

一种基于多智能体的分布式深度神经网络算法

王 闯¹, 沈苏彬²

(1. 南京邮电大学 物联网学院, 江苏 南京 210003;
2. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘 要: 深度神经网络由于其良好的非线性逼近能力与泛化能力而被应用于物联网数据的分类和预测。智能家居作为典型的物联网应用场景, 通常将家庭中采集的数据传输到云端, 使用深度神经网络单智能体集中处理。以云计算中心的数据处理方案会导致较长的网络延迟以及用户隐私数据的泄露。文中将采用多智能体模型, 在深度神经网络模型上添加分支结构, 利用分支点将神经网络分为可以部署在不同智能体的浅层部分和深层部分, 设计了基于多智能体协同的深度神经网络的数据分类算法; 基于边缘计算模型, 在边缘设备上部署浅层神经网络智能体, 云服务器设备上部署深层神经网络智能体, 以构建边缘与云端协同的多智能体, 仿真实验和测试了该算法。仿真实验表明, 该算法可以减少智能家居的数据处理时间, 有效地保护用户隐私。

关键词: 边缘计算; 多智能体; 分布式; 深度神经网络; 智能家居

中图分类号: TP391.4; TP274.2

文献标识码: A

文章编号: 1673-629X(2021)12-0045-05

doi: 10.3969/j.issn.1673-629X.2021.12.008

A Distributed Deep Neural Network Algorithm Based on Multi-agent

WANG Chuang¹, SHEN Su-bin²

(1. School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Telecommunications & Information Engineering, Nanjing University of
Posts and Telecommunications, Nanjing 210003, China)

Abstract: Deep neural networks are used to classify and predict the data collected from the Internet of Things (IoT) because of their excellent nonlinear approximation and generalization. As a typical application scenario of the IoT, smart home usually transmits the data collected in the home to the cloud, and uses a single agent of deep neural network for centralized processing. The data processing scheme based on cloud computing center will lead to long network delay and user privacy data leakage. We adopt a multi-agent model, add a branch structure to the deep neural network model, use branch points to divide the neural network into a shallow part and a deep part that can be deployed in different agents, and design a deep neural network based on multi-agent collaboration network data classification algorithm. Based on the edge computing model, a shallow neural network agent is deployed on edge devices, and a deep neural network agent is deployed on cloud server devices to form a multi-agent for edge-cloud collaboration. The algorithm is simulated and tested. Simulation experiments show that the proposed algorithm can reduce the data processing time of smart homes and effectively protect user privacy.

Key words: edge computing; multi-agent; distributed; deep neural network; smart home

0 引 言

随着物联网设备的增多, 由传感器、嵌入式设备产生的结构化、非结构化或者半结构化的数据日益增多。这些具有高容量 (Volume), 高速度 (Velocity) 和多类型 (Variety) 的“3V”特点的数据可以称之为大数据^[1]。处理这些数据以及如何从这些数据中挖掘出有价值的

知识使物联网能够智能化地提供服务 and 可靠的决策已经成为急需解决的问题与挑战^[2]。鉴于深度神经网络具有高效的数据特征提取与分析能力, 现已被广泛应用在智能家居等物联网应用中^[3-4]。

目前大多数物联网设备的计算能力有限, 通常将传感器采集到的数据传输到云计算中心进行分析处

收稿日期: 2020-12-10

修回日期: 2021-04-13

基金项目: 江苏省未来网络前瞻性研究项目 (BY20130951108)

作者简介: 王 闯 (1994-), 男, 硕士研究生, 研究方向为物联网边缘计算; 沈苏彬, 博导, 教授, CCF 高级会员 (E200005482S), 研究方向为物联网及其应用、未来网络及其应用。

理。物联网应用通常需要基于数据提供实时服务,并且通常依赖于低存储和计算能力有限的设备,以及带宽有限的网络连接,因此,需要将数据计算从云计算中心转移到物联网边缘设备,实现近传感器计算和近传感器智能,在资源受限的嵌入式设备中运行机器学习甚至深度学习程序^[5]。由于嵌入式设备资源受限,部署深度神经网络应用通常会面临以下三个主要问题:(1)嵌入式设备的计算能力有限,无法满足终端设备的实时性要求;(2)嵌入式设备的存储空间较小,不能提供足够的存储空间来存储程序运行所需的数据;(3)运行深度神经网络过程中由于需要大量的数值计算与频繁的数据读取,高额的能耗会大幅减少设备的工作时长。因此一般采用特征剪枝和权重量化等方式对模型进行压缩^[6],在嵌入式设备上使用压缩后的深度学习网络模型,但是会造成系统准确率下降;而把传感器数据输入到云端的大型深度神经网络中,还需要解决通信延迟和隐私方面的问题。

文中将神经网络进行划分,浅层部分部署到物联网边缘设备上构成边缘智能体,深层部分部署到云服务器上构成云端智能体。边缘与云端智能体构成系统合作的多智能体系统,在保证分类推理准确率的前提下,使用边缘设备承担深度神经网络的浅层部分推理,减少数据传输带来的延迟和保护用户隐私。

1 相关工作

1.1 边缘计算

边缘计算是与云计算相对的一种计算模型或服务。云计算强调依托服务器丰富的计算资源将数据进行集中式处理,而边缘计算是将物联网嵌入式设备生成的数据在物联网中进行本地处理。

作为缓解资源拥塞升级的策略,边缘计算已成为满足物联网和本地化计算需求的技术热点^[7],将计算压力从集中式数据中心卸载到分布在网络上的许多计算节点,可以减少数据传输带来的延迟^[8]。边缘计算模型不仅可降低数据所需的网络带宽,同时能较好地保护隐私数据,降低终端敏感数据隐私泄露的风险^[9]。

1.2 深度神经网络

深度神经网络成功的原因之一是它们能够在连续的非线性层上学习更高级别的特征表示。随着硬件技术和机器学习方法的进步,可以培养更深层次的网络,这些网络进一步提高了机器分类的性能^[10]。ImageNet 的挑战可以说明神经网络层数的发展趋势,因为最先进的方法已经从 8 层 (AlexNet) 发展到 19 层 (VGGNet),并且在四年的时间内推进到 152 层 (ResNet)^[11-12]。然而,向更深层网络的发展极大地增加了前馈推理所需的时延和计算资源^[13]。

使用深度神经网络处理复杂任务时,通常需要构建层数较深的网络模型来增强模型对数据特征的提取能力。虽然增加深度神经网络的网络层数可以在一定程度上提高网络模型处理数据的能力,但也会导致深度神经网络的计算量与参数数量急剧增加,由此导致应用深度神经网络的物联网设备需要配置强大的计算能力和巨大的存储空间。

1.3 物联网多智能体

多智能体系统是多个智能体组成的集合,它的目标是将大而复杂的系统建设成小的、彼此互相通信和协调的、易于管理的系统。智能体通常分为完全合作式、完全竞争式、混合关系式三种类型^[14]。

在智能物联网环境下,每个设备可视为一个智能体,智能体在单独行动的同时,也要学会与其他的智能体进行交互协作,提高物联网提供智能化服务的能力。

2 分布式深度神经网络

深度神经网络由网络层组成,因此可以在边缘设备上部署经过压缩优化的浅层神经网络,在云服务器设备上部署深层神经网络,以此构建一个如图 1 所示的混合云端与终端的基于分布式计算层级的分布式深度神经网络。物联网边缘设备和云服务器组成一个合作式的多智能体系统。分布式深度神经网络是具有中间分支结构的神经网络。通常情况下,在深度神经网络的早期阶段学习的特征可以正确地推断出数据总体的大部分,因此在主网络上设置分支让分类准确的数据提前退出可以减少时延和计算资源。分布式深度神经网络通过中间分支划分为浅层和深层两个部分。

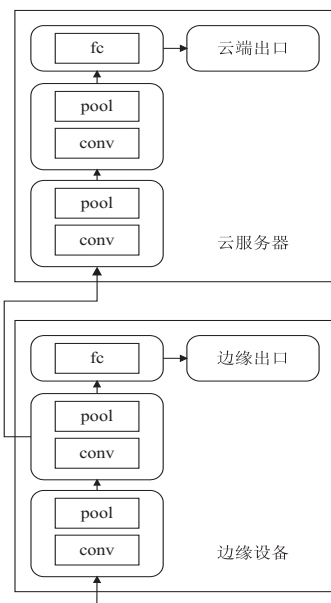


图 1 层级的分布式深度神经网络

在数据分类中,希望分类模型在训练数据上学到的预测数据分布与真实数据分布越相近越好,如果退

出点的分类结果信息熵越小,则说明该分支的退出点的分类器对于正确标记该测试样本的预测结果置信度越高。合作式多智能体整体通常需要最大化全局的期望回报,在分支点设置信息熵阈值来评估分支点的分类效果。当分支点的分类信息熵小于阈值时,深度神经网络的推理执行可以在本地终端上完成分类并退出,进行快速的局部推断;当大于所设定的阈值,需要进一步进行额外处理时,中间数据可以传输至云服务器端,进一步利用云端的深层深度神经网络进行处理,以提高系统的表现精度。

将一个在终端设备上的小型神经网络模型(更少的参数)和一个在云上的大型神经网络模型(更多的参数)组合起来。终端设备上的小型神经网络模型可以快速地初始化数据提取,并分析出这个模型是否是满足要求的。另外,终端设备还可以通过云端的大型神经网络模型执行进一步的程序并完成最终的分类。神经网络的中间层输出可以设计为比传感器输入小得多(例如,来自摄像机的原始图像),因此可以减少终端设备和云端之间所需的网络通信。此外,由于使用了从终端设备处理数据的方法代替原始传输数据的方法,该系统可以更好地保护个人隐私。

3 多智能体协同的深度神经网络

在具有 L 层的深度神经网络模型中,令 $\vec{x} = [x_1, x_2, \dots, x_i, \dots, x_m]$, $i = 1, 2, \dots, m$ 为输入向量, $\vec{y} = [y_1, y_2, \dots, y_i, \dots, y_n]$, $i = 1, 2, \dots, n$ 为输出向量, $h^l = [h_1^l, h_2^l, \dots, h_j^l, \dots, h_{sl}^l]$, $j = 1, 2, \dots, sl$ 为第 l 层神经元的输出,其中 sl 为第 l 层神经元个数。设 w_{ij}^l 为从 $l-1$ 层的第 j 个神经元与 l 层的第 i 个神经元之间的连接权重, b_i^l 为第 l 层第 i 个神经元的偏置,那么:

$$h_j^l = \varphi(\text{net}_j^l) \quad (1)$$

$$\text{net}_i^l = \sum_{j=1}^{sl-1} w_{ij}^l h_j^{l-1} + b_i^l \quad (2)$$

其中, net_i^l 为 l 层第 i 个神经元的输入, $\varphi(\cdot)$ 为神经元的激活函数。假定有 m 个训练样本 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, 为对应输入 x_i 的期望输出。BP 算法通过最优化各层神经元的输入权值以及偏置,使得神经网络的输出尽可能地接近期望输出,以达到学习的目的。

3.1 适合于多智能体的深度神经网络

通过解决与出口点相关联的损失函数的加权和的联合优化问题来训练具有分支结构的神经网络。一旦网络被训练,模型利用分支出口允许样本提前退出,从而降低推理成本。在每个分支出口,使用分类结果的信息熵作为分类的置信度的度量。如果测试样本的熵

低于学习阈值,意味着分类器在预测中是有效的,则样本在该出口点处以预测结果离开网络,并且不由较高网络层处理。如果熵值高于阈值,则该出口点处的分类器被认为是不可靠的,并且样本继续到网络中的下一个出口点。如果样本到达最后一个出口点,这是主神经网络的最后一层,它总是执行分类。

通过在整个网络的某些位置添加退出分支(简称分支)来修改标准深度网络结构。这些早期退出分支允许在网络的早期阶段准确分类的样本提前退出,对于浅层网络不能准确分类的数据则需要更深层的网络进行分类。

对于分类任务,通常使用交叉熵损失函数作为目标函数。交叉熵是香农信息论中的一个重要概念,主要用于度量两个概率分布间的差异性信息,也就是交叉熵的值越小,两个概率分布越接近。在分支网络训练过程中,先通过 SoftMax 函数对网络层的输出值进行归一化处理,再对其归一化处理后的值进行计算得到交叉熵值,作为神经网络的损失函数。

将 x 定义为一个输入样本, y 定义为该输入样本的真实标签, \hat{y} 定义为该输入样本的预测输出, S 定义为所有可能的样本标签集合, θ 定义为一个分支网络层从入口点到退出点的参数集合, f_{exit} 定义为模型的第 n 个模型分支退出点的输出。

将损失函数用公式进行如下表达:

$$L_n(\hat{y}, y; \theta) = - \frac{1}{|S|} \sum_{s \in S} y_s \log \hat{y}_s \quad (3)$$

其中,预测输出 \hat{y} 表达为:

$$\hat{y} = \text{softmax}(z) = \frac{\exp(z)}{\sum_{s \in S} \exp(z_s)} \quad (4)$$

网络层输出 z 表达为:

$$z = f_{\text{exit}}(x; \theta) \quad (5)$$

对于分支网络模型,采用集中式学习、分布式部署执行的策略,集中式的学习方法用来处理环境不稳定性问题以及考虑多智能体的联合动作效应,需要各个退出点的损失函数值乘以其权重来进行联合优化。将 ω_n 定义为每个分支模型的权重, N 为所有退出点的数量,因此其最终损失函数表达式为:

$$L(\hat{y}, y; \theta) = \sum_{n=1}^N \omega_n L_n(\hat{y}, y; \theta) \quad (6)$$

该算法在前馈过程中,训练数据集通过网络,包括主支路和支路,记录神经网络在所有出口点的输出,并计算出神经网络的误差。在反向传播中,误差通过网络传递回来,并使用梯度下降更新权重。使用随机梯度下降算法 Adam 进行模型训练。

当测试样本在训练好的分支网络模型上进行测试时,最终会经过网络层的计算,在退出点产生一个输出

结果 z , 使用 SoftMax 函数对其输出进行归一化, 生成一个 $0 \sim 1$ 之间的所有类概率集, 其中每个类标签的预测概率定义为 y_s , 所有可能的类标签集合定义为 S , 则将该退出点的样本输出信息熵定义为:

$$\text{entropy}(y) = \sum_{s \in S} y_s \log y_s \quad (7)$$

3.2 边缘与云端智能体的任务分割与协同

卷积神经网络的基本结构由输入层、卷积层、池化层、全连接层及输出层构成。卷积层由多个特征面组成, 每个特征面由多个神经元组成, 它的每一个神经元通过卷积核与上一层特征面的局部区域相连, 卷积核是一个权值矩阵。深度神经网络的卷积层通过卷积操作提取输入的不同特征, 第 l 层卷积层提取低级特征如边缘、线条、角落, 更高层的卷积层提取更高级的特征。

池化层紧跟在卷积层之后, 同样由多个特征面组成, 它的每一个特征面唯一对应于其上一层的一个特征面, 不会改变特征面的个数。池化层是神经网络的重要组成部分, 通过减少卷积层之间的连接, 降低运算复杂程度, 同时改善结果, 使结构不容易出现过拟合。

在神经网络结构中, 经多个卷积层和池化层后, 连接着一个或多个的全连接层。全连接层中的每个神经元与其前一层的所有神经元进行全连接。全连接层可以整合卷积层或者池化层中具有类别区分性的局部信息。为了提升神经网络的性能, 全连接层每个神经元的激励函数一般采用 ReLU 函数。最后一层全连接层的输出值被传递给一个输出层, 可以采用 SoftMax 逻辑回归进行分类。

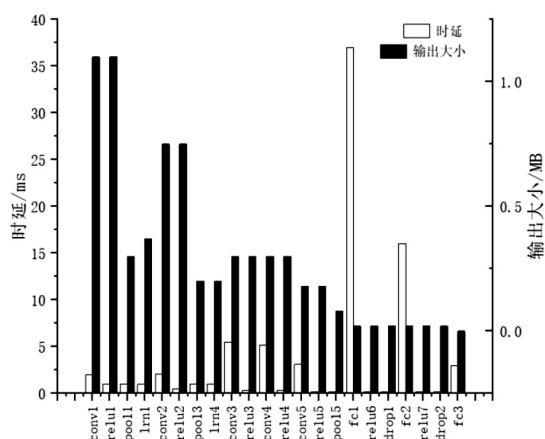


图 2 Alex Net 各层运行时延与输出

图 2 中在层级粒度上显示了 Alex Net 运行时的情况。根据每个层在网络中的类型和位置, 每一层的运行时间和输出数据大小也有所不同。卷积层和池化层的延迟相对较小, 而全连接层的延迟较高。卷积层和池化层主要位于网络的前端, 而完全连接层则位于后端, 原始数据经过卷积层和池化层到达后端远远小于

原始数据。将卷积神经网络分区执行, 在边缘设备中执行神经网络的浅层部分, 其余部分在具有更多计算资源的云服务器执行。

多智能体系统是一组自主的, 相互作用的实体, 它们共享一个共同的环境, 利用传感器感知, 并利用执行器作动。多智能体系统提供了分布式数据处理的视角, 在智能家居数据分类中为了提高识别的响应速度以及节省网络带宽, 分类任务会优先在本地边缘设备的模型上进行, 仅当本地识别结果的置信度不满足置信度阈值时, 才考虑是否请求到云服务器进行计算。当训练好任务需求的分支网络时, 同时为分支网络中的不同神经网络层的时延和输出数据量大小训练回归模型, 以此估算神经网络层在边缘设备上和云服务器上的运行时延; 回归模型将被用于寻找出符合任务时延需求的退出点以及模型切分点。

深度神经网络的最佳划分点取决于其拓扑结构, 它体现在每一层的计算延时和输出数据大小的变化上, 此外网络带宽不同导致数据从边缘段传输到云端时的传输时延不同。在特定网络带宽 B 下, 该神经网络模型有 N 个分割点, ET_i 为第 i 层在边缘设备上的运行时间, CT_i 为第 i 层在云服务器上的运行时间, O_i 为第 i 层输出计算边缘设备和云服务器的总运行时间。

$$T_i = \sum_{i=1}^j ET_i + \sum_{i=j+1}^N CT_i + \frac{O_j}{B} \quad (8)$$

在神经网络模型运行时, 遍历计算不同分割点的运行总时间, 以 $\min(T_i)$ 为边缘端智能体与云端智能体的协同合作策略, 将深度神经网络划分部署。

4 实验

4.1 实验装置

实验使用 Raspberry Pi 3b 模拟智能家居边缘设备, 具有 1 GB 运行内存, 运行 debian 9 操作系统。使用个人计算机模拟云服务器, 该计算机具有 CPU I7-8750、GPU GTX-1060Ti 以及 16 GB 内存运行 Ubuntu 18 操作系统。

Chainer 是一个专门为高效研究和开发深度学习算法而设计的开源框架。目前大多数深度学习框架都是基于 Define-and-Run 的方案, 而 Chainer 采用 Define-by-Run 的方案, 神经网络定义在运行时即时定义, 允许网络动态更改, 可以更加灵活地构建神经网络^[15]。实验使用分布式深度学习开源框架 Chainer 来构建具有分支的卷积神经网络。

4.2 实验数据集

用于测试分支神经网络分类效果的数据集为开源数据集, CIFAR-10 数据集。CIFAR-10 数据集被划分成了 5 个用于训练的数据子集和 1 个用于测试的数据

子集,每个子集均包含 10 000 张图片。测试集的图片是从每个类别中随机挑选的 1 000 张图片组成的,训练集则以随机的顺序将剩下的 50 000 张图片进行分组。不过一些训练集可能出现包含某一类图片比其他类的图片数量多的情况。训练集包含来自每一类的 5 000 张图片,一共 50 000 张训练图片。官方给出了多个 CIFAR-10 数据集的版本,文中测试实验使用 CIFAR-10 python 版本。

4.3 实验结果与分析

根据神经网络每层的运行时延,输出数据大小以及实际的网络带宽在训练好的分支神经网络模型上得出最佳分割点,将网络模型划分部署在 Raspberry Pi 3b 和个人计算机上。对比直接部署在计算机上的神经网络模型,所需时延以及准确率如图 3 所示。

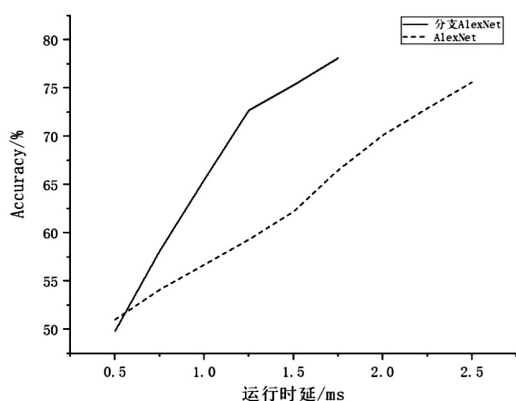


图3 分区 AlexNet 网络和 AlexNet 运行时延

当数据在 Raspberry Pi 3b 中满足提前推出点要求时,模型推理结束减少运行时延。在相同准确度要求的情况下,具有分支退出点的深度神经网络相比较于原神经网络模型所需的推理时间显著减小,提高了智能家居用户图像数据的处理效率,并且不需要将用户的原始图像数据传输到云端,保障了用户的隐私安全。

对于具有多个分支出口点的 AlexNet 网络,在不同带宽的情况下,划分点变化如图 4 所示。

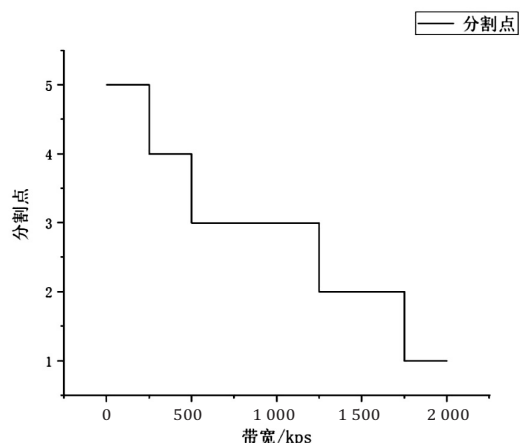


图4 最佳退出点划分点

可以看到,随着带宽的增加,最佳出口点会变得更低。在拥有足够带宽的情况下,神经网络模型的更多部分将会被划分到云服务器中运行。此外在带宽良好的情况下,由于边缘部分只执行少量的推理,大部分数据不会在边缘退出,会在云服务器中执行进一步推理。在没有足够带宽的情况下,最佳出口点会变得更低,神经网络模型的更多部分将会被划分到边缘设备中执行,由于边缘部分的神经层数增加了推理结果的准确性,大部分数据会在边缘结束推理,少部分需要在云服务器中进一步执行。

5 结束语

物联网的数据日益增加,使用深度神经网络等数据挖掘技术可以提取物联网数据中的潜在知识或模式,使物联网能够提供智能化的服务。受到网络带宽有限的影响,以云计算为中心的数据处理方法不能满足物联网实时性提供服务的要求,并且将用户数据传输到云计算中心增加了用户隐私泄露的风险。

文中使用边缘计算进行数据处理,降低云计算中心的计算负载,减缓网络带宽的压力;针对用于对图像分类处理的卷积神经网络,利用添加提前退出分支使神经网络在边缘设备对数据进行处理;搭建了具有分支结构的卷积神经网络进行仿真实验,验证了所提方案的有效性。边缘计算并非旨在完全取代集中式的基于云计算的基础架构,而是对现有云计算平台的有效补充。随着 5G 通信等相关技术的应用,高效的边缘协同数据处理值得进一步的研究和改进。

参考文献:

- [1] 沈苏彬,杨 震. 物联网体系结构及其标准化[J]. 南京邮电大学学报:自然科学版,2015,35(1):1-18.
- [2] TSAI C W, LAI C F, CHIANG M C, et al. Data mining for internet of things: a survey[J]. IEEE Communications Surveys and Tutorials, 2014, 16(1): 77-97.
- [3] 邓 芳. 大型物联网电子设备的海量数据高效挖掘方法研究[J]. 现代电子技术, 2016, 39(4): 159-162.
- [4] 孟晓丽. 物联网平台下基于云计算的智能家居系统设计[J]. 科技通报, 2016, 32(6): 67-71.
- [5] SHAFIQUE M, THEOCHARIDES T, BOUGANIS C S, et al. An overview of next-generation architectures for machine learning: roadmap, opportunities and challenges in the IoT era [C]//2018 design, automation & test in Europe conference & exhibition (DATE). Dresden, Germany: IEEE, 2018: 827-832.
- [6] 李江昀,赵义凯,薛卓尔,等. 深度神经网络模型压缩综述[J]. 工程科学学报, 2019, 41(10): 1229-1239.
- [7] 施巍松,孙 辉,曹 杰,等. 边缘计算:万物互联时代新型

(下转第 77 页)