

# TinyAirNet: TinyML Model Transmission for Energy-efficient Image Retrieval from IoT Devices

Junya Shiraishi, *Member, IEEE*, Mathias Thorsager, Shashi Raj Pandey, *Member, IEEE*,  
and Petar Popovski, *Fellow, IEEE*

**Abstract**—This letter introduces an energy-efficient pull-based data collection framework for Internet of Things (IoT) devices that use Tiny Machine Learning (TinyML) to interpret data queries. A TinyML model is transmitted from the edge server to the IoT devices. The devices employ the model to facilitate the subsequent semantic queries. This reduces the transmission of irrelevant data, but receiving the ML model and its processing at the IoT devices consume additional energy. We consider the specific instance of image retrieval in a single device scenario and investigate the gain brought by the proposed scheme in terms of energy efficiency and retrieval accuracy, while considering the cost of computation and communication, as well as memory constraints. Numerical evaluation shows that, compared to a baseline scheme, the proposed scheme reaches up to 67% energy reduction under the accuracy constraint when many images are stored. Although focused on image retrieval, our analysis is indicative of a broader set of communication scenarios in which the preemptive transmission of an ML model can increase communication efficiency.

**Index Terms**—6G IoT networks, TinyML model, energy efficiency, wireless image retrieval, semantic query

## I. INTRODUCTION

ENERGY efficiency is one of the essential Key Performance Indicator (KPIs) for Internet of Things (IoT) networking of sixth generation (6G) communication systems [1], [2], in which Artificial Intelligence (AI) technology plays an important role in supporting new emerging applications, including extended reality, connected robotics, and automated systems [3]. As indiscriminate collection of data may lead to wasteful energy consumption for the IoT devices, an Machine Learning (ML) model should be introduced in this aspect to filter out irrelevant data to the current query. Then, it becomes crucial to design an energy-efficient communication protocol for IoT devices serving 6G applications, in which one needs to consider the energy cost caused by the introduction of AI/ML models as well as that of the primary radio circuit. In this context, a framework called SEMantic DATA Soucing (SEMDAS) introduced in [4] is an attractive approach, in which only the relevant data can be collected by broadcasting the semantic query to the IoT devices and by calculating a matching score with the help of an ML model [4], [5]. Further, it is important to consider the practical constraints of IoT devices, such as

the memory size, where the typical Micro Controller Unit (MCU) for IoT devices have extremely limited on-chip Static Random Access Memory (SRAM) memory (< 512KB) and flash storage (<2MB) [6], [7]. To solve this problem, this letter proposes the transmission of the TinyML model [8] from the Edge Server (ES) to the IoT devices considering the query content and timing, which we call Tiny Neural Network transmission over the Air (TinyAirNet). Specific designs for TinyML include MCUNet [9], EtinyNet [10], etc. Fig. 1 shows an example of TinyAirNet for AI-empowered data collection. In our framework, first, the ES transmits a TinyML model relating to the task of facilitating subsequent semantic query to the IoT devices, as shown in Fig. 1-1. The IoT device that receives this TinyML model stores it in the memory and exploits it to calculate the matching score (Fig. 1-2). After the processing using the TinyML model, each IoT device decides whether it transmits data or not to the ES based on the matching score. In the example of Fig. 1, the IoT device only transmits data  $I_3$  whose level of relevance (matching score) is high and suppresses the others (Fig. 1-3), by which the IoT device can save wasteful energy consumption. From these observations, we can clearly see the advantage of introducing the TinyML model in terms of energy reduction for data transmission and new challenges in energy-efficient protocol designs, in which we need to consider the additional cost caused by the introduction of the TinyML model. The TinyAirNet can be applicable to a variety of IoT data collection scenarios to reduce wasteful energy consumption, including object detection and vehicles at the edge [11], AI empowered IoT sensing [4], and wild-life animal monitoring/tracking [12]. The use case also includes the distributed implementation of the ML model as considered in the edge learning [13], [14], in which the ML or TinyML model is deployed over the edge network/device, and the edge device uploads/downloads the model to realize goal-oriented and semantic communication, etc. Furthermore, it could contribute to managing wireless access of IoT devices belonging to different communication classes, such as the coexistence of pull-based and push-based devices [15]. Our work is related to the SEMDAS [4], but we focus on the energy efficiency of the IoT data collection and rely on the TinyML model to send the semantic query.

The ML model transmission over the wireless channel has been considered in the edge networks [16] and in IoT networks in the context of semantic communication [17]. In [18], the authors investigated the energy cost of calculation using ML model with Quantized Neural Networks (QNNs) in

This work was supported partly by the Villum Investigator Grant “WATER” from the Velux Foundation, Denmark, and partly by the Horizon Europe SNS “6G-XCEL” project with Grant 101139194. (Corresponding author is Junya Shiraishi)

J. Shiraishi, M. Thorsager, S. R. Pandey, and P. Popovski are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: {jush, mdth, srp, petarp}@es.aau.dk)

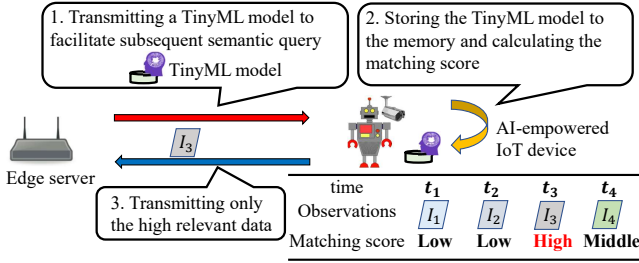


Fig. 1: An example of TinyAirNet for AI-empowered IoT data collection.

a federated learning setup. Moreover, ML model transmission has also been considered in 3GPP in the context of AI/ML model/data distribution and sharing over 5G system [19] with actual use cases/applications. To the best of our knowledge, this is the first work that proposes transmission of TinyML model to facilitate subsequent semantic query for the wireless image retrieval task [11] in IoT networks and demonstrates its effectiveness in terms of total energy consumption of IoT devices and retrieval accuracy. The main contribution of this work is TinyAirNet, an energy-efficient pull-based IoT image retrieval framework, and its associated protocol. The results show significant gains brought by the use of the TinyML model at the devices in terms of retrieval accuracy and total energy consumption, considering the communication/computation/memory costs.

## II. THE TINYAIRNET FRAMEWORK

### A. Scenario

We consider a scenario where an ES collects data from a single IoT device, such as a mobile robot, at the specific time instance based on the query requested by a user<sup>1</sup>. A mobile robot patrols a given area with a predetermined route and senses the environment with a camera at a predefined interval. The sensing data captured by the robot are stored in a storage  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ . Here,  $N$  denotes the number of stored images and  $\mathbf{x}_j \in \mathbb{R}^{M_C \times M_H \times M_W}$  is the  $j$ -th stored image, where  $M_C$  is number of channels,  $M_H$  and  $M_W$  are the height and width of images in pixel. The communication is conducted in pull-based manners, in which the ES first transmits the query to retrieve images with a specific object or entity. The mobile robot receiving its query transmits stored images to the ES through a shared wireless link. We assume that the downlink/uplink channel is error-free.

### B. Overview of TinyAirNet

TinyAirNet consists of three phases, as summarized below.

*a) TinyML Model Transmission Phase:* In this phase, the ES transmits the TinyML model related to the task to the IoT devices. Let  $\mathcal{M}$  be the ML model that the ES uses to conduct the task requested by an external user, in which  $\mathcal{M}$  is assumed to be the feature extractor for the specific task. Then, in order to share this model with the memory-constrained IoT devices, the ES first compresses  $\mathcal{M}$  to  $\hat{\mathcal{M}}$ . For instance, one can apply

a neural network transmission strategy as in [16], in which the ES compresses the model with a pruning and knowledge distillation considering the communication constraint. The IoT device receiving the TinyML model stores the parameter of its model into memory.

*b) Image Retrieval Phase:* In order to collect relevant images from the mobile robot, first, the ES transmits the threshold of similarity measure,  $V_{th}$ . The node receiving this information starts processing the observed data using the received TinyML model, i.e., feature extraction and similarity measure calculation, e.g., calculation of cosine similarity and checks whether the similarity measure of each image exceeds a threshold  $V_{th}$  or not. Note that, as we mentioned, we assume the shared TinyML is a feature extractor, which outputs a feature vector  $\mathbf{y}$  of an image  $\mathbf{x}$ , as  $\hat{\mathcal{M}} : \mathbf{x} \rightarrow \mathbf{y}$ . Let  $g(\cdot, \cdot)$  be the function to extract the similarity measure of an image. Then, the similarity measure of the  $j$ -th image  $\mathbf{x}_j$  observed by the IoT device, denoted as  $z_j \in [0, 1]$  is  $z_j = g(\hat{\mathcal{M}}(\mathbf{x}_j), \hat{\mathcal{M}}(\mathbf{q}))$ , where  $\mathbf{q}$  is a query image. Here, the subset of relevant images extracted by the TinyML model can be described as follows:

$$\mathcal{R} = \{\mathbf{x}_j | z_j \geq V_{th}, \forall \mathbf{x}_j \in \mathcal{D}\}. \quad (1)$$

If the ML model only has a smaller capability, the observed similarity value  $z_j$  is highly likely to deviate from the true one, denoted as  $\beta_j = g(\mathcal{M}(\mathbf{x}_j), \mathcal{M}(\mathbf{q})) \in [0, 1]$ . For example, the top-1 accuracy of ImageNet decreases as the model size or quantization value becomes smaller [10], [20]. In order to take this into account, we model  $z_j$  as

$$z_j = \beta_j + w_j, \quad (2)$$

where  $w_j$  is the observation noise, which follows  $w_j \sim \mathcal{N}(0, \sigma_{ML}^2)$ . Here,  $\sigma_{ML}$  is a standard deviation representing the model noise. Considering the trade-off between the ML model's size with different quantization levels and its accuracy, we set  $\sigma_{ML} = \frac{1}{b_q}$ , where  $b_q$  is the number of bits for the quantization of weights. The smaller (larger) value of  $\sigma_{ML}$  represents the high (low) bit quantization, by which the extracted feature of the ML model includes less (more) quantization error.

*c) Relevant Image Identification Phase:* After collecting images from the IoT device, the ES identifies the relevant images to the query by exploiting the larger ML model. Specifically, the ES first conducts feature extraction for each received image  $\hat{\mathbf{x}}$  using the large ML model, whose extracted feature vectors are denoted as  $\mathcal{M}(\hat{\mathbf{x}})$ , then calculates the similarity measure by comparing  $\mathcal{M}(\hat{\mathbf{x}})$  with the feature vector of query image  $\mathcal{M}(\mathbf{q})$ . The ES considers the received image as relevant if the similarity measure  $\chi_D = g(\mathcal{M}(\hat{\mathbf{x}}), \mathcal{M}(\mathbf{q}))$  is higher than the predetermined threshold  $\delta$ , i.e.,  $\chi_D \geq \delta$ .

Then, we define the retrieval accuracy  $\gamma$ , as the probability that the estimated relevant image set is the exact ones that are relevant to the query images. Formally, denoting by  $\mathcal{T}$  the subset of images whose true similarity measure is higher than  $\delta$ , and by  $\mathcal{S}$  the subset of relevant images that are successfully received by the ES, then the retrieval accuracy  $\gamma$  is:

$$\gamma = \Pr(\mathcal{T} = \mathcal{S}). \quad (3)$$

<sup>1</sup>Our proposed framework can be directly applicable to the multi-device setting, which will be kept for future work.

The retrieval accuracy  $\gamma$  becomes larger by collecting more images from the mobile robot; however, total energy consumption increases. The goal of this work is to elicit the gain of TinyML model introduction in terms of energy efficiency and retrieval accuracy.

### III. ENERGY MODEL

1) *Computation*: In this analysis, we assume QNNs [21], in which we use only fixed-point representations for both weights and activations for the calculation. The energy cost per inference using the ML model can be described as the summation of the energy cost of Dynamic Random Access Memory (DRAM) ( $E_{\text{DRAM}}$ ) and that of processing at the hardware itself ( $E_{\text{HW}}$ ) [21]. When the ML model can not be stored in the SRAM, we need to rely on off-chip DRAM, in which the larger energy cost is required for data movement [20]. This letter considers the ideal case, where the entire ML model can be stored in on-chip memory. With this assumption,  $E_{\text{DRAM}}$  can be expressed as the amount of energy consumed to access the input image for ML model, as expressed below:

$$E_{\text{DRAM}} = E_D \times (M_C \times M_H \times M_W \times \frac{b_{\text{in}}}{b_q}), \quad (4)$$

where  $E_D$  is the energy consumed per int  $b_q$  DRAM access and  $b_{\text{in}}$  is the number of bits for a single pixel [21].

The energy consumption for the read/write from/to the small local SRAM or Register file  $E_L$  is modeled to be equal to the energy of a single Multiply-accumulation (MUAC) operation  $E_{\text{MUAC}}$ , while accessing the main SRAM costs  $E_M = 2E_{\text{MUAC}}$  [21]. Then,  $E_{\text{HW}}$  can be modeled as the summation of the compute energy ( $E_C$ ), cost of the weight ( $E_W$ ), and the activation access ( $E_A$ ), as follows [21]:

$$E_{\text{HW}} = E_C + E_W + E_A, \quad (5)$$

with  $E_C = E_{\text{MUAC}} \times (N_c + 3 \times A_s)$ ,  $E_W = E_M \times N_s + E_L \times N_c / \sqrt{p}$ , and  $E_A = 2 \times E_M \times A_s + E_L \times N_c / \sqrt{p}$ , where  $N_c$  is the network complexity in the number of MUAC operations,  $N_s$  is the model size in the number of weights and biases,  $A_s$  is the total number of activations throughout the whole network, and  $p = 64 \times b_{\text{max}} / b_q$ , where  $b_{\text{max}}$  is the full-precision bit.

2) *Communication*: The IoT device consumes energy either in the transmit or the receive state. The power consumption at the receive/transmit state is denoted as  $\xi_R / \xi_T$  [W]. To simplify the analysis, we ignore the power consumed during idle periods. The energy consumed for receiving data can be calculated by considering the time required for receiving the TinyML model  $t_{\text{ML}}$  and query feature vector  $t_q^F$ . A model consists of model topology and model weight factors [19]. Then, the time required for TinyML model reception can be

$$t_{\text{ML}} = \frac{N_s b_q + b_{\text{ml}} + b_{\text{h}}^{\text{ML}}}{R_{\text{DL}}}, \quad (6)$$

where  $R_{\text{DL}}$  is the transmission rate for the downlink,  $b_{\text{ml}}$  is the number of bits required for the information of model/topology, and  $b_{\text{h}}^{\text{ML}}$  is the number of bits for a header in TinyML transmission. Here, we set  $b_{\text{ml}} = 0$ , as the size of the model topology is much smaller than the actual model size [19].

Similarly, the time for reception of the query feature vector can be described as  $t_q^F = (l_F + b_{\text{h}}^F) / R_{\text{DL}}$ , where  $b_{\text{h}}^F$  is the number of bits for header feature vector transmission and  $l_F$  is the size of feature vector. Here, for simplicity, we set  $b_{\text{h}}^{\text{ML}} = 0$  and  $b_{\text{h}}^F = 0$ . On the other hand, the energy consumed for transmitting data depends on the size of  $|\mathcal{R}| = \psi$  in Eq. (1). Then, the total energy consumed for the image retrieval task at the specific time instance from IoT device can be:

$$E_{\text{comm}} = \xi_R (t_{\text{ML}} + t_q^F) + \xi_T t_{\text{data}} \psi, \quad (7)$$

where  $t_{\text{data}}$  is the time required for data transmission of a single image and  $t_{\text{data}} = (b_{\text{in}} M_C M_H M_W + b_{\text{h}}) / R_{\text{UL}}$ , where  $R_{\text{UL}}$  is the data rate for uplink transmission.

### IV. ANALYSIS

We derive the total energy consumption and retrieval accuracy when we apply the TinyAirNet for the wireless image retrieval tasks. Here, total energy consumption is defined as the total amount of energy consumed by a single device, while the retrieval accuracy is defined by Eq. (3). In this analysis, we ignore the energy consumed for the reception of  $V_{\text{th}}$ , as it is much smaller, compared with the cost of TinyML model reception and image transmission.

Based on the energy model in Sec. III-1, the total energy consumed for computation can be expressed as:

$$E_{\text{comp}}^{\text{TinyAirNet}}(N) = N (E_{\text{DRAM}} + E_{\text{HW}} + n_F E_{\text{MUAC}}), \quad (8)$$

where  $n_F$  is the number of MUAC operations to calculate one similarity measure, which is  $n_F = l_F(l_F - 1) + 2l_F^2 + 2$ . Next, we derive the energy cost for communication based on the model described in Sec. III-2. Let  $P_{\text{th}}(V_{\text{th}})$  be the probability the similarity measure of an observed image  $x$  is equal to or higher than the threshold of  $V_{\text{th}}$ . According to the Eq. (2) the conditional probability of  $z$  given  $\beta$  can be expressed as  $p(z|\beta) = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} \exp\left(-\frac{(z-\beta)^2}{2\sigma_{\text{ML}}^2}\right) dz$ . Then,  $P_{\text{th}}(V_{\text{th}})$  can be expressed as follows:

$$P_{\text{th}}(V_{\text{th}}) = \int_0^1 Q\left(\frac{V_{\text{th}} - \beta}{\sigma_{\text{ML}}}\right) g_T(\beta) d\beta, \quad (9)$$

where  $Q(x)$  denotes the Q-function defined as  $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp(-\frac{u^2}{2}) du$  and  $g_T(\beta)$  is the distribution of true similarity value. Then, the probability of  $w$  out of  $N$  images satisfying the threshold of  $V_{\text{th}}$ , follows binomial distribution as:

$$P_D(w) = \binom{N}{w} P_{\text{th}}(V_{\text{th}})^w (1 - P_{\text{th}}(V_{\text{th}}))^{N-w}. \quad (10)$$

Since the value of  $w$  is a random variable, the expected total energy consumption for communication, including the energy cost for image transmission and for TinyML model and feature vector reception, can be expressed as follows:

$$E_{\text{comm}}^{\text{TinyAirNet}}(N) = \sum_{w=0}^N P_D(w) w \xi_T t_{\text{data}} + (t_{\text{ML}} + t_q^F) \xi_R. \quad (11)$$

With Eqs. (8) and (11), the total energy consumption of TinyAirNet can be expressed as follows:

$$E_{\text{total}}^{\text{TinyAirNet}}(N) = E_{\text{comp}}^{\text{TinyAirNet}}(N) + E_{\text{Data}}^{\text{TinyAirNet}}(N). \quad (12)$$

Now, we derive the retrieval accuracy of TinyAirNet  $\gamma_{\text{TinyAirNet}}$  defined by Eq. (3). Let us denote the probability that the true similarity measure of an image is equal to or higher than  $\delta$  as  $P_\delta$ , which can be expressed as  $P_\delta = \int_\delta^1 g_T(\beta) d\beta$ . We are then interested in the probability that an actual relevant image is successfully delivered to the ES, denoted as  $P_A$ , which is  $\text{Prob}(z \geq V_{\text{th}} | \beta \geq \delta)$ . This can be described as:

$$P_A = \frac{1}{P_\delta} \int_\delta^1 Q\left(\frac{V_{\text{th}} - \beta}{\sigma_{\text{ML}}}\right) g_T(\beta) d\beta. \quad (13)$$

Let  $\zeta$  be a random variable representing the number of actual relevant images observed by the IoT device. The probability that  $\zeta$  out of  $N$  images is actually relevant can be  $P_\zeta(\zeta) = \binom{N}{\zeta} P_\delta^\zeta (1 - P_\delta)^{N - \zeta}$ . Then, according to the definition in Eq. (3), the probability that the retrieval accuracy is one given  $\zeta$  can be described as  $P_A^\zeta$ . Finally, the expected retrieval accuracy for TinyAirNet can be described as

$$\gamma_{\text{TinyAirNet}} = \sum_{\zeta=0}^N P_A^\zeta P_\zeta(\zeta). \quad (14)$$

## V. NUMERICAL EVALUATION

### A. Simulation Setting

We conduct the simulation based on the description in Sec. II. We set  $R_{\text{UL}} = R_{\text{DL}} = 10^5$  [bps], and  $b_{\text{max}} = 16$  [21]. Assuming NB-IoT,  $\xi_T$  and  $\xi_R$  are set to be 170 mW and 160 mW, respectively [22], [23]. The values of  $E_{\text{MUAC}}$ ,  $E_L$ ,  $E_M$ , and  $E_D$  are set to  $E_{\text{MUAC}} = 3.7\text{pJ} \times (b_q/b_{\text{max}})^{1.25}$ ,  $E_L = 3.7\text{pJ} \times (b_q/b_{\text{max}})$ ,  $E_M = 2 \times 3.7\text{pJ} \times (b_q/b_{\text{max}})$ , and  $E_D = 128 \times 3.7\text{pJ} \times (b_q/b_{\text{max}})$  [21]. For simplicity of analysis, we generate  $g_T(\beta)$  based on uniform distribution with the range of  $[0, 1]$ . Here, we select the parameter of TinyML model from EtinyNet1.0 [10] as a showcase of performance evaluations, where  $N_s$  and  $N_c$  are 0.976 M and 117 M, and we set  $A_s$  to 4.309 M based on our calculation for the base EtinyNet1.0 architecture. The image size is set to  $(M_C, M_H, M_W, b_{\text{in}}) = (3, 256, 256, 8)$ , considering the STM32F746 MCUs [10] and the feature vector is set to  $l_F = 1000$  based on the size of a fully connected layer of EtinyNet [10]. We conducted a simulation  $10^4$  times.

As a baseline scheme, we consider the simple offloading scheme, in which the IoT device transmits all observations without computations. In this scheme, as the ES can collect all images without collisions, retrieval accuracy is always 1, while the total energy consumption of the baseline is

$$E_{\text{Baseline}} = N t_{\text{data}} \xi_T. \quad (15)$$

### B. Numerical Results

Fig. 2 shows the total energy consumption and retrieval accuracy of TinyAirNet against threshold of  $V_{\text{th}}$ , where we set  $b_q = 8$ ,  $\delta = 0.9$ , and  $N = 10$ . From this figure, first we can see that the results for TinyAirNet obtained by theoretical analysis coincide with that of simulation results for both total energy consumption and retrieval accuracy, which validates our analysis. Next, from Fig. 2, we can see the basic trade-off between the retrieval accuracy and the total

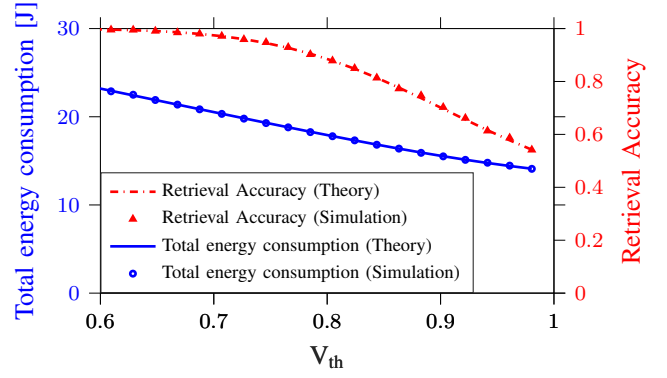


Fig. 2: Total energy consumption and retrieval accuracy of TinyAirNet against the threshold of  $V_{\text{th}}$ .

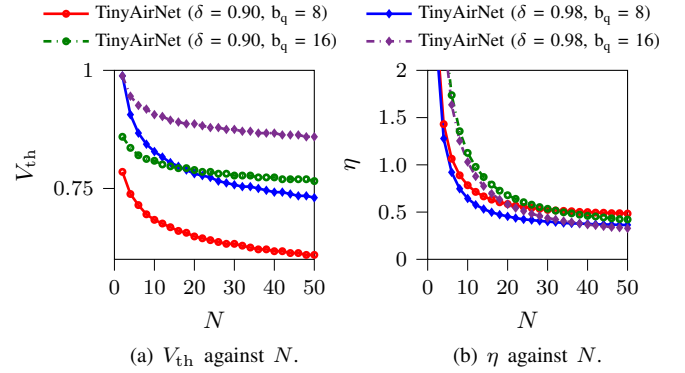


Fig. 3:  $V_{\text{th}}$  and  $\eta$  against  $N$ .

energy consumption through the value of  $V_{\text{th}}$ : the smaller (higher)  $V_{\text{th}}$  realizes higher (lower) retrieval accuracy because of the increasing (decreasing) number of available images at the ES, while it increases (decreases) the energy consumption for image transmission.

Because of this trade-off, we compare the minimum energy consumption of TinyAirNet that achieves the target retrieval accuracy  $\gamma_{\text{th}}$  with the baseline. To this end, first, we obtain the optimal parameter of  $V_{\text{th}}^{\text{opt}}$  that minimizes the total energy consumption of TinyAirNet under the constraint of  $\gamma_{\text{TinyAirNet}} \geq \gamma_{\text{th}}$ , as expressed below:

$$V_{\text{th}}^{\text{opt}}(N) = \min_{V_{\text{th}} \in [0, 1]} E_{\text{total}}^{\text{TinyAirNet}}(N) \quad (16)$$

s.t.  $\gamma_{\text{TinyAirNet}} \geq \gamma_{\text{th}}$ ,

where we vary the value of  $V_{\text{th}}$  with a step of  $\frac{1}{2^8}$ . Through these evaluations, we set  $\gamma_{\text{th}} = 0.98$ ,  $b_q = 8, 16$ , and  $\delta = 0.90, 0.98$ . Fig. 3a shows the optimal threshold of ( $V_{\text{th}}^{\text{opt}}$ ) against  $N$ . First, we can see that the value of  $V_{\text{th}}^{\text{opt}}$  is always equal to or lower than  $\delta$ , and  $V_{\text{th}}^{\text{opt}}$  becomes lower when the value of  $b_q$  is smaller and/or the value of  $N$  is larger. This is because as the value of  $b_q$  becomes smaller, the observed similarity measure is more likely to deviate from the true one, by which the ES needs to select the lower value of  $V_{\text{th}}$  in order to ensure the retrieval accuracy to exceed the threshold  $\gamma_{\text{th}}$ . Likewise, as  $N$  increases, the probability of these deviations increases, which requires the ES to set the lower threshold  $V_{\text{th}}$ .

Fig. 3b shows the energy consumption ratio, defined as  $\eta = E_{\text{total}}^{\text{TinyAirNet}}(V_{\text{th}}^{\text{opt}})/E_{\text{Baseline}}$ , against  $N$ . Note that the energy consumption ratio becomes lower than 1 if the energy consumption of TinyAirNet is smaller than that of the baseline scheme. From Fig. 3b, first, we can see that, when the value of  $N$  is smaller (e.g.,  $N < 5$ ),  $\eta$  is higher than 1, i.e., the total energy consumption of TinyAirNet scheme is larger than that of the baseline. This is because the proposed scheme consumes a large amount of energy for receiving the TinyML model and extracting the similarity measure to suppress a smaller number of image transmissions, while the baseline scheme only transmits a small amount of stored images without consuming energy for TinyML model introduction. However, as the value of  $N$  becomes larger, we can see that the proposed scheme outperforms the baseline in terms of total energy consumption while satisfying the constraint of retrieval accuracy. This is because the TinyAirNet can suppress wasteful image transmissions, which are unrelated to the query image, thanks to the introduction of the task-related TinyML model, leading to a reduction of the overall energy consumption, while the baseline scheme needs to consume substantial energy for the image transmissions. Next, we can see that  $\eta$  becomes lower as the value of  $\delta$  increases. This is because when the value of  $\delta$  becomes higher, the ES can set a relatively larger  $V_{\text{th}}$ , by which the IoT device can suppress the transmission of a large portion of images that are not relevant for the query image. Finally, compared with the results for  $b_q = 8$  and  $b_q = 16$ , we can see that the gain of TinyAirNet for  $b_q = 8$  is higher (lower) than for  $b_q = 16$  for the range of smaller (higher) value of  $N$ . For the smaller range of  $N$ , most of the energy is consumed for receiving the TinyML model and its processing, for which smaller quantization like  $b_q = 8$  is preferable to reduce these costs. On the other hand, for the larger range of  $N$ , reducing the transmission cost becomes important for the overall energy reduction, in which case  $b_q = 16$  is preferable as it enables the ES to set a relatively large  $V_{\text{th}}$ .

## VI. CONCLUSIONS

In this letter, focusing on wireless image retrieval from the single IoT device, we have investigated how we can reduce overall energy consumption when a specific query image is given. In order to reduce wasteful data transmission from the IoT devices that do not observe the desired image, we have proposed TinyAirNet, in which a TinyML model is transmitted from the ES to the IoT device, filtering out undesired image transmission. We have derived a theoretical equation expressing total energy consumption and retrieval accuracy of TinyAirNet in order to analyze the performance and effectiveness of our proposed scheme. Our numerical results have revealed that the proposed TinyAirNet achieves high energy efficiency while maintaining high retrieval accuracy, especially when the number of stored images is large.

This work demonstrated the effectiveness of our scheme based on a theoretical concept, which we deem to be sufficient at this initial stage and, based on the promising results, we will carry out experimental evaluation in our future work.

More generally, beyond image retrieval, this analysis can be generalized to encompass a broader set of scenarios in which a preemptive transmission of an ML model can make the subsequent communication more efficient.

## REFERENCES

- [1] K. Sheth *et al.*, "A taxonomy of AI techniques for 6G communication networks," *Comput. Commun.*, vol. 161, pp. 279–303, 2020.
- [2] E. C. Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Comput. Networks*, vol. 190, p. 107930, 2021.
- [3] M. Z. Chowdhury *et al.*, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 957–975, 2020.
- [4] K. Huang, Q. Lan, Z. Liu, and L. Yang, "Semantic data sourcing for 6G edge intelligence," *IEEE Commun. Mag.*, vol. 61, no. 12, pp. 70–76, 2023.
- [5] A. E. Kalor, P. Popovski, and K. Huang, "Random access protocols for correlated IoT traffic activated by semantic queries," in *Proc. 2023 21st Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*. IEEE, 2023, pp. 643–650.
- [6] S. Soro, "TinyML for ubiquitous edge AI," *arXiv preprint arXiv:2102.01255*, 2021.
- [7] Y. Abadade *et al.*, "A comprehensive survey on TinyML," *IEEE Access*, vol. 11, pp. 96 892–96 922, 2023.
- [8] L. Dutta and S. Bharali, "TinyML meets IoT: A comprehensive survey," *Internet of Things*, vol. 16, p. 100461, 2021.
- [9] J. Lin, W.-M. Chen, Y. Lin, C. Gan, S. Han *et al.*, "MCUNet: Tiny deep learning on IoT devices," *Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 11 711–11 722, 2020.
- [10] K. Xu, Y. Li, H. Zhang, R. Lai, and L. Gu, "EtinyNet: Extremely tiny network for TinyML," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 4, 2022, pp. 4628–4636.
- [11] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 89–100, 2020.
- [12] A. R. Elias, N. Golubovic, C. Krintz, and R. Wolski, "Where's the bear? automating wildlife image processing using IoT and edge cloud systems," in *Proc. 2nd Int. Conf. Internet Things Des. Implementation*, 2017, pp. 247–258.
- [13] W. Xu *et al.*, "Edge learning for 5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 9–39, 2023.
- [14] 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, 2021.
- [15] S. Cavallero *et al.*, "Coexistence of pull and push communication in wireless access for IoT devices," *arXiv preprint arXiv:2404.07650*, 2024.
- [16] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "AirNet: Neural network transmission over the air," *IEEE Trans. Wireless Commun.*, 2024.
- [17] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 142–153, 2020.
- [18] M. Kim, W. Saad, M. Mozaffari, and M. Debbah, "Green, quantized federated learning over wireless networks: An energy-efficient design," *IEEE Trans. Wireless Commun.*, 2023.
- [19] 3rd Generation Partnership Project, "5G system (5GS); study on traffic characteristics and performance requirements for AI/ML model transfer," 3GPP, Tech. Rep. 22.874, V18.2.0, 2021.
- [20] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [21] B. Moons *et al.*, "Minimum energy quantized neural networks," in *Proc. 2017 51st Asilomar Conf. Signals, Syst., Comput.* IEEE, 2017, pp. 1921–1925.
- [22] "LPWA BG96 cat m1/NB1/EGPRS module," <https://www.quectel.com/product/lpwa-bg96-cat-m1-nb1-egprs/>, accessed: June. 13, 2024.
- [23] S. M. Z. Khan *et al.*, "An empirical modeling for the baseline energy consumption of an NB-IoT radio transceiver," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14 756–14 772, 2021.