Using Imperfect Transmission in MEC Offloading to Improve Service Reliability of Time-Critical Computer Vision Applications

JIANHUI LIU (Member, IEEE), AND QI ZHANG (Member, IEEE)

DIGIT, Department of Engineering, Aarhus University, 8200 Aarhus, Denmark

Corresponding author: Qi Zhang (qz@eng.au.dk)

This work was supported by the Independent Research Fund Denmark (DFF) through the Agile Edge Intelligence for Delay Sensitive IoT (AgilE-IoT) Project under Grant 9131-00119B.

ABSTRACT The emerging time-critical Internet-of-Things (IoT) use cases, e.g., augmented reality (AR), virtual reality (VR), autonomous vehicle etc., on the one hand, involve computation intensive computer vision (CV) components in the services, on the other hand, the computation task should be completed within the stringent latency constraint, otherwise, the service reliability will deteriorate. The state-of-the-art work has shown that it is promising to tackle this challenge by offloading the computation tasks to the edge servers (ESs) using mobile edge computing (MEC). However, offloading tasks from local IoT devices to remote ESs could cause communication errors, thereby resulting in transmission failure even service timeout. The existing work mainly requires perfect transmission during task offloading at physical layer or transport layer. In fact, CV algorithms for, e.g., image classification and recognition, are able to tolerate certain level distortion of the input image to maintain the required inference accuracy. In this paper, we focus on the service reliability at application layer and study how feasible it is to improve the service reliability of the time-critical CV services in MEC system by allowing imperfect transmission. The service reliability is modeled by the transmission failure probability, service timeout probability and inference accuracy. The optimization goal is to maximize the service reliability, subject to the latency constraint. Due to the non-convexity, we solve this problem by the semi-definite relaxation based algorithm for a multi-user scenario. We evaluate the algorithm considering the practical scenarios, i.e., object detection with SSD and YOLOv2. The proposed algorithm achieves the performance closed to the exhaustive method but at a much lower complexity.

INDEX TERMS Edge intelligence, mobile edge computing, IoT, image distortion, latency and reliability, semi-definite relaxation.

I. INTRODUCTION

A variety of time-critical Internet-of-Things (IoT) use cases, e.g., augmented reality (AR), virtual reality (VR), autonomous vehicle and remote surgery etc., have attracted widespread interests in recent years. These uses cases are facing the challenges of the stringent latency and high reliability. For example, the cooperative collision avoidance service in autonomous driving requires a maximal latency of 10 ms and reliability of 99.99% [1]. However, due to limited computation capability of user equipments (UEs), it is challenging for UEs to complete the computation intensive tasks within the delay deadline merely by the local processing [2]. Mobile edge computing (MEC) is a promising paradigm to shorten service latency [3], [4], by offloading the computational tasks from UEs to powerful computation units in the proximity to UEs, i.e., edge servers (ESs). However, the communication errors during the offloading process could deteriorate the transmission reliability and latency, thereby decreasing overall service reliability.

Generally speaking, the service reliability is affected by several factors. For the use cases like AR, VR and autonomous vehicle etc., object detection and classification [5], [6] are one of the most popular computer vision (CV)
computation tasks. The overall service reliability depends on the inference accuracy of CV algorithms, service timeout probability, and transmission failure probability. The service latency and transmission reliability have significant impact on the user of experience (UoE) for the time-critical CV services in MEC system, in spite of high inference accuracy. To improve UoE of the CV services, the state-of-the-art work mainly requires perfect communication, i.e., zero bit error or packet error during task offloading [7]–[9]. When a transmission error occurs, retransmission scheme is used to guarantee reliable data delivery between UE and ES, which could exacerbate the communication delay. For example, when an AR headset transmits a captured image to an ES, transmission control protocol (TCP) is used to ensure all the packets are delivered to the ES. Due to the overall latency constraint of the time-critical IoT applications, increasing communication delay means reducing computation time budget. As a result, it could increase the service timeout probability.

In fact, CV algorithms for, e.g., image classification and recognition, are able to tolerate certain level distortion of the input image, while keeping the required inference accuracy [10]. For instance, the faces in Extended Yale B Face Database can be still recognized if 13.3% or fewer pixels of an image are corrupted [11]. Allowing a small image distortion, it can be used to relax the transmission reliability requirements. In other words, a few errors are acceptable for a task transmission, e.g., image sent by UE, which can help to reduce the service timeout probability while maintaining the service accuracy.

The main contribution in this paper is that we study how feasible it is to improve the overall reliability of the time-critical CV services in MEC system allowing imperfect transmission. The MEC system with multiple UEs and ESs is considered. We systematically model the latency and reliability of the time-critical CV services under image distortion, and formulate the optimization problem to maximize the overall service reliability, subject to the latency constraint. Due to non-convexity, the semi-definite relaxation (SDR) based algorithm is designed to achieve a sub-optimal solution for the UE-ES association, as well as communication and computation resource allocation. We evaluate the performance of the proposed algorithm using the practical object detection methods (SSD [5] and YOLOv2 [6]) in the multiple scenarios. The results show that the SDR based algorithm achieves the performance close to the exhaustive method at a much lower complexity.

The rest of the paper is organized as follows. Section II elaborates the related work. Section III describes the system model and problem formulation. Section IV explains how the SDR-based optimization algorithm solve the formulated problem. Sections V presents the numerical results. Finally, Section VI concludes the paper.

II. RELATED WORK

Many researchers have studied the reliability of computation offloading in MEC from the perspectives of different layers at the communication protocol stack. The work in [12], [13] studies the reliability in the application layer. Merluzzi et al. define an out-of-service probability as the probability that the overall latency to complete the task exceeds a threshold. They propose the algorithm to strike a balance between energy consumption and task delay for computation offloading, subject to the reliability constraint. Liu et al. [13] model the reliability of computation offloading as the probability that the queue lengths at both ES and UE do not exceed their thresholds. The work in [14]–[17] considers the reliability of wireless transmission at PHY layer. Azimi et al. [14] study the transmission failure probability using superposition coding on wireless fading channel, and optimize the energy consumption with the constraints of the latency and reliability. Liu and Zhang [15] jointly minimize the transmission failure probability and latency allowing multiple ESs cooperatively to complete a task. Zhou et al. [16] maximize the successful transmission probability within latency constraints by optimizing the transmitted data bits in each time slot. Wang et al. [17] consider the successful transmission probability under channel condition in the worse case and leverage the conditional value-at-risk approach to ensure the reliable transmission for computation offloading. Moreover, other work in [3], [18], [19] proposes to consider the computation reliability of ES, i.e., the probability that the hardware or software break down. Haber et al. [18] introduce multiple Unmanned Aerial Vehicle mounted Cloudlets as ESs and leverage redundancy to meet the reliability requirements for the computation offloading. The work in [3], [19] proposes to jointly improve both communication and computation reliability of offloading. Liu and Zhang [3] model a task as a directed acyclic graph and schedule the allocation of sub-tasks to minimize the transmission and computation failure probability. Liu and Zhang [19] propose an online algorithm to maximize the communication and computation reliability by dynamic selections of transmission rates and ESs.

For CV services, the inference accuracy of the CV algorithms could be regarded as the computation reliability. Offloading the CV tasks to MEC could allow advanced CV algorithms to be applied to the CV tasks with reduced CV service latency while guaranteeing the computation reliability. Various offloading strategies are proposed in literature, including input rescaling [7], DNN partitioning [9], [20], and early-exit points in DNN for fast inference [8], [22]. Liu et al. [7] propose to dynamically rescale the input images to adapt to the degradation of wireless channel, thereby, make a tradeoff between the task runtime and the inference accuracy. Instead of offloading the input images to ES directly, the work in [9], [20] partitions DNN computation between UE and ES at the granularity of NN layers. Kang et al. [20] study the optimal partitioning points of 8 different DNN models to minimize the overall latency and energy consumption. As the size of the intermediate layer data in DNN is large, to reduce the transmission delay on wireless channel, Esratifar et al. [9] utilize JPEG compressor to encode
the intermediate data and introduce a fine-tuning method to reduce the inference accuracy loss caused by the lossy compression. Instead of partitioning the DNN between UE and ES like [9], [20], Zhang et al. allocate the DNN task among multiple ESs and propose an algorithm to jointly optimize the task latency and resource utilization in the MEC system. Furthermore, the work in [8], [22] leverages early-exit points in DNN to reduce the runtime of CV services. Teerapittayanon et al. [22] design a new DNN architecture, i.e., BranchyNet, where several early-exit points are set in the DNN and the inference could be terminated at the early-exit point reaching high confidence. Li et al. [8] combine the early-exit points with DNN partitioning in MEC system. They partition the computation of a shallow NN obtained by early exiting the test DNN. To minimize latency, the early exit point and partitioning point are dynamically optimized, according to the channel data rate and the runtime of each DNN’s layer profiled at the offline stage.

Most of the previous work prefers to use reliable transmission scheme to offload CV services to MEC, which guarantees the inference accuracy but could result in retransmission delay and control overhead. On the contrary, Liu and Zhang [23] conduct a series of experiments to investigate how image distortion, caused by unreliable transmission, affects the inference accuracy and service reliability of CV tasks in MEC system in real-life environment. Compared to the work in [23], in this paper, we analyze the overall service reliability of CV services in the MEC system with multiple UEs and multiple ESs, and systematically model the communication reliability at PHY layer, service timeout probability, and the computation reliability, i.e., the inference accuracy of the CV services.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

As illustrated in Fig. 1, we consider an MEC system with multiple UEs and ESs. To facilitate the multiple ESs, a software-defined networking (SDN) architecture could be introduced to the MEC system [24], [25]. The SDN controller, referred to as an edge orchestrator, keeps the global status information updating in the system, e.g., channel state information (CSI) between UEs and ESs, available computing capacity of ESs, service requirements etc., and optimizes the resource allocation and offloading strategy for UEs. UEs are granted communication and computation resources through the control plane after service requests. The CV tasks are subsequently offloaded to ESs in the data plane. The physical communication channel of the control plane is independent from that of the data plane in the SDN architecture. In this paper, we mainly focus on the latency and reliability of computation offloading in the data plane.

The sets of UEs and ESs are denoted by $\mathcal{M}$ and $\mathcal{N}$, respectively, where $\mathcal{M} = \{1, 2, \ldots, M\}$ and $\mathcal{N} = \{1, 2, \ldots, N\}$. The computing capacity of ES $n$ is $f_{\text{max}}^n$ (in cycles/s). The computation task of UE $m$ is denoted by a tuple $(b_m, \alpha_m)$, $b_m$ represents the task size (in bits), while $\alpha_m$ is the computation intensity of the task, i.e., the number of the required CPU cycles to compute one-bit task (in cycles/bit). We use a vector $x_m = [x_{m,1}, x_{m,2}, \ldots, x_{m,N}]^T$. If UE $m$ is associated with ES $n$, the element $x_{m,n}$ in $x_m$ is 1; otherwise, $x_{m,n} = 0$. Note that a UE is associated with one of ESs, i.e., $\sum_{n=1}^{N} x_{m,n} = 1$ for $\forall m \in \mathcal{M}$. The wireless channels of different ESs are orthogonal. UEs offload tasks to their associated ESs with orthogonal frequency-division multiplexing access (OFDMA) scheme. $k_{\text{max}}^n$ denotes the total number of resource blocks (RBs) on the channel of ES $n$. $k_{\text{max}}^n$ RBs are shared by the associated UEs.

| Table 1. Notations used throughout the paper. |
|----------------------------------|----------------------------------|
| Notation | Definition |
| $\mathcal{N}$ | Set of the ESs |
| $\mathcal{M}$ | Set of the UEs |
| $M$ | Total number of the UEs |
| $\beta_m$ | Task size of UE $m$ (in bits) |
| $\alpha_m$ | Task computation intensity of UE $m$ (in cycles/bit) |
| $\gamma_m$ | Association indicator between UE $m$ and ES $n$ |
| $x_{m,n}$ | Available computing capacity on ES $n$ |
| $r_{\text{max}}^n$ | Total number of RBs on channel of ES $n$ |
| $k_{\text{max}}^n$ | The number of RBs assigned to UE $m$ by ES $n$ |
| $T_{\text{run}}^m$ | Transmission rate of each RB from UE $m$ to ES $n$ |
| $k_{\text{mem}}^n$ | The number of RBs assigned to UE $m$ by ES $n$ |
| $f_{\text{run}}^m$ | Computing capacity assigned to UE $m$ by ES $n$ |
| $\phi_{\text{run}}^m$ | Task completion latency of UE $m$ |
| $\delta_m$ | Task runtime of UE $m$ |
| $\delta_m$ | Total latency to complete task of UE $m$ |
| $\delta_{\text{run}}^m$ | Task execution failure probability of UE $m$ |
| $\delta_{\text{run}}^m$ | Task timeout probability of UE $m$ |
| $\gamma_{\text{run}}^m$ | Task offloading failure probability of UE $m$ |
| $\delta_{\text{run}}^m$ | Inference accuracy for the task of UE $m$ |
| $R_{\text{run}}^m$ | Quality of inference of UE $m$ |
| $N_{\text{run}}^m$ | The number of transport blocks to offload a task of UE $m$ |
| $\delta_{\text{run}}^m$ | Latency threshold to complete the task of UE $m$ |
| $\eta$ | Targeted BLER of MEC |
| $\theta$ | Maximal allowed percentage of the distorted pixels in a task |
Due to the fact that data size of a task output is normally much smaller than the input one, the latency caused by downlink transmission is negligible [7], [13], [15].

**B. LATENCY AND RELIABILITY MODEL**

The latency to complete a task is composed of the communication and computation latency. Due to the fluctuation of wireless channel, it is required to utilize the proper modulation and coding schemes (MCSs) to ensure a constant transport block error rate (BLER). The proper MCS for the task transmission could be determined based on the estimated CSI in practice. How to optimally select MCS based on CSI for different communication channels is a classical link adaptation or rate adaptation problem in wireless and mobile communication, which has been well studied in the literature [26]–[28]. Note that the selected MCS has the fixed transmission rate on each RB. Therefore, in this article, we mainly focus on the RB and computation allocation among UEs. For a given MCS, \( y_{m,n} \) is used to denote the transmission rate of the RB from UE \( m \) to ES \( n \) (in \( \text{bits/s} \)). \( k_{m,n} \) denotes the number of RBs assigned to UE \( m \) by ES \( n \), which meets \( \sum_{m \in M} k_{m,n} \leq k_{\text{max}} \). Thus, the latency to transmit the task from UE \( m \) to the associated ES is

\[
T_{m}^{\text{tran}} = \sum_{n \in N} x_{m,n} b_{m,n}/k_{m,n}. \tag{1}
\]

We assume that ES \( n \) assigns computing capacity of \( f_{m,n} \) to UE \( m \) at the associated ES

\[
T_{m}^{\text{comp}} = \sum_{n \in N} x_{m,n} a_{m,n} b_{m,n}/f_{m,n}. \tag{2}
\]

Therefore, the latency to complete the task of UE \( m \) is

\[
T_{m} = T_{m}^{\text{tran}} + T_{m}^{\text{comp}}. \tag{3}
\]

The overall service reliability is modeled as a quality of inference (QoI) which includes transmission failure probability, timeout probability and inference accuracy. For a given MCS, the number of transport blocks (TBs) to offload a task of UE \( m \), \( N_{m}^{\text{TB}} \), can be calculated by

\[
N_{m}^{\text{TB}} = \left\lceil \sum_{n \in N} x_{m,n} b_{m,n}/k_{m,n} \right\rceil, \tag{4}
\]

where \( \tau \) is the duration of one TB (in seconds), and \( \lceil \cdot \rceil \) is a ceiling function. Given the targeted BLER is \( \eta \), the conventional way to define the transmission failure probability of UE \( m \) is \( 1 - (1 - \eta)^{N_{m}^{\text{TB}}} \) [14]–[16]. However, when small distortion of an image is allowed, it means it can be regarded as a successful transmission if the number of lost TBs is not more than a threshold. Thus, considering TBs are independent with each other, the transmission failure probability can be modeled as

\[
F_{m}^{\text{tran}} = 1 - \sum_{i=0}^{[\theta N_{m}^{\text{TB}}]} \left( N_{m}^{\text{TB}} \right)^i \theta^i (1 - \eta)^{N_{m}^{\text{TB}} - i}, \tag{5}
\]

where \( [\cdot] \) is a floor function, \( \theta \) denotes the maximal allowed percentage of the distorted pixels (0 ≤ \( \theta \) ≤ 1).

Moreover, due to the latency constraint, the timeout probability of UE \( m \) is

\[
F_{m}^{\text{timeout}} = P(T_{m} > \delta_{m}), \tag{6}
\]

where \( \delta_{m} \) is the latency threshold for UE \( m \). In this way, the offloading failure probability (OFP) of UE \( m \) is defined as

\[
F_{m} = 1 - (1 - F_{m}^{\text{tran}})(1 - F_{m}^{\text{timeout}}). \tag{7}
\]

Note that, for the case that completes one task only, the OFP is equivalent to

\[
F_{m} = \begin{cases} F_{m}^{\text{tran}}, & T_{m} \leq \delta_{m}, \\ 1, & T_{m} > \delta_{m}. \end{cases} \tag{7}
\]

We assume the inference accuracy of UE \( m \), \( A_{m}^{\text{Infer}} \), can be guaranteed, if the distortion in an image is not more than \( \theta \). Hence, the QoI of a CV service is defined as

\[
K_{m}^{\text{QoI}} = A_{m}^{\text{Infer}} (1 - F_{m}). \tag{8}
\]

**C. PROBLEM FORMULATION**

We formulate an optimization problem to maximize the average QoI of multiple UEs allowing partial distortion in a task, subject to the latency constraint, which is

\[
\text{P1:} \quad \max_{x_{m,k,n},a_{m,n},f_{m,n}} \frac{1}{M} \sum_{m \in M} K_{m}^{\text{QoI}}
\]

\[
s.t. \quad T_{m} \leq \delta_{m}, \quad \forall m \in M, \tag{9a}
\]

\[
\sum_{n \in N} x_{m,n} = 1, \quad \forall m \in M, \tag{9b}
\]

\[
\sum_{m \in M} k_{m,n} \leq k_{\text{max}}, \quad \forall n \in N, \tag{9c}
\]

\[
\sum_{m \in M} f_{m,n} \leq f_{\text{max}}, \quad \forall n \in N, \tag{9d}
\]

\[
\delta_{m,n} \in [0, 1], \quad k_{m,n} \geq 0, f_{m,n} \geq 0, \quad \forall m \in M, \quad n \in N. \tag{9e}
\]

Constraint (9a) denotes the latency constraints for each UE. (9b) presents one UE is associated with one of ESs. (9c) and (9d) show the limit of communication and computation resource. P1 is non-convex since the expressions of \( N_{m}^{\text{TB}} \) and \( F_{m}^{\text{tran}} \) in the objective function.

**IV. SDR-BASED OPTIMIZATION ALGORITHM**

In this section, we propose the algorithm based on SDR to solve P1. The key idea is to relax the problem into a quadratic constrained quadratic programming (QCQP) problem. Then the transformed problem is homogenized and solved by dropping the rank-one constraint.

The required inference accuracy \( A_{m}^{\text{Infer}} \) in (8) is a constant. As a result, according to (7), maximizing average QoI in P1 is equivalent to minimize average \( F_{m}^{\text{tran}} \). To simplify the problem further, we derive the upper bound of \( F_{m}^{\text{tran}} \). It can be regarded as a random variable for the number of failed TBs to transmit a task, where the mean is \( \eta N_{m}^{\text{TB}} \) and the variance is \( \eta (1 - \eta) N_{m}^{\text{TB}} \). According to the Chebyshev’s inequality,
we have \( F_{m}^{\text{trans}} \leq \frac{\eta(1-\eta)}{[N_{m}^{TB}(\theta - \eta)^2]} \). Hence, P1 is relaxed into

\[
P2: \min_{x_{m},k_{m,n},f_{m,n},t_{m}} \frac{1}{M} \sum_{m \in M} [\eta(1-\eta)]/[N_{m}^{TB}(\theta - \eta)^2]
\]

s.t. (9a) – (9e).

P2 is a mixed-integer nonlinear programming problem. It can be solved by the exhaustive method, which searches all the possible association of UE-ES pair and obtains the optimal \( k_{m,n} \) and \( f_{m,n} \) via the interior-point method. The computation complexity of the exhaustive method is \( O \left( N^M (MN)^{5/2} \right) \), which increases exponentially with the increasing number of UEs growing. It is thus impractical to solve the problem with the exhaustive method.

To solve P2 efficiently, we relax the ceiling function in \( N_{m}^{TB} \). Due to the constraint (9b), we have \( k_{m,n} = 0 \) and \( f_{m,n} = 0 \), if the UE \( m \) is not associated with the ES \( n \). Therefore, according to the expression of \( T_{m}^{\text{trans}} \) and \( T_{m}^{\text{comp}} \), P3 can be obtained as

\[
P3: \min_{x_{m},T_{m}^{\text{trans}},T_{m}^{\text{comp}}} \frac{\Theta}{M} \sum_{m \in M} \sum_{n \in N} x_{m,n,t} T_{m}^{\text{trans}}
\]

s.t. \( T_{m}^{\text{trans}} + T_{m}^{\text{comp}} \leq \hat{d}_{m} \), \( \forall m \in M \),

\[
\sum_{n \in N} x_{m,n} = 1, \quad \forall m \in M,
\]

\[
\sum_{m \in M} x_{m,n} b_{m} \leq k_{m}^{\text{max}}, \quad \forall n \in N,
\]

\[
\sum_{m \in M} x_{m,n} c_{m} b_{m} \leq f_{m}^{\text{max}}, \quad \forall n \in N,
\]

\[
x_{m,n} \in [0, 1], \quad T_{m}^{\text{trans}} > 0, \quad T_{m}^{\text{comp}} > 0,
\]

\[
\forall m \in M, \quad \forall n \in N,
\]

where \( \Theta = \frac{\eta(1-\eta)}{(\theta - \eta)^2} \). The constraints (11a)-(11e) are equivalent to the constraints (9a)-(9e), respectively. Compared with P2, P3 reduces the number of the variables from \( 3MN \) to \( M(N+2) \). Subsequently, to transform P3 into a QCQP problem, we relax the binary variable \( x_{m,n} \) to real numbers \( x_{m,n} \in [0, 1] \), and substitute the \( T_{m}^{\text{trans}} \) and \( T_{m}^{\text{comp}} \) with \( 1/f_{m,n} \) and \( 1/\epsilon_{m,n} \), respectively. As a result, we have

\[
P4: \min_{x_{m},\epsilon_{m,n},t_{m}} \frac{\Theta}{M} \sum_{m \in M} \sum_{n \in N} x_{m,n} \epsilon_{m,n} t_{m}
\]

s.t. \( t_{m} + \epsilon_{m,n} - t_{m} \epsilon_{m,n} \leq 0, \quad \forall m \in M \),

\[
\sum_{n \in N} x_{m,n} = 1, \quad \forall m \in M,
\]

\[
\sum_{m \in M} b_{m} x_{m,n} t_{m} \leq k_{m}^{\text{max}}, \quad \forall n \in N,
\]

\[
\sum_{m \in M} a_{m} b_{m} x_{m,n} \epsilon_{m,n} \leq f_{m}^{\text{max}}, \quad \forall n \in N,
\]

\[
x_{m,n} \in [0, 1], \quad t_{m} > 0, \quad \epsilon_{m,n} > 0,
\]

\[
\forall m \in M, \quad \forall n \in N.
\]

To homogenize P4, we define a vector of the variables in P4 as \( y = [x_{1}^{T}, x_{M}^{T}, t_{1}, \ldots, t_{M}, e_{1}, \ldots, e_{M}]^{T} \), of which the length is \( M(N + 2) \). Then a new variable is introduced by \( Z = \frac{y}{\eta} \) \([y^{T}, 1]\), which is rank-one symmetric positive semidefinite matrix. Assuming \( Q = (N + 2) \), the \((Q+1, Q+1)\)-th element in \( Z \) is equal to 1, i.e. \( Z_{Q+1,Q+1} = 1 \). Therefore, the equivalent formulation of P4 is presented as

\[
P5: \min_{Z} \frac{\Theta}{M} \text{Trace}(\tilde{H}Z)
\]

s.t. \( \text{Trace}(\tilde{A}_{m}Z) \leq 0, \quad \forall m \in M \),

\[
\text{Trace}(\hat{B}_{m}Z) = 1, \quad \forall m \in M,
\]

\[
\text{Trace}(\tilde{C}_{n}Z) \leq \frac{k_{m}^{\text{max}}}{\eta^{2}}, \quad \forall n \in N,
\]

\[
\text{Trace}(\hat{D}_{n}Z) \leq \frac{f_{m}^{\text{max}}}{\eta^{2}}, \quad \forall n \in N,
\]

\[
Z \geq 0, \quad \text{Rank}(Z) = 1, \quad Z_{Q+1,Q+1} = 1,
\]

where

\[
\tilde{H} = \begin{bmatrix} 0_{M \times MN} & H & 0_{MN \times (M+1)} \\ H^{T} & 0_{M \times M} & 0_{M \times (M+1)} \\ 0_{(M+1) \times MN} & 0_{(M+1) \times M} & 0_{(M+1) \times (M+1)} \end{bmatrix},
\]

\[
\tilde{A}_{m} = \begin{bmatrix} A_{m} & 0 \\ 0 & \frac{1}{2} \tilde{a}_{m} \end{bmatrix}, \quad \tilde{B}_{m} = \begin{bmatrix} 0_{Q \times Q} & \frac{1}{2} \tilde{b}_{m} \\ \frac{1}{2} \tilde{b}_{m} & 0 \end{bmatrix},
\]

\[
\tilde{C}_{n} = \begin{bmatrix} 0_{MN \times MN} & C_{n} \\ C_{n}^{T} & 0_{M \times M} \end{bmatrix},
\]

\[
\tilde{D}_{n} = \begin{bmatrix} 0_{(Q-M) \times (Q-M)} & D_{n} \\ D_{n}^{T} & 0_{M \times M} \end{bmatrix},
\]

\[
H = \text{diag}(h_{1}, \ldots, h_{M}), \quad C_{n} = \text{diag}(c_{n,1}, \ldots, c_{n,M}),
\]

\[
D_{n} = \text{diag}(d_{n,1}^{T}, \ldots, d_{n,M}^{T}), \quad 0_{M \times M}^{T}.
\]

Moreover, \( A_{m} \) is a \( Q \times Q \) matrix, where the \((Q - M + MN - m)\)-th and \((MN + m - Q - M + m)\)-th elements are \(-\delta_{m,2}\) and the other elements are zero. The constraints (13a)-(13d) are equivalent to the constraints (12a)-(12d), respectively.

It is noticed that, except for the constraint \( \text{Rank}(Z) = 1 \) in (13e), the objective function and the other constraints in P5 are convex. To solve the problem, we relax P5 by dropping the rank-one constraint. The relaxed problem can be solved by a convex programming solver, e.g., CVX [29], of which the solution is denoted by \( \hat{Z} \). If \( \hat{Z} \) is a rank-one matrix, it is obviously the solution of P5. Otherwise, if \( \text{Rank}(\hat{Z}) > 1 \), it is necessary to convert \( \hat{Z} \) to a feasible solution of the problem.

Based on [29], the randomization way is utilized to obtain the feasible solution. We first extract \( Z^{\delta} \), the upper-left \( Q \times Q \) sub-matrix of \( \hat{Z} \). Then a random column vector \( \xi \) is generated, following a Gaussian distribution with zero mean and covariance \( Z^{\delta} \), i.e. \( \xi \sim N(0_{Q \times 1}, Z^{\delta}) \). Here \( \xi \)
is a random sample in the solution space of P4. However, it may not meet the constraints of the problem. According to (11a), $\hat{\xi}$ is rescaled to approximate the feasible solution by $\tilde{\xi} = \hat{\xi} \min \left\{ \frac{-\epsilon A_{n,\tilde{m}}}{a_{n,\tilde{m}}} \right\}$. We subsequently utilize a sigmoid function $\text{sig}(x) = \frac{1}{1 + \exp(-\mu x)}$ to map $\tilde{\xi}$ into a non-negative vector space and obtain the vector $\hat{y}$. Based on $\hat{y}$, we extract the first $MN$ elements, the last $M$ elements, and the rest of the elements as the three vectors, i.e., $[\hat{x}_1^T, \ldots, \hat{x}_M]^T$, $[\hat{\epsilon}_1, \ldots, \hat{\epsilon}_M]^T$, and $[\hat{f}_1, \ldots, \hat{f}_M]^T$, respectively. To obtain the association of UE $m$, we set the biggest element in $\hat{x}_m$ as 1 and the rest $(N - 1)$ elements as 0. Meanwhile, $\hat{b}_m$ and $\hat{e}_m$ are mapped to $\tilde{b}_m$ and $\tilde{e}_m$ by $\tilde{b}_m = \hat{b}_m b_m / y_{\hat{m},\hat{\hat{m}}}$ and $\tilde{e}_m = \hat{e}_m e_m b_m$, respectively. Note that $y_{\hat{m},\hat{\hat{m}}}$ is the transmission rate of one RB from UE $m$ to the associated ES based on $x_m$. Finally, to meet (9c) and (9d), $\hat{b}_m$ and $\hat{f}_m$ are normalized according to the maximal RBs and computing capacity at the associated ES.

Algorithm 1 SDR-Based Algorithm

\textbf{Input:} $\Theta$, $b_m$, $e_m$, $\alpha_m$, $\gamma_{m,n}$, $\forall m \in \mathcal{M}$, $n \in \mathcal{N}$;

1: solve P5 to obtain $\tilde{Z}$;
2: if $\text{Rank}(\tilde{Z}) = 1$ then
3: extract the first $Q$ elements from the diagonal of $\tilde{Z}$;
4: calculate square roots of extracted diagonal elements;
5: obtain optimal $\tilde{x}_m$, $\tilde{b}_m$ and $\tilde{f}_m$ based on the square roots;
6: else
7: extract the upper-left $Q \times Q$ sub-matrix of $\tilde{Z}$ as $Z$;
8: for $l = 1$ to $L$ do
9: generate Gaussian sample $\xi^{(l)} \sim N(0, \text{cov}(Z))$;
10: $\tilde{\xi}^{(l)} = \xi^{(l)} \min \left\{ \frac{-\epsilon A_{n,\tilde{m}}}{a_{n,\tilde{m}}} \right\}$;
11: calculate $\text{sig}(\tilde{\xi}^{(l)})$, and extract $\hat{x}_m$, $\hat{b}_m$ and $\hat{f}_m$;
12: for $m = 1$ to $M$ do
13: $\check{\hat{n}} \leftarrow \arg\max \{\hat{x}_{m,n}\}$;
14: $x_m \leftarrow 0, \forall n \neq \check{\hat{n}}$, and $x_m, \check{\hat{n}} \leftarrow 1$;
15: end for
16: $k_m^{(l)} \leftarrow 0$ and $f_m^{(l)} \leftarrow 0, \forall m \in \mathcal{M}$, $n \in \mathcal{N}$;
17: for $n = 1$ to $N$ do
18: find $\hat{m}$ that meet $x_m, n = 1$;
19: $k_m^{(l)} \leftarrow \max \left\{ \frac{l}{n_{\hat{m},\hat{m}}} \right\}$;
20: $f_m^{(l)} \leftarrow \max \left\{ \frac{\sum_{n=1}^{N} m_{\hat{m},\hat{m}} \hat{m}}{n_{\hat{m},\hat{m}}} \right\}$;
21: end for
22: end for
23: Among $L$ sets of $x_m$, $k_m$ and $f_m$, the one maximizing the average QoI is the optimum, i.e., $\hat{x}_m$, $\hat{b}_m$ and $\hat{f}_m$;
24: end if

\textbf{Output:} $\hat{x}_m$, $\hat{b}_m$, $\hat{f}_m$.

In order to achieve the effective approximation, we randomly generate $L$ samples, $\xi^{(l)}$, and obtain $x_m^{(l)}$, $k_m^{(l)}$ and $f_m^{(l)}$ ($l \leq L$) following the above-mentioned steps, respectively. During the $L$ set of results, the one that maximizes the objective function of P1 will be the final solution. We summarize the proposed SDR-based Algorithm in Algorithm 1.

| Parameter       | Value      | Parameter       | Value      |
|-----------------|------------|-----------------|------------|
| $[\text{SNR}_{\text{min}}, \text{SNR}_{\text{max}}]$ | [0, 30] dB | $k_{\text{max}}$ | 50         |
| $[f_{\text{min}}, f_{\text{max}}]$ | [12, 30] Cycle/s | $\eta$ | $10^{-7}$ |
| $\alpha_m$ | 238 cycles/bit | $L$ | 100       |

We provide the analysis of the computational complexity for the proposed algorithm as follows. A worst-case complexity to solve the semi-definite programming problem is $\mathcal{O}(\max\{g, h\} h^{0.5} \log (1/\epsilon))$ [30], where $g$ denotes the number of the constraints, $h$ denotes the dimension of the symmetric matrix $Z$, and $\epsilon$ is the solution accuracy ($\epsilon > 0$).

In our case, the number of the constraints is $2M + 2N + 1$, while the dimension of $Z$ is $MN + 2M + 1$. Moreover, the maximal number of loops to obtain the feasible solution of P5 is $2LMN$. Therefore, for the MEC system with multiple UEs, the overall complexity of the proposed algorithm in the worst case is $\mathcal{O}((MN + 2M + 1)^{4.5} \log (1/\epsilon) + 2LMN)$.

V. NUMERICAL RESULTS

In this section, we evaluate the proposed SDR-based algorithm in different scenarios. The number of the ESs in the system is set to 3. The channel bandwidth of each ES is 10 MHz with 50 RBs. The signal-to-noise ratio (SNR) between the UEs and the ESs follows a uniform distribution. Based on the SNRs, the UEs select the proper MCSs to guarantee that the BLER does not exceed $10^{-7}$. The SNR-MCS mapping is obtained by the link abstraction model according to the LTE networks [28]. Our algorithm can still work if we use the SNR-MCS mapping from a different wireless communication system. It is possible to have various MCSs by combining different modulations (QPSK, 16QAM, 64QAM) and coding rates (from 1/9 to 9/10). We use object detection as an example of CV algorithms used in IoT applications. The maximal tolerable image distortion $\theta$ is set to $13\%$ [11].

Table 2 summarizes the default simulation parameters.

We compare the SDR-based algorithm with the three schemes: 1) Exhaustive scheme; 2) Distance-based algorithm, 3) the scheme with perfect input image, i.e., the case $\theta = 0\%$. For the Distance-based algorithm, a UE is always associated with the closest ES. Based on the associations, the RBs and computing capacity are allocated by the interior-point method. We evaluate the QoI and OFP of the algorithms. The numerical results are based on an averaged value over 5000 Monte Carlo simulations. Note that the timeout probability could be caused by both the resource limit in the system and the unfeasible solution of the algorithms.

In Fig. 2 and Fig. 3, we present the performance comparisons with the different task sizes, i.e., the input image size for object detection. The tick labels in x-axis mean the dimension of the images, e.g., $300 \times 300$ image. The ticks 300x, 375x, 416x, 512x and 600x correspond to the task size of 0.27 Mbits, 0.42 Mbits, 0.52 Mbits, 0.78 Mbits and 1.08 Mbits (8 bits/pixel in gray scale), respectively. The number of the UEs is set to 5. Fig. 2 shows that the OFP of
the proposed algorithm is approximated to that of the Exhaustive scheme but with much lower computational complexity. The Distance-based algorithm has much higher OFP than the SDR-based algorithm, which is mainly caused by higher timeout probability. Compared with the 3rd scheme, i.e., using perfect input image, the proposed algorithm allowing 13% distortion improves the OFP by one order of magnitude, when the task size is not larger than 416x. For the case2 $\delta_m = 33$ ms, when offloading images of 512x and 600x, the SDR-based algorithms have similar OFP, no matter distortion is allowed or not. The reason lies in that the timeout probability dominates the OFP due to the limited resource to process the tasks of large size. When $\delta_m$ is relaxed to 40 ms, compared with the scheme with perfect input image, it can reduce OFP nearly one order of magnitude by allowing 13% distortion.

Fig. 3 illustrates how task size impacts QoI of the schemes. The inference accuracy of different image sizes is modeled by $A_{\text{Infer}} = 1 - 1.578e^{-6.5\times10^{-3}d_{\text{img}}}$ [7], where $d_{\text{img}}$ denotes the dimension of the image. Note that $A_{\text{Infer}}$ increases with $d_{\text{img}}$ increasing. It shows that the SDR-based algorithm has the similar QoI as the Exhaustive scheme. It achieves the maximal QoI when the dimension of image is 512x, at which it strikes a balance between the OFP and inference accuracy. The proposed algorithm clearly outperforms the Distance-based scheme. In addition, we only show the performance of the SDR-based algorithm allowing 13% distortion in Fig.3, as the algorithm with perfect input image does not improve the QoI performance. Note that the 13% distortion here is the upper bound that the CV algorithms can allow. On the one hand, once the distortion of the input image exceeds 13%, the accuracy of object detection declines rapidly, thereby significantly decreasing the QoI. On the other hand, the SDR-based algorithm allowing distortion of less 13% will achieve similar OFP and QoI to the case that $\delta = 13\%$, since the service latency is same and inference accuracy is very close.

Fig. 4 and Fig. 5 show how the computational intensity of a task, $\alpha_m$, affects the performance of the algorithms. Here, the number of UEs is 5, and the image size is 512x. The values on the x-axis, i.e., 167 cycles/bit, 238 cycles/bit,
that when $\delta_m$ proposed algorithm is capable to execute YOLOv2 (When is much lower than that of the Distance-based algorithm. $\delta$ use MobileNetV1-based SSD ($\delta$ based SSD, while the Distance-based scheme is better to which is the maximum inference accuracy of MobileNetV2- $\alpha$ SSD ($\alpha$ SSD, YOLOv2, and VGG16-based SSD is 0.725, 0.736, 0.786 and 0.799, respectively, which is tested on the dataset of Pascal VOC2007. It shows the proposed algorithm is able to run the advanced object detection algorithms to achieve high QoI under the different time constraints. The reason is that the proposed algorithm optimizes the radio resource allocation and shortens the communication latency, thereby, obtaining more time budget to carry out the computation. For example, when $\delta_m = 40$ ms, although the tested two schemes can both run MobileNetV2-based SSD ($\alpha_m = 238$ cycles/bit) to detect objects in the image; however, the proposed scheme can achieve the QoI of 0.736 which is the maximum inference accuracy of MobileNetV2-based SSD, while the Distance-based scheme is better to use MobileNetV1-based SSD ($\alpha_m = 167$ cycles/bit) instead achieving the QoI of 0.69. When relaxing $\delta_m$ to 66 ms, the proposed algorithm is capable to execute YOLOv2 ($\alpha_m = 500$ cycles/bit) to achieve the maximal QoI of 0.786. In Fig. 6, we compare the change of OFP of the tested algorithms under different number of UEs. The task size is set to 0.78 Mbits. The OFP of the proposed SDR-based algorithm is much lower than that of the Distance-based algorithm. When $\delta_m = 33$ ms, the OFP of the SDR-based algorithms increases monotonically with the increasing number of UEs, no matter distortion is allowed or not. This is due to the fact that when $\delta_m = 33$ ms, the timeout probability dominates the OFP. However, when $\delta_m$ is relaxed to 40 ms, it achieves lower OFP if distortion is allowed. Furthermore, the differences of the OFP between the proposed algorithms and the Distance-based scheme are enlarged significantly.

VI. CONCLUSION

The CV algorithms for, e.g., object detection and classification, are able to tolerate certain level of image distortion while keeping the required inference accuracy. Motivated by this feature, we investigate the impact of image distortion on the overall service reliability of the time-critical IoT using edge intelligence. We systematically model the reliability as QoI, and formulate the optimization problem to maximize the average QoI for the multi-UE scenarios. The problem is solved by the SDR-based algorithm, which sub-optimally allocates communication and computation resources in the MEC system. The numerical results show that the proposed algorithm achieves similar performance as the Exhaustive scheme but at lower complexity. By allowing small distortion in the image, it can reduce the OFP by one order of magnitude, without any compromise in the inference accuracy. The proposed algorithm is applicable to determine the optimal deep learning method for the time-critical IoT service by striking a balance between the OFP and the inference accuracy.

In the current work, we assume the pixels have equal importance for pattern recognition or objective detection tasks. In practice, it might not be the case. Therefore, it is interesting to study different resource allocation schemes and offloading strategies that jointly taking into account image encoding.

REFERENCES

[1] "5G, Service Requirements for Enhanced V2X Scenarios," document 3GPP TS 22.186, Version 15.3.0 Release 15, 2018.
[2] Q. Zhang and F. H. Fitzek, "Mission critical IoT communication in 5G," in Future Access Enablers of Ubiquitous and Intelligent Infrastructures. Cham, Switzerland: Springer, 2015, pp. 35–41.
[3] J. Liu and Q. Zhang, "Code-partitioning offloading schemes in mobile edge computing for augmented reality," IEEE Access, vol. 7, pp. 11222–11236, 2019.
[4] Z. Zhou, H. Liao, B. Gu, S. Mumtaz, and J. Rodriguez, "Resource sharing and task offloading in IoT fog computing: A contract-learning approach," IEEE Trans. Emerg. Topics Comput. Intell., vol. 4, no. 3, pp. 227–240, Jun. 2020.
[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.
[6] I. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7263–7271.
[7] Q. Liu, S. Huang, J. Opadere, and T. Han, "An edge network orchestrator for mobile augmented reality," in Proc. IEEE Conf. Comput. Commun. (INFOCOM), Apr. 2018, pp. 756–764.
[8] E. Li, Z. Zhou, and X. Chen, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in Proc. Workshop Mobile Edge Commun. (MECOMM), 2018, pp. 31–36.
[9] A. Erfan Eshratifar, A. Esmaili, and M. Pedram, "BottleNet: A deep learning architecture for intelligent mobile cloud computing services," 2019, arXiv:1902.01000.[Online].Available: http://arxiv.org/abs/1902.01000
[10] S. Dodge and L. Karam, “Understanding how image quality affects deep neural networks,” in Proc. 8th Int. Conf. Qual. Multimedia Exp. (QoMEX), Jun. 2016, pp. 1–6.
