Multimodal Learning for Integrated Sensing and Communication Networks

Xiaonan Liu*, Tharmalingam Ratnarajah*, Mathini Sellathurai†, and Yonina C. Eldar‡
* School of Engineering, The University of Edinburgh, UK, {xliu8, t.ratnarajah}@ed.ac.uk
† Department of Engineering and Physical Science, Heriot-Watt University, Edinburgh, UK, m.sellathurai@hw.ac.uk
‡ Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel, yonina.eldar@weizmann.ac.il

***Abstract*—Integrated sensing and communication (ISAC) is a promising technique for beyond 5G networks. In ISAC networks, the sensed environmental data may be multimodal data, which may result in high computation and communication latency due to the large size of data modalities and limited computation capability of mobile devices. To solve the problem, in this paper, we propose multimodal learning in ISAC networks. Simulation results show that the proposed multimodal learning design significantly outperforms several benchmarks without considering multimodal data sensing and communication.**

***Index Terms*—Integrated sensing and communication, multimodal learning.**

I. Introduction

The increasing demands of various emerging services in beyond fifth-generation (5G) wireless networks, such as vehicular networks, virtual reality, and unmanned aerial vehicles, facilitate wireless networks to provide sensing with high accuracy and high communication capabilities [1]. By providing wireless sensing and simultaneous data transmission services in the same spectrum, the integrated sensing and communication (ISAC) technique is envisioned to effectively increase spectral and hardware efficiency, which has attracted extensive attention and discussion recently [2]–[4].

For a specific sensing task in wireless networks with the ISAC technique, the base station (BS) first determines a set of devices based on task requirements. Then, the selected devices sense the environment, collect the data, and transmit the sensed data to the BS for data processing. Besides, to decrease wireless resource consumption, such as bandwidth and energy, and traffic burden caused by sensed data transmission, mobile devices can also be used for sensed data processing [5].

Although the ISAC technique enables numerous applications and advantages, it still has some challenges in terms of high communication and computation latency and high computational complexity. In particular, the sensed environmental data may be multimodal, resulting in high computation latency because of the large size of multiple data modalities and the limited computation capability of mobile devices. Fortunately, the mobile edge computing (MEC) technique [6], by allowing mobile devices to offload sensing tasks to the edge of the network, is able to decrease the computational burden of mobile devices. Generally, MEC has two computation offloading models: binary and partial offloading [7]. In the binary offloading model, each sensing task is either processed locally or directly transmitted to the MEC server for execution. However, partial offloading requires the sensing task to be partitioned and processed both locally and at the MEC server. In this paper, we consider binary computation offloading, which is commonly used in ISAC systems for processing non-partitionable simple tasks.

On the other hand, multimodal fusion networks offer distinct advantages over unimodal counterparts in various applications, such as task-oriented semantic communications [8], beam selection in vehicular networks [9], power prediction in mmWave networks [10], video quality of experience prediction over wireless networks [11], and massive multiple-input multiple-output (MIMO) channel prediction [12]. In [8], a unified deep learning-enabled semantic communication system was developed, where a unified end-to-end framework could serve many different tasks with multiple modalities of data. In [9], a novel approach for machine learning-aided fast beam selection technique was proposed for vehicular networks, where multimodal non-radio frequency sensor data was exploited to reduce the search space for identifying best-performing beams. In [10], a distributed multimodal machine learning framework, coined multimodal split learning, in which a large neural network (NN) was split into two wirelessly connected segments, was proposed to improve the accuracy of millimetre wave received power prediction by utilizing the camera images and radio frequency (RF) signals. In [11], a novel and end-to-end framework called Deep quality of experience (QoE) was developed to predict video QoE by using a combination of deep learning techniques, such as word embedding and 3D convolutional neural network. In [12], a systematic framework on deep multimodal learning-based wireless communications was developed to predict channels in massive MIMO systems by leveraging multi-source sensing information. However, the multimodal learning algorithms in [8]–[12] are static, namely, they process and fuse multimodal data with identical computation, without accounting for diverse computational demands of different multimodal data, and how to effectively combine multimodal data to better exploit information remains a challenge. Therefore, in this paper, we consider dynamic multimodal learning in resource-limited ISAC networks.

The rest of this paper is organized as follows. Section II presents the system model. Multimodal learning is presented in Section III. The simulation results and conclusions are detailed in Section IV and Section V, respectively.

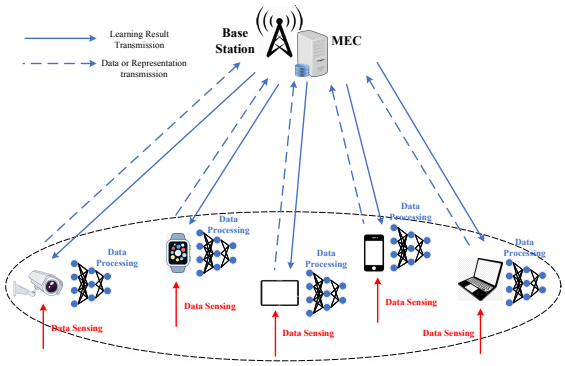

Fig. 1. Illustration of proposed ISAC system.

II. System Model

We consider a multimodal integrated sensing and communication (ISAC) system with a cellular base station (BS) and K mobile devices, as shown in Fig. 1. Mobile edge computing (MEC) is deployed at the BS to guarantee high computation capability. The BS is equipped with M antennas, and each edge device has one antenna. Multimodal data is sensed by mobile devices and processed by themselves or delivered to BS for processing.

A. Multimodal ISAC Operation

In practical multimodal ISAC systems, various sensors are embedded in mobile devices and used in multiple sensing applications, generating multimodal data with different sampling rates, where the sampling rate is the number of data samples obtained per second [13]. Each mobile device reports its states, such as channel state information (CSI), power, and computation capability, to BS periodically, so that BS can periodically update these states affected by the mobility and activity of mobile devices [14], [15]. Sensed multimodal data processing in ISAC systems is modelled as a function $f(\boldsymbol{x}_k)$, where $\boldsymbol{x}_k$ is the multimodal data sensed by the kth mobile device.

B. Multimodal Data Sensing Model

The multimodal data sensing rate m_k of the kth edge device or the amount of data sensed by the kth edge device per second is equal to the sampling rate multiplied by the number of bits required to represent one data sample. The value of m_k is determined by the sensing context. For example, for vision/audio-based data, m_k is usually at the level of millions of bits per second. While for text-based data, m_k is at the level of a few bits per second. We set a binary variable $a_k \in \{0, 1\}$ to denote whether or not the kth edge device is selected for multimodal sensing. When the kth edge device is selected, $a_k = 1$, otherwise, $a_k = 0$.

We set a binary variable $a_k \in \{0, 1\}$ to denote whether or not the kth edge device is selected for multimodal sensing. When the kth edge device is selected, $a_k = 1$, otherwise, $a_k = 0$. The relationship between a_k and multimodal data sensing latency T_k^s is presented as

$$T_k^s \leq a_k T, \tag{1}$$

which means that $T_k^s = 0$ if $a_k = 0$ and $T_k^s \leq T$ if $a_k = 1$. The multimodal data sensing duration T_k^s decides the amount of multimodal data that needs to be sensed, processed/transmitted by the kth device. Actually, the amount of multimodal data sensed by the kth device is calculated as

$$n_k = m_k T_k^s. \tag{2}$$

C. Communication Model

When the kth edge device is selected, namely, $a_k = 1$, proper bandwidth B_k should be allocated to transmit the sensed multimodal data or the raw sensed data, where we consider the OFDMA technique for transmission. Since it is impractical to allocate arbitrary bandwidth to each edge device, we assume a minimum bandwidth unit $B_{\min}$ and the bandwidth selection of the kth edge device is a multiple of $B_{\min}$, namely, $B_k = b_k B_{\min}$, where $b_k = 0, 1, 2, ..., \lfloor B/B_{\min} \rfloor$. The value of b_k determines the number of bandwidth units allocated to the kth edge device, and the relationship between b_k and a_k is denoted as

$$a_k = \lceil b_k / B \rceil, \tag{3}$$

which means that $a_k = 1$ if $b_k > 0$, and $a_k = 0$ if $b_k = 0$. Based on the bandwidth allocation model, the bandwidth constraint is rewritten as

$$\sum_{k \in \mathcal{K}} b_k B_{\min} \leq B. \tag{4}$$

Given the allocated bandwidth, the achievable uplink transmission rate between the kth device and MEC is denoted as

$$R_k^{\text{up}} = b_k B_{\min} \log_2 \left(1 + \frac{P_k h_k}{\sigma^2}\right), \tag{5}$$

where P_k is the transmit power of the kth edge device, h_k is the channel gain between the kth edge device and the BS, and σ^2 is additive white Gaussian noise (AWGN) power. With the uplink transmission R_k^{up} and transmission latency T_k^{comm}, the amount of raw multimodal data transmitted to MEC for processing is calculated as

$$A_k^{\text{comm}} = T_k^{\text{comm}} R_k^{\text{up}}. \tag{6}$$

D. Optimization Problem

Based on the multimodal data sensing, communication, and computation model mentioned above, we consider an optimization problem to minimize the loss function of the multimodal learning, which is formulated as

$$\min \ \mathcal{L} = \mathcal{L}_{\text{task}}, \tag{7}$$

where $\mathcal{L}_{\text{task}}$ is the loss of the learning task in multimodal learning, which can be text classification, vision recognition, and audio classification.

III. Multimodal Learning

In this section, we consider multimodal learning. We first introduce a modality-level selection scheme to select proper modalities for training and inference. Next, we transform the optimization problem into a new one.

A. Task Description

For multimodal learning in ISAC networks, we consider multimodal data from different domains, including text classification, vision recognition, and audio classification.

- **Text Classification:** The text classification task aims to classify whether the sentiment of a given sentence is positive or negative, which is denoted a binary classification problem. Thus, binary accuracy, namely, positive or negative sentiment, is considered as the performance metric for text classification, and the cross-entropy is the loss function to train the model.
- **Vision Recognition:** The vision recognition task is about leveraging sensed videos to recognize the sentiment of users effectively [16]. Similar to the text classification task, binary accuracy and cross-entropy are selected as the evaluation criterion and loss function, respectively
- **Audio Classification:** The main purpose of audio classification is to identify the sentiment of sensed audio signals of uses. Similarly, binary classification accuracy is selected as the metric for the evaluation criterion, and the binary cross-entropy is deployed for the loss function.

Considering emotion recognition as an example, traditional multimodal learning fuses text, vision, and audio data to predict emotion, which leads to high sensing, computation, and communication latency. Motivated by this example, we consider a dynamic multimodal fusion, which can adaptively fuse input data from multiple modalities.

B. Modality Selection

We assume that the sensed data in ISAC networks has M modalities, denoted as $\boldsymbol{x} = (x_1, x_2, ..., x_M)$. According to the typical Mixture-of-Experts (MoE) framework [17], we design a set of expert networks, where each expert specializes in a subset of all M modalities, and the number of expert networks is up to $2^M - 1$. For example, if $M = 2$, there are at most 3 expert networks, denoted as $E_1(x_1)$, $E_1(x_2)$, and $E_1(x_1, x_2)$. In a practical ISACnetwork, the candidate expert networks can be narrowed down with domain expertise.

To determine which expert should be activated, we consider a gating network $G(\boldsymbol{x})$, which takes multimodal inputs $\boldsymbol{x}$ to form a global view and then produces a N-dimensional sparse vector $\boldsymbol{g}$ as output. The final output y takes the form of

$$y = \sum_{i=1}^{N} g_i E_i(\boldsymbol{x}_i), \tag{8}$$

where $\boldsymbol{x}_i$ is the subset of modalities that the ith expert takes as input, and N denotes the number of selected expert networks.

Unlike traditional MoEs where the output is a weighted summation of expert networks, and each expert network should be processed, in our formulation, the gating network $\boldsymbol{g}$ output is a one-hot encoding, namely, only one expert network is selected for learning, and the computations required for other expert networks are saved. Since our expert network already covers a broad range of modality combinations, we only select one expert network during each forward pass for maximum computational savings. As shown in Fig. 2, it provides an illustration of the proposed design with $M = 2$ modalities and $N = 3$ expert networks. Only one expert network $E_1(x_1)/E_2(x_1, x_2)/E_3(x_2)$ is selected. The design of the gating networks $G(\boldsymbol{x})$ follows two basic requirements: (1) it should have high computation efficiency; (2) it needs to be sufficiently expressive to make effective decisions on which expert network to select.

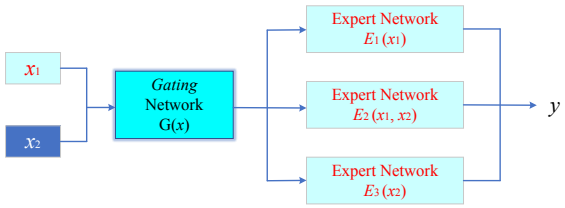

Fig. 2. An illustration of modality-level DynMM, where input data has two modalities.

Although acquiring alignment information for mulitmodal data across different devices is challenging, our assumption is that data from different modalities, such as sensory data and visual data, on a multimodal device, inherently have some alignment information, based on which we can train models to extract multimodal representations from the data. We consider multimodal autoencoders to encode the data into shared or correlated hidden representations.

C. Optimization Problem Transformation

Based on the proposed modality selection scheme, not all data modalities need to be sensed, calculated or communicated. Suppose the learning network is directly trained by minimizing a task-specific loss. In that case, the gating network may learn a trivial solution that always selects the expert network with high computation capability, which results in a heavy burden on devices with limited computation capability. To address this issue, we consider a resource-aware loss function in the training objective, and the optimization problem in (7) is transformed as

$$\min \ \mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{i=1}^{N} g_i C(E_i), \tag{9}$$

where $C(E_i)$ is the computation cost of processing the expert network $E_i(x_i)$, $\mathcal{L}_{\text{task}}$ is the task loss, such as cross-entropy between the true label and prediction output, and λ is a value determining the relative importance of the two loss terms.

The transformed optimization problem in (9) enables dynamic multimodal fusion to achieve a tradeoff between learning accuracy and computation efficiency, which can select the value of λ to adapt to the constraints. A small value of λ

explores expert networks with high computation capabilities and leads to high learning accuracy. While a large value of λ guarantees lightweight computations for high computation efficiency.

D. Parameterization Techniques

One problem in dynamic multimodal fusion is the training of gating network $G(\boldsymbol{x})$. Because of the non-differentiability of the discrete decisions given by $G(\boldsymbol{x})$, the learning network cannot be directly trained by back-propagation. To solve the problem in dynamic multimodal fusion, Gumbel-softmax and parameterization techniques are considered in the training process [18]. Since $\boldsymbol{g}$ denotes the desired one-hot J-dimensional decision vector created by a gating network $G(\boldsymbol{x})$, denoted as

$$\boldsymbol{g} = \text{one-hot}(\arg\max_j G(\boldsymbol{x})_j). \tag{10}$$

We use a real-valued soft vector $\hat{\boldsymbol{g}}$ with the following form

$$\hat{g}_j = \frac{\exp((\log G(\boldsymbol{x})_j + e_j)/\tau)}{\sum_{j=1}^{J} \exp((\log G(\boldsymbol{x})_j + e_j)/\tau)}, \tag{11}$$

where $e_1, e_2, ..., e_J$ are samples independently selected from Gumbel (0,1) and τ is the softmax temperature. The distribution of $\hat{\boldsymbol{g}}$ is more uniform with large τ and resembles a categorical distribution with small τ. The real-valued soft vector $\hat{\boldsymbol{g}}$ serves as a continuous, differentiable approximation of $\boldsymbol{g}$. In the training process, $\boldsymbol{g}$ is replaced by soft $\hat{\boldsymbol{g}}$ to enable back-propagation. Through annealing τ, $\hat{\boldsymbol{g}}$ gradually converges to a desired one-hot vector. According to the straight-through technique [18], we use $\boldsymbol{g}$ in the forward pass and soft $\hat{\boldsymbol{g}}$ in the backward propagation with the gradient approximation $\nabla \boldsymbol{g} \approx \nabla \hat{\boldsymbol{g}}$. In this way, the gating network still outputs a discrete decision during training. In the inference process, we directly use trained $\boldsymbol{g}$ for computational benefits. Further, we propose a two-stage training of dynamic multimodal fusion that jointly optimizes the multimodal and gating networks.

Stage I: *Pre-training.* The sparse decisions of the gating network in the early stage of training may lead to biased optimization. Expert networks that are rarely selected have fewer weight updates, resulting in poor learning performance. Therefore, the goal of a pre-training stage is to ensure that each expert is fully optimized before the gating network gets involved.

Stage II: *Fine-tuning.* After the pre-training stage, gating networks are integrated into our optimization process at this stage. Based on the aforementioned parameterization technique, we jointly optimize the dynamic network along with gating networks.

IV. Simulation Results

We conduct experiments on sentiment analysis on CMU-MOSEI [19]. To demonstrate the wide applicability of our proposed multimodal learning, we select three tasks that include different modalities (i.e. video, audio, and text).

CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is the largest dataset of sentiment analysis and emotion recognition. It contains 3,228 real-world online videos from more than 1000 speakers and 250 topics. Each video is split into short segments of 10-20 seconds. Each segment is annotated for a sentiment from -3 (strongly negative) to 3 (strongly positive). The task is to predict the sentiment scores from video, audio and text. Following [20], we use 16,265 data for training, 1,869 data for validation and 4,643 data for testing. The feature extraction steps are the same as [20].

TABLE I
Learning performance on CMU-MOSEI sentiment analysis. Modalities V, A, and T represent video, audio, and text, respectively.

Method	Modality	Accuracy (%)
Video Network	V	69.02
Audio Network	A	67.68
Text Network	T	78.35
Multimodal	V+A+T	79.54

As text is the best-performing modality in this task, we adopt an unimodal network that takes textual features as input to be the expert network E_1. The second expert network E_2 of our multimodal learning receives inputs from three modalities. The gating network is designed as a lightweight transformer network with hidden dimensions equal to 512 and 2 attention heads, followed by a linear layer. The gating network receives concatenated features from three modalities and generates sample-wise decisions on which expert network to activate during inference time. Results are shown in Table 1. We can find that our multimodal learning achieves a higher learning accuracy.

V. Conclusions

In this paper, multimodal learning for ISAC networks was proposed. Particularly, devices were able to sense multimodal data and transmit them to BS to make predictions. Simulation results have demonstrated that our proposed framework could achieve higher learning accuracy compared to the framework only sensing unimodal data.

References

[1] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6g and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[2] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 6, pp. 1295–1315, Nov. 2021.

[3] J. Zhang, N. Garg, and T. Ratnarajah, "In-band-full-duplex integrated sensing and communications for iab networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 12 782–12 796, Dec. 2022.

[4] F. Liu, C. Masouros, A. Li, T. Ratnarajah, and J. Zhou, "Mimo radar and cellular coexistence: A power-efficient approach enabled by interference exploitation," *IEEE Trans. Signal Process.*, vol. 66, no. 14, pp. 3681–3695, Jul. 2018.

[5] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3834–3862, Jun. 2020.

[6] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surv. Tut.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart. 2017.

[7] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surv. Tut.*, vol. 19, no. 3, pp. 1628–1656, 3rd. Quart. 2017.
[8] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multitask semantic communication system for multimodal data," *IEEE Trans. Commun.*, pp. 1–1, 2024.
[9] B. Salehi, G. Reus-Muns, D. Roy, Z. Wang, T. Jian, J. Dy, S. Ioannidis, and K. Chowdhury, "Deep learning on multimodal sensor data at the wireless edge for vehicular network," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7639–7655, Jul. 2022.
[10] Y. Koda, J. Park, M. Bennis, K. Yamamoto, T. Nishio, M. Morikura, and K. Nakashima, "Communication-efficient multimodal split learning for mmwave received power prediction," *IEEE Commun. Lett.*, vol. 24, no. 6, pp. 1284–1288, Jun. 2020.
[11] H. Zhang, L. Dong, G. Gao, H. Hu, Y. Wen, and K. Guan, "Deepqoe: A multimodal learning framework for video quality of experience (qoe) prediction," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3210–3223, Dec. 2020.
[12] Y. Yang, F. Gao, C. Xing, J. An, and A. Alkhateeb, "Deep multimodal learning: Merging sensory data for massive mimo channel prediction," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 1885–1898, Jul. 2021.
[13] F. Wang and J. Liu, "Networked wireless sensor data collection: Issues, challenges, and approaches," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 4, pp. 673–687, Fourth Quarter 2011.
[14] H. Li, T. Li, W. Wang, and Y. Wang, "Dynamic participant selection for large-scale mobile crowd sensing," *IEEE Trans. Mobile Comput.*, vol. 18, no. 12, pp. 2842–2855, Dec. 2019.
[15] X. Li, C. You, S. Andreev, Y. Gong, and K. Huang, "Wirelessly powered crowd sensing: Joint power transfer, sensing, compression, and transmission," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 391–406, Feb. 2019.
[16] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[17] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: a literature survey," *Artif. Intell. Rev.*, vol. 42, pp. 275–293, 2014.
[18] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017.
[19] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
[20] P. P. Liang, Y. Lyu, X. Fan, Z. Wu, Y. Cheng, J. Wu, L. Chen, P. Wu, M. A. Lee, Y. Zhu *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *arXiv preprint arXiv:2107.07502*, 2021.