文章编号:2096-1472(2024)01-0027-05
DOI:10.19644/j.cnki.issn2096-1472.2024.001.007

面向直播的边缘计算任务卸载方案研究

谢 松,王 薇

(长春大学网络安全学院,吉林 长春 130022)
✉ 1248904025@qq.com; wangwei@ccu.edu.cn

摘 要:文章研究分析了直播中多用户、多服务器场景下存在的直播用户的体验质量(Quality of Experience,QoE)不高的问题,为提升用户的 QoE,将能耗和时延作为决策目标,设计一种经改进的 NSGA-Ⅱ(非支配排序遗传算法),即 L-NSGA-Ⅱ,用线性排名的方式进行父代的选择加速算法收敛。实验表明,与 LUA、NSGA-Ⅱ和 Random 算法策略相比,所提方案的平均延迟降低约 9.1%,用户 QoE 提升约 4.39%。该方案应用于直播场景中,在减少延迟、提升吞吐量和降低能源开销方面表现出较好的效果。

关键词:直播;边缘计算;卸载;用户体验;用户分配;L-NSGA-Ⅱ
中图分类号:TP393 **文献标志码**:A

Research on Edge Computing Task Offloading Scheme for Live Broadcasting

XIE Song[1], WANG Wei[2]

(*College of Network Security, Changchun University, Changchun* 130022, *China*)
✉ 1248904025@qq.com; wangwei@ccu.edu.cn

Abstract: This paper studies and analyzes the problem of low Quality of Experience (QoE) for live streaming users in multi-user and multi-server scenarios. To improve the QoE of users, this paper proposes to design an improved NSGA-Ⅱ (Non-Dominant Sorting Genetic Algorithm), namely L-NSGA-Ⅱ, taking energy consumption and delay as decision-making objectives. The algorithm uses linear ranking to accelerate the convergence of the parent selection algorithm. The experiment shows that compared with the LUA, NSGA-Ⅱ, and Random algorithm strategies, the proposed scheme reduces the average latency by about 9.1% and improves user QoE by about 4.39%. This scheme has shown good performance in reducing latency, improving throughput, and reducing energy consumption when applied to live streaming scenarios.

Key words: live streaming; edge computing; offloading; user experience; user assignment; L-NSGA-Ⅱ

0 引言(Introduction)

随着互联网技术的飞速发展,网络直播成为人们娱乐、交流、传播信息的重要方式。直播应用通过将实时的音视频数据流传输到广大观众面前,实现了信息的即时传播。然而,由于直播应用本身对实时性的要求较高,传统的云计算模式无法满足其延迟敏感的要求,严重影响用户体验。基于此,本研究致力于解决直播流的边缘计算卸载问题,提出一种基于直播流的边缘计算卸载方案,以期实现降低延迟、减少带宽消耗及提升直播质量和用户体验的目标;其意义主要体现在以下几个方面:首先,将计算任务从云端卸载到边缘设备上进行处理,可以降低延迟和带宽消耗,从而提升直播质量和用户体验;其次,明确边缘计算在解决多用户、多服务场景下合理分配用户,提升直播用户体验方面的优势和潜力;最后,提出一种基于 L-NSGA-Ⅱ算法的边缘服务卸载方案,并在真实场景下进行模

收稿日期:2023-05-23
基金项目:2024 年度吉林省教育厅科学研究重点课题项目(JJKH20240749KJ);中国职业技术教育学会 2023 年度重点课题立项(ZJ2023A022)

拟实验，然后将该方案与 LUA、NSGA-Ⅱ和 Random 算法进行对比，以验证其在减少延迟、提升用户体验及降低能源开销方面的优势，从而为直播应用的发展和边缘计算领域的研究提供有价值的参考。

1 相关工作(Related work)

在边缘计算领域，卸载算法是一项重要的研究内容，其主要目标是将计算任务从中央服务器卸载到边缘设备上进行处理，以减轻中央服务器的负载和网络带宽的压力，提高系统的性能和效率。目前，已经有大量学者在此研究领域取得一些成果。LOGANATHAN 等[1]分析了云存储数据库在各种网络环境下的边缘设置性能，使用 Continum 框架在多个边缘节点上部署端点、节点和云存储服务 Cassandra，并测量写请求所占用的平均往返时间(RTT)，探讨不同节点集群(3、4 和 5 个节点)的延迟、数据包丢失和复制因子(RF)变化对 Cassandra 云存储服务性能的影响。WENG 等[2]利用认知移动边缘计算(MEC)实现对 MPEG-DASH 视频流缓存的管理，强调了对视频流媒体服务的需求日益增长的必然趋势，以及提高用户体验质量(QoE)的必要性。郭嵩[3]讨论了边缘计算和内容传递网路(CDN)之间的合作技术，突出了两个网络的异同点，提出三种合作模式：情景合作、资源合作和能力合作。总的来说，边缘计算在直播领域的应用，尤其是如何应用于多用户、多边缘服务器场景下提升用户的观看体验，仍是一个需要深入研究的问题。

2 系统模型(System model)

在边缘计算环境中，为了提高直播流的传输效率和用户体验质量，需要设计有效的卸载策略。本小节针对直播流的特点，考虑了直播流的实时性、带宽和节点利用率等。首先，视频的连续播放和低延迟，通常对直播流具有较高的实时性要求[4]，因此，在设计卸载策略时，需要考虑选择具有较低延迟和高带宽的边缘节点处理直播流。其次，直播流通常需要配置较大的带宽传输视频数据，因此，设计卸载策略时需要选择具有高带宽的边缘节点进行卸载。最后，边缘节点计算能力和存储资源有限，因此在设计卸载策略时，需要考虑边缘节点的资源利用率，并根据节点的负载情况选择合适的节点进行卸载。基于以上考虑，提出如图 1 所示的卸载策略设计方案。假设用户 A 期望得到更高质量的直播视频，则可以将用户 A 分配到剩余资源多的边缘服务器 1；或者用户 D 期望直播可以有更低的延迟，则可以将用户 D 分配到总延迟低的边缘服务器 4。卸载策略设计方案要解决的问题是为每位用户提供适配需求的边缘服务器，使所有用户的 QoE 最优。

在图 1 中，根据多边缘服务器覆盖能力范围内的用户需求差异，可以为他们分配适合的视频码率，以便最大限度地提升用户的 QoE，在边缘计算环境中可用用户集合用 $U=\{1,2,\cdots,u,\cdots,|U|\}$ 表示，可用边缘服务器集合用 $M=\{1,2,\cdots,m,\cdots,|M|\}$ 表示，在边缘计算的直播场景中，用户侧边缘服务器到主播侧边缘服务器的设备传输速率为 V_u，表示如下：

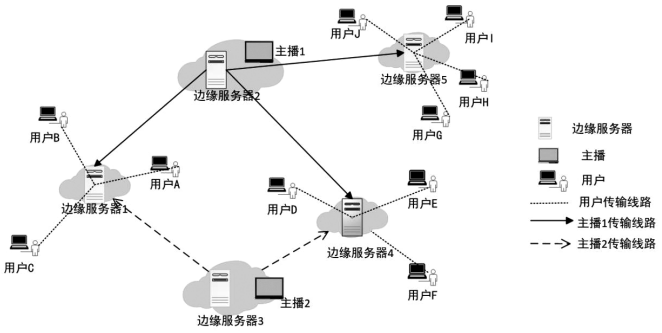

图 1 卸载策略设计方案

Fig. 1 Design scheme of unloading strategy

$$V_u = W_k \log_2\left(1+\frac{p_u h_{u,k}}{w_0+\sum_{s\in\frac{U}{u}} p_u h_{u,k}}\right) \tag{1}$$

其中：p_u 是链路的传输功率；W_k 是信道的带宽；w_0 是信道白噪音；$h_{u,k}$ 是信道增益，它是一个独立的随机变量，它与两个服务器之间的距离相关。

边缘服务器与用户之间的传输时延 $T_{\text{cloud}}^{\text{off}}$ 表示如下：

$$T_{\text{cloud}}^{\text{off}} = \frac{\lambda_m}{r_{mn}} \tag{2}$$

其中：λ_m 表示数据的大小，r_{mn} 表示节点间的传输速率。

因为原视频的码率和用户码率之间存在差别，所以存在一个转码的延迟，用户 u 连接边缘服务器 m 的直播视频延迟 $L_{u,m}$ 表示如下：

$$L_{u,m} = T_{m_{u_y},m} + T_{u_v,m_{b_y}} + T_{\text{cloud}}^{\text{off}} + (M-M_v)\cdot G \tag{3}$$

其中：$T_{m_{u_y},m}$ 表示边缘服务器之间的延迟，$T_{u_v,m_{b_y}}$ 表示直播人员到边缘服务器的延迟，$T_{\text{cloud}}^{\text{off}}$ 表示用户到边缘服务器的延迟，G 表示转码产生的延迟。

直播视频的质量是影响直播观看体验的主要因素之一，直播视频的质量与视频直播时的码率相关，用户 u 的直播视频质量函数 Q_u 表示如下：

$$Q_u = \ln\frac{b_1\min(R_u,R_u^*)}{R_u^*}, b_1=100 \tag{4}$$

其中：R_u^* 是用户 u 期望的视频直播时的码率，用户可以根据自己的网络情况自主选择需要的分辨率；R_u 是用户观看直播时的实际码率，当观看直播时的实际码率不超过用户期望的码率时，直播视频的质量达到最优。

抖动是视频流中的一个相当大的问题，视频直播时码率的突然升高或降低，会降低服务质量[5-9]。抖动要求因应用而异，数值范围在 10～50 ms，视频质量的抖动 E_u 表示如下：

$$E_u = k\cdot\max\{0,[R_u^- - R_u]\} \tag{5}$$

其中：R_u^- 是上一时间段的平均码率，k 表示实际码率每降低一个单位所引起的用户观看体验的损失值。边缘计算服务器的切换会造成直播的卡顿，影响用户观看体验的质量，边缘服务器切换对用户体验造成的影响 L_u^* 表示如下：

$$L_u^* = k\cdot[x_{u,m} - x_{u,m}^-]^+ \tag{6}$$

其中：$x_{u,m}^-$ 表示边缘服务器 m 在前一段时间的连接状态，如果用户 u 在前一个时间段连接到边缘服务器，则 $x_{u,m}^-=1$，否则

$x^{-}_{u,m}=0$;k 是切换边缘服务器对用户观看直播体验造成的损失值。

在直播体验中,延迟越低,用户观看直播体验越好,直播延迟映射到直播体验质量的函数 $Q(L_{u,m})$ 表示如下:

$$Q(L_{u,m})=\left(\frac{\beta}{L_{u,m}}\right)^{h} \tag{7}$$

其中:β 是延迟的阈值;$L_{u,m}$ 是用户观看直播的延迟;h 是一个权值,根据具体情况设定。

当用户 u 选择连接边缘服务器 m 时,用户的 $QoE_{u,m}$ 表示如下:

$$QoE_{u,m}=aQ(L_{u,m})+bQ_u-cE_u-dL^{*}_u \tag{8}$$

用户的观看体验质量与视频直播体验质量、直播质量、直播抖动、切换边缘服务器相关,其中 a、b、c、d 为权值,根据实际情况设置。直播时,用户选择边缘服务器问题是一个NP-hard问题,即用户选择哪台边缘服务器,使连接到边缘服务器 m 的用户的整体 QoE 最大化。问题目标函数表示如下:

$$\mathrm{Max}\sum_{x_v,R_v}QoE_u\cdot\gamma_u \tag{9}$$

其中,γ_u 表示用户与边缘服务器的连接关系:

$$\gamma_u=\begin{cases}1,\text{用户 } u \text{ 与边缘服务器 } m \text{ 连接}\\0,\text{用户 } u \text{ 与其他边缘服务器连接}\end{cases} \tag{10}$$

因为边缘服务器的资源是有限的,因此一个边缘服务器的用户码率不能超过服务器的总吞吐量,假设边缘服务器 m 的计算资源为 P_m、带宽为 F_m,则问题的约束条件如下:

$$\sum_{u\in U}G\cdot(R-R_u)\cdot\gamma_u\leqslant F_m \tag{11}$$

其中,G 为转码的计算开销。

$$\sum_{u\in U}R_u\cdot\gamma_u\leqslant P_m \tag{12}$$

为求解问题,本文提出使用改进型算法 L-NSGA-Ⅱ。NSGA-Ⅱ通过遗传算法进化以及非支配排序思想实现多目标优化,而 L-NSGA-Ⅱ则是在此基础上进行了改进,即加入线性排名的父代选择,从而加快了收敛速度。

3　改进型 NSGA-Ⅱ算法卸载决策方案(Unloading decision scheme of improved NSGA-Ⅱ algorithm)

3.1　NSGA-Ⅱ基本原理

NSGA-Ⅱ,又称为非支配排序遗传算法Ⅱ,可以用来解决多目标优化问题,它根据非支配的 $Rank$ 值和拥挤度进行排序以保留选择的个体。为指引搜索往帕累托最优解集的方向前进,通过个体的非劣解对种群进行分层。例如,把找到的种群中的非支配解集记为第一非支配层 $f1$,赋予所有的个体非支配序 $rank=1$,然后从种群中去除,继续寻找剩余的非支配解集并记为第二非支配层 $f2$,个体的非支配序为 $rank=2$,以此类推,直到将整个群体都分层,在同一层的种群个体都有相同的 $rank$,在分层后相同 $rank$ 的个体能够进行选择性排序,利用个体拥挤距离优先选择拥挤距离大的个体,以维持种群的多样性,然后通过精英选择策略保留父代中的优秀个体进入子代,从而防止帕累托最优解的丢失[10-14]。

3.2　改进型 NSGA-Ⅱ算法

NSGA-Ⅱ算法采用锦标赛选择方式进行父代选择,它随机从种群中选择指定数量的个体并使之通过竞争得到父代个体,这种方式的缺点是父代的选择近似随机,会影响整个种群的进化速度,无法保证搜索能力。为了保证算法的收敛速度,需要在算法前期扩大搜索范围,避免陷入局部最优解,后期再提高父代选择能力,保证算法的收敛。本文不使用锦标赛方式进行父代的选择,而是采用线性排名的方式,将改进后的 NSGA-Ⅱ算法称为 L-NSGA-Ⅱ。L-NSGA-Ⅱ按照适应值的大小把个体从小到大排序,把拥挤距离升序、等级降序。适应值越大,表示排名越低,那么选择概率就越高。假设 M 为种群规模,$x_1\sim x_n$ 个体排名逐渐降低,则个体 x_i 的选择概率表示如下:

$$P_i=\frac{1}{M}\left[\mathrm{a}+(\mathrm{b}-\mathrm{a})^{\frac{i-1}{M-1}}\right] \tag{13}$$

其中:$i=1,2,3,\cdots,M$,a 和 b 为常数,并且 $0\leqslant a\leqslant 1$。根据选择概率,若 $b=2$、$a=0$,种群选择压力最大;若 $a=b=1$,父代选择变为随机选择,种群选择压力最小。为了逐渐增大父群的选择压力,可以逐步改变 a 和 b 的值。其父代选择的伪代码如下所示:

算法:基于线性排序的亲本选择
输入:pop,ud,up,parent_size;
输出:parent_pop;
[sizepop,td]←Size(pop)
pop←SortRows(pop,[−(td−1),td])
parent_pop←{}
π←Zeros(1,sizepop)
for i←1to sizepop **do**
$\pi[1,i]\leftarrow\frac{1}{\mathrm{sizepop}}\left(\mathrm{ud}+(\mathrm{up}-\mathrm{ud})\frac{i-1}{\mathrm{sizepop}-1}\right)$
π←CumulativeSum(π)
r←Sort(Rand(1,parent_size))
fitin←1
newin←1
while newin≤parent_size **do**
if r(newin)≤π(fitin) **then**
parent_pop←Concatenate(parent_pop,pop(fitin,:))
newin←newin+1
else
fitin←fitin+1
return parent_pop

4　实验与结果分析(Experiment and result analysis)

4.1　实验设置

本文实验使用基站数据集[15]来构建用户和边缘服务器之间的关系,该数据集包含墨尔本商务区的 125 个基站和基站周围的 816 个用户的位置信息,边缘服务器之间的连接随机生成。实验环境为 Ubuntu16.04 Linux 系统、处理器 AMD Ryzen 7 with Radeon Graphics(3.20 GHz)、磁盘容量 2 TB、内存 32 GB 的 PC 机,实验中的总网络吞吐量为 30 Mbps,总计算资源为 18 vCPUs,转码延迟为 40 ms/Mbps,转码需要的边缘服务器

计算资源开销为 0.7 vCPU/Mbps。用户 QoE 的权重 $a=7$、$b=1$、$c=1$、$d=1$。用户到边缘服务器的传输延迟为 5～20 ms，边缘服务器到边缘服务器的传输延迟为 50～200 ms，延迟阈值为 200 ms，实验的主要参数如表 1 所示。

表 1 实验的主要参数表

Tab.1 Main parameters table of the experiment

参数	数值
总网络吞吐量/Mbps	30
总计算资源/vCPUs	18
转码需要的计算资源开销 vCPU/Mbps	0.7
用户到边缘服务器的传输延迟/ms	[5,20]
边缘服务器之间的传输延迟/ms	[50,200]
边缘服务器的覆盖范围/m	[50,100]

为评估 L-NSGA-Ⅱ算法在直播分配上的性能，选取边缘服务器数量、用户数量、用户 QoE 三个指标作为实验参数。L-NSGA-Ⅱ算法的主要参数如表 2 所示。

表 2 L-NSGA-Ⅱ算法的主要参数表

Tab.2 Main parameters table of L-NSGA-Ⅱ algorithm

参数	数值
种群数量/个	[50,200]
最大迭代数/次	200

4.2 性能分析

本文首先分析了不同边缘服务器数量对能耗的影响，由图 2 可以看出，随着边缘服务器数量的增加，直播用户分配算法（LUA）、随机分配算法（Random）、非支配排序遗传算法（NSGA-Ⅱ）和改进型非支配排序遗传算法（L-NSGA-Ⅱ）的能耗都在降低，服务器数量增加提供了充足的计算资源，使得每个边缘服务器所分担处理的任务随之减少，所以能耗也就降低了[16]。从图 2 数据可以看出，由于随机分配算法采用随机的任务分配，没有进行任何的优化，因此它的性能是最差的，而 L-NSGA-Ⅱ算法不仅提高了收敛速度，而且加快了节点的选择和降低了能耗。

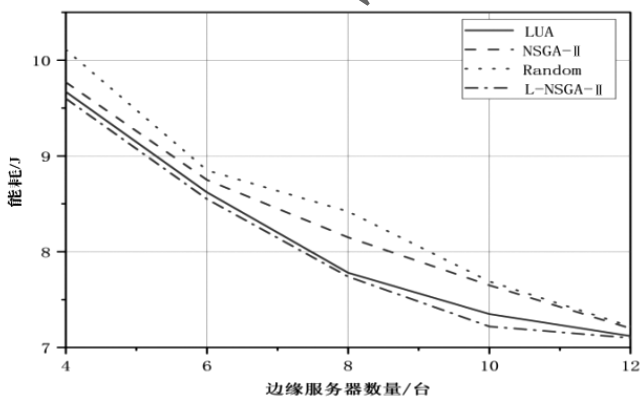

图 2 同边缘服务器数量对能耗的影响

Fig. 2 The impact of different numbers of edge servers on energy consumption

如图 3 所示，随着边缘服务器的增加，可用的计算资源随之增加，使得平均延迟有所降低，整体各个算法的延迟呈现下降趋势，而当边缘服务器数量达到 12 个及以上时，系统的计算资源达到饱和，平均延迟趋于稳定。L-NSGA-Ⅱ相较于随机分配算法，平均延迟降低了约 9.1%，相较于优化前的 NSGA-Ⅱ算法，平均延迟降低了 1.32%。

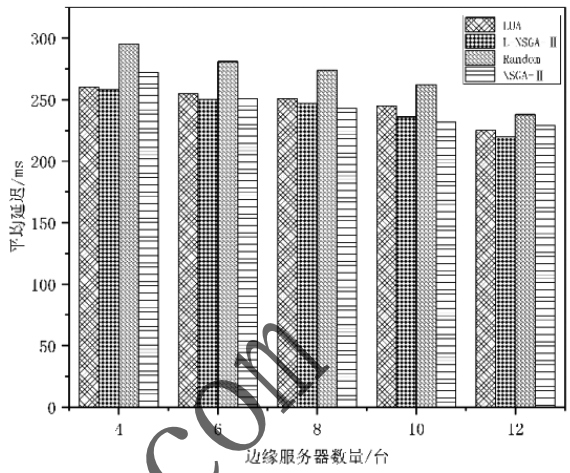

图 3 边缘服务器数量对平均延迟的影响

Fig. 3 The impact of different numbers of edge servers on average latency

如图 4 所示，随着用户数量的增加，因为总资源是有限的，所以用户分配到的资源减少，导致用户的体验质量下降。在平均用户体验质量中，L-NSGA-Ⅱ表现最优，比 NSGA-Ⅱ高出 2.52%，比 LUA 高出约 4.39%。据此可知，L-NSGA-Ⅱ算法能合理分配用户需要的带宽和资源，从而避免不必要的资源浪费。

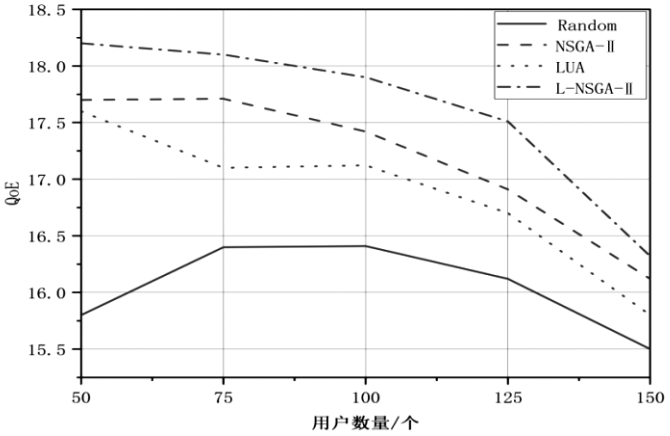

图 4 用户数量对 QoE 的影响

Fig. 4 The impact of user numbers on QoE

5 结论(Conclusion)

本文提出一种直播流的边缘计算卸载方案，旨在解决传统云计算无法满足实时直播应用需求的问题。研究通过将计算任务从云端卸载并转移到边缘设备上来进行处理，有效降低了延迟和带宽消耗，提升了直播质量和用户体验质量。本文提出

的改进算法 L-NSGA-Ⅱ是通过线性排名和加速父代选择的方式提高了算法的收敛速度。通过实验,验证了 L-NSGA-Ⅱ算法与 LUA、NSGA-Ⅱ、Random 算法相比,在降低能耗、减少延迟和提升用户 QoE 等方面均具有显著优势,该方案在减少延迟、提升吞吐量和降低能源开销方面具有现实意义。本文的研究还存在一些不足:首先,实验仅在真实场景下进行了模拟,缺乏大规模实际应用的验证;其次,基于边缘计算的卸载方案如何应对网络变化的复杂环境,尚需要进一步研究。此外,本研究未对边缘设备的资源分配和管理进行深入探究,这也是未来研究的方向之一。

参考文献(References)

[1] LOGANATHAN P, RAUTHAN D, TRIVEDI A, et al. Performance measurement of distributed storage on edge devices[C]//IEEE. Proceedings of the 2023 15th International Conference on Communication Systems & Networks (COMSNETS). Piscataway: IEEE, 2023: 841-846.

[2] WENG H Y, HWANG R H, LAI C F. Live MPEG-DASH video streaming cache management with cognitive mobile edge computing[J]. Journal of ambient intelligence and humanized computing, 2020, 20(12): 1-18.

[3] 郭嵩. 边缘计算与 CDN 协同技术[J]. 电信科学, 2019, 35(增刊 2): 65-70.

[4] 刘伟,张骁宇,杜薇,等. 边缘计算中面向互动直播的用户分配策略[J]. 计算机研究与发展, 2023, 60(8): 1858-1874.

[5] 李莉. 基于遗传算法的多目标寻优策略的应用研究[D]. 无锡:江南大学, 2008.

[6] 张海波,栾秋季,朱江,等. 基于移动边缘计算的 V2X 任务卸载方案[J]. 电子与信息学报, 2018, 40(11): 2736-2743.

[7] XU J, HU Z, ZOU J. Computing offloading and resource allocation algorithm based on game theory for IoT devices in mobile edge computing[J]. International journal of innovative computing, information and control, 2020, 16(6): 1895-1914.

[8] 袁培燕,蔡云云. 移动边缘计算中一种贪心策略的内容卸载方案[J]. 计算机应用, 2019, 39(9): 2664-2668.

[9] 刘星星. 基于移动边缘计算的计算卸载与能效优化研究[D]. 兰州:兰州理工大学, 2020.

[10] 李季. 基于深度强化学习的移动边缘计算中的计算卸载与资源分配算法研究与实现[D]. 北京:北京邮电大学, 2019.

[11] 章雨鹏. 移动边缘计算场景下的移动性支持和资源分配研究[D]. 成都:电子科技大学, 2019.

[12] DEB K, PRATAP A, AGARWAL S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-Ⅱ[J]. IEEE transactions on evolutionary computation, 2002, 6(2): 182-197.

[13] 刘建华,罗荣鑫,刘佳嘉,等. 基于 NSGA2 的车联网边缘计算任务卸载方案[J/OL]. 西安理工大学学报, (2023-01-31)[2023-03-17]. https://kns-cnki-net.webvpn.ccu.edu.cn/kcms/detail/61.1294.n.20230130.1410.002.html.

[14] 崔玉亚. 面向移动边缘计算的任务调度的关键技术研究[D]. 天津:天津理工大学, 2021.

[15] HE Q, CUI G M, ZHANG X Y, et al. A game-theoretical approach for user allocation in edge computing environment[J]. IEEE transactions on parallel and distributed systems, 2020, 31(3): 515-529.

[16] 天津理工大学. 一种面向移动边缘计算应用的多用户细粒度任务卸载调度方法:CN202011258509.8[P]. 2022-04-15.

作者简介:

谢　松(1997-)男,硕士生。研究领域:边缘计算。

王　薇(1975-)女,硕士,教授。研究领域:边缘计算,人工智能与智能计算。本文通信作者。

(上接第 26 页)

[9] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//IEEE. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[10] KIM B K, DONG S Y, ROH J, et al. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach[C]//IEEE. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2016: 1499-1508.

[11] TURAN C, LAM K M, HE X J. Soft locality preserving map (SLPM) for facial expression recognition[DB/OL]. (2018-11-11)[2023-04-18]. https://arxiv.org/abs/1801.03754.

[12] OJALA T, PIETIKÄINEN M, HARWOOD D. A comparative study of texture measures with classification based on featured distributions[J]. Pattern recognition, 1996, 29(1): 51-59.

[13] RODRIGUEZ P, CUCURULL G, GONZALEZ J, et al. Deep pain: exploiting long short-term memory networks for facial expression classification[J]. IEEE transactions on cybernetics, 2022, 52(5): 3314-3324.

[14] WANG W X, SUN Q, CHEN T, et al. A fine-grained facial expression database for end-to-end multi-pose facial expression recognition[DB/OL]. (2019-07-25)[2023-04-18]. https://arxiv.org/abs/1907.10838.

作者简介:

张中华(1997-),男,硕士生。研究领域:深度学习。

杨慧炯(1972-),男,硕士,教授。研究领域:图像处理,机器学习。