Decentralized Over-the-Air Computation for Edge AI Inference with Integrated Sensing and Communication

Zeming Zhuang*†‡, Dingzhu Wen*, and Yuanming Shi*
*School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
†Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, China
‡University of Chinese Academy of Sciences, Beijing 100049, China
E-mail: {zhuangzm, wendzh, shiym}@shanghaitech.edu.cn

***Abstract*—Collaborative artificial intelligent (AI) inference has been an effective approach to deploying well-trained AI models at the network edge for empowering immersive intelligent services such as autonomous driving and smart cities. In this paper, we propose an integrated sensing-computation-communication (ISCC) scheme for decentralized collaborative inference systems. In the proposed scheme, multiple devices connect to each other via device-to-device (D2D) links. Each device first extracts a homogeneous feature vector from the raw sensory data obtained from the same wide view of the source target and then aggregates all local feature vectors using the over-the-air computation technique. To further enhance the spectrum efficiency, the full-duplex technology is utilized to allow all devices to transmit and receive in the same frequency band. This, however, introduces significant self-interference and coupling among different tasks. To address these challenges, a multi-objective optimization-based ISCC approach is proposed.**

I. Introduction

The rapid development of communication technology and computing capability has been leading emerging scenarios of intelligent services such as autonomous robots or vehicles, smart factory infrastructure, etc [1]–[5]. An emerging technique named collaborative artificial intelligent (AI) inference prompts services providing partitioned pre-trained machine learning models at network edge devices for collaboratively making decisions. Collaborative inference can utilize the computing resources on several devices to reduce the latency and energy consumption of inference tasks while maintaining high accuracy and protecting local data privacy.

A primary research focus in collaborative inference is balancing the trade-off between communication and computation. Previous works have proposed several techniques to reduce the communication and computation overhead, such as network pruning and early exiting [6], [7]. However, these works did not consider the task-oriented property of edge inference, where the accuracy and efficiency of the inference task is the ultimate goal rather than reducing communication distortion [8]. As pointed out by [9], different feature elements with the same size and distortion level may impact the inference accuracy differently. Moreover, existing works only considered the data transmission stage and neglected the impact of the data acquisition process on inference performance. [10]–[12] proposed an integrated sensing-communication-computation (ISCC) scheme for collaborative inference to jointly optimize the cooperation of sensing, computation, and communication at edge devices.

In the conventional ISCC-based collaborative inference systems mentioned above, local features extracted from radar sensing signals are transmitted to a central server and input to the machine learning models for the inference task [11], [12]. In some scenarios, however, the central server is not always available and steadily connected to all devices (e.g., swarms of drones or cooperative automated driving [13]). This necessitates the devices to connect and communicate via device-to-device (D2D) links to share their features and reach a consensus for inference tasks. Additionally, the decentralized network offers scalability for completing more tasks on different devices. This approach can overcome the limitations of traditional centralized co-inference schemes. However, sequentially aggregating local features to all devices from others causes a high communication overhead. To improve communication efficiency in decentralized networks, the technique of full-duplex communication [14], [15] and over-the-air computation are introduced to transmit and receive signals with different antennas simultaneously.

In this paper, we propose a decentralized ISCC system for multitask collaborative inference. Each device is equipped with both receive and transmit antennas. First, all devices sense the target in the same wide view and derive noised sensory data. Then, a singular value decomposition (SVD) based filter is applied for clutter signal elimination and a low-dimension local feature vector is extracted by principal component analysis (PCA). Local feature vectors on all devices are shared through full-duplex communication and over-the-air computation (AirComp) [15]. By adopting the criterion of maximum minimum pair-wise discriminant gain which reflects the inference accuracy [12], we propose a multi-objective joint sensing power allocation, multicast, and receive beamforming problem. The challenge to solve this problem arises from three aspects: the impact of the self-interference (SI) channel imposed by simultaneously transmitting and receiving features with full-duplex communication, the design of pre-coding of each device needing to fit multiple coupled tasks, and the coupling among the sensing, computation, and communication processes. These challenges lead to a high-complexity problem difficult to solve. To address these issues, we first design the multicast beamforming vectors to utilize the SI channel and aggregate its local feature with features transmitted from other devices. Then, we jointly designed the precoders of all devices

The work of Dingzhu Wen was supported by Shanghai Sailing Program under Grants No. 23YF1427400. The work of Yuanming Shi was supported in part by the Natural Science Foundation of Shanghai under Grant No. 21ZR1442700, the National Nature Science Foundation of China under Grant 62271318, and the Shanghai Rising-Star Program under Grant No. 22QA1406100.

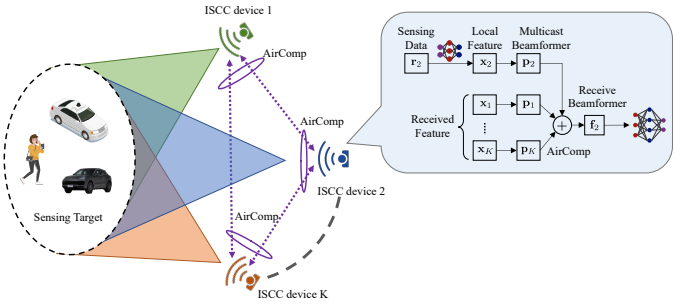


Fig. 1. The system architecture of the proposed ISCC framework.

for fitting all tasks and formulated a multi-objective problem by summarizing the minimum pair-wise discriminant gains of all tasks. Finally, we applied a successive approximation method and proposed an alternating algorithm to derive a sub-optimal solution to this complex problem.

II. System Model and Problem Formulation

A. Network Model

Consider a decentralized network of K ISAC devices for co-inference between devices with a central server not always available for devices, as illustrated in Fig. 1. Each device is equipped with a dual-functional-radar-communication (DFRC) system containing N_t transmit antennas and N_r receive antennas. It is assumed that $N_t >> K$ and $N_r >> K$ to ensure a suffcient degree of freedom (DoF) in communication. All devices obtain the same wide-view sensory data from a sensing target via signals on orthogonal sensing frequencies where local feature vectors of M dimensions are extracted. Utilizing full-duplex communication and the AirComp technique [15], every device broadcasts its local features to all other devices and aggregates features from other devices simultaneously to derive a denoised global feature vector. To transmit all feature elements at the same time, the OFDM technique is leveraged. All M dimensions of local feature vectors are transmitted in M orthogonal subcarriers. Given that the duration to transmit one feature element is significantly shorter than the channel coherence time [16], channels are assumed to be static within a single time slot. All devices are assumed to have the channel state information (CSI) of links connecting to all other devices. Finally, the aggregated feature vector is fed into a pre-trained AI model to carry out the inference task.

B. Sensing Signal Processing and Feature Extraction

We adopt the models for sensing signal process and feature extraction as proposed in [11]. During the radar sensing stage, each device transmits a frequency modulation continuous wave (FMCW) signal $s_k(t)$ in total sensing time T_s, and the received signal of ISAC device k reflected from the target is given by

$$r_k(t) = u_k(t) + \sum_{j=1}^{J} v_{k,j}(t) + n_r(t), \tag{1}$$

where $u_k(t) = H_{s,k}(t)s_k(t-\tau)$ is the desired signal for completing the inference task with $H_{s,k}(t)$ being the reflection matrix of the target and τ being the round-trip delay, $v_{k,j}(t) = C_{r,k,j}(t)s_k(t-\tau_j)$ is the clutter of j-th indirect reflection path with $C_{s,k,j}(t)$ being the round-trip coefficient of path j and τ_j being the delay of the j-th path, and $n_r(t)$ is the white Gaussian noise. It is assumed that $H_{s,k}(t)$ and $C_{s,k,j}(t)$ are estimated before the inference task.

In (1), the desired signal $u_k(t)$ is polluted by the additive sensing clutter and noise. Subsequently, we introduce the clutter cancellation procedure as detailed in [11]. The received signal of device k is sampled at a frequency of f_s into a complex feature vector $\mathbf{r}_k \in \mathbb{C}^{NT_0f_s}$ and replaced into a complex matrix $\mathbf{R}_k \in \mathbb{C}^{T_0f_s \times N}$, the column dimension of which is usually used for ranging and the row dimension contains the feature in the Doppler spectrum shift. To utilize the SVD-based linear filter for clutter cancellation, $\mathbf{R}_k$ is decomposed into $\sum_{i=1}^{I} \mathbf{u}_i\sigma_i\mathbf{v}_i^H$ where $I = \min\{T_0f_s, N\}$, $\mathbf{u}_i$, σ_i and $\mathbf{v}_i$ are the i-th left singular vector, singular value and right singular vector of $\mathbf{R}_k$. Then the principal and least dimensions of $\mathbf{R}_k$ are deprecated, resulting in $\tilde{\mathbf{R}}_k = \sum_{i=r_1}^{r_2} \mathbf{u}_i\sigma_i\mathbf{v}_i^H$. Here r_1 and r_2 are empirical parameters that vary based on different types of radar sensors. Since only the information in row dimension is useful to the inference task, $\tilde{\mathbf{R}}_k$ is compressed vertically into a vector $\tilde{\mathbf{r}}_k$

$$\tilde{\mathbf{r}}_k = \left[\sum_{j=1}^{T_0f_s} \tilde{R}_k^{j,1}, \ldots, \sum_{j=1}^{T_0f_s} \tilde{R}_k^{j,N}\right]. \tag{2}$$

Following [9], [11], the PCA-based linear extractor is used to extract the local feature vector from clutter-canceled sensory data $\tilde{\mathbf{r}}_k \in \mathbb{C}^{1\times N}$. The PCA is pre-performed at a server before the inference task using the training dataset. Then, the template of the M principal eigen-subspace is sent to all devices for extracting the local feature vectors $\{\bar{\mathbf{r}}_k \in \mathbb{R}^M\}$ with M being the number of extracted feature elements. Since the clutter cancellation and feature extraction processes are linear and based on (1), the m-th feature element of $\bar{\mathbf{r}}_k$ is given by

$$\tilde{r}_k(m) = \tilde{u}_k(m) + \sum_{j=1}^{J} \tilde{v}_{k,j}(m) + n_r(m), \tag{3}$$

where $\tilde{u}_k(m)$ is the ground-truth of feature m, $\tilde{v}_{k,j}(m)$ is the clutter of j-th path in J paths, $n_r(m)$ is the noise in Gaussian distribution $\mathcal{N}\left(0, \sigma_r^2\right)$. Next, each feature element of device k is normalized by its sensing power $P_{s,k}$ and the normalized feature element m is given by

$$x_k(m) = \frac{\tilde{r}_k(m)}{\sqrt{P_{s,k}}} = x(m) + c_{s,k}(m) + \frac{n_r(m)}{\sqrt{P_{s,k}}}, \tag{4}$$

where $x(m) = \tilde{u}_k(m)/\sqrt{P_{s,k}}$ is the normalized ground-truth feature and $c_{s,k}(m) = \sum_{j=1}^{J}\left(\tilde{v}_{k,j}(m)/\sqrt{P_{s,k}}\right)$ denotes the normalized clutter. Since clutter is rich scattering, J is very large, and $c_{s,k}(m)$ follows a zero-mean Gaussian distribution $\mathcal{N}\left(0, \sigma_{s,k}^2\right)$ according to the central limit theorem.

Consider a classification task with L classes. Following [9], [11], the ground-truth feature vector $\mathbf{x}$ is assumed to follow a Gaussian mixture distribution. Since PCA is performed, different elements of the ground-truth feature vector are independent. Specifically, the distribution of element $x(m)$ is given as

$$f(x(m)) = \frac{1}{L}\sum_{\ell=1}^{L} f_\ell(x(m)), \tag{5}$$

where $f_\ell(x(m)) = \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2\right)$ is the probability density function of the Gaussian component corresponding to the ℓ-th class, $\mu_{\ell,m}$ is the centroid of class ℓ and σ_m^2 is the variance. These parameters are pre-estimated using the training dataset. Based on (5) and the distributions of clutters and noise, we can derive the distribution of local feature element $x_k(m)$ as

$$x_k(m) \sim \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2 + \sigma_{s,k}^2 + \frac{\sigma_r^2}{P_{s,k}}\right). \quad (6)$$

C. Broadband Decentralized AirComp

In the decentralized co-inference system shown in Fig. 1, every device needs to broadcast and receive local features with all other devices to collect a denoised feature. The technique of AirComp (see e.g., [17], [18]) has been introduced to aggregate data symbols through transmitting over the same carrier by exploiting the waveform superposition property. Conventional methods like sequentially aggregating features at each device will cause the communication delay to increase linearly with the number of devices. To address this issue, full-duplex communication (see e.g., [14], [15]) is adopted for feature aggregation. In this scheme, all devices broadcast their local feature simultaneously using multicast beamforming and, at the same time, receive the over-the-air aggregated signals from other devices. All devices are assumed to be synchronized via a common clock.

Specifically, consider an arbitrary subcarrier to aggregate an arbitrary feature dimension m. For device j, the local feature element $x_j(m)$ is first modulated with a multicast beamformer $\mathbf{p}_{j,m}$ and then the signal is transmitted over a multiple-input-multiple-output (MIMO) channel to all other devices. The aggregated received signal at device k is given by

$$\mathbf{y}_k(m) = \sum_{j\neq k}\mathbf{H}_{j,k}\mathbf{p}_{j,m}x_j(m) + \mathbf{H}_{k,k}\mathbf{p}_{k,m}x_k(m) + \mathbf{w}_k(m), \quad (7)$$

where $\mathbf{H}_{j,k} \in \mathbb{C}^{N_r\times N_t}$ is the channel gain from device j to device k, particularly when $j = k$, $\mathbf{H}_{k,k}$ represents the channel gain of device k's self-interference channel, $\mathbf{p}_{j,m}$ is the multicast beamformer of device j and $\mathbf{w}_k$ is the additive white Gaussian noise following the distribution of $\mathcal{N}(\mathbf{0}, N_0\mathbf{I})$. To aggregate the features $x_j(m)$ transmitted from other devices with its local feature $x_k(m)$, the multicast beamformer is designed to exploit the self-interference channel. As mentioned, the channel matrix $\mathbf{H}_{j,k}$ remains static for aggregating all feature elements. After receiving the feature $\mathbf{y}_k(m)$, a receive beamforming vector $\mathbf{f}_{k,m} \in \mathbb{C}^{N_r}$ is applied to extract the feature vector $\hat{x}_k(m)$

$$\begin{aligned}\hat{x}_k(m) &= \mathbf{f}_{k,m}^H\mathbf{y}_k(m)\\ &= \sum_{j=1}^{K}\mathbf{f}_{k,m}^H\mathbf{H}_{j,k}\mathbf{p}_{j,m}x_j(m) + \mathbf{f}_{k,m}^H\mathbf{w}_k(m).\end{aligned} \quad (8)$$

It is worth noticing that all beamformers are computed at the server and then transmitted to all devices. Similar to (6), the distribution of $\hat{x}_k(m)$ can be further derived as

$$f\left(\hat{x}_k(m)\right) = \frac{1}{L}\sum_{\ell=1}^{L} f_\ell(\hat{x}_k(m)),\ 1 \le m \le M. \quad (9)$$

Then, all dimensions of the local feature vectors from all devices are aggregated over M orthogonal subcarriers in the same way and used for the k-th inference task.

III. Problem Formulation

In this work, the metric of maximum minimum pair-wise discriminant gain [12] is adopted to achieve a balanced inference accuracy based on the received feature distribution in (9). For a classification task, a pair-wise discriminant gain $G_{\ell,\ell'}(\hat{x}_k(m))$ measures the distance of two classes ℓ and ℓ' in feature space. A larger pair-wise discriminant gain leads to a better separation of the corresponding pair of classes in the feature space and results in an improved achievable inference accuracy. Thus, maximizing the minimum pair-wise discriminant gain guarantees the closest class pair can be well separated. Thereby, since different feature elements $\hat{x}_k(m)$ are independent, the minimum pair-wise discriminant gain of $\hat{\mathbf{x}}_k = [\hat{x}_k(1), \ldots, \hat{x}_k(m), \ldots, \hat{x}_k(M)]^T$ is written as

$$G_{\min}\left(\hat{\mathbf{x}}_k\right) = \min_{1\le\ell\neq\ell'\le L}\sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}_k(m)\right). \quad (10)$$

Thus, the objective is to maximize the weighted sum of the minimum pair-wise discriminant gains of all devices

$$\max \quad \sum_{k=1}^{K}\eta_k G_{\min}\left(\hat{\mathbf{x}}_k\right), \quad (11)$$

where η_k is the weight of minimum pair-wise discriminant gain $G_{\min}\left(\hat{\mathbf{x}}_k\right)$.

ISAC devices are usually designed for easy deployment and suffer the drawback of limited energy and computation resources (see e.g., [11], [12]). Consider an arbitrary device k, the energy consumption is comprised of three aspects, the sensing energy consumption $P_{s,k}T_{s,k}$ with sensing power $P_{s,k}$ and fixed sensing time $T_{s,k}$, the constant energy consumption for local feature extraction denoted as $E_{p,k}$, and energy consumption to transmit the m-th feature element through AirComp with power $P_{c,k}(m) = \mathbf{p}_{k,m}^H\mathbb{E}\left[x_k(m)x_k(m)^H\right]\mathbf{p}_{k,m}$.

Since the distribution of $x_k(m)$ is known in (5), its variance is determined and is denoted as $X_k(m) = \mathbb{E}\left[x_k(m)x_k(m)^H\right]$. It follows that the energy consumption constraint of device k can be derived as

$$P_{s,k}T_{s,k} + E_{p,k} + T_c\sum_{m=1}^{M}\|\mathbf{p}_{k,m}\|^2 X_k(m) \le E_k, \quad (12)$$

where E_k is the energy threshold of device k and T_c is the duration time of AirComp. Also, due to the energy limitation, the receive beamforming vector $\mathbf{f}_{k,m}$ is constrained with $\|\mathbf{f}_{k,m}\|^2 = 1$ to only control the angle of arrival (AoA).

Accordingly, the problem of maximizing the minimum pair-wise discriminant gain under the energy consumption constraint can be formulated as

$$\mathbf{P1} \quad \max_{\substack{\{P_{s,k}\},\\ \{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\}}} \sum_{k=1}^{K}\eta_k\left\{\min_{1\le\ell\neq\ell'\le L}\sum_{m=1}^{M}\frac{\left(\hat{\mu}_{\ell,k,m} - \hat{\mu}_{\ell',k,m}\right)^2}{\hat{\sigma}_{k,m}^2}\right\}, \quad (13a)$$

$$\text{s.t.} \quad P_{s,k}T_s + E_p + T_c\sum_{m=1}^{M}\|\mathbf{p}_{k,m}\|^2 X_k(m) \le E_k, \quad (13b)$$

$$\|\mathbf{f}_{k,m}\|^2 = 1, \quad (13c)$$

where

$$\begin{cases} \hat{\mu}_{\ell,k,m} = \left(\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k} \mathbf{p}_{j,m}\right) \mu_{\ell,m}, \\ \hat{\sigma}_{k,m}^2 = \sigma_{k,m}^2 \left(\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k} \mathbf{p}_{j,m}\right)^2 \\ \quad + \sum_{j=1}^{K} \left(\sigma_{s,j}^2 + \frac{\sigma_r^2}{P_{s,j}}\right) \left(\mathbf{f}_{k,m}^H \mathbf{H}_{j,k} \mathbf{p}_{j,m}\right)^2 + N_0 * \|\mathbf{f}_{k,m}\|^2. \end{cases}$$

are the mean and variance of the distribution of $\hat{x}_k(m)$.

IV. Joint Sensing Power Allocation, Multicast and Receive Beamforming

The problem $\mathbf{P1}$ formulated in Section III has a non-convex minimax form, which makes the problem difficult to solve. To solve this complex problem, first, auxiliary variables are introduced to decouple the objective, and approximation and relaxation are adopted to convert the non-convex constraints. Then, an alternating method is utilized to optimize the multicast beamforming vector $\mathbf{p}_{j,m}$ and the receive beamforming vector $\mathbf{f}_{k,m}$ in turns, resulting in two convex sub-problems with respect to $\mathbf{p}_{j,m}$ and $\mathbf{f}_{k,m}$.

A. Variable Substitution

First, α_k is defined as the minimum pair-wise discriminant gain of device k to decouple the objective

$$\alpha_k = \min_{1 \le \ell \ne \ell' \le L} \sum_{m=1}^{M} G_{\ell,\ell'}\left(\hat{x}_k(m)\right), \ \forall (k, \ell, \ell'), \tag{14}$$

where as a result, the original problem is equivalent to the problem that maximizes the weighted sum of α_k under the constraints of all pair-wise discriminant gain being no less than α_k and the energy consumption from $\mathbf{P1}$

$$\mathbf{P2} \quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\{\mathbf{f}_{k,m}\},\\ \{\mathbf{p}_{j,m}\},\{\alpha_k\}}} \quad & \sum_{k=1}^{K} \eta_k \alpha_k, \\ \text{s.t.} \quad & \text{(13b), (13c)}, \\ & \sum_{m=1}^{M} \frac{\left(\hat{\mu}_{\ell,k,m} - \hat{\mu}_{\ell',k,m}\right)^2}{\hat{\sigma}_{k,m}^2} \ge \alpha_k. \end{aligned}$$

Then, to simplify the non-convex ratio in the form of pair-wise discriminant gain, we denote $\beta_{\ell,\ell',k,m}$ as the discriminant gain of class pair (ℓ, ℓ') of element m in device k

$$\begin{aligned} &Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}\right) \\ &\quad = Q_{\ell,\ell',k,m}\left(\{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right), \end{aligned} \tag{15}$$

where

$$\begin{cases} Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}\right) \triangleq \hat{\sigma}_{k,m}^2, \\ Q_{\ell,\ell',k,m}\left(\{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right) \\ \quad \triangleq \frac{\left(\mu_{\ell,k,m} - \mu_{\ell',k,m}\right)^2}{\beta_{\ell,\ell',k,m}} \left(\sum_{j=1}^{K} \mathbf{f}_{k,m}^H \mathbf{H}_{j,k} \mathbf{p}_{j,m}\right)^2. \end{cases}$$

$$\sum_{m=1}^{M} \beta_{\ell,\ell',k,m} - \alpha_k \ge 0 \tag{16}$$

The feasible region of this new equality constraint can be extended as in (17) while keeping the same optimal solution to $\mathbf{P2}$ [12, Lemma 2].

$$\begin{aligned} &Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}\right) \\ &\quad \le Q_{\ell,\ell',k,m}\left(\{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right), \end{aligned} \tag{17}$$

Thus, $\mathbf{P2}$ can be equivalently reformulated as follows

$$\mathbf{P3} \quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\},\\ \{\alpha_k\},\{\beta_{\ell,\ell',k,m}\}}} \quad & \sum_{k=1}^{K} \eta_k \alpha_k, \\ \text{s.t.} \quad & \text{(13b), (13c), (16), (17)}. \end{aligned}$$

B. Approximation and Alternating Algorithm

The new problem $\mathbf{P3}$ is still non-convex due to the constraint (13c) and (17). First, the convex relaxation technique is utilized to relax the unit modulus constraint (13c) to an inequality $\|\mathbf{f}_{k,m}\|^2 \le 1$. Then a successive approximation method is adopted to iteratively approximate $\mathbf{P3}$ with a new problem $\mathbf{P4}$ and then solve $\mathbf{P4}$ to update the reference point. Consider the received signal power of an arbitrary feature element $c_{j,k,m} = \mathbf{f}_{k,m}^H \mathbf{H}_{j,k} \mathbf{p}_{j,m}$ and convert the parameter of $Q_{\ell,\ell',k,m}\left(\{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right)$ into $Q_{\ell,\ell',k,m}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right)$, which is a convex function. Thus, it can be proved that, for an arbitrary round t, $Q_{\ell,\ell',k,m}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right)$ is no less than its first-order Taylor expansion at $\left(\{c_{j,k,m}^{[t]}\}, \beta_{\ell,\ell',k,m}^{[t]}\right)$

$$\begin{aligned} &Q_{\ell,\ell',k,m}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right) \\ &\quad \ge Q_{\ell,\ell',k,m}^{[t]}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right), \end{aligned} \tag{18}$$

where $c_{j,k,m}^{[t]} = \mathbf{f}_{k,m}^{[t]\,H} \mathbf{H}_{j,k} \mathbf{p}_{j,m}^{[t]}$.

By substituting the right-hand side function in (17) with $Q_{\ell,\ell',k,m}^{[t]}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right)$, an approximated problem of $\mathbf{P3}$ can be derived as

$$\mathbf{P4} \quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\{\mathbf{f}_{k,m}\},\{\mathbf{p}_{j,m}\},\\ \{\alpha_k\},\{\beta_{\ell,\ell',k,m}\}}} \quad & \sum_{k=1}^{K} \eta_k \alpha_k, \\ \text{s.t.} \quad & \text{(13b), (16)}, \\ & \|\mathbf{f}_{k,m}\|^2 \le 1, \\ & Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}, \mathbf{f}_{k,m}\right) \\ & \quad \le Q_{\ell,\ell',k,m}^{[t]}\left(\{c_{j,k,m}\}, \beta_{\ell,\ell',k,m}\right). \end{aligned}$$

Since $\mathbf{P4}$ is convex with respect to $\mathbf{f}_{k,m}$ and $\mathbf{p}_{j,m}$, its sub-optimal solution can be derived by alternatively solving the sub-problem of fixing $\mathbf{f}_{k,m}$ and $(\mathbf{p}_{j,m}, P_{s,k})$ with the reference point $\mathbf{f}_{k,m}^{[t]}$ and $(\mathbf{p}_{j,m}^{[t]}, P_{s,k}^{[t]})$, respectively, until convergence.

- *Sub-problem* $\mathbf{P5}$: Fix $\mathbf{p}_{j,m}$, $P_{s,k}$ and update $\mathbf{f}_{k,m}$:

$$\mathbf{P5} \quad \begin{aligned} \max_{\substack{\{\mathbf{f}_{k,m}\},\{\alpha_k\},\\ \{\beta_{\ell,\ell',k,m}\}}} \quad & \sum_{k=1}^{K} \eta_k \alpha_k, \\ \text{s.t.} \quad & \|\mathbf{f}_{k,m}\|^2 \le 1, (16), \\ & Z_{\ell,\ell',k,m}\left(\mathbf{f}_{k,m}\right) \\ & \quad \le Q_{\ell,\ell',k,m}^{[t]}\left(\mathbf{f}_{k,m}, \beta_{\ell,\ell',k,m}\right). \end{aligned}$$

Algorithm 1 Joint Sensing Power Allocation, Multicast and Receive Beamforming Design

Input: Channel gain $\{\mathbf{H}_{j,k}\}$, Device energy E_k
Output: $\{P_{s,k}^*\}$, $\{\mathbf{p}_{j,m}^*\}$, $\{\mathbf{f}_{k,m}^*\}$

Initialize $t=0$, $\{P_{s,k}^{[0]}\}$, $\{\mathbf{p}_{j,m}^{[0]}\}$, $\{\mathbf{f}_{k,m}^{[0]}\}$, $\alpha_k^{[0]}$, $\beta_{\ell,\ell',k,m}^{[0]}$ in feasible region of $\mathbf{P3}$;
Initialize function $Q_{\ell,\ell',k,m}^{[0]}\left(\{\mathbf{p}_{j,m}^{[0]}\}, \mathbf{f}_{k,m}^{[0]}, \beta_{\ell,\ell',k,m}^{[0]}\right)$;
repeat
  Derive the approximated problem $\mathbf{P4}$;
  repeat
    Update $\mathbf{f}_{k,m}$ by solving $\mathbf{P5}$;
    Update $\mathbf{p}_{j,m}$, $P_{s,k}$ by solving $\mathbf{P6}$;
  until convergence
  Update function $Q_{\ell,\ell',k,m}^{[t+1]}\left(\{\mathbf{p}_{j,m}^{[t+1]}\}, \mathbf{f}_{k,m}^{[t+1]}, \beta_{\ell,\ell',k,m}^{[t+1]}\right)$;
  $t \leftarrow t+1$;
until convergence
Optimal solution $P_{s,k}^* \leftarrow P_{s,k}^{[t]}$, $c_{k,m}^* \leftarrow c_{k,m}^{[t]}$, $\mathbf{f}_m^* \leftarrow \mathbf{f}_m^{[t]}$;

- *Sub-problem* $\mathbf{P6}$: Fix $\mathbf{f}_{k,m}$ and update $\mathbf{p}_{j,m}$, $P_{s,k}$:

$$\mathbf{P6} \quad \begin{aligned} \max_{\substack{\{P_{s,k}\},\{\mathbf{p}_{j,m}\},\\ \{\alpha_k\},\{\beta_{\ell,\ell',k,m}\}}} \quad & \sum_{k=1}^{K} \eta_k \alpha_k, \\ \text{s.t.} \quad & (13b), (16), \\ & Z_{\ell,\ell',k,m}\left(\{P_{s,j}\}, \{\mathbf{p}_{j,m}\}\right) \\ & \quad \leq Q_{\ell,\ell',k,m}^{[t]}\left(\{\mathbf{p}_{j,m}\}, \beta_{\ell,\ell',k,m}\right). \end{aligned}$$

Since the received signal power $c_{j,k,m}$ is only relevant to $\mathbf{f}_{k,m}$ or $\mathbf{p}_{j,m}$ while fixing the other, the parameters of function $Q_{\ell,\ell',k,m}^{[t]}$ is changed into $\mathbf{f}_{k,m}$ or $\mathbf{p}_{j,m}$, respectively. The sub-optimal solution $\mathbf{f}_{k,m}^{[t+1]}$, $\mathbf{p}_{j,m}^{[t+1]}$, $P_{s,k}^{[t+1]}$ is then used to derive the new received signal power $c_{j,k,m}^{[t+1]} = \mathbf{f}_{k,m}^{[t+1]H}\mathbf{H}_{j,k}\mathbf{p}_{j,m}^{[t+1]}$, which is in the feasible region of $\mathbf{P4}$. Then $\mathbf{P4}$ is updated for the next round $t+1$.

Based on the approximation and alternating method described before, the solution procedure to $\mathbf{P3}$ is summarized in Algorithm 1.

V. Performance Evaluation

A. Simulation Settings

1) Network settings: A decentralized network consisting of 3 ISAC devices is simulated for inference tasks. Each device is equipped with 8 transmit antennas and 8 receive antennas. Distances between devices are in the range of [0.4km, 0.45km]. The channel gains of the link between devices are modeled as $\mathbf{H}_{j,k} = |\varphi_{j,k}\boldsymbol{\rho}_{j,k}|^2$. $[\varphi_{j,k}]_{\mathrm{dB}} = -[\mathbf{PL}_{j,k}]_{\mathrm{dB}} + [\zeta_{j,k}]_{\mathrm{dB}}$ is the large-scale fading channel coefficient, where $[\mathbf{PL}_{j,k}]_{\mathrm{dB}} = 128.1 + 37.6\log_{10} d_{j,k}$ is the path loss in dB, $d_{j,k}$ is the distance between device j and device k, and $[\zeta_{j,k}]_{\mathrm{dB}} \sim \mathcal{N}(0, \sigma_\zeta^2)$ is the shadowing in dB. On the other hand, $\boldsymbol{\rho}_{j,k} \sim \mathcal{CN}(0, \mathbf{I})$ stands for the Rayleigh small-scale fading channel coefficient. The variances of sensing noise σ_r^2 and clutter signal $\sigma_{s,k}^2$ are both set to 0.2. The channel noise variance N_0 is set to 1 and the variance of shadow fading $\sigma_\zeta^2 = 8$ dB. The sensing time T_s and communication time T_c of devices are set to 1 second and the computation energy $E_{p,k}$ is set to 0.1 Joule.

2) Inference tasks and models: In this experiment, the University of Glasgow Radar Signature dataset [19] is used to evaluate the performance of the proposed algorithm. This dataset contains the radar echo signals of different motions of 103 people in all age groups at nine different locations. Data corresponding to five motions are selected for the recognition task: walking, sitting down, standing up, picking up an object, and drinking water. Based on the dataset, each device is assigned a different task along with a machine learning model that demonstrated the best performance during the training phase:

- The task of device 1 is to identify the target's motion from 5 motions with a K-Nearest Neighbour (KNN) model where $K=5$.
- The task of device 2 is to distinguish the gender of the target with a support vector machine (SVM) model.
- The task of device 3 is to separate the age group of the target (e.g., [0, 30] or [30, 50] or [50,]) with a multi-layer perceptron (MLP) neural network where the numbers of neurons in the hidden layers of MLP set to 80 and 40.

The dataset contains 1500 samples in total and is divided into 90% training dataset and 10% testing dataset. The training dataset is considered the ground-truth data when training all ML models at a powerful server and then the trained model are transmitted to each device and used for inference. The testing dataset is distorted by sensing and communication noise determined by the two schemes mentioned below.

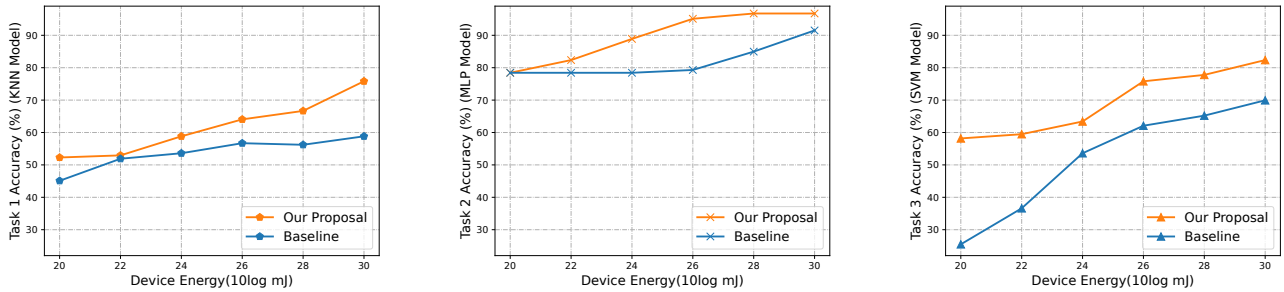
3) Inference algorithms: The proposed algorithm and baseline scheme are described below

- *Our proposal*: All parameters are allocated by the proposed scheme in Algorithm 1.
- *Baseline*: The sensing power is allocated to a constant, and the multicast and receive beamforming vectors are set to a constant during all elements' transmission.

B. Performance Comparison

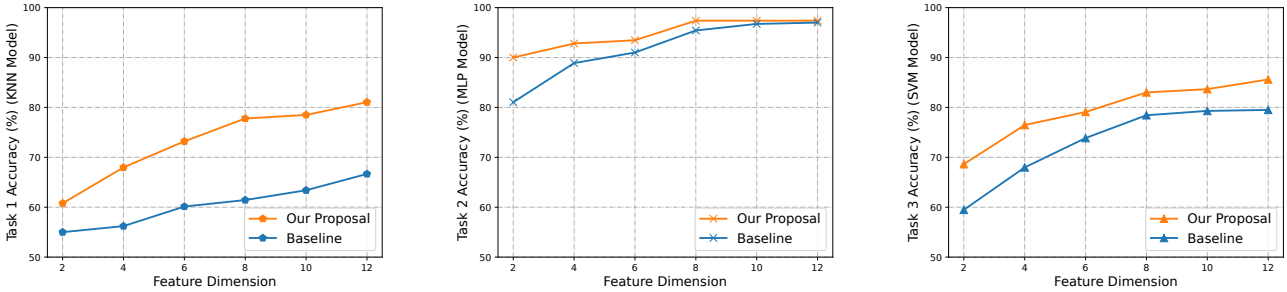
1) Inference accuracy v.s. device energy: Fig. 2 shows the impact of the device's total energy on the accuracy of the inference task on three devices. The results indicate that a higher upper limit of energy leads to better inference accuracy. This is because a higher device energy threshold indicates that the devices can allocate more sensing power and communication power to suppress the clutter distortion and resist channel fading and channel noise. Moreover, our proposal outperforms the baseline since the sensing power and beamforming vectors are not adjusted to different feature elements in the baseline scheme. Besides, the accuracy of task 2 is much higher than the other two tasks, which is might because task 2 is much simpler than task 1 and task 3.

2) Inference accuracy v.s. feature dimensions: Fig. 3 illustrates the inference accuracy of three tasks under different feature dimensions. The performance of all three tasks increases as the feature dimension increases. This is because different feature dimensions are orthogonal and independent from the algorithm of PCA. It follows that more feature dimensions can store more data about the target, which improves the performance of inference tasks. In addition, the proposed scheme reaches a higher accuracy on all three tasks.



(a) Task 1 inference acc. v.s. device energy (b) Task 2 inference acc. v.s. device energy (c) Task 3 inference acc. v.s. device energy

Fig. 2. Inference accuracy of three tasks on corresponding devices versus different device energy.



(a) Task 1 inference acc. v.s. feature dimension (b) Task 2 inference acc. v.s. feature dimension (c) Task 3 inference acc. v.s. feature dimension

Fig. 3. Inference accuracy of three tasks on corresponding devices versus different feature dimensions.

VI. Conclusion

This paper proposed a decentralized AirComp based ISCC system tailored for multitask collaborative inference. Our proposed scheme facilitates simultaneous multicast and AirComp aggregation of local features of all devices through full-duplex communication, which makes the communication overhead irrelevant to the number of devices. We exploited the self-interference channel in full-duplex communication to aggregate features from one device itself with others, reducing the cost of computation resources. Leveraging these benefits, our proposed scheme shows a better inference performance in experiments. This new scheme paves the way for broader applications with massive ISCC devices for distributed learning.

References

[1] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2167–2191, 2020.

[2] Y. Shi, Y. Zhou, D. Wen, Y. Wu, C. Jiang, and K. B. Letaief, "Task-oriented communications for 6G: Vision, principles, and technologies," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 78–85, 2023.

[3] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2022.

[4] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[5] D. Li, Y. Gu, H. Ma, Y. Li, L. Zhang, R. Li, R. Hao, and E.-P. Li, "Deep learning inverse analysis of higher order modes in monocone TEM cell," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 12, pp. 5332–5339, 2022.

[6] S. H. Shabbeer Basha, S. N. Gowda, and J. Dakala, "A simple hybrid filter pruning for efficient edge inference," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 3398–3402.

[7] Z. Liu, Q. Lan, and K. Huang, "Resource allocation for multiuser edge inference with batching and early exiting," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1186–1200, 2023.

[8] G. Zhu, Z. Lyu, X. Jiao, P. Liu, M. Chen, J. Xu, S. Cui, and P. Zhang, "Pushing AI to wireless network edge: an overview on integrated sensing, communication, and computation towards 6G," *Sci. China Inf. Sci.*, vol. 66, no. 3, p. 130301, 2023.

[9] D. Wen, X. Jiao, P. Liu, G. Zhu, Y. Shi, and K. Huang, "Task-oriented over-the-air computation for multi-device edge AI," *IEEE Trans. Wireless Commun.*, 2023, early access.

[10] X. Li, F. Liu, Z. Zhou, G. Zhu, S. Wang, K. Huang, and Y. Gong, "Integrated sensing, communication, and computation over-the-air: MIMO beamforming design," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5383–5398, 2023.

[11] D. Wen, P. Liu, G. Zhu, Y. Shi, J. Xu, Y. C. Eldar, and S. Cui, "Task-oriented sensing, computation, and communication integration for multi-device edge AI," *IEEE Trans. Wireless Commun.*, 2023.

[12] Z. Zhuang, D. Wen, Y. Shi, G. Zhu, S. Wu, and D. Niyato, "Integrated sensing, communication, and computation for edge AI inference with over-the-air computation," *IEEE Trans. Wireless Commun.*, 2023, early access.

[13] A. Sarker, C. Qiu, and H. Shen, "Connectivity maintenance for next-generation decentralized vehicle platoon networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 4, pp. 1449–1462, 2020.

[14] S. Datta, D. N. Amudala, E. Sharma, R. Budhiraja, and S. S. Panwar, "Full-duplex cell-free massive mimo systems: Analysis and decentralized optimization," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 31–50, 2022.

[15] Z. Lin, Y. Gong, and K. Huang, "Distributed over-the-air computing for fast distributed optimization: Beamforming design and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 274–287, 2023.

[16] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[17] Y. Shi, S. Xia, Y. Zhou, Y. Mao, C. Jiang, and M. Tao, "Vertical federated learning over cloud-RAN: Convergence analysis and system optimization," *IEEE Trans. Wireless Commun.*, 2023.

[18] Z. Wang, Y. Zhao, Y. Zhou, Y. Shi, C. Jiang, and K. B. Letaief, "Over-the-air computation: Foundations, technologies, and applications," *arXiv preprint arXiv:2210.10524*, 2022.

[19] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. Le Kernec, "Radar signatures of human activities," 2019.