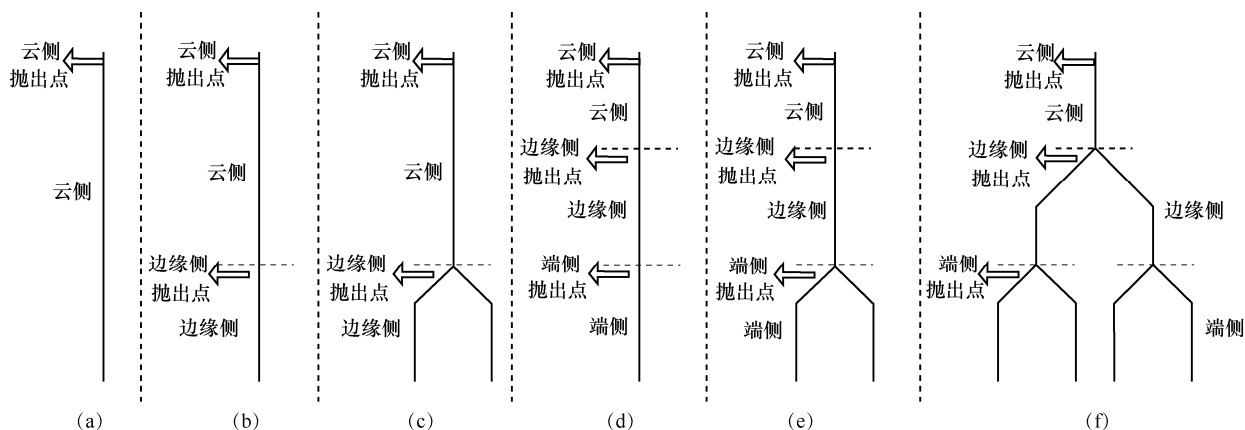


图1 云边协同分布式网络形式^[30]



的数据, 分别是最大池化 (max pooling, MP)、平均池化 (average pooling, AP) 和串联 (concatenation, CC)。其中, 边缘侧使用 MP 聚合、云端使用 CC 聚合得到最佳分类准确率。参考文献[30]虽然提出了一种提高边缘侧模型准确率的方案, 但是并未对模型进行优化, 同时边缘侧需要进行全部推理计算, 某些不通过的结果需要在云侧进行重新计算, 增加了整体计算量。

3.2 基于模型分割的分布式云边协同

根据不同的场景需求, 边缘侧往往使用查询的方法从云侧下载与场景相对应的完整网络模型自动部署, 然后进行边缘推理 (如 fog nodes^[31]、cloudlet^[32])。此种方式耗费大量的网络带宽, 更长的模型载入时间导致边缘侧的启动效率更慢。因此提出将模型进行分割, 部分模型部署在边缘侧, 另一部分部署在云侧的分布式协同网络。

其中一种以网络层为单位的分割, 确定网络分割点的过程一般可以分为 3 个步骤^[33]:

- (1) 测量和建模不同 DNN 层的资源消耗成本以及中间 DNN 层之间的传输数据;
- (2) 通过特定的层配置和网络带宽预测总成本;
- (3) 根据时延、能量需求等从候选划分点中选择最优的划分区域。

另一种模型分割是垂直分割, 与水平分区相反, 垂直分区将图层融合在一起, 并以网格的方式在垂直方向上对其进行分区, 从而将 CNN 图层划分为独立可分配的计算任务。

Jeong 等^[34]提出 IONN 结构, 不同于传统的边

缘服务器查询云端算法将网络模型从云侧发送至边缘节点的部署方式, IONN 将网络划分为多个分区逐个发送至边缘侧。在执行推理需求时, 优先选取边缘侧原有的模型分区进行复用。为了最优化网络分区上传顺序, 提出使用图结构的启发式算法, 如图 2 所示, 以获得最佳的分区网络查询和边缘加载性能。具体做法是分别预测每层网络分区在云侧的执行时间, 和边缘侧执行时间与云边数据传输的网络延迟时间总和, 两者对比得到每层网络分区执行的最优策略。如果云侧网络层运行时间短, 则数据直接在云侧运行。如果边缘侧复用或加载并运行网络层时间短, 则网络层在边缘侧执行。参考文献[34]也提出了网络分层策略。首先确定边缘侧网络层的最优加载偏好, 即网络分区效率高但开销小的期望。通过迭代网络分区, 分别计算网络在云侧和边侧的运行效率, 提取在运行效率高的网络分区部署在边缘侧。使用 AlexNet、ResNet、GoogleNet 和 MobileNet 分别进行测试。最终结果显示在刚开始执行 IONN 时, 边缘侧加载全部网络, 整体运行速率较慢, 但经过几次加载后, 由于复用边缘侧原有网络结构, 上述模型加载时间全部下降至 0.2 s 内, 加速效果较好。但是由于移动端侧可能在短时间内穿越不同网络网关, 系统会经常遇到运行时长峰值, 不利于系统稳定运行。

随后 Jeong 等^[35]提出 PerDNN 结构, 为了解决 IONN 资源占用波峰问题, 提出预测并提前加载模型的机制, 以减少移动端更换网关导致边缘

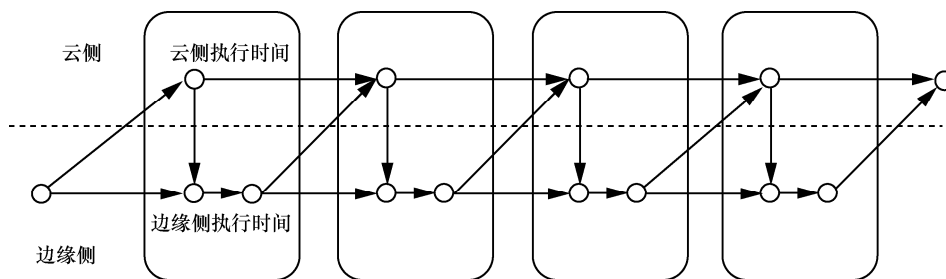


图2 图结构启发式算法

服务器冷启动对系统运行效率的影响。云侧服务器使用移动端侧返回的 GPS 或 Wi-Fi 地理位置信息, 预测移动至下一个未来网关的区域。未来网关区域的边缘服务器将提前加载当前边缘服务器中的模型资源。由于边缘侧服务器需要同时处理多个任务, 参考文献[35]提出使用 GPU 信息来估计资源占用的方法。通过对每个类型层训练随机森林模型, 预测部署在边缘侧不同层组合的运行时间。由于需要边缘侧进行不断的模拟预测, 因此预测对象需要具备轻量化特点。参考文献[35]提出将模型数据分为两部分: 一部分是不具备权重数据只包含模型结构和超参数的模型网络结构文件, 另一部分是模型权重。使用模型结构文件可以快速地预测模型推理效率, 然后由云端根据边缘侧 GPU 资源使用量配置运行策略。

Li 等^[36]提出网络分层部署边缘侧以减少云侧压力, 使用 INT 网络量化^[37]以减少存储量、传输和算力需求, 并提出了一些典型网络的分区策略, 如在 Inception 中将并行结构作为分区, 并输出 INT8 和 FP32 两种量化标准在云端进行处理。ResNet 中将残差块作为分区等。记录每个分区的性能并找到最佳的分区策略。最终得到网络平均速度增加 1.2 倍, 且精度下降小于 0.1%。

Tian 等^[38]提出 MDPO(mobility-included DNN partition offloading) 结构, 用于处理移动端侧的模型加载问题。作者将神经网络看作有向无环图(directed acyclic graph, DAG)进行描述, 记录边缘侧网络运行时间和传输时间作为图的边, 将模型最优分割问题转化为链拓扑的最小切割问题, 并使用 Yuri Boykov 算法^[39]解决, 最终得到模型最优分区。同时考虑端侧移动导致端侧切换网关而产生的边缘侧模型冷启动问题, 参考文献[38]还提出一种边缘模型在无线网关间的切换策略。由于并不能预先得知新场景下边缘侧和端侧的数据传输速度, 因此当端侧设备移动至新边缘侧网关时, 系统会同时执行边缘对边缘传输和边缘传

输至端侧再上传至边缘侧两种策略。取最短传输路径以确保整体效率最高。

3.3 分布式模型云边部署的优化策略

使用边缘计算虽然更接近数据收集端, 减少数据传输时延, 但是如何有效地分配设备集群资源, 充分调度分布式算力进行推理仍然是一个充满挑战的问题。Zhao 等^[40]在 DeepThings 中提出融合切片分区方法, 不同于将卷积层按层划分, 参考文献[40]使用纵向划分将卷积层划分为自上而下的可分区的任务, 并提出一种边缘资源分配策略以优化边缘侧的算力使用和存储复用策略。其论证基于网络层进行分区时会增加内存的资源占用, 因此使用跨层分区结构可以有效减少资源占用和通信开销。基于 YOLOv2 的检测原理设置单阶目标检测网络, 首先将特征图划分成不重叠的网格块, 然后在运算过程中递推延伸网格边缘宽度, 每个网格块作为切片的分区。由于在推理中延伸出的网格宽度在不同分区存在重叠部分, 因此参考文献[40]设计了一种重叠区域重用机制, 计算完成的重叠区域特征将被其他计算节点分享共用, 以此来减少由于重复计算而产生的算力浪费。系统通过云侧确定每个边缘节点的资源占用情况, 使用动态负载均衡分配每个分块的执行节点。经过实验, DeepThings 最终将边缘侧的运行内存减少 69%, 权重参数减少 10.3%。但此种方法在使用多边缘设备时, 每增加 1 台边缘侧设备, 通信开销呈线性增长 8%。因此当有足够带宽时, 系统可以在大型异构集群网络中提供更好的拓展性和负载均衡性能。

Huang 等^[41]提出利用分布式网络减少权重传输时间的调度优化算法。使用式(2)来评价云侧和边缘侧的模型精度和硬件空间使用关系:

$$A = a \ln(R + b) + c \quad (2)$$

其中, A 是模型精度比例, R 是空间剩余率, a 、 b 、 c 是模型在设备中运行的时间常数。并提出系统运行时间和模型精度比 W_i , 使整体 W 最小来



最优化系统性能:

$$W_i(R_i, x_i) = \frac{R_i}{a \ln(R_i + b) + c} \sum_{j=0}^m \frac{x_{i,j}}{B_j} D_i \quad (3)$$

其中, B 是约束带宽, D 是数据量。为使 $W_i(R_i, x_i)$ 达到最小, 参考文献[41]使用模拟退火重排算法寻求最优方案。首先以每个边缘节点的数据处理总量最小为标准部署节点集群, 然后将每个节点模型进行重排, 每经过一次重排便判定系统状态, 如果总消耗减小则继续进行当前状态转移, 反之减少该种状态转移权重。使用模拟退火算法系统最终减少 30%~50% 的系统存储资源, 并且减少 66% 的数据传输量。以上方法虽然效果较好, 但云侧服务器只作为任务分配功能, 边缘缺乏和云侧服务器的数据交互, 并未充分利用云侧计算资源。

2018 年 Ko 等^[42]提出一种将卷积层部署在边缘侧的特征空间编码方式。参考文献[42]指出, 越靠近输入层, 网络特征图中特征信息越稀疏, 特征信息密度越低。由于特征图具有较高的零占比, 因此使用行程长度压缩算法和哈夫曼编码相结合的方式, 即 JPEG 编码的最后一部分压缩算法对特征图进行编码。最终在没有精度损失的情况下得到了 3~10 倍的压缩率, 在 1% 精度损失情况下实现 5~50 倍的压缩效率, 有效减少了网络特征图在设备间的传输压力。但是并未对使用联合训练、训练校正等方法对高压压缩率导致的精度损失进行改进。同时并未对较深层网络的分区策略提出解决方案。

参考文献[43]对云边缘协同推理网络的时延和计算强度进行分析, 将推理物理层分为云侧、边缘侧和端侧, 分别由上一层为下一层分配任务提供模型。不同基于任务启动时间来决定优先级的资源分配方式, 参考文献[43]使用基于蚁群优化的启发式算法来优化边缘节点的任务调度。将路径定义为:

$$P_{v_j}^{a_i} = \frac{[\tau_{i,j}]^\alpha [\eta_{i,j}]^\beta}{\sum_{v_j \in V} [\tau_{i,j}]^\alpha [\eta_{i,j}]^\beta}, 1 \leq \alpha, \beta \leq 10 \quad (4)$$

其中, $\tau_{i,j}$ 是 v_j 节点的任务 a_i 的信息素, $\eta_{i,j}$ 是由 $\delta(1 - \mu_{v_j})$ 确定的蚁群启发信息。

$$\mu_{v_j} t + 1 = \mu_{v_j}(t) + P \quad (5)$$

其中, P 是当前任务计算资源占总任务比例。算法通过迭代蚁群最佳路径得到最佳的任务分配策略。并使用拉格朗日算子对网络分层的延迟、能耗、运行时间等方面进行分析, 得到模型加载算法。但使用蚁群迭代算法计算量大, 算法延迟高。同时要求网络延迟低, 在迭代过程中如有数据丢包情况出现, 则迭代处于等待状态, 这会降低整个切换模型效率。

4 挑战与展望

由于边缘计算与云计算优势互补, 云边协同的模型推理是许多智能应用落地工作的焦点。尽管业界做出了巨大努力, 但是仍然存在许多尚未解决的问题和挑战。更高的推理精度, 往往需要使用更复杂的网络。但是在产业落地过程中需要考虑能耗、性能、成本等诸多问题, 具有最佳精度的网络和部署策略可能并不具备最佳的落地价值, 需要权衡效率、性能等多方面的问题。

4.1 数据协同的挑战

(1) 基于模型数据优化的协同方式

目前, 云边协同网络更多的是将云侧模型加载至边缘侧, 虽然有部分文献提到使用边缘侧自动备份转移的方式, 但是根据不同场景优化的备份转移方式并未过多披露, 且无行业统一标准, 这致使在不同硬件和网络环境中无法进行规模性应用。因此, 还需结合场景研究更具体的分配方案, 并对其进行深度优化和兼容性处理。

(2) 基于推理数据流的优化协同方式

越来越多的应用场景将数据安全置于首位, 如何在实现功能的同时考虑数据不出域和加密性

传输是研究重点。而现有方案中虽然有诸如 JPEG 编码的方式,但是缺少统一高效的加密和传输协议,这些协议将是未来智能云边协同监控全面推广面临的重要问题。

4.2 网络分区协同的挑战

现有网络的分区方案实现的是从网络层级的横向或纵向的分区,最终达到和完整网络接近或相似的性能和准确率。因此,从结果来看,现有方案只相当于把计算资源从云侧下放到边缘侧,并未实现性能优化的跨越式发展。近年深度学习中诸如显著性目标^[44]、注意力机制^[45]、元学习^[46]等方向的研究快速发展,未来如何将上述新方法应用于云边协同网络,同时这些新的结合会带来哪些性能提升,都还需进一步研究。

4.3 业务功能分区的挑战和机遇

现有方案多数从深度学习的图像分类领域作为切入点,但对监控领域(如目标检测、属性分析、关键点检测、行为分析、运动状态、重识别等)的研究较少。监控领域更具有商业应用价值,并且具备更高的研究投入回报比。希望未来可以有更多不同场景的云边协同分布式算法框架出现。

5 结束语

本文从传统的计算机视觉深度学习网络切入,以卷积神经网络为基础,对云侧和边缘侧的网络推理机制进行了调研和阐述。综合云侧网络和边缘侧网络的各自劣势,预测云边协同推理方式的发展前景,并分别从云边互补、网络分区和策略优化 3 个主要方向描述云边协同深度学习推理的发展现状。云边协同的算法推理机制可以有效提高云边资源调用率、减少重复运算导致的算力开支、减少数据带宽。最后从数据、网络分区策略和业务功能 3 个方面提出未来可能面临的机遇和挑战。

参考文献:

[1] VÉSTIAS M P, DUARTE R P, DE SOUSA J T, et al. Moving

deep learning to the edge[J]. *Algorithms*, 2020, 13(5): 125.

- [2] LE CUN Y, JACKEL L D, BOSER B, et al. Handwritten digit recognition: applications of neural network chips and automatic learning[J]. *IEEE Communications Magazine*, 1989, 27(11): 41-46.
- [3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. *Advances in neural information processing systems*, 2012(25): 1097-1105.
- [4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. *arXiv preprint arXiv: 1409.1556*, 2014.
- [5] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE Press, 2015: 1-9.
- [6] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//*Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2016: 770-778.
- [7] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2017: 4700-4708.
- [8] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2018: 7132-7141.
- [9] HAN S, POOL J, TRAN J, et al. Learning both weights and connections for efficient neural networks[J]. *arXiv preprint arXiv:1506.02626*, 2015.
- [10] HAN S, MAO H, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding[J]. *arXiv preprint arXiv:1510.00149*, 2015.
- [11] GUO Y, YAO A, CHEN Y. Dynamic network surgery for efficient dnns[J]. *arXiv preprint arXiv:1608.04493*, 2016.
- [12] MAO H, HAN S, POOL J, et al. Exploring the regularity of sparse structure in convolutional neural networks[J]. *arXiv preprint arXiv:1705.08922*, 2017.
- [13] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. *arXiv preprint arXiv:1503.02531*, 2015.
- [14] CHEN J, RAN X. Deep Learning With Edge Computing: A Review[J]. *Proceedings of the IEEE*, 2019, 107(8): 1655-1674.
- [15] LACEY G, TAYLOR G W, AREIBI S. Deep learning on fpgas: Past, present, and future[J]. *arXiv preprint arXiv:1602.04283*, 2016.
- [16] NING Z, FENG Y, COLLOTTA M, et al. Deep learning in edge of vehicles: Exploring trirelationship for data transmission[J]. *IEEE Transactions on Industrial Informatics*, 2019, 15(10): 5737-5746.



- [17] 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.
- [18] GHOLAMI A, KWON K, WU B, et al. Squeezenext: Hardware-aware neural network design[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2018: 1638-1647.
- [19] ZHANG X, ZHOU X, LIN M, et al. Shufflenet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.
- [20] HUANG G, LIU S, VAN DER MAATEN L, et al. Condensenet: an efficient densenet using learned group convolutions[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2752-2761.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1314-1324.
- [27] ZOPH B, LE Q V. Neural architecture search with reinforcement learning[J]. arXiv preprint arXiv:1611.01578, 2016.
- [28] SHI W, CAO J, ZHANG Q, et al. Edge computing: vision and challenges[J]. IEEE Internet of Things Journal, 2016, 3(5): 637-646.
- [29] 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.
- [30] TEERAPITTAYANON S, MCDANEL B, KUNG H T. Distributed deep neural networks over the cloud, the edge and end devices[C]//Proceedings of 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2017: 328-339.
- [31] BONOMI F, MILITO R, ZHU J, et al. Fog computing and its role in the internet of things[C]//Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing. Piscataway: IEEE Press, 2012: 13-16.
- [32] SATYANARAYANAN M, BAHL P, CACERES R, et al. The case for VM-based cloudlets in mobile computing[J]. IEEE Pervasive Computing, 2009, 8(4): 14-23.
- [33] WANG X, HAN Y, LEUNG V C M, et al. Convergence of edge computing and deep learning: A comprehensive survey[J]. IEEE Communications Surveys & Tutorials, 2020, 22(2): 869-904.
- [34] JEONG H J, LEE H J, SHIN C H, et al. IONN: incremental offloading of neural network computations from mobile devices to edge servers[C]//Proceedings of the ACM Symposium on Cloud Computing. Piscataway: IEEE Press, 2018: 401-411.
- [35] JEONG H J, LEE H J, SHIN K Y, et al. PerDNN: offloading deep neural network computations to pervasive edge servers[C]//Proceedings of 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS). Piscataway: IEEE Press, 2020: 1055-1066.
- [36] LI G, LIU L, WANG X, et al. Auto-tuning neural network quantization framework for collaborative inference between the cloud and edge[C]//Proceedings of International Conference on Artificial Neural Networks. Heidelberg: Springer, 2018: 402-411.
- [37] WU J, LENG C, WANG Y, et al. Quantized convolutional neural networks for mobile devices[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4820-4828.
- [38] TIAN X, ZHU J, XU T, et al. Mobility-included DNN partition offloading from mobile devices to edge clouds[J]. Sensors, 2021, 21(1): 229.
- [39] BOYKOV Y, KOLMOGOROV V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision[J]. IEEE Transactions on Pattern Analysis and Computer Vision Intelligence, 2004, 26(9): 1124-1137.
- [40] ZHAO Z, BARIJOUGH K M, GERSTLAUER A. Deepthings: distributed adaptive deep learning inference on resource-constrained IoT edge clusters[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(11): 2348-2359.
- [41] HUANG Y, ZHU Y, FAN X, et al. Task scheduling with optimized transmission time in collaborative cloud-edge learning[C]//Proceedings of 2018 27th International Conference on Computer Communication and Networks (ICCCN). Piscataway: IEEE Press, 2018: 1-9.
- [42] KO J H, NA T, AMIR M F, et al. Edge-host partitioning of deep neural networks with feature space encoding for resource-constrained internet-of-things platforms[C]//Proceedings

- of 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Piscataway: IEEE Press, 2018: 1-6.
- [43] XU S, ZHANG Z, KADOCH M, et al. A collaborative cloud-edge computing framework in distributed neural network[J]. EURASIP Journal on Wireless Communications and Networking, 2020, 2020(1): 1-17.
- [44] WANG W, LAI Q, FU H, et al. Salient object detection in the deep learning era: an in-depth survey[J]. IEEE Transactions on Pattern Analysis and Computer vision Intelligence, 2021.
- [45] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7794-7803.
- [46] 朱应钊, 李嫚. 元学习研究综述[J]. 电信科学, 2021, 37(1): 22-31.
- ZHU Y Z, LI M. Review on meta-learning[J]. Telecommunications Science, 2021, 37(1): 22-31.

[作者简介]



唐博恒（1992- ），男，中国移动通信有限公司研究院工程师，主要研究方向为智慧城市、视频监控、计算机视觉、图像算法与算法的应用落地。



柴鑫刚（1976- ），男，中国移动通信有限公司研究院高级工程师，主要研究方向为视频云、计算机视觉、智慧城市等相关的关键技术与产品创新。