Task-oriented Over-the-air Computation for Device-edge Co-inference with Balanced Classification Accuracy

Abstract—**Device-edge co-inference has been a promising technique for enabling various kinds of intelligent services at the network edge, including auto-driving, remote health, etc. In this paradigm, the design target shifts from the traditional communication throughput to the effective and efficient execution of the task, e.g., the inference accuracy and latency. In this paper, a task-oriented over-the-air computation (AirComp) scheme is proposed for a multi-device AI system. Particularly, a novel inference accuracy metric is proposed for classification tasks, which is called minimum pair-wise discriminant gain and measures the smallest distance of all class pairs in the feature space. By maximizing the minimum pair-wise discriminant gain, all classes can be well separated in the feature space, and thus leading to a balanced and enhanced inference accuracy for all classes. Besides, this paper jointly optimizes the minimum discriminant gain of all feature elements instead of separately maximizing that of each element in the existing designs. As a result, the transmit power can be adaptively allocated to the elements with different contributions on the inference accuracy, resulting in an extra degree of freedom to improve inference performance. Extensive experiments are conducted based on a human motion recognition task, which verifies our theoretical analysis.**

Index Terms—**Device-edge co-inference, Task-oriented communication, Over-the-air computation.**

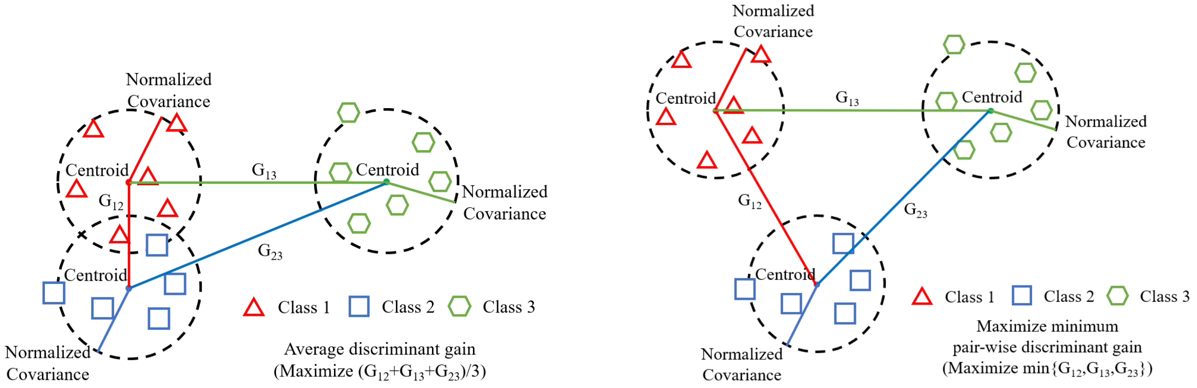
I. Introduction

Edge artificial intelligence (AI) has been an emerging technique to provide various kinds of intelligent services to support many applications like auto-driving and unmanned aerial vehicles (UAV) [1], [2]. The realization of these intelligent services depends on the deployment of well-trained AI models at the network edge and utilizes their inference capability for making intelligent decisions. This gives rise to a new research topic, called edge inference [3]–[5]. There has been extensive work targeting the efficient implementation of edge inference (see, e.g., [6]–[8]). Among others, the technique of device-edge co-inference, or called edge split inference, has been the most popular architecture [4], [8]–[12]. It divides an AI model into two parts. The former part with light size is deployed at the device for extracting low-dimensional feature vector from the high-dimensional raw data. The other computation-intensive part is deployed at the server and utilizes the received feature vector from the device to finish the downstream inference task. As a result, benefiting from avoiding transmitting high-dimensional raw data and offloading the intensive computation to the server, it can enjoy the advantages of enhanced communication and computation efficiency as well as preserving data privacy. Hence, the device-edge co-inference structure is adopted here.

As stated by [10], [11], the design of device-edge co-inference calls for task-oriented communication techniques, since traditional techniques, which target throughput maximization or receive-data distortion minimization, fail to distinguish the data samples with a same size and a same distortion level but with different contributions on the inference task. To tackle this problem, a task-oriented scheme of integrated sensing, computation, and communication is proposed in [10] for multi-device AI inference, where each device senses a target area from disjoint narrow views. Furthermore, for the case where different devices observe the same wide view of a target area and each device obtains a noise-corrupted local version of the ground-true sensory data, a task-oriented over-the-air computation (AirComp) is proposed in [11] to efficiently aggregate the local features for suppressing the sensing and channel noise. As the instantaneous inference accuracy is unknown at the design stage, where the input data of the AI model is not obtained, authors in [10], [11] adopt an approximate but tractable metric as a surrogate for classification tasks, called discriminant gain. It is proposed in [4] based on the well-known Kullback-Leibler (KL) divergence [13]. Specifically, the discriminant gain of an arbitrary class pair is the distance of their centroids in the Euclidean feature space normalized by their covariance, as shown in Fig. 1. With a larger discriminant gain, the two classes are better separated in the feature space, which results in a greater achievable inference accuracy. The system discriminant gain is the average discriminant gain of all pairs.

However, the average discriminant gain adopted in [4], [10], [11] has a drawback of unbalanced inference accuracy of different classes, as shown in Fig. 1a. To be specific, with maximized average discriminant gain, one class may be well separated to other classes but other class pairs are very close in the feature space. As a result, only samples of one specific class can be accurately determined but samples of other classes cannot be well distinguished. This leads to an unbalanced and a low inference accuracy. To address this problem, this work proposes a novel metric based on the pair-wise discriminant gain, which maximizes the minimum discriminant gain of all class pairs. Thereby, the worst separated class pair can be well distinguished in the feature space, as shown in Fig. 1b. Hence, the inference accuracy of all classes is guaranteed to be high, leading to an enhanced inference performance.

In this work, a device-edge co-inference system is considered with multiple single-antenna devices and a single-antenna edge server. All devices observe a same wide-view target source and obtain a noise-corrupted local version of the ground-true sensory data. Then, each device extracts a low dimensional local feature vector from the high dimensional local sensory data. The AirComp technique is adopted to aggregate


Normalized Covariance
Centroid
G$_{12}$
G$_{13}$
G$_{23}$
Class 1
Class 2
Class 3
Average discriminant gain
(Maximize (G$_{12}$+G$_{13}$+G$_{23}$)/3)
Maximize minimum
pair-wise discriminant gain
(Maximize min{G$_{12}$,G$_{13}$,G$_{23}$})

(a) Maximize average discriminant gain

(b) Maximize minimum pair-wise discriminant gain

Fig. 1: The difference between maximizing overall discriminant gain and maximizing minimum discriminant gain.

all local feature vectors to suppress the sensing and channel noise. Specifically, all devices transmit an element of a same feature dimension over the same resource block. At the server, they are aggregated using the waveform superposition property and used to derive an estimate of the ground-true element. All elements are sequentially transmitted over multiple time slots. The criterion of minimum discriminant gain maximization is adopted for improving the inference accuracy. The detailed contribution of this work is summarized as follows.

- **Inference Accuracy Enhancement via Maximizing Minimum Pair-wise Discriminant Gain**: This paper proposes a new metric which maximizes the minimum pair-wise discriminant gain to tackle the issue of unbalanced inference accuracy in the case of average discriminant gain maximization. Under this criterion, the class pair with smallest distance in the feature space is well separated, leading to a balanced and enhanced achievable inference accuracy.
- **Joint Optimization of All Feature Elements**: Instead of separately optimizing the average discriminant gain of each feature element in the exisiting design [11], this paper jointly optimizes the minimum pair-wise discriminant gain of all feature elements by using multi-slot AirComp. Thereby, an extra degree of freedom, that the transmit power can be adaptively allocated to different feature elements with different contributions on the inference accuracy, is enabled. Under the criterion of jointly maximizing the minimum discriminant gains of all feature elements, a transmit power allocation problem for AirComp over multiple slots is proposed. To tackle this problem, variables transformation is first applied to derive an equivalent difference-of-convex (d.c.) problem, which is then solved using the typical method of successive convex approximation (SCA) [14].
- **Performance Evaluation**: Extensive experiments are made on a high-fidelity wireless sensing simulator provided in [15] by considering a wide-view human motion recognition task. Two inference models, i.e., a support vector machine (SVM) and a multi-peceptron layer (MLP) neural network are used. The experimental results show that the proposed scheme outperform the task-oriented scheme under the criterion of average discriminant gain maximization and the method based on traditional minimum mean square error (MMSE) metric. This verifies our theoretical analysis.

II. System Model

A. Network Model

Fig. 2 depicts the device-edge co-inference system, where there are K single-antenna ISAC devices sensing the same wide-view target and obtaining the real-time noise-corrupted sensory data. Then, local low-dimensional feature spaces are extracted from the raw sensory data at each device using principal component analysis (PCA) [16], which are further simultaneously transmitted to a single-antenna server for suppressing the sensing noise using AirComp. The number of extracted feature elements is denoted as M. In practice, the edge server may represent a highly mobile vehicle in a road and the ISAC devices represent radar sensors deployed at the road side.

Time-division multiple access (TDMA) is employed. At each time slot, an element of the same feature dimension is transmitted by each device and is aggregated at the server for denoising. The overall feature vector are sequentially transmitted over M time slots. Since the transmission time of one element is far less than the coherence time [17], the channels are assumed to be static during the transmission of all M elements. Particularly, the uplink channel gain of the k-th ISAC device is denoted as h_k. Without loss of generality, the server works as the central coordinator and has the ability to acquire the channel state information of all involved links.

B. Feature Distribution

Following [11], the PCA-extracted local feature vector is a noise-corrupted version of the ground-true one, as

$$\mathbf{x}_k = \mathbf{x} + \mathbf{d}_k, \ \ 1 \le k \le K, \tag{1}$$

where $\mathbf{x} = \{x_1, ..., x_m, ..., x_M\}$ is the ground-true feature vector, $\mathbf{d}_k = \{d_{k,1}, ..., d_{k,m}, ...d_{k,M}\}$ is the Gaussian sensing noise. As PCA is applied, different elements of $\mathbf{x}$ and $\mathbf{d}_k$ are independent. Consider a classification task with L classes, according to [11], the m element of $\mathbf{x}$ has a distribution of

$$x_m \sim \frac{1}{L}\sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,m}, \sigma_m^2\right), \tag{2}$$

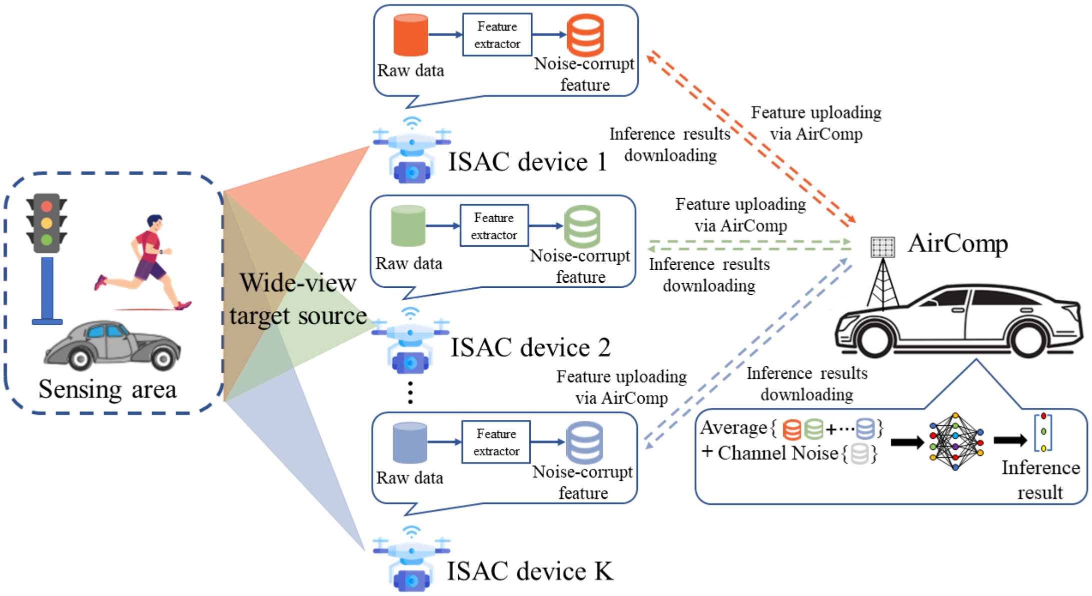


Fig. 2: Inference accuracy versus discriminant gain.

where $\mathcal{N}\left(\mu_{\ell,m},\sigma_m^2\right)$ represents the Gaussian distribution corresponding to the ℓ-th class, $\mu_{\ell,m}$ is the centroid of the ℓ-th class, σ_m^2 is the variance of m-th dimension. The m element of $\mathbf{d}_k$ has a distribution of

$$d_{k,m} \sim \mathcal{N}\left(0,\delta_{k,m}^2\right), \tag{3}$$

where $\delta_{k,m}^2$ is the sensing noise power. Thereby, the m-th element of $\mathbf{x}_k$, given by

$$x_{k,m} = x_m + d_{k,m}, \tag{4}$$

has a distribution of

$$x_{k,m} \sim \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}\left(\mu_{\ell,m},\sigma_m^2+\delta_{k,m}^2\right). \tag{5}$$

C. AirComp

To simultaneously access multiple devices to aggregate the local feature vectors for noise suppression, the technique of AirComp is adopted for reducing the communication overhead. For an arbitrary time slot m, all devices pre-code and transmit the m-th element of their local feature vectors, i.e., $\{x_{k,m}\}$ over the same resource block. Thereby, the received element at the server is given by

$$\hat{x}_m = \sum_{k=1}^{K} h_k b_{k,m} x_{k,m} + n, \quad 1 \le m \le M, \tag{6}$$

where $b_{k,m}$ is the precoding scalar at device k, n is the channel noise with a distribution of $n \sim \mathcal{N}\left(0,\delta_0^2\right)$, and δ_0 is the channel noise power.

Then, by substituting the distribution of $x_{k,m}$ in (5) into $\hat{x}_m$, its distribution can be derived as

$$\hat{x}_m \sim \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}\left(\hat{\mu}_{\ell,m},\hat{\sigma}_m^2\right), \tag{7}$$

where

$$\begin{cases} \hat{\mu}_{\ell,m} = \sum\limits_{k=1}^{K} h_k b_{k,m}\mu_{\ell,m}, \\ \hat{\sigma}_m^2 = \left(\sum\limits_{k=1}^{K} h_k b_{k,m}\right)^2 \sigma_m^2 + \sum\limits_{k=1}^{K} h_k^2 b_{k,m}^2 \delta_{k,m}^2 + \delta_0^2. \end{cases} \tag{8}$$

D. Inference Accuracy Measured by Discriminant Gain

In this work, the inference accuracy is adopted as the design criterion instead of the traditional MMSE in existing designs. Since the latter targets minimizing the distortion between the received feature vector and the ground-true one, it ignores that the same distortion level on different feature elements has different impacts on the inference accuracy [11]. However, the instantaneous inference accuracy is unknown in the design stage before the feature vector is inputted into the AI model. To address this issue, discriminant gain is adopted as the surrogate. Specifically, consider a classification task with L classes, whose feature distribution is in (7). For an arbitrary class pair, say classes ℓ and ℓ', the discriminant gain between the two classes is

$$G_{\ell,\ell'}(\mathbf{x}) = \sum_{m=1}^{M}\frac{\left(\hat{\mu}_{\ell,m}-\hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}. \tag{9}$$

The pair-wise discriminant gain in (9) represents the discrenibility of the two classes in the Euclidean feature space. With larger discriminant gain, the two classes are well separated, leading to a higher achievable inference accuracy.

III. Problem Formulation

Existing work targeting maximizing the average discriminant gain of all class pairs [10], [11]. This, however, causes an unbalanced and low accuracy as mentioned before and as shown in Fig. 1a. To address this issue, we propose a novel design criterion, which maximizes the minimum pair-wise discriminant gain over all class pairs. As a result, the calss pair with smallest distance can be well separated, resulting in an enhanced achievable inference accuracy for all classes, as shown in Fig. 1b. Besides, there are two kinds of constraints. One is the transmit power constraint for each device in each time slot. The other is the overall energy constraint of each device to transmit the whole feature vector. In summary, the problem is given by

$$
\text{(P1)}\qquad
\begin{aligned}
\max_{\{b_k\}} \ \min_{(\ell,\ell>\ell')} \ & G_{\ell,\ell'}(\mathbf{x}) = \sum_{m=1}^{M} \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}, \\
\text{s.t.} \ & b_{k,m}^2 \le P_k, \quad \forall (k,m), \\
& \sum_{m=1}^{M} b_{k,m}^2 \le \hat{P}_k, \ \forall k,
\end{aligned}
$$

where P_k is transmit power threshold in each time slot and $\hat{P}_k$ is total power constraint of all time slots.

IV. Minimum Discriminant Gain Maximization

(P1) is non-convex due to the complicated form of the objective function. To this end, the following variables transformation is applied to simplify (P1). For all class pairs $(\ell, \ell > \ell^{'})$, we introduce

$$
T_{m,\ell,\ell'} \le \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}, \quad \forall m. \tag{10}
$$

Then, denote

$$
T = \sum_{m=1}^{M} T_{m,\ell,\ell'}, \quad \forall (\ell, \ell > \ell^{'}). \tag{11}
$$

According to (11) and (12), it is easy to infer that

$$
T \le \sum_{m=1}^{M} \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}, \quad \forall (\ell, \ell > \ell^{'}). \tag{12}
$$

Next, by substituting T above into (P1), it can be equivalently derived as

$$
\text{(P2)}\qquad
\begin{aligned}
\max_{T,\{b_k\}} \ & T \\
\text{s.t.} \ & b_{k,m}^2 \le P_k, \quad \forall (k,m), \\
& \sum_{m=1}^{M} b_{k,m}^2 \le \hat{P}_k, \forall\, k, \\
& T = \sum_{m=1}^{M} T_{m,\ell,\ell'}, \quad \forall (\ell, \ell^{'}), \\
& T_{m,\ell,\ell'} \le \frac{\left(\hat{\mu}_{\ell,m} - \hat{\mu}_{\ell',m}\right)^2}{\hat{\sigma}_m^2}, \quad \forall (\ell, \ell > \ell^{'}, m).
\end{aligned}
$$

(P2) is still a non-convex problem. In the sequel, an equivalent d.c. form is first derived and the problem is addressed by the typical method of SCA [14]. To begin with, by substituting $\hat{\mu}_{\ell,m}$, $\hat{\mu}_{\ell',m}$ in (8) and $\hat{\sigma}_m^2$ in (9) into the fourth constrains of (P2) and with some simple derivations, we can get

$$
\begin{aligned}
& \left(\sum_{k=1}^{K} h_k b_{k,m}\right)^2 \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{T_{m,\ell,\ell'}} \\
& \ge \left(\sum_{k=1}^{K} h_k b_{k,m}\right)^2 \sigma_m^2 + \sum_{k=1}^{K} h_k^2 b_{k,m}^2 \delta_{k,m}^2 + \delta_0^2,
\end{aligned} \tag{13}
$$

for all $(\ell, \ell > \ell^{'}, m)$. In (13), it is observed that both sides of the inequality are convex. Define the left side term as

$$
Q_{m,l,l'}\left(\{b_{k,m}\}, T_{m,\ell,\ell'}\right) = \left(\sum_{k=1}^{K} h_k b_{k,m}\right)^2 \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{T_{m,\ell,\ell'}},
$$

for all $(\ell, \ell > \ell^{'}, m)$. It is no less than its first-order Taylor expansion, i.e., $Q_{m,\ell,\ell'}\left(\{b_{k,m}\}, T_{m,\ell,\ell'}\right) \ge \hat{Q}_{m,\ell,\ell'}^{[t]}\left(\{b_{k,m}\}, T_{m,\ell,\ell'}\right)$, where

$$
\begin{aligned}
& \hat{Q}_{m,\ell,\ell'}^{[t]}\left(\{b_{k,m}\}, T_{m,\ell,\ell'}\right) \\
& = \left(\sum_{k=1}^{K} h_k b_{k,m}^{[t]}\right)^2 \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{T_{m,\ell,\ell'}^{[t]}} \\
& + \sum_{k=1}^{K} \left[\frac{2\left(\sum\limits_{k=1}^{K} h_k b_{k,m}^{[t]}\right) h_k \left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{T_{m,\ell,\ell'}^{[t]}} \left(b_{k,m} - b_{k,m}^{[t]}\right) \right] \\
& - \left[\frac{\left(\sum\limits_{k=1}^{K} h_k b_{k,m}^{[t]}\right)\left(\mu_{\ell,m} - \mu_{\ell',m}\right)}{T_{m,\ell,\ell'}^{[t]}} \right]^2 \left(T_{m,l,l'} - T_{m,\ell,\ell'}^{[t]}\right).
\end{aligned} \tag{14}
$$

Thereby, the method of SCA can be utilized to solve (P2) via iteratively solving the following approximated convex problem, say (P3), by using solution in the last iteration as the reference point of the first-order Taylor expansion.

$$
\text{(P3)}\qquad
\begin{aligned}
\max \ & T \\
\text{s.t.} \ & b_{k,m}^2 \le P_k, \quad \forall (k,m), \\
& \sum_{m=1}^{M} b_{k,m}^2 \le \hat{P}_k, \forall\, k, \\
& T = \sum_{m=1}^{M} T_{m,\ell,\ell'}, \quad \forall (\ell, \ell^{'}), \\
& \left(\sum_{k=1}^{K} h_k b_{k,m}\right)^2 \sigma_m^2 + \sum_{k=1}^{K} h_k^2 b_{k,m}^2 \delta_{k,m}^2 + \delta_0^2 \\
& \le \hat{Q}_{m,\ell,\ell'}^{[t]}\left(\{b_{k,m}\}, T_{m,\ell,\ell'}\right), \quad \forall (\ell, \ell > \ell^{'}, m).
\end{aligned}
$$

Besides, (P3) can be solved by the common convex algorithms using e.g., the cvx toolbox [18].

TABLE I: Simulation Parameters

Parameter	**Value**
Number of ISAC devices, K	3
Channel noise variance, δ_0^2	150
feature noise variance, $\delta_{k,m}^2$	0.4
Number of dimension after PCA, M	12
Number of classes, L	4
Training data sizes, B	6400
Transmit power, P_k	$12mdB$
Variance of shadow fading, σ_ζ^2	$8dB$

V. Pefermance Evaluation

A. Experiment Setup

Consider a single-cell network with a radius of 450 meters, where the server equipped with a single-antenna BS is located at the center and K single-antenna devices are randomly distributed. The channel gain model for each device k is $H_k = |\phi_k h_k|^2$, where ϕ_k and h_k denote the large-scale and small-scale fading propagation coefficients, respectively. The large-scale propagation coefficient contains path loss and shadowing which is Gaussian distribution. The small-scale fading propagation coefficients is Rayleigh fading.

A human motion recognition inference task is considered, where there are four classes including child walking, child pacing, adult walking, and adult pacing. The model is trained using 6400 data samples and tested using 1600 data samples. The dataset is generated using the high-fidelity wireless sensing simulator proposed in [15]. Two models are adopted for the task. One is a SVM. The other is a MLP neural network with two hidden layers, each with 80 and 40 neurons respectively. In addition, Python 3.8 was used to implement the experiment on a Linux server. Table I lists the parameters' setting.

For comparison, we consider three schemes as follows.

- *Baseline*: All the parameters are allocated following the existing task-oriented AirComp scheme in [11], whose design criterion is to maximize the average discriminant gain.
- *Weighted subspace centroid*: All the parameters are allocated following the traditional AirComp scheme in [19], whose design criterion is MMSE.
- *Joint optimization of all feature elements (our proposal)*: All parameters are set follow the proposed scheme in Table I.

B. Experimental Results

In this part, the experimental results are shown. The relation between inference accuracy and the maximized minimum discriminant gain is first presented, followed by the comparison among the three schemes.

1) Inference accuracy v.s. discriminant gain: Fig. 3 illustrates the relation between the maximized minimum discriminant gain and inference accuracy for both SVM and MLP models. It shows that the inference accuracy rises as the minimum discriminant gain grows for both models. That's because larger minimum discriminant gain leads to a larger distance for arbitrary class pairs in the feature space, hence resulting in a larger inference accuracy. It is also shown that the SVM model outperforms the MLP model, as the training of the latter model is overfitting.

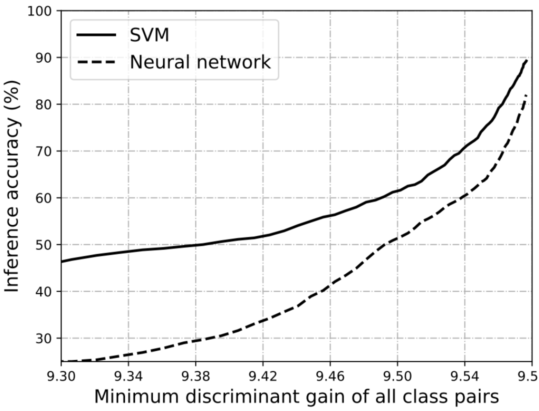


Fig. 3: Inference accuracy versus minimum discriminant gain of all class pairs

2) Inference accuracy v.s. number of devices: Fig. 4 shows the inference accuracy of the SVM and MLP models in terms of various device numbers. From the figure, the performance of the MMSE based scheme, say the scheme of weighted subspace centroid, decreases with an increasing number of devices, since larger devices leads to a lower MMSE due to the requirement of channel equalization for all devices, as well as MMSE cannot characterize the inference accuracy. The inference accuracy of both task oriented AirComp schemes grows as the number of devices increases. That's because the diversity of data can be fully utilized and the sensing noise of different devices are adaptively suppressed by allowing different receive powers of different devices. More importantly, our proposed scheme has the best performance due to the novel metric of maximizing the minimum discriminant gain of all class pairs and the joint optimization of all feature elements.

3) Inference accuracy v.s. transmit power: Fig. 5 presents the inference accuracy of the SVM and MLP models under various transmit power levels. For both models, the inference accuracy of all schemes increases as transmit power rises, since more transmit power can suppress the channel noise without doubt. For similar reasons, our proposed scheme has the best performance.

The experimental findings show that the proposed scheme has the best performance and verify our theoretical analysis.

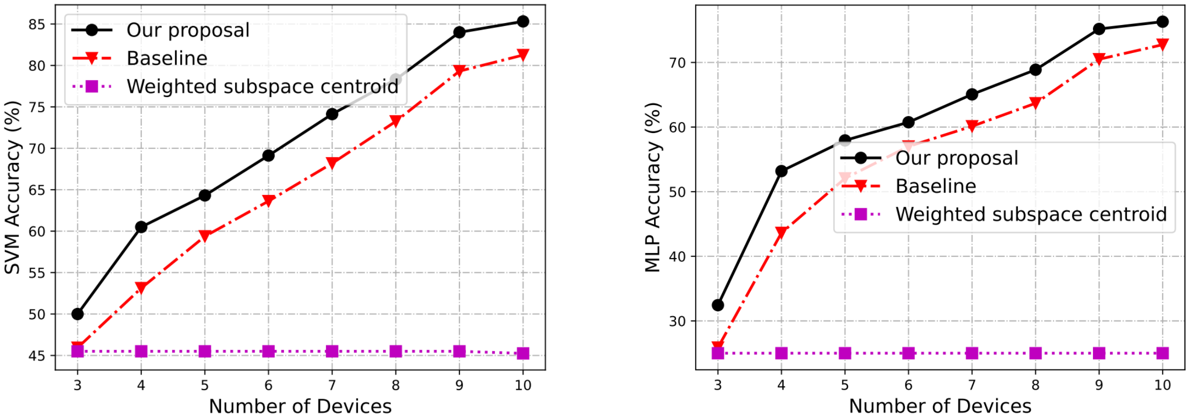
VI. Conclusion

We recommend the best task-oriented ISCC strategy for edge AI in this paper. For real-time inference tasks, accuracy is improved by distributing the communication power over various time slots to maximize the minimum discriminant gain of the received features.

For inference-task-oriented designs, this study offers up a number of promising new approaches. When there aren't enough resources to transfer all the features, one option is feature scheduling. Another is to increase the MIMO systems' inference accuracy.

References

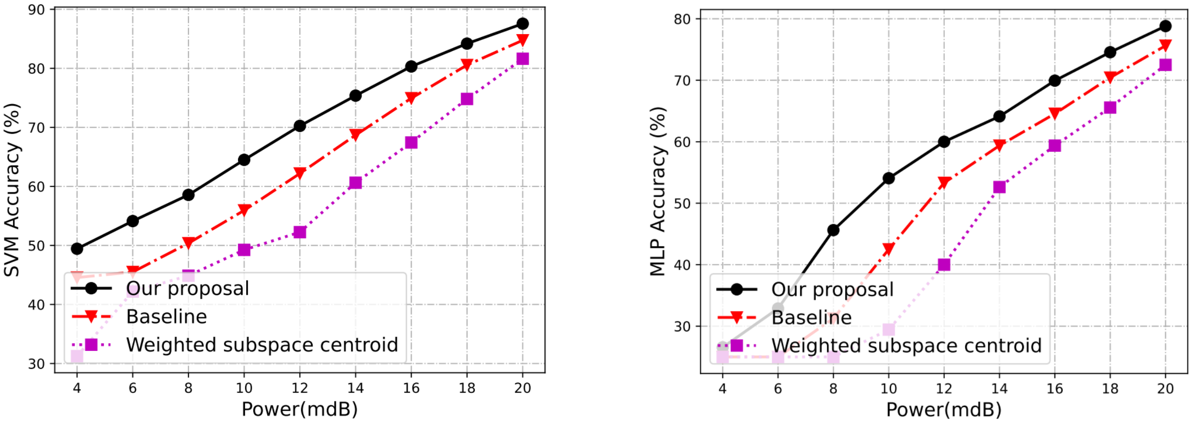
[1] J. Betz, H. Zheng, A. Liniger, U. Rosolia, P. Karle, M. Behl, V. Krovi, and R. Mangharam, "Autonomous vehicles on the edge: A survey on



(a) Inference accuracy with SVM versus number of devices

(b) Inference accuracy with MLP versus number of devices

Fig. 4: Inference accuracy comparison among different models under differenct number of devices.



(a) Inference accuracy with SVM versus transmit power

(b) Inference accuracy with MLP versus transmit power

Fig. 5: Inference accuracy comparison among different models under different transmit power.

autonomous vehicle racing," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 458–488, 2022.

[2] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 5031–5044, 2019.

[3] W. Zhang, D. Yang, H. Peng, W. Wu, W. Quan, H. Zhang, and X. Shen, "Deep reinforcement learning based resource management for dnn inference in industrial iot," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 7605–7618, 2021.

[4] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive feature transmission for split classification at the wireless edge," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[5] S. Yun, J.-M. Kang, S. Choi, and I.-M. Kim, "Cooperative inference of dnns over noisy wireless channels," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 8298–8303, 2021.

[6] M. Lee, G. Yu, and H. Dai, "Decentralized inference with graph neural networks in wireless communication systems," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[7] S. Hua, Y. Zhou, K. Yang, Y. Shi, and K. Wang, "Reconfigurable intelligent surface for green edge inference," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 2, pp. 964–979, 2021.

[8] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 20–26, 2020.

[9] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.

[10] 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," *arXiv e-prints*, p. arXiv:2207.00969, Jul. 2022.

[11] D. Wen, X. Jiao, P. Liu, G. Zhu, Y. Shi, and K. Huang, "Task-oriented over-the-air computation for multi-device edge ai," *arXiv preprint arXiv:2211.01255*, 2022.

[12] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. IEEE, May 2020, pp. 1–5.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[14] M. Razaviyayn, "Successive convex approximation: Analysis and applications," Ph.D. dissertation, University of Minnesota, 2014.

[15] G. Li, S. Wang, J. Li, R. Wang, X. Peng, and T. X. Han, "Wireless sensing with deep spectrogram network and primitive based autoregressive hybrid channel model," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 481–485.

[16] A. Alkandari and S. J. Aljaber, "Principle component analysis algorithm (pca) for image recognition," in *2015 Second International Conference on Computing Technology and Information Management (ICCTIM)*, 2015, pp. 76–80.

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

[18] M. Grant, S. Boyd, and Y. Ye, "Cvx users' guide," *[Online]. Available: http://www. stanford. edu/boyd/software. html*, 2009.

[19] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multimodal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Sep. 2019.