Adaptable Semantic Compression and Resource Allocation for Task-Oriented Communications

Chuanhong Liu, Caili Guo, Senior Member, IEEE, Yang Yang, Member, IEEE, and Nan Jiang

Abstract—Task-oriented communication is a new paradigm that aims at providing efficient connectivity for accomplishing intelligent tasks rather than reception of every transmitted bit. This paper proposes task-oriented communication architecture for end-to-end semantics transmission, where extracted semantics is compressed by the proposed adaptable semantic compression (ASC) method. However, accommodating multiple users in a delay-intolerant system poses a challenge. Higher compression ratios conserve channel resources but cause semantic distortion, while lower ratios demand more resources and may lead to transmission failure due to delay constraints. To address this, we optimize both compression ratio and resource allocation to maximize task success probability. Specifically, we propose a compression ratio and resource allocation (CRRA) algorithm that separates the problem into two subproblems and solving them iteratively. Furthermore, for scenarios with varying service levels among users, a compression ratio, resource allocation, and user selection (CRRAS) algorithm is proposed, adaptively selecting users through branch and bound method. Simulation results show that ASC approach can reduce the size of transmitted data by up to 80% without compromising task success probability. Furthermore, numerical results clearly demonstrate that both the proposed CRRA and CRRAS algorithms lead to substantial improvements in terms of success gains when compared to the baselines.

Index Terms—Semantic communication, task-oriented, semantic compression, resource allocation.

I. INTRODUCTION

TO SUPPORT the rapid development of artificial intelligence (AI), providing connectivity for intelligent tasks performed on the edge is one of the key applications in future wireless communication systems [1]. These tasks are deemed as machine understanding and performing tasks automatically in a fashion close to human cognition, such as autonomous driving, mobile virtual reality (VR)/augmented reality (AR) online games, and other distributed signal processing tasks [2]. However, taking VR as an example, achieving fully-immersive VR requires the delivery of massive data (in gigabytes) and ultra-low latency (20 ms or less), thus placing stringent requirements on ultra-reliable and low-latency communications (URLLC) [3]. To provide connectivity for such tasks and significantly alleviate the scarcity of communication resources, the goal of communication is no longer the accurate reception of every transmitted bit but to transmit the meaningful content of raw data to accomplish the tasks. This communication paradigm refers to task-oriented or semantic communication, which has attracted extensive attention from industry and academia and has been identified as one of the core techniques in the next-generation wireless communication systems [4]. Recent research shows that, with this paradigm, the semantics of the source information with respect to the requirements of tasks are exchanged between transceivers, which greatly reduces the network traffic [5].

Semantic information refers to any context, semantics, or features that are relevant to the decision-making for goals at the receiver [6], [7], [8], which is abstract and subjective. More specifically, the same data may have different semantics in various intelligent tasks. Correspondingly, the manner of semantic compression can highly depend on its application itself, which can be challenging due to the lack of a unified compression criterion. In addition, employing semantic communications in wireless networks faces several challenges, including the semantic theory, semantic extraction method, semantic-oriented resource allocation, and the performance metrics, which motivates us to investigate more in this area.

Some prior studies have been dedicated to designing fundamental frameworks for semantic communication from the information theoretic perspective [9], [10], [11], [12]. An end-to-end (E2E) semantic communication framework was proposed in [9], which introduces the semantic sampling that allows each smart device to control its traffic via semantic-aware active sampling. In [10], the framework of task-oriented semantic communication was proposed. The authors in [11] proposed an E2E learning-driven architecture of semantic communication to integrate the semantic inference and physical layer communication problems, where the transceiver is optimized jointly to reach Nash equilibrium while minimizing the average semantic errors. The authors in [12]...
introduced the concept of a semantic-effectiveness (SE) plane, which augmented the protocol stack by providing standardized interfaces that enable information filtering and direct control of functionalities at all layers of the protocol stack. Recently, deep learning (DL) has emerged as a popular solution for signal processing tasks and semantic communications due to its powerful feature extraction capability [13], [14], [15], [16], [17], [18], [19], [20], [21], [22]. The authors in [13] proposed a series-constellation multi-modal feature network (SC-MFNet) to recognize the modulation types of orthogonal frequency-division multiplexing (OFDM) subcarriers in multiple-input multiple-output (MIMO) wireless channels. Liang et al. [14] adopted echo state network to the real-time symbol detection task in MIMO-OFDM systems, which is implemented with a software-defined radio (SDR) transceiver and tested in various real-world scenarios. He et al. [15] proposed deep learning-based MIMO detection method with estimated channel, which improved the performance of MIMO receiver by considering the characteristics of channel estimation error and channel statistics. The authors in [16] proposed a semantic communication system based on Transformer, which explained the concept of semantic information at the sentence level. Based on [16], the authors in [17] further proposed a lite distributed semantic communication system, making the model easier to deploy on the Internet of Things (IoT) devices. For image data transmission, the authors in [18] presented a JSCC scheme based on convolutional neural networks (CNN) to transmit image data over wireless channel, which can jointly optimize various modules of the communication system.

More recently, DL-driven communication architecture considering the semantics of specific tasks has been proposed [19], [20], [21], [22]. Lee et al. [19] designed a joint transmission-classification system for images, in which the receiver outputs image classification results directly. It has been verified that such a joint design achieved higher classification accuracy than performing image recovery and classification separately. Jankowski et al. [20] considered image-based re-identification for persons or cars as the communication task, where two schemes were proposed to improve the retrieval accuracy. In our prior works [21], an intelligent task-oriented communication method has been proposed for AI of Things (AIoT), in which semantics can be further compressed without performance penalty. For multimodal data transmission, Xie et al. [22] developed MU-DeepSC for the visual question answering task, where one user transmits text-based questions about images, and the inquiry images are transmitted from another user. In summary, the existing works on semantic communications are focused on the implementation of semantic communication systems, in which extracted semantics is extracted via a fixed neural network and directly transmitted without further compression. None of them can adaptively compress the semantics and optimize the compression ratio.

In response to the potential enormity of extracted semantics, semantic compression endeavors to mitigate subsequent computing overhead and minimize the volume of transmitted data, thereby alleviating the communication burden. In addition, there is a compelling need to devise more efficient and appropriate resource allocation schemes that allocate limited communication resources to data with richer semantic information in a task-oriented manner. Considering semantic-aware resource allocation, the authors in [23] optimized resource allocation in terms of channel assignment and the number of transmitted semantic symbols for text semantic communication. A novel quality-of-experience (QoE) model has been proposed in [24], based on which the number of transmitted semantic symbols, channel allocation, and user power are allocated. The authors in [7] proposed a deep reinforcement learning-based resource allocation algorithm to maximize the text semantic similarity. The authors in [25] proposed a long-term resource allocation scheme to realize effective link transmission, in which conventional performance indicators (user satisfaction, queue stability, and communication delay) are considered. While previous studies have delved into resource allocation for text transmission-based semantic communication [7], [23], [24], [25], none of them have taken the communication task itself into consideration or established a connection between wireless resources and task performance. This limitation could potentially hinder the effectiveness of future communication systems, particularly when the communication tasks are well-defined. Therefore, there is a compelling need to further investigate resource allocation policies to enhance the performance of task-oriented communication. Furthermore, since users may require transmission service for various intelligent tasks [26], appropriate resource allocation in a semantic aware manner is crucial that guarantees the transmission with prioritized reliabilities. To optimize the performance of semantic communications, the following major issues remain to be solved: 1) How to adaptively compress semantics with respect to the intelligent tasks? 2) How to appropriately allocate communication resources (including bandwidth and transmit power) for compressed data?

In this paper, we investigate the semantic compression and resource allocation for task-oriented multi-user semantic communication systems. To our best knowledge, this is the first work that proposes a model of semantic compression and resource allocation for task-oriented multi-user semantic communications. The main contributions of this paper are summarized as follows:

- We first design a gradient-based semantic importance evaluation method. Based on the evaluated semantic importance, an adaptable semantic compression (ASC) approach is proposed to compress extracted semantics so as to save communication resources. The mathematical relationship between the performance of intelligent tasks and semantic compression ratios is then investigated.

- Due to wireless resource limitations, users must adaptively determine the optimal semantic compression ratios, and wireless resources must be appropriately allocated to satisfy the transmission delay constraint. The problem is formulated as an optimization problem whose goal is to maximize the success probability of tasks in terms of resource allocation, user selection, and semantic compression ratios. To solve this nonconvex problem, a compression ratio and resource allocation (CRRA)
algorithm is proposed for scenarios where users have the same service levels, in which the problem is separated into two subproblems and solved iteratively.

- Considering that users may have various service levels, we further propose a CRRA with dynamic user selection (CRRAUS) algorithm, in which compression ratio, user selection, and resource allocation are simultaneously optimized. Specifically, the branch and bound method is used to select users based on resources and service levels adaptively.

The remainder of this paper is organized as follows. The system model and problem formulation are described in Section II. Section III details the E2E semantic communication and ASC approach. The proposed CRRA algorithm and CRRAUS algorithm are presented in Sections IV and V, respectively. Simulation and numerical results are analyzed in Section VI. Section VII draws some important conclusions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a multi-user semantic communication system composed of an edge server and a set $U$ of $U$ users as illustrated in Fig. 1. Taking IoT scenario as an example, the IoT user aims at gathering data locally and performing an inference task with the assistance of the edge server. To do so, semantics of raw data is extracted and compressed locally by users before uploaded. Then, the semantics is transmitted to the edge server, where the edge server allocates channel resource according to channel state information as well as the prior knowledge of semantic compression. Finally, the edge server performs intelligent computing according to the received semantics and returns the result of tasks to users. The users and edge server are equipped with a certain knowledge base to facilitate semantic extraction and compression, where the knowledge base could be different for various applications. It is noteworthy that the system can be applied in various communication scenarios, especially those with resource constraints, and is not limited to IoT.

A. System Model

We consider an E2E semantic communication system constructed by a neural network architecture as shown in Fig. 2. Specifically, the transmitter consists of a semantic encoder to extract the semantic features from the source data, a semantic compression model to compress the semantics to reduce the amount of the transmitted data based on semantic importance, and a channel encoder to generate symbols to facilitate the transmission subsequently. The receiver is composed of a channel decoder for symbol detection and a semantic decoder with the output of semantic concepts with respect to the tasks.

At semantic transmitter, neural networks are first utilized to extract the semantic information from source data $I$, which can be denoted by

$$A = S_o(I),$$

where $S_o(\cdot)$ denotes the semantic encoder network with parameter $\alpha$. The extracted semantics can be a series of semantic features, which is hard to understand due to the deep neural networks’ poor explainability.

Then, the semantic features are compressed by

$$X = C_o(A),$$

where $C_o(\cdot)$ denotes ASC, and $o$ is the compression ratio. Semantic compression possesses the capability to further adaptively condense the extracted semantics, offering two significant advantages. First, it alleviates the demands on subsequent computing resources. Second, it substantially diminishes the volume of transmitted data, effectively reducing the strain on communication resources and minimizing transmission delay. Inevitably, semantic compression leads to a decline in task performance, and thus how to find the right compression ratio to achieve an optimal trade-off between transmission costs and semantic correctness is the most critical issue in the wireless resource allocation of semantic communication. The compression procedure is illustrated in the following definition.

Definition 1: To unify the ASC expressions for various semantic communication systems, we define the process of ASC as

$$X^k = \begin{cases} A^k, & \omega^c_k \geq \omega_0 \\ \emptyset, & \omega^c_k < \omega_0, \end{cases}$$

where $A^k$ is the $k$-th semantic feature and $\omega^c_k$ is the importance weight of $k$-th feature for semantic concept $c$, which will be detailed in Section III-B. $\omega_0$ is the importance weight threshold.

Equation (3) indicates that if a semantic feature’s importance weight exceeds the threshold, it will be transmitted; otherwise, it will be discarded. In order to keep the input size of receiver consistent with the output size of transmitter, the receiver will add zero to the discarded position, which is shared via knowledge base. Existing semantic communication systems can be regarded as a special case when $\omega_0 = 0$. It is worth noting that the main difference between the proposed semantic communication method and existing semantic communication systems lies in the adaptability to compress the extracted semantics in response to resource constraints.
Next, the compressed semantics is encoded by channel encoder to generate symbols for transmission, which can be denoted by

\[ M = Q_\sigma(X), \]  

where \( Q_\sigma(\cdot) \) denotes the channel encoder network with parameter \( \sigma \).

Then, the encoded symbols are transmitted via a wireless channel, and the received signal is expressed as

\[ Y = hM + n, \]  

where \( h \) represents the Rayleigh fading channel with \( CN(0, 1) \), and \( n \) is a vector sampled from additive white Gaussian noise.

We consider the transmitted symbols to be mapped into bits by binary quantization, and thus transmission in the physical layer still follows Shannon’s classic information theory, and the transmission rate of user \( i \) is

\[ R_i = B_i \log_2 \left( 1 + \frac{|h_i|P_i}{N_0B_i} \right), \]  

where \( B_i \) is the bandwidth of user \( i \), \( P_i \) is the transmission power of user \( i \), \( |h_i| \) is the channel gain between user \( i \) and edge server, and \( N_0 \) is the noise power spectral density.

Without loss of generality, we denote initial data size of semantic information that users extract as \( d^0 \), and semantic compression ratio of user \( i \) as \( o_i \). Therefore, the amount of data transmitted by user \( i \) is \( d_i = d_0 \times (1 - o_i) \), and the transmission delay of user \( i \) is

\[ t_i = \frac{d_i}{R_i}. \]  

In actual scenarios (e.g., Internet of Vehicles (IoV)), a large number of tasks are latency-sensitive and thus there is always a strict transmission delay constraint, which can be denoted by \( t_0 \). Thus, the success transmission probability of user \( i \) is \( P(t_i \leq t_0) \). To calculate \( P(t_i \leq t_0) \), we have the following lemma.

**Lemma 1:** The success transmission probability of user \( i \) is

\[ P(t_i \leq t_0) = 2Q \left( \frac{2^{o_i(1-o_i)} - 1}{b_i \delta} \right) \]  

where \( a_i = \frac{d_0}{B_i t_0}, b_i = \frac{P_i}{N_0 \delta B_i}, \) and \( \delta^2 \) is variance of the channel gain. The \( Q \)-function is the tail distribution function of the normal distribution.

**Proof:** Based on (6) and (7), we have

\[ P(t_i \leq t_0) = P \left( \frac{(1 - o_i)d_0}{B_i \log_2 \left( 1 + \frac{|h_i|P_i}{N_0B_i} \right)} \leq t_0 \right) \]

\[ = P \left( \frac{2^{o_i(1-o_i)} - 1}{b_i} \right) \leq |h_i| \]

\[ = 2Q \left( \frac{2^{o_i(1-o_i)} - 1}{b_i \delta} \right), \]  

where the last equality follows from \( h_i \sim N(0, \delta^2) \). This completes the proof of Lemma 1.

**Remark 1:** As we observe from Lemma 1, the success transmission probability is mainly affected by power, bandwidth, and semantic compression ratio. Therefore, the success transmission probability can be improved by optimizing the semantic compression ratio and resource allocation.

Then, received symbols are decoded to recover semantics via channel decoder, which can be expressed as

\[ X' = Q_\chi^{-1}(Y), \]  

where \( Q_\chi^{-1}(\cdot) \) denotes the channel decoder network with parameter \( \chi \).

Finally, the semantic receiver inputs the recovered semantics \( X' \) into semantic decoder to complete the intelligent tasks. Specifically, the output is

\[ p = S_\mu^{-1}(X'), \]  

where \( p \) is the task result, which will be returned to the transmitter and \( S_\mu^{-1}(\cdot) \) denotes the semantic decoder with the parameter \( \mu \). Note that the parameters of semantic codec are not unique, and they depend on the training dataset and the initial settings during the training process.

In task-oriented semantic communications, conventional communication metrics that ignore the underlying meaning of the source are no longer applicable, and thus new performance metrics need to be investigated at the semantic level. To simultaneously evaluate the impact of transmission and ASC on the performance of semantic communications, we define a novel metric, namely the success probability of tasks, which is expressed in the following definition.
Definition 2: Task success depends on two parts: one is successful transmission, and the other is successful understanding at receiver. Therefore, the success probability of tasks of \( i \)-th user can be expressed as

\[
\Phi_i = \eta(o_i) \times P(t_i \leq t_0)
\]  

(12)

where \( \eta(o_i) \) is the probability that a task is successfully executed under success transmission, while the compression ratio is \( o_i \).

From Eq. (12), we see that the proposed success probability of tasks used to evaluate the semantic communication performance can control the tradeoff between the semantic transmission and the semantic understanding. To facilitate this, we have to investigate the intelligent task performance model \( \eta(o) \) first, which draws the relationship between success probability of task and compression ratio \( o \) under success transmission. However, deriving a closed-form expression for \( \eta(o) \) is intractable due to the unexplainability of neural networks. Fortunately, we find the relationship between the semantic compression ratio and task performance by approximating a function to the statistics of the model evaluation. Note that, in practice, training the semantic communication model is executed in the server, thus calculating such function using a numerical approach is possible.

To obtain a point set \( D \) of \( D \) points reflecting the mapping between task performance \( \eta \) and compression ratio \( o \), we can remove the unimportant feature maps in turn and calculate the corresponding task performance and compression ratio. Inspired by the ideas in [27], the points of \( D \) can be estimated by an exponential function, which can be modeled as

\[
\eta(o) = \zeta_1 e^{\zeta_2 o} + \zeta_3 e^{\zeta_4 o}.
\]  

(13)

Then, we learn the parameters \( \zeta = [\zeta_1, \zeta_2, \zeta_3, \zeta_4] \) via a Levenberg-Marquardt method, which vary with the adopted neural networks. Based on this, we can obtain the relationship between semantic compression and task performance statistically. It is worth mentioning that once the training of the semantic communication network is completed, the task performance model calculation only need to be performed once.

We consider service level agreement (SLA), where users are prioritized with different service levels according to their objectives. For example, in the IoV scenario, users who perform pedestrian detection are typically characterized by a higher service level due to the requirements of high-reliable and low-latency communication [28]. Note that SLA is a general method existing in most modern cellular systems, such as 5G [28]. Consider a set \( N \) of \( N \) service levels according to users’ tasks. Let the weight of service level \( n \) be \( \varepsilon_n \), which can be obtained considering the quality of service (QoS) requirements (e.g., rate and delay) [29]. Let \( r_{in} \in \{0, 1\} \) denote the user association index, i.e., \( r_{in} = 1 \) means that the user \( i \) belongs to service level \( n \); otherwise, we have \( r_{in} = 0 \). Assuming that a user can execute one service at a time [30], and thus the importance weight of user \( i \) is expressed as

\[
w_i = \sum_{n=1}^{N} \varepsilon_n r_{in}.
\]  

(14)

From (12), we can observe that the importance of a user may vary depending on the service it performs.

Considering SLA, the weighted sum success probability of tasks of the whole semantic communication system is expressed as

\[
\Phi = \sum_{i=1}^{U} \beta_i w_i \Phi_i
\]  

(15)

where \( \beta_i \in \{0, 1\} \) denotes the user selection index, i.e., \( \beta_i = 1 \) indicates user \( i \) is selected; otherwise, we have \( \beta_i = 0 \). It is necessary to ensure the performance of users with higher priority, and thus only a subset of users may be selected to complete intelligent tasks due to the limited wireless resources. Based on (12) and (15), we can conclude that the performance of the multi-user semantic communication system is mainly affected by resource allocation, compression ratios, and user selection.

B. Problem Formulation

We aim to optimize the resource allocation, compression ratios, and user selection simultaneously to maximize the weighted sum success probability of tasks of the semantic system under the resource constraints. Mathematically, we formulate the optimization problem as

\[
\max_{B, P, o, \beta} \Phi
\]  

(16)

s.t. \[
\begin{align}
\beta_i B_i & \leq B_{\text{max}}, \quad (16a) \\
B_i & \geq 0, \forall i \in U, \quad (16b) \\
\sum_{i=1}^{U} \beta_i P_i & \leq P_{\text{max}}, \quad (16c) \\
P_i & \geq 0, \forall i \in U, \quad (16d) \\
0 < o_i & < 1, \forall i \in U, \quad (16e) \\
\beta_i & \in \{0, 1\}, \forall i \in U, \quad (16f) \\
\sum_{n} r_{in} & = 1, \forall i \in U. \quad (16g)
\end{align}
\]

where \( \beta = [\beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_U], U \) is the set of selected users, \( B_{\text{min}} \) is the minimum bandwidth allocated to users, \( B_{\text{max}} \) is the maximum total bandwidth, \( P_{\text{min}} \) is the minimum transmit power allocated to users, and \( P_{\text{max}} \) is the maximum total transmit power. Constraint (16a) indicates that the sum bandwidth of selected users cannot exceed the overall system bandwidth. Constraint (16c) indicates that the sum transmit power of selected users cannot exceed a given value, which guarantees that the energy consumption of the whole system is limited. Constraint (16e) is the compression ratio constraint. Constraint (16g) indicates that each user can only perform one service at a time.

III. E2E SEMANTIC COMMUNICATION AND ASC

In this section, we first design a loss function and introduce the training process, which are the foundation of the E2E semantic communication. Next, we illustrate the method to
evaluate the importance of semantic features, which is the basis of the proposed ASC approach.

**A. Loss Function Design**

The main goal of designing a task-oriented semantic communication system is to maximize the intelligent task performance and the capacity or the data transmission rate simultaneously. Compared with the bit error rate, the mutual information can provide extra information to train a transceiver [16]. The mutual information of the transmitted semantics, $X$, and the received semantics, $Y$, can be computed by

$$I(X; Y) = E_{p(x, y)} \left[ \log \left( \frac{p(x, y)}{p(x)p(y)} \right) \right]$$

$$= E_{p(x,y)}[\log p(y|x) - \log p(y)], \quad (17)$$

where $(X, Y)$ is a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the spaces for $X$ and $Y$. $p(X)$ and $p(y)$ are the marginal probability of sent $X$ and received $Y$, respectively, and $p(x, y)$ is the joint probability of $X$ and $Y$. To effectively estimate the mutual information of $X$ and $Y$, the following theorem is presented.

**Theorem 1:** The upper bound of the mutual information can be expressed as

$$I_{up}(X; Y) := E_{p(x, y)}[\log p(y|x)] - E_{p(x)}E_{p(y)}[\log p(y|x)] \quad (18)$$

**Proof:** In order to prove that $I_{up}(X; Y)$ is the upper bound of mutual information, it is only necessary to prove that $I_{up}(X; Y)$ is greater than the true mutual information $I(X; Y)$. The gap between the true mutual information and the upper bound can be denoted by

$$\Delta := I_{up}(X; Y) - I(X; Y)$$

$$= E_{p(x,y)}[\log p(y|x)] - E_{p(x)}E_{p(y)}[\log p(y|x)] - E_{p(x,y)}[\log p(y|x)] + E_{p(x)}E_{p(y)}[\log p(y|x)]$$

$$= E_{p(x)}E_{p(y)}[\log p(y)] - E_{p(x)}E_{p(y)}[\log p(y|x)]$$

$$= E_{p(x)}\left[ \log E_{p(y)}[p(y|x)] \right] - E_{p(x)}[\log p(y|x)] \geq 0 \quad (19)$$

The last step is derived from Jensen’s inequality. This completes the proof of Theorem 1.

However, the conditional relation $p(y|x)$ between variables in Theorem 1 is unavailable, and a variational distribution $q_\theta(y|x)$ with parameter $\theta$ is used to approximate $p(y|x)$. According to the [31, Th. 3.2], minimizing the upper bound on mutual information is equivalent to minimizing $-E_{p(x,y)}[\log q_\theta(y|x)]$. With samples $\{(x_i, y_i)\}_{i=1}^L$, we can minimize the log-likelihood function $L_{\text{ML1}}(\theta) := -\frac{1}{L} \sum_{i=1}^L \log q_\theta(y_i|x_i)$, which is the unbiased estimation of $-E_{p(x,y)}[\log q_\theta(y|x)]$. In this paper, the variational distribution $q_\theta(y|x)$ is implemented with neural networks and minimized via gradient-descent method.

The parameters of the semantic communication network are optimized via the following loss function

$$L(y_l; p, \alpha, \mu) = L_{\text{T1}}(y_l, p) - \kappa I_{up}(X; Y) \quad (20)$$

where $y_l$ is the task label. $\alpha$ and $\mu$ are the parameters of the semantic encoder and semantic decoder, respectively. The first term $L_{\text{T1}}(y_l, p)$ is the loss function related to the task (i.e., cross-entropy for classification task, triplet loss for object detection task, etc.), which aims to maximize the task performance by training the whole system. The second one $I_{up}(X; Y)$ is the estimation of mutual information between $X$ and $Y$, which maximizes the achieved data rate during the transmitter training. Parameter $\kappa$, between 0 and 1, is the weight for the mutual information.

The training process of the proposed task-oriented semantic communication network consists of two phases due to the two parts of loss functions. After initializing the parameters, the first phase is to train the mutual information model by unsupervised learning to estimate the achieved data rate for the second phase. The second phase is to train the whole system with (20) as the loss function. Each phase aims to minimize the loss by gradient descent with mini-batch until the stop criterion is met, the max number of iterations is reached, or none of the terms in the loss function is decreased anymore. Based on the trained semantic communication system, we next introduce how to obtain the semantic importance weights used for ASC. Notably, ASC is only performed during inference and based on the offline trained semantic communication network, which is energy efficient and is compatible with existing semantic communication systems without further retraining.

**B. Semantic Importance Evaluation**

In task-oriented semantic communications, different semantic features may be of different importance for completing intelligent tasks, and thus there are still semantic redundancies that are irrelevant to the intelligent tasks, which can be further compressed [32]. Here, the importance of semantics is defined as the correlation between semantics and the task. The way to measure the importance of semantic features can vary with different semantic communication systems, and here we employ a gradient-based approach. Based on the semantic communication system trained in Section III-A, we first compute the gradient of the activation value for semantic concept $c$ (such as objects, properties, and actions) [33], $\partial y^c / \partial A^k_j$ (before Softmax layer), with respect to $k$-th semantic feature activations $A^k_j$, i.e., $\frac{\partial y^c}{\partial A^k_j}$. These gradients flowing back are global-average-pooled over the width and height dimensions (indexed by $i$ and $j$ respectively) to obtain the semantic importance weights

$$\omega^c_k = \frac{1}{W \times H} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_j} \quad (21)$$

where $W$ and $H$ are the width and height of $A^k$, and $A^k_{ij}$ is the activation value at the $i$-th row and the $j$-th column of the feature map. During computation of $\omega^c_k$ while backpropagating gradients with respect to activations, the exact computation amounts to successive matrix products of the weight matrices and the gradient with respect to activation functions till the final convolution layer that the gradients are being propagated to. Hence, this weight $\omega^c_k$ represents a partial linearization of the deep network downstream from $A$, and captures the

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
‘semantic importance’ of semantic feature \( k \) for a semantic concept \( c \) [34]. Based on this, we can verify that different semantic features are of different importance for completing intelligent tasks, which can guide semantic compression.

Since the importance weights are only related to network parameters, these weights can be regarded as shared knowledge and stored in the knowledge base of the sender and receiver, where the knowledge base could be different for various tasks. Consequently, there is no need to transmit the indices corresponding to the transmitted feature maps in the subsequent semantic communication process. In this work, we only calculate the semantic importance weights via a gradient-based method. One can easily extend the proposed ASC method to other calculation methods such as attention-based mechanisms [7]. Based on the obtained semantic importance weights, ASC can be performed consequently.

**IV. CRRA ALGORITHM**

Based on the above analysis and results, we then focus on solving the problem (16) in the following two sections. In this section, we consider the scenarios where users perform the tasks with similar priority (e.g., pedestrian and vehicle detection in IoV), and thus users have the same service levels, and all of them should be selected, i.e., \( w_i = 1, \beta_i = 1, \forall i \in \mathcal{U} \).

In the considered scenarios, the optimization problem (16) is first simplified to a resource allocation and compression ratios optimization problem to maximize the total success probability of tasks. Then, the CRRA algorithm is proposed to solve the optimization problem.

Based on the approximation of \( Q \)-function \( Q(x) \approx \frac{1}{2} e^{-\frac{x^2}{2}} \) [35], problem (16) can be reformulated as

\[
\max_{B,P,\alpha} \sum_{i=1}^{U} g_i \times \eta(o_i) \\
\text{s.t. } (16a) - (16f),
\]

where \( g_i = \exp\{-\frac{1}{2}\left[-\frac{N_0 B_i (v_i - 1)}{\delta P_i}\right]^2\} \) and \( v_i = 2^\left[-\frac{\delta_0 (1 - o_i)}{\delta_0} \right] \).

Since the objective function is not concave, the total success probability of tasks maximization problem (22) is non-convex, hence, it is generally hard to optimize resource allocation and compression ratios directly. To solve the problem (22), we divide it into two subproblems and then solve these two subproblems iteratively. In particular, we first fix the resource allocation and calculate the optimal compression ratio for each user. Then, resource allocation problem is formulated and solved with the obtained compression ratios. The two subproblems are iteratively solved until a convergent solution is obtained.

### A. Optimal Compression Ratios

Given the resource allocation policy, (22) can be simplified as

\[
\max_{\alpha} \sum_{i=1}^{U} g_i \times \eta(o_i) \\
\text{s.t. } 0 < o_i < 1, \forall i \in \mathcal{U}.
\]

We can observe from (23) that if the resource allocation policy is fixed, the optimal semantic compression ratio of each user is independent. Thus, our goal transforms into maximizing each user’s success probability of tasks. For user \( i \), the problem is

\[
\max_{o_i} g_i \times \eta(o_i) \\
\text{s.t. } 0 < o_i < 1, \forall i \in \mathcal{U}.
\]

Consider the range of \( o_i \) is within 0 and 1, we here employ the one-dimension enumeration method to obtain the optimal semantic compression ratio.

### B. Optimal Resource Allocation

With the obtained semantic compression ratios, we then optimize the bandwidth and power of the considered semantic communication systems. Note that given \( o_i, \eta(o_i) \) can be seen as a constant, which is denoted by \( o_i \). Thus, the resource allocation problem can be reformulated as

\[
\min_{B,P} \sum_{i=1}^{U} -\alpha_i \times g_i \\
\text{s.t. } (16a) - (16d).
\]

To solve problem (25), we first convert the non-convex problem into a convex optimization problem. In particular, by introducing slack variables \( f = [f_1, f_2, \ldots, f_U], l = [l_1, l_2, \ldots, l_U], x = [x_1, x_2, \ldots, x_U], m = [m_1, m_2, \ldots, m_U] \) and \( q = [q_1, q_2, \ldots, q_U] \), problem (25) can be transformed into

\[
\min_{B,P,f,l,x,m,q} \sum_{i=1}^{U} -\alpha_i \times f_i \\
\text{s.t. } f_i \leq e^{h_i}, \forall i \in \mathcal{U},
\]

where \( h_i = \frac{1}{2} \sum_{j=1}^{q_i} l_{ij}^2, \forall i \in \mathcal{U} \).

### (26)

\[
x_i \leq \frac{N_0 B_i m_i}{\delta P_i}, \forall i \in \mathcal{U},
\]

\[
m_i \geq 2q_i - 1, \forall i \in \mathcal{U},
\]

\[
q_i \geq \frac{d_0 (1 - \sigma)}{B_i t_i}, \forall i \in \mathcal{U},
\]

\[
(16a) - (16d).
\]

However, constraints (26a) and (26e) are still non-convex. For constraint (26c), we use the successive convex approximation (SCA) method to transform it into a convex constraint. Performing a first-order Taylor expansion of \( e^{h_i} \) at \( e^{h_i} \), then we have

\[
f_i \leq e^{h_i} + \left( l_i - l_i^j \right) e^{h_i},
\]

while the superscript \( j \) represents the value obtained after \( j \)-th iteration of the variable.

For constraint (26c), slack variable \( z_i = [z_1, z_2, \ldots, z_U] \) is introduced, and have

\[
z_i \geq B_i m_i.
\]
Thus, constraint (26c) can be transformed into

\[ x_i P_i \geq \frac{N_0 z_i}{\delta_i}. \] (29)

Equation (28) can be rewritten as

\[ z_i \geq B_i m_i = \frac{1}{4} \left( (B_i + m_i)^2 - (B_i - m_i)^2 \right) \] (30)

By performing a first-order Taylor expansion of \((B_i - m_i)^2\) at point \((B_i^0, m_i^0)\) and using SCA, we have

\[ z_i \geq \frac{1}{4} ((B_i + m_i)^2 - 2(B_i - m_i)(B_i^j - m_i^j) + (B_i^j - m_i^j)^2). \] (31)

Similarly, (29) is equivalent to

\[ x_i P_i = \frac{1}{4} \left( (x_i + P_i)^2 - (x_i - P_i)^2 \right) \geq \frac{N_0 z_i}{\delta_i}. \] (32)

By performing a first-order Taylor expansion of \((x_i + P_i)^2\) and \((x_i - P_i)^2\) at point \((x_i^j, P_i^j)\) and using SCA, we can obtain

\[ 4 N_0 z_i \delta_i \leq 2 (x_i + P_i) \times (x_i^j + P_i^j) - (x_i^j + P_i^j)^2 \]

\[ - 2 (x_i - P_i) \times (x_i^j - P_i^j) + (x_i^j + P_i^j)^2. \] (33)

So far, all constraints are transformed into convex, and the optimization problem can be reformulated as

\[
\min_{B, P, f, l, e, m, q, z} \sum_{i=1}^{U} -\alpha_i \times f_i \\
\text{s.t.} \quad f_i \leq e^l_i + \left( l_i - \frac{P_i}{l_i} \right) e^U_i, \forall i \in U, \quad (34a) \\
\quad l_i \leq \frac{1}{2} z_i^2, \forall i \in U, \quad (34b) \\
\quad m_i \geq 2^{q_i} - 1, \forall i \in U, \quad (34c) \\
\quad q_i \geq d_0 (1 - \sigma) + \frac{B_i l_i}{B_i l_i}, \forall i \in U, \quad (16a) - (16d), \quad (31), \quad (33). \quad (34d)
\]

The slack variables \(f, l, x, m, q, z\) have been introduced to reformulate the optimization problem into a convex form. They are closely associated with the original problem but generally lack specific physical interpretations.

Problem (34) is a convex optimization problem, and can be solved via the dual method [36]. Optimal results can be obtained by setting the initial value of \(l_i^0, B_i^0, m_i^0, z_i^0\) and \(P_i^0\), updating variables, and performing iterations until the problem converges.

Finally, we can iteratively solve (23) and (34) until a convergent solution is obtained.

C. Convergence and Complexity Analysis of CRRA

This subsection analyzes the convergence and computational complexities of CRRA. The convergence of CRRA mainly depends on the resource allocation subproblem, while the first subproblem is solved by the one-dimension enumeration method. Thus, we focus on analyzing the convergence of solving (34), which is illustrated by the following lemma.

**Lemma 2:** The total success probability of task is monotonically non-decreasing, and the sequence \((B^{(n)}, P^{(n)})\) converges to a point fulfilling the KKT optimal conditions of the original non-convex problem (23).

**Proof:** Since Lemma 2 directly follows from [37, Proposition 3], the proof of Lemma 2 is omitted.

The major complexity in each iteration lies in solving the semantic compression ratios subproblem (23) and the resource allocation subproblem (34). First, subproblem (23) is addressed using a one-dimensional enumeration method. The complexity of the enumeration method for each user is constant and independent of the number of users. However, since we perform this enumeration for each of the \(U\) users, the overall complexity for this subproblem scales linearly with the number of users. Additionally, considering the number of enumerations denoted as \(K\), the complexity becomes proportional to the product of \(K\) and \(U\). Therefore, the complexity for this subproblem is \(O(KU)\). Second, the complexity of solving (34) is \(O(U^{3.5})\) [38]. As a result, the total complexity of CRRA is given by \(O(T_0 KU + T_0 U^{3.5})\), where \(T_0\) is the number of iterations in CRRA.

V. CRRAUS ALGORITHM

In this section, we consider the scenarios where users perform the tasks with different priorities (e.g., face and fire detection in the smart factory), and thus users have various service levels, and only part of them can be selected due to the wireless resource constraints. To deal with the problem (16), the CRRAUS algorithm is proposed, which is able to adaptively adjust the user selection based on the wireless resources and service levels. Then, the convergence and complexity of CRRAUS are analyzed.

A. Algorithm Design

It is hard to obtain the optimal solutions to the problem (16) due to non-concave objective function and nonconvex constraints. To obtain a suboptimal solution to problem (16), we propose a CRRAUS algorithm, in which problem (16) is separated into three subproblems and solved iteratively. In particular, we first fix the resource allocation and user selection scheme to calculate the optimal compression ratio for each user. Second, we fix the resource allocation and compression ratios to solve the optimal user selection scheme. Finally, the problem of resource allocation is formulated and solved with the obtained compression ratios and user selection scheme.

Similar to (23) and (24), the first subproblem can be simplified into maximizing each user’s weighted success probability of tasks by optimizing the semantic compression ratio

\[
\max_{\alpha_i} \lambda_i \Phi_i \\
\text{s.t.} \quad 0 \leq \alpha_i \leq 1, \quad (35a)
\]

where \(\lambda_i = \beta_i w_i\) is a constant only relevant to user \(i\). Problem (35) can also be solved by the one-dimension enumeration method, and the solution process is omitted.
Algorithm 1 User Selection With Branch and Bound Method

1: Find the optimal solution to the linear programming problem with the 0-1 integer restrictions relaxed.

2: At node 1, let the relaxed solution be the upper bound \( Q^U \) and the rounded-down integer solution be the lower bound \( Q^L \). Select the variable with the greatest fractional part for branching.

3: Create two new nodes, one is for the \( \beta_j = 0 \) and the other is for the \( \beta_j = 1 \).

4: Solve the relaxed linear programming problem with the new constraint added at each of these nodes, and obtain the relaxed solution \( Q \) and \( \beta \).

5: Let the upper bound \( Q^U = Q \) at each node, and the existing maximum integer solution \( Q^L \) is the lower bound.

6: if All elements in \( \beta \) are integers then

7: The optimal integer solution \( \beta^* = \beta \).

8: else

9: Branch from the node with the greatest upper bound and return to step 3.

10: end if

Given the resource allocation policy and compression ratios, the user selection subproblem can be simplified as

\[
\max_{\beta} \sum_{i=1}^{U} c_i \beta_i \\
\text{s.t.} \quad (16a), (16c), (16f), \tag{36}
\]

where \( c_i = w_i \Phi_a \) can be regarded as a constant only related to user \( i \). Problem (36) is a 0-1 integer programming problem, which can be solved by branch and bound method [39]. The algorithm for user selection is summarized in Algorithm 1.

With the obtained compression ratios and the user selection scheme, the resource allocation subproblem can be reformulated as

\[
\max_{B,P} \sum_{i=1}^{U} \mu_i g_i \\
\text{s.t.} \quad (16a) - (16d), \tag{37}
\]

where \( \mu_i = \beta_i w_i t(\theta) \) is a constant independent of the resource allocation scheme. Problem (37) has the same form as the problem (25), and both of them are non-convex and have linear constraints. Therefore, we can also use the SCA approach to transform (37) into an approximated convex problem and solve it via the dual method. The detailed solution process is omitted here.

Finally, the three subproblems are iteratively solved until a convergent solution is obtained.

B. Convergence and Complexity Analysis of CRRAUS

This subsection analyzes the convergence and computational complexities of CRRAUS.

Since the convergence analysis of CRRAUS is similar to that of CRRR, the detailed analysis is omitted. The computational complexity of CRRAUS consists of three parts. The first part is for solving the problem (35) by one-dimension enumeration method, the second part is for solving 0-1 integer programming problem (36) via branch and bound method, and the third part is for solving the problem (37) by SCA. The complexity of solving problem (35) is \( O(KU) \). The complexities of solving problems (36) and (37) are \( O(U^3) \) and \( O(U^{3.5}) \), respectively. Therefore, the total complexity of CRRAUS is \( O((U^3 + U^{3.5} + KU)T_1) \), where \( T_1 \) is the number of iterations. CRRAUS completes user selection at the expense of higher algorithm complexity than CRRR. In addition, it can be observed that the complexity of the two algorithms increases sharply with the increase in the number of users, which can be further optimized in future work.

VI. SIMULATION RESULTS AND ANALYSIS

In our simulation, a circular network is considered with one edge server and \( U = 10 \) users. Unless specifically stated, the simulation parameters are listed in Table I. In the experiments, we take the image classification task as an example to illustrate. STL-10 dataset [40] is used as training and testing data, which contains images of 10 categories of objects, corresponding to 10 semantic concepts. It contains 5,000 training and 8,000 testing images, each captured from natural scenes. STL-10 dataset is well-suited for tasks like image classification and object recognition. To verify the applicability of the proposed semantic communication system to different neural networks, experiments are conducted based on the backbone of EfficientNet [41] and Resnet [42] networks. The simulations are based on the assumption of perfect CSI. However, in order to validate the performance of semantic communication in the presence of imperfect channel estimation, we also conducted an experiment with imperfect CSI. The hyperparameters during training are listed in Table I. The parameters of the intelligent task performance model proposed.

| Simulation Parameter | Value |
|----------------------|-------|
| Initial data size, \( d_0 \) | 24.5 MB |
| Delay constraint of users, \( t_0 \) | 1-10 ms |
| Noise power spectral density, \( N_0 \) | -174 dBm/Hz |
| Minimum bandwidth, \( B_{\min} \) | 0.01 MHz |
| Minimum transmit power, \( P_{\min} \) | -20dBm |
| The number of users, \( U \) | 10 |
| Compression ratio, \( \sigma \) | 0-1 |
| Maximum bandwidth, \( B_{\max} \) | 1 MHz-30 MHz |
| Maximum transmit power, \( P_{\max} \) | 1 mW-1 W |
| Weights of service levels, \( \varepsilon \) | [0.2,0.4,0.6,0.8] |

| Hyper Parameter | Value |
|-----------------|-------|
| Epoch | 50 |
| Batchsize | 32 |
| Optimizer | Adam |
| Learning rate | 0.01 |
| Momentum | 0.9 |
| \( \kappa \) | 10^{-3} |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
in system model are listed in Table II. The simulations and experiments are performed by the computer with Ubuntu 16.04 + CUDA 11.0, and the selected deep learning framework is Pytorch.

In the following, the proposed semantic communication system with ASC (labeled as “ASC”) is first compared with the standard semantic communication method [19] (labeled as “SSC”) and the traditional communication method (labeled as “TCM”). In the standard semantic communication method, semantics is extracted and then directly transmitted without further compression. The traditional communication method use JPEG for image compression followed by LDPC channel coding, and then reconstruct image at receiver to complete the task. Then, we compare the proposed CRRA and CRRAUS algorithms with three baselines: resource allocation scheme with fixed compression ratios (labeled as “FCR”), compression ratios optimization scheme with fixed resource allocation (labeled as “FRA”), and conventional resource allocation scheme to maximize the system sum-rate [43] (labeled as “MSR”). The computational complexity of FCR and FRA is $O(U^{3.5})$ and $O(KU)$, respectively. In accordance with [43], the complexity of MSR can be expressed as $O(U^2)$.

Fig. 3 shows the classification accuracy versus channel SNR under different communication systems, where Fig. 3(a) is conducted based on the backbone of EfficientNet and Fig. 3(b) is conducted based on the backbone of Resnet. “ASC=65%” and “ASC=80%” refer to the user employing the proposed ASC with the semantic compression ratio equaling to 65% and 80%, respectively. “SSC imperfect CSI” refers to the semantic communication under the imperfect CSI scenario, and the CSI estimation error follows $CN(0,0.025)$ in line with [22]. As shown in Fig. 3, the performance of DL-based semantic communications is much better than that of TCM, especially in low SNR regimes. Even when using imperfect CSI, the semantic communication system still outperforms TCM with slight performance degradation, which verifies its high robustness to physical transmission. In addition, compared with TCM, SSC can reduce the amount of transmitted data by up to 99%, which means that semantic communication can save communication resources more than 100 times, at the expense of being unable to reconstruct the source. Moreover, “ASC=65%” and “ASC=80%” only necessitate the transmission of 35% and 20% of the data compared to SSC while incurring slight loss in classification performance, particularly in low SNR regions. With the increase of SNR, the performance loss becomes negligible even when the compression ratio is 80%. It is worth noting that this comparison is grounded in the assumption of abundant resources. However, when communication resources become limited, and stringent delay constraints are considered, transmitting the complete set of semantics becomes impractical, especially considering the potential size of the semantics due to the neural network-based semantic encoder. The key advantage of the proposed ASC approach lies in its ability to adaptively curtail the volume of transmitted data while maintaining task performance, making it particularly well-suited for resource-constrained IoT scenarios.

Figs. 4 and 5 illustrate the average success probability of tasks versus different compression ratios under different maximum transmit power and different maximum bandwidth, respectively. It should be noted that the system is equivalent to SSC when abscissa equals to 0. However, the communication performance is very poor due to the limited resources, which verifies the necessary of adaptively adjusting the compression ratio. Besides, we can observe that the average success
probability of tasks increases first and then decreases as the compression ratio increasing. This is because the choice of compression ratios is a trade-off between communication transmission and task performance, reflecting that choosing the optimal compression ratios is highly significant for semantic communications. When the maximum transmit power and maximum bandwidth change, it can be observed that the optimal compression ratios are variable. For example, when the maximum bandwidth is 10 MHz, the optimal compression ratio is 0.9, and when the maximum bandwidth is 20 MHz, the optimal compression ratio is 0.8. This verifies that resources will affect the optimal compression ratios and the necessity of optimizing the compression ratios. In addition, the maximum bandwidth and the maximum transmit power of subsequent simulations are set to 20 MHz and 1 W, respectively, and thus the fixed compression ratio of FCR in subsequent simulations is set to 0.8 for a fair comparison.

Fig. 6 illustrates the average success probability of tasks versus the number of iterations at various maximum bandwidth settings. We set $P_{\text{max}} = 1 \text{ mW}$, $U = 10$ and $\text{SNR} = -5 \text{ dB}$ in Fig. 6. It can be observed that the convergent solution is achieved within 5 iterations at each bandwidth setting, which verifies the convergence and efficiency of the proposed algorithm. Note that for a fixed user set, the convergence of CRRAUS is similar to that of CRRA. Therefore, for the rest of simulations in both CRRA and CRRAUS, we set the number of iterations $T_0$ and $T_1$ to 5.

The average success probability of tasks versus the maximum bandwidth under different maximum transmit power is shown in Fig. 7. As shown in this figure, the average success probability of tasks increases with the maximum bandwidth and gradually converges to a certain threshold. This is because large bandwidth can decrease the transmission delay and tolerate a small semantic compression ratio, which consequently increases the probability of successful transmission and the average success probability of tasks. In addition, it can be observed that the proposed CRRA can improve 14.3%, 43.1% and 2-fold average success probability of tasks compared to baselines FRA, FCR, MSR, respectively, when maximum bandwidth is set to 10 MHz. The superior performance of the proposed algorithm is attributed to its capability to jointly optimize resource allocation and compression ratios. This
Fig. 8. Average success probability of tasks versus the maximum sum transmit power.

Fig. 9. Total success probability of tasks versus number of users.

Fig. 10. Weighted success probability of tasks versus the maximum bandwidth.

result underscores the value of this joint optimization, as it enables a more harmonious trade-off between compression and task performance. Besides, conventional resource allocation scheme MSR is found to be inadequate for semantic communication scenarios. This is because the conventional resource allocation scheme only optimizes the transmission rate and lacks the consideration of semantics and subsequent intelligent tasks. It is worth noting that these improvements come at the expense of increased computational complexity.

Fig. 8 shows the average success probability of tasks versus the maximum transmit power. As expected, the average success probability of tasks increases as the maximum transmit power. This is due to the fact that large transmit power can increase the transmission rate, consequently increasing the amount of transmitted semantics. Similar observations are achieved in Fig. 7, e.g., the proposed algorithm harvests significant performance gains compared with the baselines. The reason is that the proposed algorithm is able to simultaneously optimize compression ratios and resource allocation policy to improve the resource utilization. Furthermore, from Fig. 8, we can find that the MSR has little improvement in semantic performance. This is because the MSR method only focuses on technical performance, which may not transmit the semantic information required for intelligent tasks well. Remarkably, the proposed algorithm can perform well even in very low transmit power regions, which shows that our algorithm is very suitable for low-power IoT scenarios, emphasizing its versatility across various deployment contexts.

The total success probability of tasks versus the number of users is given in Fig. 9. Clearly, the proposed CRRA is always better than FCR, FRA, and MSR, especially when the number of users is large. This is because CRRA can effectively determine the compression ratios and the resource allocation scheme to meet the delay constraint, while FCR and FRA only take one of them into consideration. MSR has the worst performance because only maximizing the sum rate cannot guarantee an accurate understanding of semantic information. When the number of users is large, the multi-user gain is more apparent by the proposed CRRA compared to conventional FCR and FRA. This is because the resources are relatively tight when there are a large number of users, and CRRA can make full use of resources and find the optimal trade-off between compression and transmission. CRRA achieves better performance than FCR and FRA at the cost of additional computational complexity.

Fig. 10 shows how the weighted success probability of tasks changes as the maximum bandwidth, considering that users have different task priorities. This figure presents a comparison between the proposed CRRAUS, CRRA, and three other baseline approaches: “RAUS”, which involves a resource allocation and user selection scheme with fixed compression ratios, and “CRUS,” which incorporates compression ratio optimization and user selection with fixed resource allocation, and “MSR_RAUS”, which involves a resource allocation and user selection scheme to maximize the system sum-rate. From Fig. 10, we can see that the proposed CRRAUS improve 6.4%, 8.6%, 40.3%, and 1.5-fold weighted success probability of tasks compared to CRRA, CRUS, RAUS, and MSR_RAUS, respectively, when maximum bandwidth is set to 10 MHz.
This is because CRRAUS can simultaneously optimize the resource allocation, compression ratios, and user selection, while the CRRA, CRUS, and RAUS can only optimize two of them separately. Conventional resource allocation and user selection scheme MSR_RAUS yields the poorest performance. This deficiency arises from MSR_RAUS not factoring in task-oriented semantics and instead prioritizing high bit rates. Furthermore, it becomes apparent that CRRA surpasses CRUS, and CRUS outperforms RAUS. This finding offers valuable insights, particularly in resource-constrained situations where users possess varying levels of task priorities. It suggests that, in such contexts, optimizing the semantic compression ratio may yield more significant benefits compared to resource allocation and user selection alone.

The weighted success probability of tasks versus maximum transmit power is given in Fig. 11. The figure demonstrates that the weighted success probability of tasks increases across all schemes as maximum transmit power varies. This improvement is attributed to the augmented amount of transmitted semantics facilitated by higher transmit power, subsequently enhancing the performance of semantic communication. Consistent with the observations in Fig. 10, the proposed CRRAUS achieves the best performance under different maximum transmit power, while the conventional scheme delivers the worst performance. Through the comprehensive simulations, it is evident that optimizing resource allocation, semantic compression ratios, and user selection can significantly enhance the performance of semantic communication. This multifaceted optimization is crucial and cannot be overlooked in practical semantic communication scenarios. It is worth noting that CRRA may offer larger gains at the cost of compromising high-service-level users’ performance in scenarios characterized by diverse service levels. In modern communication environments, such as smart factories, the need to cater to varying service levels for distinct user tasks is a common occurrence. Therefore, the choice between these two algorithms should be made judiciously, taking into account the specific requirements and context of the application.

VII. Conclusion

In this paper, we first develop a task-oriented multiuser semantic communication system, in which an ASC approach is proposed to compress semantics to reduce the communication burden adaptively. Then, we have formulated a resource allocation and compression ratios optimization problem under bandwidth and power constraints to maximize the success probability of tasks, which is defined to measure the performance of semantic communications. For scenarios where users have the same service levels, we have proposed a CRRA algorithm to optimize resource allocation and compression ratios, where the nonconvex problem is decomposed into two subproblems and solved iteratively. Furthermore, considering that users have various service levels, a CRRAUS algorithm has been proposed, in which users are adaptively selected based on the branch and bound method. Simulation results have shown that the proposed ASC approach can significantly reduce the size of transmitted data, and both CRRA and CRRAUS algorithms achieve higher success probability of tasks than the benchmarks, especially when communication resources are tight. Compared with CRRA, the CRRAUS is more suitable for scenarios with significant differences in service levels at the expense of higher complexity. Future extensions of this work will delve into the development of low-complexity resource allocation schemes and explore a broader range of advanced intelligent tasks.

REFERENCES

[1] K. Chen, T. Zhang, R. D. Gitlin, and G. Fettweis, “Ultra-low latency mobile networking,” IEEE Netw., vol. 33, no. 2, pp. 181–187, Mar./Apr. 2018.
[2] C. Wang, Y. Xu, L. Xu, Z. Wang, and W. Wang, “Multimodal semantic communication accelerated bidirectional caching for 6G MEC,” Future Gener. Comput. Syst., vol. 140, pp. 225–237, Mar. 2023.
[3] C. She et al., “A tutorial on ultra-reliable and low-latency communications in 6G: Integrating domain knowledge into deep learning,” Proc. IEEE, vol. 109, no. 3, pp. 204–246, Mar. 2021.
[4] S. Yao, K. Niu, S. Wang, and J. Dai, “Semantic coding for text transmission: An iterative design,” IEEE Trans. Cogn. Commun. Netw., vol. 8, no. 4, pp. 1594–1603, Dec. 2022.
[5] Z. Qin, X. Tao, J. Lu, and G. Y. Li, “Semantic communications: Principles and challenges,” 2022, arXiv:2201.01389.
[6] D. Gündüz et al., “Beyond transmitting bits: Context, semantics, and task-oriented communications,” IEEE J. Sel. Areas Commun., vol. 41, no. 1, pp. 5–41, Jan. 2023.
[7] Y. Wang et al., “Performance optimization for semantic communications: An attention-based reinforcement learning approach,” IEEE J. Sel. Areas Commun., vol. 40, no. 9, pp. 2598–2613, Sep. 2022.
[8] Y. Zhang, H. Zhao, J. Wei, J. Zhang, M. F. Flanagan, and J. Xiong, “Context-based semantic communication via dynamic programming,” IEEE Trans. Cogn. Commun. Netw., vol. 8, no. 3, pp. 1453–1467, Sep. 2022.
[9] P. Kountouris and N. Pappas, “Semantics-empowered communication for networked intelligent systems,” IEEE Commun. Mag., vol. 59, no. 6, pp. 96–102, Jun. 2021.
[10] J. Chen et al., “Semantic communications with AI tasks,” 2021, arXiv:2109.14170.
[11] B. Guler, A. Yener, and A. Swami, “The semantic communication game,” IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 4, pp. 787–802, Sep. 2018.
[12] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, “Semantic-effectiveness filtering and control for post-5G wireless connectivity,” J. Indian Inst. Sci., vol. 100, pp. 435–443, May 2020.
[13] Z. An et al., “Series-constellation feature based blind modulation recognition for beyond 5G MIMO-OFDM systems with channel fading,” IEEE Trans. Cogn. Commun. Netw., vol. 8, no. 2, pp. 793–811, Jun. 2022.
