Resource Allocation for Semantic-Aware Mobile Edge Computing Systems

Yihan Cang, Ming Chen, *Member, IEEE*, Zhaohui Yang, *Member, IEEE*, Yuntao Hu, Yinlu Wang, Zhaoyang Zhang, *Senior Member, IEEE*, and Kai-Kit Wong, *Fellow, IEEE*

***Abstract*—In this paper, a semantic-aware joint communication and computation resource allocation framework is proposed for mobile edge computing (MEC) systems. In the considered system, each terminal device (TD) has a computation task, which needs to be executed by offloading to the MEC server. To further decrease the transmission burden, each TD sends the small-size extracted semantic information of tasks to the server instead of the large-size raw data. An optimization problem of joint semantic-aware division factor, communication and computation resource management is formulated. The problem aims to minimize the maximum execution delay of all TDs while satisfying energy consumption constraints. The original non-convex problem is transformed into a convex one based on the geometric programming and the optimal solution is obtained by the alternating optimization algorithm. Moreover, the closed-form optimal solution of the semantic extraction factor is derived. Simulation results show that the proposed algorithm yields up to 37.10% delay reduction compared with the benchmark algorithm without semantic-aware allocation. Furthermore, small semantic extraction factors are preferred in the case of large task sizes and poor channel conditions.**

***Index Terms*—Mobile edge computing, semantic-aware, compute-then-transmit, resource management.**

I. Introduction

Due to the rapid development of Internet of things, the desire of numerous terminal devices for a huge amount of emerging computation services has drastically increased the traffic of core networks. In order to release the burden of core networks, mobile edge computing (MEC), deploying the network functions at the network edge, provides users with nearby real-time computing services [1]. Therefore, as a new network architecture, MEC is considered as a promising paradigm to remedy the problem of computational resources and energy shortage of mobile equipments in future wireless networks [2]. In MEC networks, communication and computation resources can be jointly optimized to improve a certain system utility, such as energy consumption [3], delay [4], throughput [5], computation efficiency [6], etc.

Motivated by Shannon's classic information theory, existing works including [3]–[6] are dedicated to research on data-oriented communications. As a novel paradigm that involves the meaning of messages in communication, semantic communications, which concentrate on transmitting the meaning of source information, have revealed the significant potential to reduce the network traffic and thus alleviate spectrum shortage [7], [8]. Different from conventional communications, semantic information needs to be extracted from raw data before transmitted in semantic-aware networks. The development of semantic information discipline provides a foundation for semantic communications [9], [10]. Nowadays, the realization of semantic communications has been demonstrated under different types of transmitted contents such as text [11], image [12], speech [13] transmissions, etc. These works studied the feasibility of semantic communications under different scenarios from the viewpoint of technical levels. However, the problem of how to implement resource management in a semantic communication system also needs to be investigated so as to explore the potential of practical semantic communication networks. Moreover, when the network loads are heavy, communication loads can be converted to computation amounts with the help of semantic aware technology in MEC systems. Thus a tradeoff between communications and computations can be achieved, improving the quality of service (QoS) of terminal devices.

Motivated by above observations, we attempt to investigate a novel resource management framework, which flexibly orchestrates communication and computation resources for semantic-aware MEC systems. The main contributions of this paper are summarized as follows:

- We propose a semantic-aware MEC system to achieve low execution latency. In the considered system, each terminal device (TD) has a computation task to be executed through offloading to the MEC server. To further decrease the transmission burden, each TD sends the small-size extracted semantic information of tasks to the server instead of the large-size raw data. With the help of semantic extraction, the amount of data uploading to the MEC server is reduced, and thus lower transmission delay is achieved.
- To coordinate the operations on TDs and MEC server for minimizing the maximum execution delay of all users, the problem of joint communication and computation resource management is studied. In particular, this problem is constructed as an optimization framework aiming to acquire

This work was supported by the National Natural Science Foundation of China (NSFC) under grant 61871128, 61960206005 and 61960206006, by the National Key Research and Development Program under grant 2018YFB1801905, and by Fundamental Research on Foreword Leading Technology of Jiangsu Province under grant BK20192002.

Y. Cang, M. Chen, Y. Hu and Y. Wang are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mails: yhcang@seu.edu.cn, chenming@seu.edu.cn, huyuntao@seu.edu.cn, yinluwang@seu.edu.cn). Ming Chen is also with the Purple Mountain Laboratories, Nanjing 211100, China.

Z. Yang and Z. Zhang are with College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and with International Joint Innovation Center, Zhejiang University, Haining 314400, China, and also with Zhejiang Provincial Key Laboratory of Info. Proc., Commun. & Netw. (IPCAN), Hangzhou 310027, China (e-mails: yang_zhaohui@zju.edu.cn, ning_ming@zju.edu.cn).

K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London, UK (e-mail: kai-kit.wong@ucl.ac.uk).

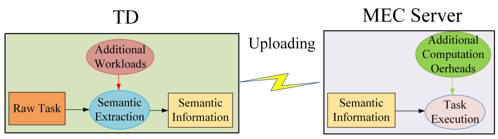


Fig. 1: Flow chart of the semantic-aware MEC networks.

the optimal local computing rate and transmit power, remote computing capacity, and semantic extraction factors while satisfying the energy consumption constraints. Besides, this optimization framework can be applied to general semantic tasks.

- Since the variables are highly coupled with each other, the formulated problem is non-convex which makes it intractable. To this end, a geometric programming based resource management algorithm is proposed to transform the original problem into a convex one and the optimal solution is obtained by the alternating optimization algorithm. Simulation results verify the outstanding performance of the proposed algorithm in terms of execution delay. Compared with the benchmark algorithm without semantic-aware, the proposed algorithm can reduce up to 37.10% delay reduction.

II. System Model and Problem Formulation

A. System Model

Consider a semantic-aware MEC network which consists of a set $\mathcal{N}$ of N TDs and a MEC server attached to a base station (BS). Suppose that TD n has a task with A_n bits to be executed. All the TDs wirelessly access the MEC server for task offloading. Assume that the MEC server is equipped with N core CPUs such that offloaded tasks from different TDs can be executed in parallel.

To further decrease the transmission burden, each TD equipped with a semantic processing unit sends the extracted semantic information of tasks to the server instead of the raw data. With the help of semantic extraction, the amount of data uploading to the MEC server is reduced, thus releasing the traffic loads. Meanwhile, additional workloads for semantic extraction are brought to each TD and the computation intensity of tasks gets large for the MEC server in order to process semantic information, as demonstrated in Fig. 1. Compute-then-transmit protocol is adopted. Specifically, TDs extract and transmit the semantic information of raw data to the server. For examples in Fig. 2, we can utilize the picture cutout method for image transmission, where a pre-trained deep neural network is utilized to cut out the part of important character while discarding the irrelevant background [14]. For text transmission, we can discard meaningless words and use abbreviations to replace the raw text without losing valid information by constructing knowledge graphs [15]. In the considered scenario, only simple background knowledge is required to implement semantic extraction. Hence, the pre-trained model is lightweight and universal. Compared with task processing, the computation resource required for semantic extraction is much less [16].

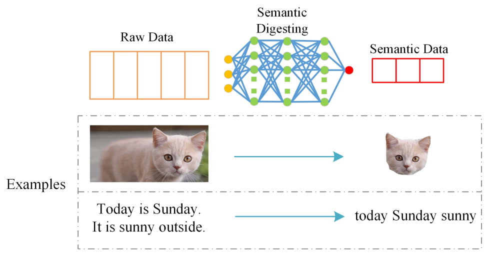
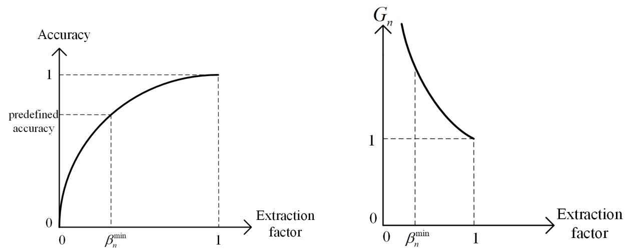


Fig. 2: The schematic diagram of semantic extraction and two examples.



(a) Accuracy versus extraction factor.

(b) Additional computation overhead versus extraction factor.

Fig. 3: Performance demonstration of semantic aware model.

Denote $\beta_n \in [\beta_n^{\min}, 1]$ as the extracting factor of the raw data to be uploaded by TD n, where $\beta_n^{\min}$ is the minimum extracting factor to maintain the most important information from TD n. As shown in Fig. 3(a) [11], [12], [14], the task accuracy of the computed results decreases when the extraction factor becomes small. In order to guarantee the predefined accuracy, β_n should not be smaller than $\beta_n^{\min}$. Therefore, the amount of semantics to be uploaded for TD n is equivalent to $A_n\beta_n$ and the transmission delay is given by

$$T_n^T = \frac{A_n\beta_n}{R_n}, \quad \forall n \in \mathcal{N}, \tag{1}$$

where R_n denotes the achievable rate for TD n. Orthogonal frequency division multiple access (OFDMA) is adopted. Total bandwidth is divided into N sub-channels each with a bandwidth of B (Hz) for each TD. Denote the channel gain between TD n and MEC server by h_n, which keeps stable during the task offloading period. Therefore, R_n is given as

$$R_n = B\log_2\left(1 + \frac{h_n p_n^T}{\sigma^2}\right), \quad \forall n \in \mathcal{N}. \tag{2}$$

where σ^2 denotes the additive noise power at the MEC server and p_n^T represents the transmit power of TD n.

As shown in Fig. 1, at the transmitter, semantic extraction brings additional workloads. Meanwhile, at the receiver, additional computation overhead is required to process the semantic data rather than raw data. This indicates that communication loads are converted to computation amounts through semantic transmission. Denote C_n as the workloads for semantic extraction in CPU cycles. Note that C_n should increase with the amount of extracted raw data A_n, while decreases with extraction factor β_n. Without loss of generality, we assume that[1]

$$C_n = \frac{aA_n}{\beta_n^k}, \quad \forall n \in \mathcal{N}, \tag{3}$$

where $a > 0, k > 0$ are constants relevant to the tasks. Therefore, the additional delay caused by semantic extraction in the local is expressed as

$$T_n^L = \frac{C_n}{f_n^L} = \frac{aA_n}{\beta_n^k f_n^L}, \quad \forall n \in \mathcal{N}, \tag{4}$$

where f_n^L represents the local CPU frequency of TD n. Therefore, the local computing power consumption is given as

$$p_n^C = \kappa_n (f_n^L)^3, \quad \forall n \in \mathcal{N}, \tag{5}$$

where κ_n denotes the energy coefficient of TD n.

Denote I_n as the computation intensity of TD n in units of CPU cycles per raw data bit. Since only extracted semantic data of the raw data can be received by the MEC server, the computation intensity of the extracted semantic increases accordingly. As shown in Fig. 3(b), $G_n \in [1, +\infty]$, which denotes the ratio of computation intensity of semantic data to that of raw data about TD n, increases monotonously with extraction factor β_n getting small. The increase is caused by computations for processing semantic data and compensations for enhancing accuracy. Besides, we should have $G_n = 1$ for the case that raw data are processed, i.e., $\beta_n = 1$. Without loss of generality, we assume that

$$G_n = \frac{1}{\beta_n^p}, \quad \forall n \in \mathcal{N}, \tag{6}$$

where $p > 0$ is a constant relevant to specific task computation. Thus the remote semantic recovery and tasks execution delay can be given by

$$T_n^O = \frac{A_n \beta_n I_n G_n}{f_n^O}, \quad \forall n \in \mathcal{N}, \tag{7}$$

where f_n^O denotes the computation resource of MEC server allocated to the TD n. Note that the precise formulation parameters of C_n and G_n can be obtained by fitting the corresponding data points of abundant prior experiments. We will show that the proposed framework is well applicable to the general formulations of C_n and G_n in the simulations.

B. Problem Formulation

Due to the high transmit power of the BS, it is assumed that the downlink transmission time is negligible. In this case, the total delay for TD n, which contains delay for semantic extraction, transmission delay, and remote processing delay, is given as

$$T_n = T_n^L + T_n^T + T_n^O, \quad \forall n \in \mathcal{N}. \tag{8}$$

[1]Note that the additional workload formulation for semantic extraction is provided in (3) as an example, which can be verified via simulations. Our method can also be applied to different types of workload formulation.

Let $\mathbf{\Phi} = \{\boldsymbol{f}^L, \boldsymbol{p}^U, \boldsymbol{f}^O, \boldsymbol{\beta}\}$, where $\boldsymbol{f}^L = [f_1^L, \cdots, f_N^L]^T$ and $\boldsymbol{f}^O = [f_1^O, \cdots, f_N^O]^T$ respectively denote local and remote computation capacity vectors, $\boldsymbol{p}^T = [p_1^T, \cdots, p_N^T]^T$ is the transmission power vector, $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_N]^T$ denotes extraction factor vector of all TDs. We aim at minimizing the maximum delay of all TDs under the energy consumption of each TD which consists of semantic extraction and transmission energy consumption. Mathematically, the optimization problem is posed as

$$\begin{aligned}
&\min_{\mathbf{\Phi}} \quad \max_{n \in \mathcal{N}} \; T_n, && \text{(9a)}\\
&\text{s.t.} \quad p_n^C T_n^L + p_n^T T_n^T \le E_n, \quad \forall n \in \mathcal{N}, && \text{(9b)}\\
&\qquad 0 \le f_n^L \le f_n^{\max}, \quad \forall n \in \mathcal{N}, && \text{(9c)}\\
&\qquad \sum_{n=1}^{N} f_n^O \le F_{MEC}, && \text{(9d)}\\
&\qquad 0 \le p_n^T \le p_n^{\max}, \quad \forall n \in \mathcal{N}, && \text{(9e)}\\
&\qquad \beta_n^{\min} \le \beta_n \le 1, \quad \forall n \in \mathcal{N}, && \text{(9f)}
\end{aligned}$$

where $f_n^{\max}$ and $p_n^{\max}$ respectively denote the maximum local computation capacity and maximum transmission power of TD n, F_{MEC} is the maximum computation capacity of MEC server, and E_n represents the maximum energy consumption of TD n. In problem (9), constraint (9b) reflects that the energy consumption of TD n should not be larger than its predefined energy consumption threshold. Constraints (9c) and (9e) enforce the local computation capacity and transmission power to be non-negative and should not exceed the predefined budget. The remote computation capacity allocation is constrained in (9d); and constraint (9f) specifies the semantic extraction factor limitations. Different from previous works [3]–[5], [17], [18], semantic extraction factor β_n is considered in problem (9).

III. Optimal Algorithm Design

A. Problem Transformation

The objective function (9a) is in a max-min form, which is complicated. To handle this issue, we introduce an auxiliary variable t. Hence, problem (9) is transformed into the following equivalent problem:

$$\begin{aligned}
&\min_{\mathbf{\Phi}, t} \; t, && \text{(10a)}\\
&\text{s.t.} \; T_n^L + T_n^T + T_n^O \le t, \quad \forall n \in \mathcal{N}, && \text{(10b)}\\
&\qquad (9\text{b}) - (9\text{f}).
\end{aligned}$$

Due to the intractability complex term $p_n^T T_n^T$ in constraint (9b), we introduce t_n^T as the transmission delay for TD n and $\boldsymbol{t}^T = [t_1^T, \cdots, t_N^T]^T$ as the transmission delay vector. Thus, problem (10) is reformulated as

$$\begin{aligned}
&\min_{\mathbf{\Phi}, \boldsymbol{t}^T, t} \; t, && \text{(11a)}\\
&\text{s.t.} \; \frac{aA_n}{\beta_n^k f_n^L} + t_n^T + \frac{A_n \beta_n I_n}{f_n^O \beta_n^p} \le t, \quad \forall n \in \mathcal{N}, && \text{(11b)}\\
&\qquad p_n^C T_n^L + p_n^T t_n^T \le E_n, \quad \forall n \in \mathcal{N}, && \text{(11c)}\\
&\qquad t_n^T R_n \ge \beta_n A_n, \quad \forall n \in \mathcal{N}, && \text{(11d)}\\
&\qquad (9\text{c}) - (9\text{f}),
\end{aligned}$$

where constraint (11d) implies that all the extracted semantic information should be uploaded to the MEC server. Since constraint (11d) is non-convex, we introduce $e_n^T = p_n^T t_n^T$ as the transmission energy consumption variable for TD n. Therefore, (11d) is transformed into

$$t_n^T B \log_2(1 + \frac{h_n e_n^T}{t_n^T \sigma^2}) \geq \beta_n A_n, \quad \forall n \in \mathcal{N}. \tag{12}$$

Since function $f(e_n^T) = B \log_2(1 + \frac{h_n e_n^T}{\sigma^2})$ is concave with respect to e_n^T, its perspective $t_n^T f(e_n^T/t_n^T)$ is also concave with (e_n^T, t_n^T). Therefore, constraint (12) is a convex set. Moreover, due to the coupleness among β_n, f_n^L, and f_n^O, the geometric programming (GP) algorithm can be utilized to handle this issue. In specific, denote $\beta_n = e^{\tilde{\beta}_n}$, $f_n^L = e^{\tilde{f}_n^L}$, $f_n^O = e^{\tilde{f}_n^O}$, $(\forall n \in \mathcal{N})$. Problem (11) is transformed into the following problem

$$\min_{\tilde{\boldsymbol{f}}^L, \boldsymbol{e}^T, \boldsymbol{t}^T, \tilde{\boldsymbol{f}}^O, \tilde{\boldsymbol{\beta}}, t} \quad t, \tag{13a}$$

$$\text{s.t. } aA_n e^{-k\tilde{\beta}_n - \tilde{f}_n^L} + t_n^T + A_n I_n e^{(1-p)\tilde{\beta}_n - \tilde{f}_n^O} \leq t, \forall n \in \mathcal{N}, \tag{13b}$$

$$aA_n \kappa_n e^{2\tilde{f}_n^L - k\tilde{\beta}_n} + e_n^T \leq E_n, \quad \forall n \in \mathcal{N}, \tag{13c}$$

$$t_n^T B \log_2(1 + \frac{h_n e_n^T}{t_n^T \sigma^2}) \geq e^{\tilde{\beta}_n} A_n, \quad \forall n \in \mathcal{N}, \tag{13d}$$

$$\tilde{f}_n^L \leq \ln(f_n^{\max}), \quad \forall n \in \mathcal{N}, \tag{13e}$$

$$\sum_{n=1}^{N} e^{\tilde{f}_n^O} \leq F_{MEC}, \tag{13f}$$

$$0 \leq e_n^T \leq p_n^{\max} t_n^T, \quad \forall n \in \mathcal{N}, \tag{13g}$$

$$\ln(\beta_n^{\min}) \leq \tilde{\beta}_n \leq 0, \quad \forall n \in \mathcal{N}, \tag{13h}$$

which is convex.

Theorem 1. *For problem* (13)*, we have the following properties:*

a) *Problem* (13) *is equivalent to problem* (9)*.*
b) *For the optimal solution of problem* (13)*, the equality in (13b) holds for $\forall n \in \mathcal{N}$.*
c) *The equality in (13e) holds for the optimal solution of problem* (13)*.*

Proof. a) can be obtained according to the transformation procedures from (10) to (13). b) and c) can be derived by contradiction. Detailed proof is omitted due to space limitations. □

B. Optimal Algorithm

To further decrease the complexity of solving problem (13) with utilizing the interior-point method, an alternating optimization algorithm is proposed. In the proposed alternating algorithm, three subproblems requires to be solved, i.e., local and remote computation rate optimization, transmission delay and energy consumption optimization, as well as semantic extraction factor optimization.

1) Optimal Local and Remote Computation Rate: According to (13b), the optimal t decreases with $\tilde{f}_n^L$. Since the left hand side (LHS) of (13c) increases with $\tilde{f}_n^L$, the optimal $\tilde{f}_n^L$ should satisfy either (13c) or (13e). Due to that $f_n^L = e^{\tilde{f}_n^L}, \beta_n = e^{\tilde{\beta}_n}$ and setting (13c) with equality, the optimal f_n^L is given by

$$f_n^L = \min\left\{ f_n^{\max}, \sqrt{\frac{\beta_n^k (E_n - e_n^T)}{aA_n \kappa_n}} \right\}, \quad \forall n \in \mathcal{N}. \tag{14}$$

According to Theorem 1, setting (13b) with equality yields

$$f_n^O = \frac{A_n I_n \beta_n^{1-p}}{t - \frac{aA_n}{\beta_n^k f_n^L} - t_n^T}. \tag{15}$$

Therefore, the bisection method can be adopted to find the optimal t and f_n^O. The upper bound and lower bound of t can be respectively provided as

$$t^{\max} = \max_{n \in \mathcal{N}} \frac{aA_n}{\beta_n^k f_n^L} + t_n^T + \frac{A_n I_n \beta_n^{1-p} N}{F_{MEC}}, \tag{16}$$

$$t^{\min} = \max_{n \in \mathcal{N}} \frac{aA_n}{\beta_n^k f_n^L} + t_n^T + \frac{A_n I_n \beta_n^{1-p}}{F_{MEC}}. \tag{17}$$

The detailed steps are summarized in Algorithm 1.

Algorithm 1 Optimal Local and Remote Computing Capacity Algorithm

---

Initialize: $t^{\max}$ and $t^{\min}$ respectively in (16) and (17), bisection accuracy ϵ_1.
Calculate the optimal local computation rate f_n^L according to (14).
Repeat:
  Set $t \leftarrow \frac{t^{\max} + t^{\min}}{2}$;
  For $n = 1 : N$:
    Calculate f_n^O according to (15).
  If $\sum_{n=1}^{N} f_n^O(t) - F_{MEC} < 0$:
    $t^{\max} \leftarrow t$;
  Else:
    $t^{\min} \leftarrow t$.
  End
Until: $t^{\max} - t^{\min} \leq \epsilon_1$.
Output: optimal f_n^L and f_n^O.

---

2) Optimal Transmission Delay and Energy Consumption: With fixed $\tilde{\boldsymbol{f}}^L$, $\tilde{\boldsymbol{f}}^O$, and $\tilde{\boldsymbol{\beta}}$, problem (13) is reduced to the following N parallel subproblems:

$$\min_{e_n^T, t_n^T} \quad t_n^T, \tag{18a}$$

$$\text{s.t.} \quad e_n^T \leq E_n - aA_n \kappa_n (f_n^L)^2 / \beta_n^k, \tag{18b}$$

$$t_n^T B \log_2(1 + \frac{h_n e_n^T}{t_n^T \sigma^2}) \geq \beta_n A_n, \tag{18c}$$

$$0 \leq e_n^T \leq p_n^{\max} t_n^T. \tag{18d}$$

Similarly, bisection method can be used to solve problem (18). Specifically, given t_n^T, e_n^T should satisfy the following condition according to (18b) and (18d):

$$0 \leq e_n^T \leq \nu_n, \tag{19}$$

where $\nu_n = \min\{E_n - aA_n \kappa_n (f_n^L)^2 / \beta_n^k, p_n^{\max} t_n^T\}$. If $t_n^T B \log_2(1 + \frac{h_n \nu_n}{t_n^T \sigma^2}) \geq \beta_n A_n$, problem (18) is feasible; otherwise, it is infeasible. The detailed procedures are summarized in Algorithm 2.

Algorithm 2 Optimal Transmission Delay and Energy Consumption Algorithm

Initialize: $ub \leftarrow$sufficiently large, $lb \leftarrow 0$, bisection accuracy ϵ_2.
For $n = 1 : N$:
  Repeat:
    Set $t_n^T \leftarrow \frac{ub+lb}{2}$;
    If $t_n^T B \log_2(1 + \frac{h_n \nu_n}{t_n^T \sigma^2}) \geq \beta_n A_n$:
      $ub \leftarrow t_n^T$;
    Else:
      $lb \leftarrow t_n^T$.
    End
  Until: $ub - lb \leq \epsilon_2$.
End
Output: the optimal t_n^T and e_n^T.

3) Optimal Semantic Extraction Factor: With fixed $\tilde{\boldsymbol{f}}^L, \boldsymbol{t}^T, \boldsymbol{e}^T, \tilde{\boldsymbol{f}}^O$, problem (13) is reduced to

$$\min_{\boldsymbol{\beta}, t} \quad t, \tag{20a}$$

$$\text{s.t.} \quad \frac{aA_n}{f_n^L \beta_n^k} + t_n^T + \frac{A_n I_n \beta_n^{1-p}}{f_n^O} \leq t, \quad \forall n \in \mathcal{N}, \tag{20b}$$

$$\frac{aA_n \kappa_n (f_n^L)^2}{\beta_n^k} + e_n^T \leq E_n, \quad \forall n \in \mathcal{N}, \tag{20c}$$

$$\beta_n A_n \leq t_n^T B \log_2\left(1 + \frac{h_n e_n^T}{t_n^T \sigma^2}\right), \quad \forall n \in \mathcal{N}, \tag{20d}$$

$$\beta_n^{\min} \leq \beta_n \leq 1, \quad \forall n \in \mathcal{N}. \tag{20e}$$

Problem (20) can be further equivalent to

$$\min_{\beta_n} \quad \frac{aA_n}{f_n^L \beta_n^k} + t_n^T + \frac{A_n I_n \beta_n^{1-p}}{f_n^O}, \tag{21a}$$

$$\text{s.t.} \quad \eta_1 \leq \beta_n \leq \eta_2, \tag{21b}$$

where $\eta_1 = \max\left\{\beta_n^{\min}, \sqrt[k]{\frac{aA_n\kappa_n(f_n^L)^2}{E_n - e_n^T}}\right\}$ and $\eta_2 = \min\left\{1, \frac{t_n^T}{A_n} B \log_2\left(1 + \frac{h_n e_n^T}{t_n^T \sigma^2}\right)\right\}$. We consider the following two cases for parameter p.

Case 1: $p \geq 1$. In this case, (21a) decreases with β_n. Therefore, the optimal $\beta_n = \eta_2$.

Case 2: $0 < p < 1$. Since (21a) is convex with respect to β_n and the stationary point can be derived by $\mu = \sqrt[k+1-p]{\frac{akf_n^O}{f_n^L I_n (1-p)}}$. Therefore, the closed-form optimal solution is given as

$$\beta_n = \begin{cases} \eta_1, & \text{if } \mu \leq \eta_1, \\ \mu, & \text{if } \eta_1 < \mu \leq \eta_2, \\ \eta_2, & \text{if } \eta_2 < \mu. \end{cases} \tag{22}$$

Remark: According to (22), semantic extraction factor β_n decreases with task size A_n since the data amount to be uploaded can be greatly reduced with a small β_n. Moreover, bad channel conditions results in a small β_n as a small β_n can alleviate the negative impacts incurred by bad channel conditions. Furthermore, according to (19), energy consumption decreases with β_n decreasing. This is due to the fact that transmit power is reduced with the aid of semantic transmission.

Algorithm 3 Overall Semantic-Aware MEC Resource Allocation Algorithm

Initialize: N, A_n, κ_n, $f_n^{\max}$, $p_n^{\max}$, $\beta_n^{\min}$, F_{MEC}, E_n.
Repeat:
  With given transmission delay, energy consumption and semantic extraction, obtain local and remote computation rate according to Algorithm 1.
  With given local, remote computation rate, and semantic extraction, obtain transmission delay and energy consumption according to Algorithm 2.
  Update semantic extraction factors through solving problem (21).
Until: the objective value (13a) converges.
Output: the optimal $\boldsymbol{f}^L$, $\boldsymbol{e}^T$, $\boldsymbol{t}^T$, $\boldsymbol{f}^O$, $\boldsymbol{\beta}$, and t.

TABLE I: Simulation parameters

$\sigma^2 = -174$ dBm/Hz	$B = 1$ MHz	$I_n = 70$ cycles/bit
$\kappa_n = 10^{-26}$	$A_n = 3$ Mbits	$f_n^{\max} = 1$ GHz
$p_n^{\max} = 1$ Watts	$\beta_n^{\min} = 0.6$	$F_{MEC} = 13$ GHz
$a = 10^{-5}$	$k = 4$	$p = 3$
$E_n = 0.5$ Watts	$\epsilon_1 = 10^{-7}$	$\epsilon_2 = 10^{-7}$

IV. Simulation Results

In this section, numerical results are conducted to evaluate the performance of the proposed framework. There are $N = 10$ TDs with each equipped with a semantic extracting processing unit to implement semantic-aware edge computing. The distances between TDs and BS are evenly set in $[120, 255]$ meters. The channel model follows in [20]. The other parameters are set as in Table I unless otherwise mentioned.

Fig. 4 illustrates the total delay performances versus the predefined energy threshold. The following benchmark algorithms are provided to compare with the proposed semantic-aware MEC framework: 1) MEC without semantic, i.e., $\beta_n(t) = 1$ for all TDs. This algorithm corresponds to the conventional MEC scenario that TDs directly upload raw data to the MEC server [4]. 2) Local Execution Algorithm, where the tasks are executed locally under the energy and power constraints. As can be seen in Fig. 4, maximum delay of all users decreases with energy thresholds. Moreover, the proposed framework significantly outperforms benchmark algorithms under different tasks sizes. On average, the proposed framework yields up to 37.10% and 69.35% delay reduction compared with MEC without semantic algorithm and local execution algorithm, respectively. This is due to the fact that semantic-aware MEC can extract key data, thus efficiently reducing transmission delay while maintaining accuracy.

To test the framework in applications to various semantic-aware scenarios, in Fig. 5, we plot the performance comparisons between different workload expressions C_n and additional computation overhead expressions G_n under various minimum extraction factors. Note that the case when $\beta_n^{\min} = 1$ is equivalent to the traditional MEC without semantic-aware. As can be seen, as the minimum extraction factor decreases from 1.0 to 0.5, the total delay becomes small. This can be explained by that a smaller extraction factor significantly reduces the

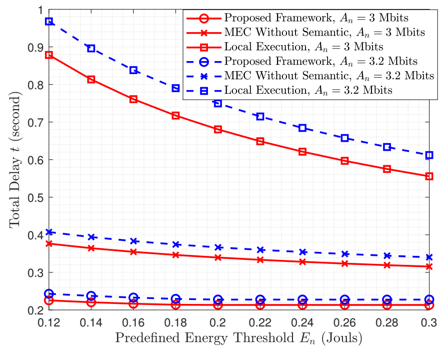


Fig. 4: Performance comparisons between different algorithms under different energy thresholds and tasks sizes.

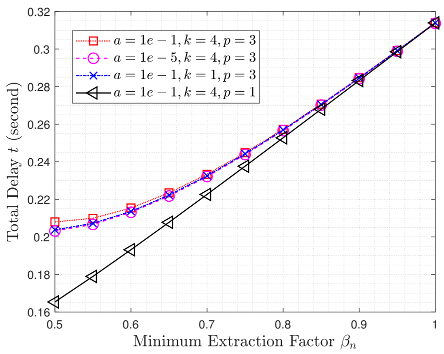


Fig. 5: Performance comparisons between different C_n and G_n expressions under different minimum extraction factors.

amount of data to be uploaded, thus the transmit delay is reduced. Furthermore, the performance of the proposed algorithm is similar even under different workload expressions. This shows its ability to apply to different semantic-aware scenarios. Besides, larger a and k can result in a larger delay. This is because as a and k become large, more workloads for semantic extraction shall be finished at TDs according to (3), which incurs a larger processing delay. Similarly, a larger p leads to a higher delay in Fig. 5. This is due to the fact that G_n becomes larger when p gets bigger, which indicates that processing a bit of semantic data requires more CPU cycles at the MEC server. Thus a higher execution delay is needed for semantic recovery.

V. Conclusions

In this paper, we have proposed a joint communication and computation resource allocation framework for MEC systems with the aid of prevalent semantic transmission technology. An optimization problem has been formulated to optimize semantic extraction factors, communication and computation resources to minimize the maximum delay of all TDs by taking into account energy consumption constraints. The original non-convex problem is transformed into a convex one based on GP, which can be efficiently solved through the proposed alternating optimization algorithm. Simulation results demonstrate the superiority of the proposed framework over the benchmark schemes with respect to delay. Besides, the maximum delay of all users is significantly reduced as the semantic extraction factor gets small.

References

[1] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 3, pp. 1657–1681, thirdquarter 2017.

[2] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," *IEEE Access*, vol. 8, pp. 116 974–117 017, June 2020.

[3] X. Hu, K.-K. Wong, and K. Yang, "Wireless Powered Cooperation-Assisted Mobile Edge Computing," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 4, pp. 2375–2388, Apr. 2018.

[4] Y. Yu, Y. Yan, S. Li, Z. Li, and D. Wu, "Task Delay Minimization in Wireless Powered Mobile Edge Computing Networks: A Deep Reinforcement Learning Approach," in *Proc. International Conference on Wireless Communications and Signal Processing (WCSP)*, Changsha, China, Oct. 2021, pp. 1–6.

[5] Z. Zhu, J. Peng, X. Gu, H. Li, K. Liu, Z. Zhou, and W. Liu, "Fair Resource Allocation for System Throughput Maximization in Mobile Edge Computing," *IEEE Access*, vol. 6, pp. 5332–5340, Jan. 2018.

[6] Y. Cang, M. Chen, J. Zhao, T. Gong, J. Zhao, and Z. Yang, "Fair Computation Efficiency for OFDMA-Based Multiaccess Edge Computing Systems," *IEEE Commun. Lett.*, vol. 27, no. 3, pp. 916–920, Mar. 2023.

[7] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.

[8] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge Learning for B5G Networks With Distributed Signal Processing: Semantic Communication, Edge Computing, and Wireless Sensing," *IEEE J. Sel. Topics Sig. Proces.*, vol. 17, no. 1, pp. 9–39, Jan. 2023.

[9] R. Carnap, Y. Bar-Hillel *et al.*, "An outline of a theory of semantic information," 1952.

[10] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, "Towards a theory of semantic communication," in *Proc. IEEE Network Science Workshop*, West Point, NY, USA, June 2011, pp. 110–117.

[11] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep Learning Enabled Semantic Communication Systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, Apr. 2021.

[12] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, "Deep Joint Source-Channel Coding for Wireless Image Transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, May 2019.

[13] Z. Weng and Z. Qin, "Semantic Communication Systems for Speech Transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Aug. 2021.

[14] L. Cao and S. Lin, "Target Detection Algorithm of Optimized Convolutional Neural Network under Computer Vision," in *Proc. Int. Conf. Unmanned Systems (ICUS)*, Harbin, China, Nov. 2020, pp. 923–930.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[16] Z. Yang, M. Chen, Z. Zhang, and C. Huang, "Energy Efficient Semantic Communication Over Wireless Networks With Rate Splitting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 5, pp. 1484–1495, May 2023.

[17] F. Zhou and R. Q. Hu, "Computation Efficiency Maximization in Wireless-Powered Mobile Edge Computing Networks," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 5, pp. 3170–3184, Feb. 2020.

[18] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access*, vol. 4, pp. 5896–5907, Aug. 2016.

[19] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[20] S. Bi, L. Huang, H. Wang, and Y.-J. A. Zhang, "Lyapunov-Guided Deep Reinforcement Learning for Stable Online Computation Offloading in Mobile-Edge Computing Networks," *IEEE Trans. Wirel. Commun.*, vol. 20, no. 11, pp. 7519–7537, Nov. 2021.