

Attention-aware Semantic Communications for Collaborative Inference

Jiwoong Im¹, Nayoung Kwon¹, Taewoo Park¹, Jiheon Woo¹, Jaeho Lee¹, and Yongjune Kim¹

¹Affiliation not available

March 04, 2024

Abstract

We propose a communication-efficient collaborative inference framework in the domain of edge inference, focusing on the efficient use of vision transformer (ViTs) models. The partitioning strategy of conventional collaborative inference fails to reduce communication cost because of the inherent architecture of ViTs maintaining consistent layer dimensions across the entire transformer encoder. Therefore, instead of employing the partitioning strategy, our framework utilizes a lightweight ViT model on the edge device, with the server deploying a complicated ViT model. To enhance communication efficiency and achieve the classification accuracy of the server model, we propose two strategies: 1) attention-aware patch selection and 2) entropy-aware image transmission. Attention-aware patch selection leverages the attention scores generated by the edge device's transformer encoder to identify and select the image patches critical for classification. This strategy enables the edge device to transmit only the essential patches to the server, significantly improving communication efficiency. Entropy-aware image transmission uses min-entropy as a metric to accurately determine whether to depend on the lightweight model on the edge device or to request the inference from the server model. In our framework, the lightweight ViT model on the edge device acts as a semantic encoder, efficiently identifying and selecting the crucial image information required for the classification task. Our experiments demonstrate that the proposed collaborative inference framework can reduce communication overhead by 68 % with only a minimal loss in accuracy compared to the server model.

Attention-aware Semantic Communications for Collaborative Inference

Jiwoong Im, Nayoung Kwon, Taewoo Park, Jiheon Woo, Jaeho Lee, and Yongjune Kim

Abstract—We propose a communication-efficient collaborative inference framework in the domain of edge inference, focusing on the efficient use of vision transformer (ViTs) models. The partitioning strategy of conventional collaborative inference fails to reduce communication cost because of the inherent architecture of ViTs maintaining consistent layer dimensions across the entire transformer encoder. Therefore, instead of employing the partitioning strategy, our framework utilizes a lightweight ViT model on the edge device, with the server deploying a complicated ViT model. To enhance communication efficiency and achieve the classification accuracy of the server model, we propose two strategies: 1) attention-aware patch selection and 2) entropy-aware image transmission. Attention-aware patch selection leverages the attention scores generated by the edge device’s transformer encoder to identify and select the image patches critical for classification. This strategy enables the edge device to transmit only the essential patches to the server, significantly improving communication efficiency. Entropy-aware image transmission uses min-entropy as a metric to accurately determine whether to depend on the lightweight model on the edge device or to request the inference from the server model. In our framework, the lightweight ViT model on the edge device acts as a semantic encoder, efficiently identifying and selecting the crucial image information required for the classification task. Our experiments demonstrate that the proposed collaborative inference framework can reduce communication overhead by 68% with only a minimal loss in accuracy compared to the server model.

Index Terms—Collaborative inference, edge computing, edge inference, Internet of Things (IoT), semantic communications, split inference, vision transformer.

I. INTRODUCTION

The rapid advancement of computational resources, coupled with the proliferation of massive datasets, has significantly enhanced the practicality of artificial intelligence (AI) services. Integrating AI techniques with edge devices, including smartphones, wearable devices, and Internet of things (IoT) devices, seeks to seamlessly incorporate AI services into a wide range of daily life. This effort to advance AI technologies in the domain of edge computing is commonly known as edge AI [1]–[3].

An important research theme in edge AI is *edge inference*, focused on efficiently executing inference tasks within the edge network [2]–[5]. Traditionally, raw data is sent from edge devices (clients) to a server, where a complicated model

conducts the inference task, i.e., *server-based inference*. However, this method incurs significant communication overhead, particularly in scenarios dealing with large volumes of raw data [3], [6]. An alternative is *on-device inference*, which executes the inference task directly on resource-constrained devices, thereby minimizing communication costs. However, this approach often leads to lower performance due to the limited computational capabilities of edge devices [3], [4].

To address the dual challenges of excessive communication overhead and limited computational resources, the concept of *collaborative inference* has been introduced [2]–[7]. This strategy involves dividing a deep neural network (DNN) model into separate parts for the edge device and the server. Within this framework, the edge device first uses its component to extract features from the raw data and then transmits them to the server. As these extracted features are typically more compressed than the raw data, the communication cost can be aggressively reduced. The server then utilizes these features and its portion of the model to determine the final inference result, which is sent back to the device [5]. The selection of the split point is critical as it significantly impacts the computational load on the device and the communication overhead [3]. This approach is also known as *split inference* [5], [8] and *device-edge co-inference* [3], [7]. Notably, collaborative inference is closely connected to *semantic communications* [9]–[14], considering that the extracted features are essentially semantic information tailored for the inference task.

Transformers, originally developed for natural language processing (NLP) [15], have been widely adopted across multiple domains. Particularly, the vision transformer (ViT) [16], [17] has demonstrated superior performance and efficiency in image classification tasks. However, the deployment of ViTs on resource-constrained edge devices is challenging due to their substantial model size and intensive computational requirements [18].

In collaborative inference scenarios, the strategy of partitioning ViT models fails to effectively reduce communication overhead. This limitation stems from the inherent architecture of ViTs, which maintains consistent layer dimensions across the entire transformer encoder [16], in contrast to convolutional neural network (CNN) models whose intermediate layer dimensions can be significantly smaller than the raw data dimensions. Consequently, partitioning a ViT model for collaborative inference still requires the transmission of significant data volumes, thereby failing to substantially mitigate the communication load.

We propose an efficient collaborative inference framework utilizing pre-trained ViT models. Instead of partitioning a

J. Im, N. Kwon, T. Park, J. Woo, J. Lee, and Y. Kim are with the Department of Electrical Engineering, Pohang University of Science and Technology (POSTECH), Pohang 37673, South Korea (e-mail: {jw3562, kwonna, parktaewoo, jhwoo1997, jaeho.lee, yongjune}@postech.ac.kr).

TABLE I
THE COMPARISON OF DEiT MODEL PARAMETERS AND CLASSIFICATION ACCURACY ON IMAGENET DATASET [17]

Model	Parameters (million)	Classification Accuracy (%)
DeiT-Tiny	5	72.2
DeiT-Small	22	79.8
DeiT-Base	86	81.8

single model, the edge device operates a lightweight ViT model (e.g., DeiT-Tiny), while the server utilizes a complex ViT model (e.g., DeiT-Base). As shown in Table I, DeiT-Tiny (DeiT-Ti) is notably lightweight, making it suitable for edge deployment. However, its classification accuracy is approximately 10% lower than that of DeiT-Base (DeiT-B) [17].

Our objective is to develop a collaborative inference strategy that achieves classification accuracy comparable to the server model while also minimizing communication overhead between the edge device and the server. This strategy is designed to leverage the strengths of both models: the efficiency and low resource demand of the tiny model at the edge, and the higher classification accuracy of the base model at the server. In our proposed framework, the edge device utilizes its tiny model to conduct an initial inference without transmitting the image to the server. Subsequently, the edge device assesses whether to accept this initial inference or to send the image to the server for a more accurate inference using the base model. By doing so, we aim to achieve an optimal trade-off between classification accuracy and communication cost in edge-server collaborative systems.

To enhance communication efficiency in our framework, we propose two primary strategies: 1) attention-aware patch selection, which involves selectively transmitting only the most relevant patches of the image, and 2) entropy-aware image transmission, where the decision to transmit the image to the server is determined by the level of uncertainty or confidence in the edge device's initial inference.

- *Attention-aware patch selection*: In cases where the edge device needs to transmit the image to the server, our strategy is to transmit only the essential patches that are crucial for classification, rather than the entire image. This selective transmission approach is guided by the *attention scores* from the *class token* [16] to the image patches, as processed by the tiny model. These attention scores indicate the relevance of each patch to the classification task. We investigate several techniques to effectively select these key patches based on the attention scores, showing that the tiny model is capable of accurately identifying the most informative patches, even in cases where it may not classify the image correctly. Our experimental results validate that this approach enables the base model to maintain its classification accuracy, although it processes only selectively transmitted patches. Hence, attention-aware patch selection can effectively reduce the communication cost between the edge device and the server while maintaining classification accuracy.
- *Entropy-aware image transmission*: It is critical for the edge device to make an accurate determination between

relying on the tiny model's inference and requesting the inference of the more accurate base model. This decision is particularly important due to the communication overhead involved in transmitting images. Ideally, the edge device should transmit only those images incorrectly classified by the tiny model while avoiding the transmission of images where the initial inference is correct. To facilitate this decision, we utilize the *min-entropy* derived from the softmax output values of the classification head in the tiny ViT model. Our experimental findings reveal that the decisions based on the min-entropy yield higher accuracy compared to those based on the Shannon entropy.

In the context of semantic communications, our framework is aptly characterized as *attention-aware semantic communications*. The tiny model's *transformer encoder* on the edge device acts as a *semantic encoder*, particularly when the decision is made to transmit the image to the server. The edge device leverages attention scores generated by the transformer encoder to identify the most essential image patches, which are critical for accurate classification. Interestingly, our findings reveal that the tiny model functions effectively as a semantic encoder, in spite of its lower classification accuracy compared to the base model.

Moreover, the proposed collaborative inference framework offers the advantage of reducing the computational complexity on the server model, as the server's inference is conducted solely on the selected patches. While our primary goal focuses on minimizing communication overhead between the edge device and the server, this framework also yields the ancillary benefit of server-side computational efficiency.

The rest of this paper is organized as follows. Section II offers a brief overview of the ViT and related work. Section III details our collaborative inference framework. Section IV and V present our main contributions, which include attention-aware patch selection and entropy-aware image transmission, respectively. Section VI provides experimental results, followed by conclusions in Section VII.

II. BACKGROUNDS

A. Vision Transformer

The ViT [16] is a transformer-based model for computer vision tasks, setting a standard in vision models. A simplified overview of the ViT model is shown in Fig. 1. An input image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ is reshaped into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) , C , and (P, P) denote the resolution of the original image, the number of channels, and the resolution of each image patch, respectively. Note that $N = \frac{HW}{P^2}$ is the resulting number of patches. These patches are then linearly projected to a consistent dimension D across the transformer layers via $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$. The input embedding of the ViT's transformer encoder $\mathbf{z}_0 \in \mathbb{R}^{(N+1) \times D}$ is given by

$$\mathbf{z}_0 = [\mathbf{x}_{\text{cls}}; \mathbf{x}_p^1 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad (1)$$

where \mathbf{E}_{pos} denotes the standard learnable position embedding. The *class token* $\mathbf{z}_0^0 = \mathbf{x}_{\text{cls}} \in \mathbb{R}^{1 \times D}$ is particularly prepended

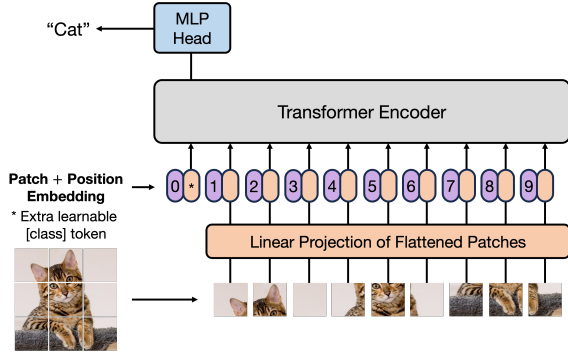


Fig. 1. Overview of the ViT model [16].

to the sequence of embedded patches [16]. This class token is crucial in classification tasks, serving as a key element in aggregating the information from the entire sequence of patches for the final classification output.

The transformer encoder is composed of alternating layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks as follows:

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1, \dots, L; \quad (2)$$

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad l = 1, \dots, L; \quad (3)$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0), \quad (4)$$

where LN represents the layer normalization. In particular, the image representation \mathbf{y} is the encoder output of the class token \mathbf{z}_0^0 . This image representation \mathbf{y} then serves as the input for the MLP head as shown in Fig. 1.

In the MSA block for an input sequence $\mathbf{z} \in \mathbb{R}^{(N+1) \times D}$, a weighted sum of all values \mathbf{v} is computed using query \mathbf{q} , key \mathbf{k} , and value \mathbf{v} , where $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{(N+1) \times D_h}$. The standard self-attention (SA) is formalized as follows [15], [16]:

$$[\mathbf{q}, \mathbf{k}, \mathbf{v}] = \mathbf{z} \mathbf{U}_{\mathbf{qkv}}, \quad (5)$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{qk}^\top}{\sqrt{D_h}}\right), \quad (6)$$

$$\text{SA}(\mathbf{z}) = \mathbf{A} \mathbf{v}, \quad (7)$$

where $\mathbf{U}_{\mathbf{qkv}} \in \mathbb{R}^{D \times 3D_h}$ and $\mathbf{A} \in \mathbb{R}^{(N+1) \times (N+1)}$ denote the projection matrix and the attention weight matrix, respectively. The MSA extends the standard self-attention (SA) by performing H parallel SA operations (i.e., heads) and concatenating their outputs [16]:

$$\text{MSA}(\mathbf{z}) = [\text{SA}_1(\mathbf{z}), \dots, \text{SA}_H(\mathbf{z})] \mathbf{U}_{\text{MSA}}, \quad (8)$$

where \mathbf{U}_{MSA} is the projection matrix for the MSA output.

B. Related Work

Prior work on collaborative inference has primarily focused on CNN architectures [2]–[7]. The exploration of collaborative inference with transformer models has been limited because of the consistent dimension of the MSA blocks in the transformer encoder.

Recent studies have delved into on-device inference utilizing lightweight ViT models [18]–[20], yet achieving the accuracy

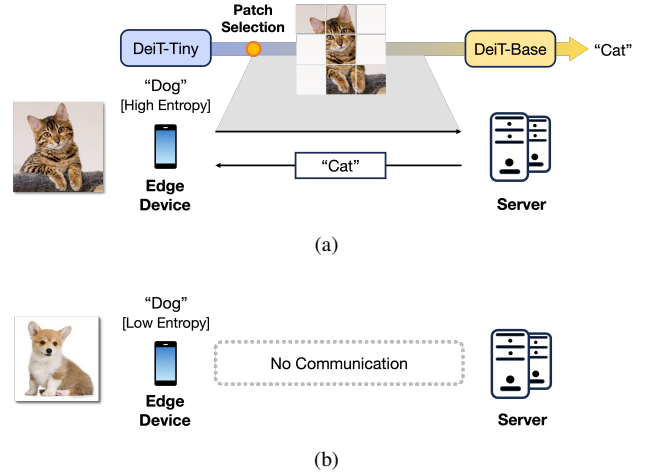


Fig. 2. The proposed collaborative inference framework utilizing pre-trained ViT models (DeiT-Tiny on client and DeiT-Base on server). (a) Case 1: The edge device (client) transmits the selected image patches to the server when the entropy of client’s inference is high. (b) Case 2: The edge device confirms its tiny model’s classification outcome when the entropy of edge device’s inference is low.

of server models is challenging. To improve classification accuracy of the edge device, an ensemble strategy employing multiple edge devices was proposed in [21]. In this approach, a central edge device disseminates raw data to nearby edge devices, which then conduct inference using compact ViT models and return their intermediate inference results to the central edge device. The central edge device obtains the final inference outcome by aggregating the received intermediate results. However, this approach leads to increased communication overhead among the edge devices.

In [22], the authors proposed a collaborative inference by leveraging the masked autoencoder (MAE) [23]. The edge device selects image patches *randomly* and transmits them to the server. The server then reconstructs the entire image using the decoder of MAE and performs classification on this reconstructed image. In spite of the benefit of reducing computational demands on the edge devices, the classification accuracy is compromised by the random selection of image patches.

Transformers have been utilized in semantic communications for the transmission of text [24] and images [25]. Nonetheless, these works do not pertain to classification tasks. Our approach distinctively employs attention scores, particularly for the class token, setting our work apart from existing research in semantic communications.

III. COLLABORATIVE INFERENCE FRAMEWORK BASED ON TRANSFORMER MODELS

We propose a collaborative inference framework that utilizes pre-trained ViT models. This framework is designed to achieve server-level classification accuracy with minimized communication overhead between the edge device and the server.

Due to the consistent layer dimensions of ViTs, conventional methods of collaborative inference [2]–[7], which typically partition a single CNN model, are ineffective at reducing communication costs for ViT models. As a solution, we employ

Algorithm 1 Proposed Collaborative Inference Framework

Input: Images $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$.

Output: Classification results $\{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}$.

```

1: for  $i = 1 : n$  do
2:    $y_c^{(i)} := f_c(\mathbf{x}^{(i)})$            ▷ Inference on edge device
3:   Client computes entropy  $g(\mathbf{x}^{(i)})$ 
4:   if  $g(\mathbf{x}^{(i)}) \geq \eta$  then
5:      $\tilde{\mathbf{x}}^{(i)} := \text{patch-selection}(\mathbf{x}^{(i)})$ 
6:     Client transmits  $\tilde{\mathbf{x}}^{(i)}$  to server
7:      $y_s^{(i)} := f_s(\tilde{\mathbf{x}}^{(i)})$        ▷ Inference on server
8:     Server transmits  $y_s^{(i)}$  to client
9:      $y_c^{(i)} := y_s^{(i)}$ 
10:  end if
11: end for

```

a lightweight ViT model (e.g., DeiT-Tiny) at the edge device, instead of splitting a complex ViT model (e.g., DeiT-Base), as depicted in Fig. 2. The proposed inference framework establishes an efficient collaborative protocol between the edge device and the server, aiming to achieve high classification accuracy of DeiT-Base model while significantly reducing communication overhead.

In our collaborative inference framework, the edge device (client) first performs inference with its tiny model. The edge device then evaluates the entropy level of this initial inference. High entropy (low confidence) necessitates transmitting the image to the server since it indicates that the tiny model’s inference would be unreliable. In such instances, only essential patches for classification are transmitted instead of the entire image patches to minimize communication costs. The server, utilizing its complex ViT model, conducts inference based on these selected patches and sends its classification results back to the edge device, as shown in Fig. 2(a). This process of selecting critical patches is governed by the proposed *attention-aware patch selection* rule, elaborated in Section IV.

If the initial inference’s entropy is low, the edge device confirms its classification result without further interaction with the server, as shown in Fig. 2(b). Reducing reliance on the server to reduce communication costs is achieved through *entropy-aware image transmission* rule, detailed in Section V. By integrating these rules, our framework significantly lowers communication costs while maintaining classification accuracy comparable to the server model.

The steps of the proposed collaborative inference are outlined in Algorithm 1. Here, Step 2 and Step 3 involve computing the initial inference result $f_c(\mathbf{x}^{(i)})$ and its entropy $g(\mathbf{x}^{(i)})$, respectively. If the entropy is below a given threshold η , then $f_c(\mathbf{x}^{(i)})$ is deemed the final classification outcome. In cases of higher entropy, as identified in Step 4, the client selects and transmits only essential patches to the server at Step 5 and 6, effectively lowering communication costs by ensuring $\dim(\tilde{\mathbf{x}}^{(i)}) < \dim(\mathbf{x}^{(i)})$. At Step 7, the server conducts inference on these selected patches, producing the result $f_s(\tilde{\mathbf{x}}^{(i)})$, which is then sent back to the client at Step 8.

The proposed collaborative inference framework can reduce the computational complexity for the server model by limiting

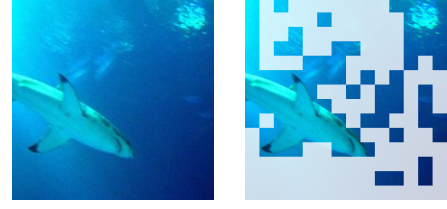


Fig. 3. The visualization of attention-aware patch selection. The left is an image of ImageNet dataset that the client model (DeiT-Tiny) inaccurately classifies as ‘Hammerhead Shark’. The right shows the selected patches by *attention-aware patch selection*. These selected patches allow the server model (DeiT-Base) to correctly classify the image as ‘White Shark’.

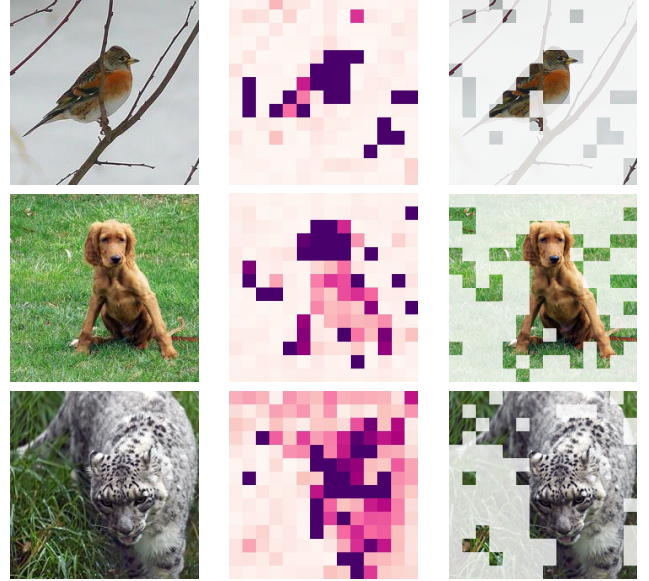


Fig. 4. The visualization of images and attention scores. In the left column, we have the original images of ImageNet dataset. The middle column displays the attention score maps generated by DeiT-Tiny. The right column shows the patches selected by the attention-sum threshold selection. The images are labeled as ‘Brambling’, ‘Irish Setter’, and ‘Snow Leopard’, respectively.

the inference process to only the selected patches. The computational complexity of DeiT-Base, according to [26], [27], is specified by the formula:

$$144ND^2 + 24N^2D, \quad (9)$$

where N is the number of patches. Assuming the number of patches of $\tilde{\mathbf{x}}^{(i)}$ is represented by \tilde{N} such that $\tilde{N} < N$, our framework not only reduces communication overhead but also enhances computational efficiency on the server side. This results in a significant secondary benefit of our collaborative inference framework.

IV. ATTENTION-AWARE PATCH SELECTION

This section introduces our attention-aware patch selection method, motivated by an intriguing observation: the tiny ViT model is capable of identifying the essential patches for classification, even when its classification is incorrect (see Fig. 3). Consequently, the tiny model on the edge device acts as a *semantic encoder*, effectively extracting essential information for the classification task.

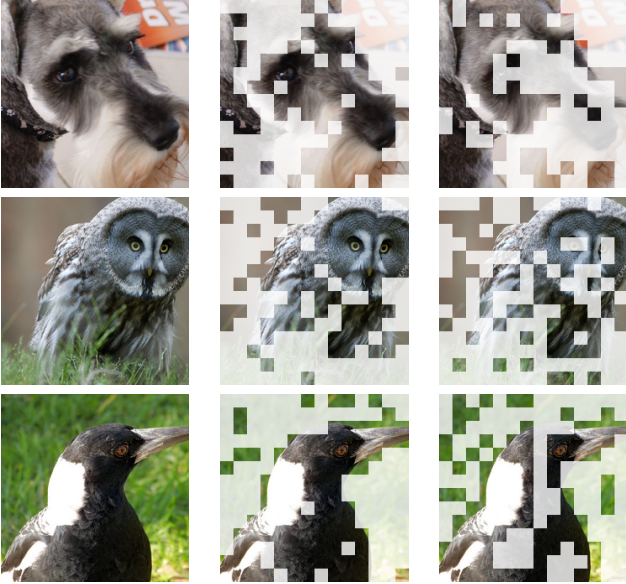


Fig. 5. The comparison of image patches selected by DeiT-Tiny (middle column) and DeiT-Base (right column). The same number of patches are selected according to the mean attention scores. The left column displays the original images of ImageNet dataset. The images are labeled as ‘Miniature Schnauzer’, ‘Great Gray Owl’, and ‘Magpie’, respectively.

To enhance communication efficiency and classification accuracy, we address two key questions: 1) how to accurately quantify the importance of each patch and 2) how to determine the optimal number of selected patches.

A. Quantifying Patch Importance

To quantify the importance of each patch for classification, we utilize the *attention scores* generated by the SA mechanism. The attention score for the class token in a single-head attention is calculated as follows:

$$\mathbf{a} = \text{softmax} \left(\frac{\mathbf{q}_{\text{cls}} \mathbf{k}_p^T}{\sqrt{D_h}} \right), \quad (10)$$

where $\mathbf{q}_{\text{cls}} \in \mathbb{R}^{1 \times D_h}$ represents the query for the class token of the last layer and $\mathbf{k}_p \in \mathbb{R}^{N \times D_h}$ denotes the keys corresponding to the image patches in the last layer. The *mean attention score* is then obtained by averaging the attention scores from all multi-heads.

Our experimental findings indicate that the mean attention scores, as computed by the tiny model, effectively assess the significance of each patch in contributing to the classification task. Fig. 4 presents a side-by-side comparison of ImageNet dataset images (left column) and their corresponding attention score maps (middle column). These maps clearly reveal that patches crucial for classification are distinguished by higher attention scores, setting them apart from less critical areas, such as background patches, which receive lower attention scores.

This observation supports that the tiny model on the edge device is adept at identifying and selecting the most informative patches for classification. Within ViT models, the class token aggregates information from other tokens (image



Fig. 6. The comparison of attention score maps computed by DeiT-Tiny (middle column) and DeiT-Base (right column). The left column displays the original images of ImageNet dataset. The images are labeled as ‘Miniature Schnauzer’, ‘Great Gray Owl’, and ‘Magpie’, respectively.

patches) via the attention mechanism. For the final classification, the ViT relies on the MLP head, which considers only the information associated with the class token from the last layer, disregarding any other inputs. Thus, the attention score as defined in (10) serves as a key metric for quantifying the contributions of individual image patches to the class token. In the context of semantic communications, the tiny model of the edge device acts as a *semantic encoder*, tasked with extracting essential information for the classification task. This role aligns with the broader objectives of semantic communications to emphasize meaning and relevance in the transmission of information [10]–[12].

An interesting finding is that DeiT-Tiny can act more effectively as a semantic encoder than DeiT-Base in spite of its inferior classification accuracy. Fig. 5 compares the patches selected by DeiT-Tiny (middle column) and those by DeiT-Base (right column), showing DeiT-Tiny’s superior ability to discard irrelevant image patches. This seemingly contradictory finding can be elucidated by the insights from recent work [28], which shows that large ViT models tend to allocate high attention scores to less informative background areas. It is because the large models adeptly identify patches with minimal information, such as background areas, repurposing the corresponding tokens to assimilate global image information while neglecting their spatial information. High attention scores are allocated to these repurposed patches containing global information, particularly in the background areas, as shown in Fig. 6. Although this strategy enhances the classification accuracy of larger ViT models, it compromises their effectiveness as semantic encoders. The experimental results on how model complexity affects the efficacy of semantic encoders are presented in Section VI-E.

The mean attention score from the last layer has previously been leveraged for purposes such as model interpretation [29],

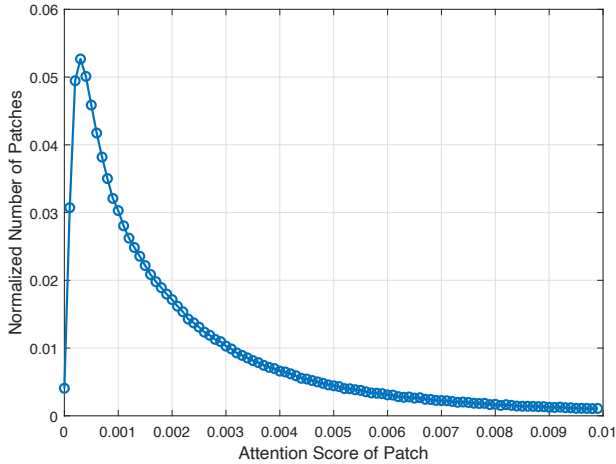


Fig. 7. The normalized histogram of the mean attention scores obtained by DeiT-Tiny on the ImageNet dataset.

[30] and knowledge distillation [27]. Our approach distinctively utilizes the mean attention score to reduce communication costs within our collaborative inference framework, differentiating our methodology from previous applications. Additionally, while *attention rollout* [31] is an established technique for interpreting transformer models, we opt for the mean attention score. This decision is based on the observation that attention rollout tends to produce more uniform attention scores, which do not align well with our objective of attention-aware patch selection. The experimental evidence supporting this decision is provided in Section VI-C.

B. Patch Selection Rule

In this subsection, we investigate the patch selection rule utilizing mean attention scores. Selecting an appropriate number of patches for transmission to the server is crucial, as this directly impacts both communication costs and classification accuracy. Our goal is to transmit the fewest possible patches to the server to minimize communication overhead. However, this approach poses a trade-off, as reducing the number of transmitted patches can limit the information available to the server model, potentially lowering classification accuracy.

The distribution of attention scores reveals that most patches are assigned low values, as shown in Fig. 7. This property enables a significant reduction in the number of transmitted patches without affecting classification accuracy. By taking advantage of this beneficial property, we explore the following patch selection rules:

- *Top- k selection*: Selecting the top- k patches that have the highest attention scores.
- *Attention threshold selection*: Selecting the patches whose attention scores exceed a predefined threshold δ .
- *Attention-sum threshold selection*: Selecting the patches with the highest attention scores until their cumulative attention sum reaches a predetermined threshold δ_{sum} .

The top- k selection method selects a fixed number of patches based on the highest attention scores, resulting in consistent communication costs for all input images. Nonetheless, it overlooks the variation in attention score distribution

among different images. Fig. 4 shows that the quantity of patches essential for classification can vary across images, with important information typically concentrated on the class object, where the highest attention scores are located. Consequently, to achieve optimal classification accuracy, the number of selected patches should be tailored to the size of the object within each image.

Both the attention threshold selection and attention-sum threshold selection methods provide the capability to adjust the number of selected patches for transmission, making them more adaptable than the top- k selection method. For instance, as shown in Fig. 4, the numbers of selected patches for ‘Brambling’ and ‘Snow Leopard’ are 35 and 124, respectively, by using the attention-sum threshold selection with $\delta_{\text{sum}} = 0.94$. Adopting this method ensures the transmission of a consistent sum of attention scores, effectively lowering the risk of omitting crucial information. The experimental results show that the attention threshold selection and attention-sum threshold selection methods outperform the top- k selection method in achieving an optimal trade-off between classification accuracy and communication efficiency, as detailed in Section VI-C.

V. ENTROPY-AWARE IMAGE TRANSMISSION

This section delves into entropy-aware image transmission, a strategy aimed at reducing communication overhead by considering the varied classification difficulty inherent to different images. For less complex images, the edge device’s initial inference may be accurate enough, eliminating the need for further interaction with the server. In contrast, more intricate images necessitate more accurate classification from the server model, leading to increased communication overhead. It is critical for the edge device to make an accurate decision between relying on its initial inference and requesting more accurate classification from the server model.

Even though the edge device cannot ascertain the correctness of its initial inference, it can estimate the inference’s confidence through the softmax output values of the MLP classification head. This softmax output can be interpreted as the posterior probability $p_{\theta}(y|\mathbf{x})$, where y denotes the class label and θ denotes the tiny model. Then, we set an entropy function $g : \mathbb{R}^L \rightarrow \mathbb{R}$, where L denotes the number of class labels. The client requests more accurate inference results from the server if:

$$g(\mathbf{x}) \geq \eta, \quad (11)$$

where η denotes a predetermined threshold. To access the confidence of the client’s inference, we consider two exemplary entropy measures: 1) Shannon entropy and 2) min-entropy, with their respective thresholds.

The Shannon entropy, a widely used metric for quantifying uncertainty [32], is calculated by

$$g_s(\mathbf{x}) = - \sum_{y \in \mathcal{Y}} p_{\theta}(y|\mathbf{x}) \log_2 p_{\theta}(y|\mathbf{x}), \quad (12)$$

where \mathcal{Y} denotes the set of all possible class labels. High Shannon entropy indicates that the given image \mathbf{x} is challenging for the tiny model to classify accurately. Therefore, if $g_s(\mathbf{x}) \geq \eta_s$,

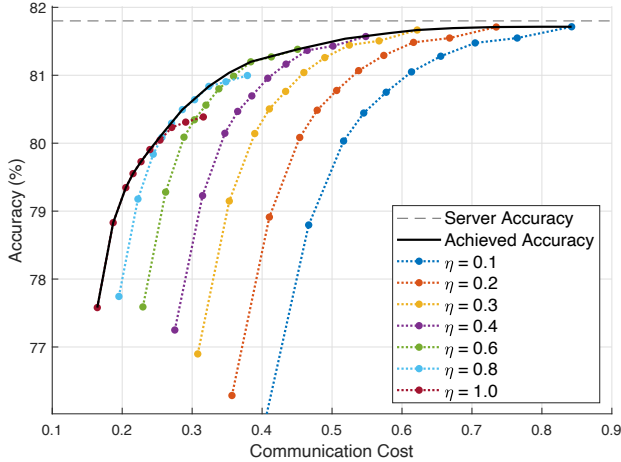


Fig. 8. Trade-off between communication cost and classification accuracy, with the edge device employing DeiT-Tiny and the server employing DeiT-Base. The black line represents the achievable curve by the proposed collaborative inference. We utilize the attention-sum threshold selection method for attention-aware patch selection. For entropy-aware image transmission, the min-entropy serves as the entropy metric, using a threshold value $\eta = \eta_m$.

the edge device transmits the selected patches to the server for an inference from the base model.

Another key metric, min-entropy, evaluates uncertainty in the most conservative manner [33]. The min-entropy is defined as

$$g_m(\mathbf{x}) = -\log_2 \max_{y \in \mathcal{Y}} p_\theta(y|\mathbf{x}), \quad (13)$$

which is directly associated with the confidence level of the initial inference. If $g_m(\mathbf{x}) \geq \eta_m$, the edge device transmits the selected patches to the server for an inference from the base model.

Our experimental results in Section VI-D show that the min-entropy serves as a better metric within our collaborative inference framework. Entropy-aware image transmission utilizing the min-entropy improves communication efficiency for a given level of classification accuracy when compared to using the Shannon entropy.

Shannon entropy has been utilized in diverse applications, such as prioritizing unlabeled data inputs in active learning [34] and wireless data acquisition [35]. Our work leverages min-entropy as a metric to evaluate the quality of initial inferences made by edge devices, setting our application apart from its previous utilization.

VI. EXPERIMENTAL RESULTS

A. Experiment Settings

Our experiment employs the ImageNet validation dataset and resizes each image to a resolution of 224×224 pixels by center cropping. An image is flattened to $N = 196$ patches before the inference. We utilize the DeiT models pre-trained on the ImageNet dataset, as presented in Table I.

We evaluate the impact of varying the number of transmitted patches on communication cost and classification cost. The communication cost in our collaborative inference system is quantified by the ratio of the number of transmitted patches to the total number of patches. Consequently, if the edge device

TABLE II
ATTENTION-SUM PATCH SELECTION THRESHOLD VALUES AND EXPECTED NUMBER OF TRANSMITTED PATCHES

δ_{sum}	Expected Number of Transmitted Patches
0.87	88.40
0.9	100.84
0.92	110.82
0.93	116.51
0.94	122.80
0.95	129.86
0.96	137.89
0.97	147.24
0.98	158.47
0.99	172.76

TABLE III
MIN-ENTROPY THRESHOLD VALUES AND EXPECTED RATIO OF TRANSMITTED IMAGES

η_m	Expected Rate of Transmitted Images
1	0.3567
0.8	0.4290
0.6	0.5116
0.4	0.6246
0.3	0.7109
0.2	0.8445
0.1	0.9714

sends all image patches to the server, the communication cost is quantified as 1.

When the edge device transmits the selected patches, it is required to transmit the position information of these selected patches. For instance, assigning one bit per patch as a marker of its selection status is a practical solution. The overhead of these additional bits for position information is negligible in comparison to the size of the image itself, given that only one bit is appended for each image patch containing 6,144 bits.

B. Communication Cost vs. Classification Accuracy

In our collaborative inference framework, we assess the trade-off between communication cost and classification accuracy. We utilize the attention-sum threshold selection for attention-aware patch selection and the min-entropy for entropy-aware image transmission.

Fig. 8 shows the trade-off between communication cost and classification accuracy, employing DeiT-Tiny on the edge device and DeiT-Base on the server. We achieve a 68% reduction in communication cost while attaining a classification accuracy of 80.84%, with only a minimal loss in accuracy compared to DeiT-Base's accuracy of 81.8%. This communication cost reduction is achieved by the attention-sum threshold selection method with a threshold $\delta_{\text{sum}} = 0.97$ and the min-entropy with the threshold $\eta_m = 0.8$. The black line indicates the optimized trade-off curve achieved by selecting the optimized threshold values of δ_{sum} and η_m . Table II and Table III detail the resulting communication costs depending on threshold values.

Additionally, we investigate a case where the server employs DeiT-Small instead of DeiT-Base. Fig. 9 shows the trade-off

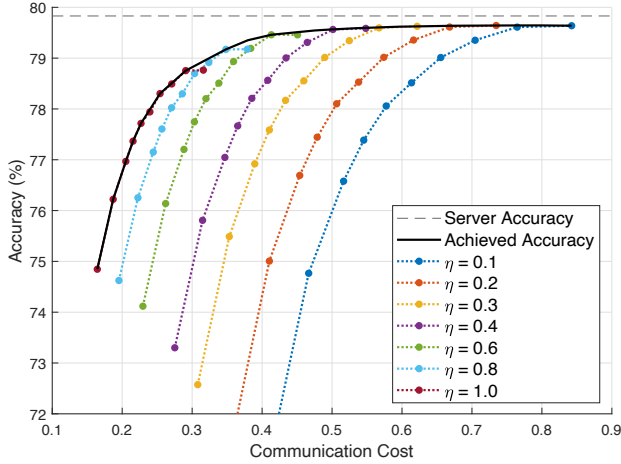


Fig. 9. Trade-off between communication cost and classification accuracy, with the edge device employing DeiT-Tiny and the server employing DeiT-Small. The black line represents the achievable curve by the proposed collaborative inference. We utilize the attention-sum threshold selection method for attention-aware patch selection. For entropy-aware image transmission, the min-entropy serves as the entropy metric, using a threshold value $\eta = \eta_m$.

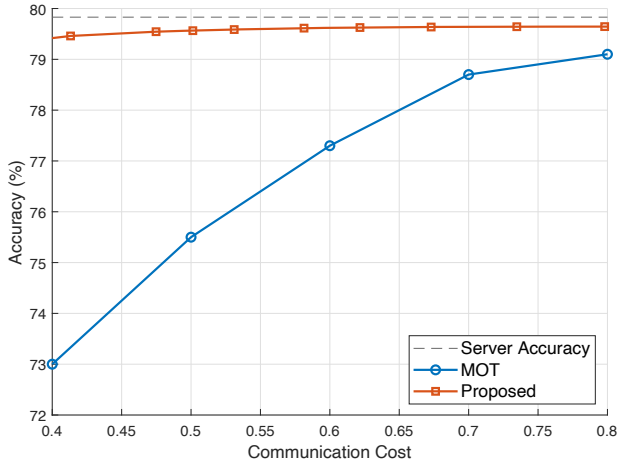
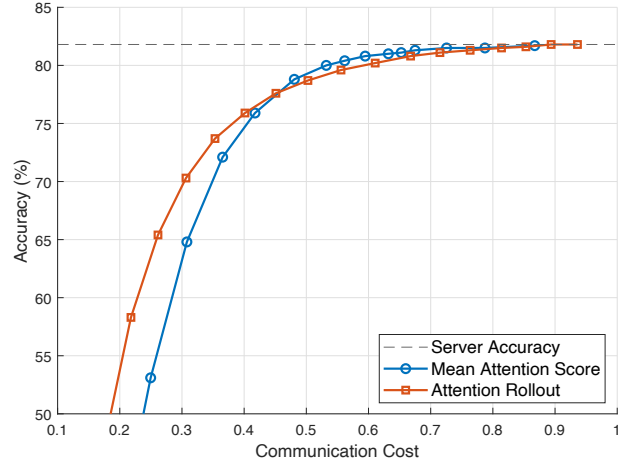
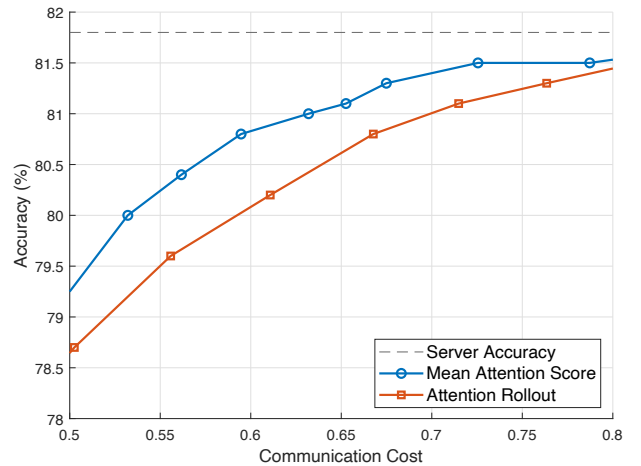


Fig. 10. Comparison of the MAE-based offloading for transformer inference (MOT) [22] and our proposed collaborative inference, with the edge device employing DeiT-Tiny and the server employing DeiT-Small.

between communication cost and classification accuracy. In this case, we achieve a 71% reduction in communication cost while attaining a classification accuracy of 78.8%, with only a minimal loss in accuracy compared to DeiT-Small’s accuracy of 79.8%. It is observed that greater communication cost reductions can be realized when the accuracy gap between the edge device and the server is reduced. Furthermore, Fig. 10 presents a comparison between two different inference systems: MAE-based offloading for transformer inference (MOT) [22] and our collaborative inference framework. The experimental results show that our proposed collaborative inference framework significantly improves performance by employing the tiny model on the edge device, which adeptly transmits essential image patches for classification. In contrast, the edge device of MOT [22] transmits randomly selected image patches to the server.



(a)



(b)

Fig. 11. Comparison of mean attention score and attention rollout, with the edge device employing DeiT-Tiny and the server employing DeiT-Base. The patch selection rule is the attention-sum threshold selection. (a) shows the overall trade-off between communication cost and classification accuracy, while (b) zooms in on the region of interest, focusing on areas near the server model’s classification accuracy.

C. Attention-aware Patch Selection

In this subsection, we delve into identifying the most effective metrics for determining patch importance and the optimal rules for patch selection. First, we examine and compare the efficacy of mean attention scores versus attention rollout. Next, we evaluate different patch selection methodologies, including top- k selection, attention threshold selection, and attention-sum threshold selection. We maintain a consistent environment with DeiT-Tiny on the edge device and DeiT-Base on the server. For this analysis, we specifically focus on attention-aware patch selection, excluding considerations of entropy-aware image transmission.

Fig. 11(a) reveals that both the mean attention score and attention rollout have distinct advantages. In situations demanding substantial reductions in communication cost, attention rollout is better than the mean attention score. However, as shown in Fig. 11(b), the mean attention score is a better metric for attaining accuracy comparable to that of the server. Consequently, we adopt the mean attention score for our

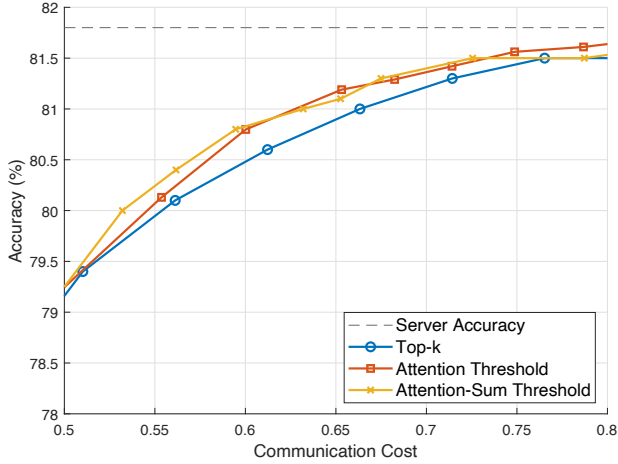


Fig. 12. Comparison of the patch selection rules: Top- k selection, attention threshold selection, and attention-sum threshold selection. The edge device and the server employ DeiT-Tiny and DeiT-Base, respectively. The patch importance is quantified by the mean attention score.

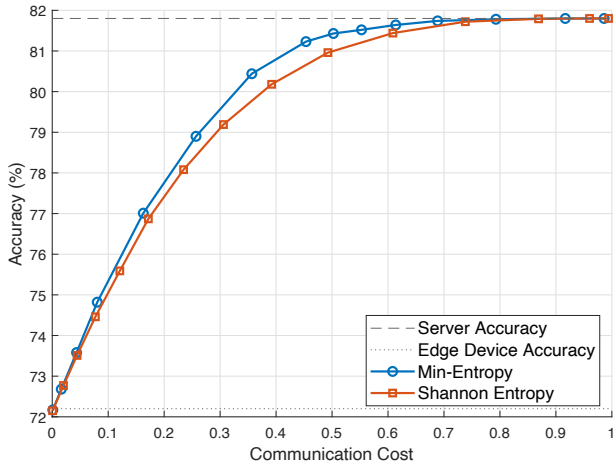


Fig. 13. Comparison of min-entropy and Shannon entropy with the edge device employing DeiT-Tiny and the server employing DeiT-Base.

primary experiments.

Fig. 12 shows that both attention threshold selection and attention-sum threshold selection outperform the performance of the top- k selection. These methods offer the capability to adjust the number of selected patches for transmission, facilitating the maintaining classification accuracy while minimizing communication costs.

D. Entropy-aware Image Transmission

To determine the most effective entropy measure, we compare the min-entropy and the Shannon entropy within our collaborative inference framework. For this analysis, we specifically focus on entropy-aware image transmission, deliberately setting aside the aspect of attention-aware patch selection. The entropy values are derived from the softmax output of the MLP classification head in the DeiT-Tiny model used on the edge device. As shown in Fig. 13, the min-entropy is a better metric for our collaborative inference, demonstrating a more effective distinction between correctly and incorrectly inferred

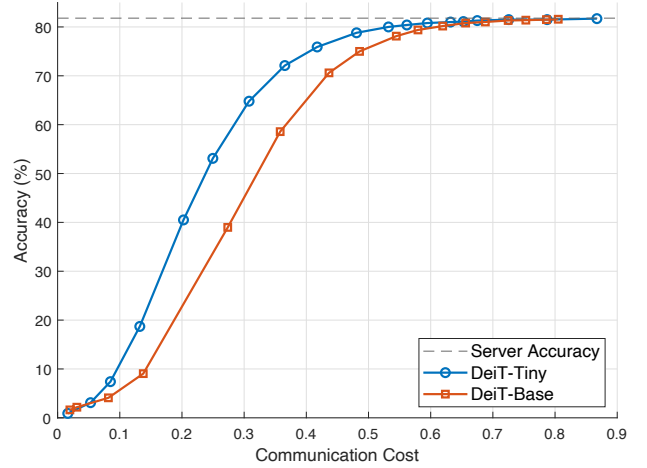


Fig. 14. Comparison of DeiT-Tiny and DeiT-Base as semantic encoders. The classification accuracy is obtained by DeiT-Base.

images compared to the Shannon entropy. Consequently, min-entropy is chosen as the preferred metric for entropy-aware image transmission.

E. Model Complexity of Semantic Encoder

As discussed in Section IV-A, DeiT-Tiny can act more effectively as a semantic encoder than DeiT-Base in spite of its inferior classification accuracy, which is supported by Fig. 5 and Fig. 6. In our experiments, we concentrate solely on attention-aware patch selection to evaluate and compare DeiT-Tiny and DeiT-Base. Both models select only important patches based on the mean attention scores, disregarding initial inference outcomes. The classification accuracy is obtained by DeiT-Base, which processes only these selected image patches. Fig. 14 shows that DeiT-Tiny more effectively identifies the essential image patches for classification better than DeiT-Base.

VII. CONCLUSION

We presented a novel attention-aware collaborative inference system using pre-trained ViT models. The edge device employs a compact ViT model as a semantic encoder, selecting pivotal patches that focus on objects crucial for classification. Our results confirm that this framework not only significantly lowers communication costs but also preserves accuracy comparable to that of server models. Furthermore, it offers the added benefit of diminishing the server model's computational complexity. Extending this approach to encompass a broader range of tasks beyond classification provides an intriguing direction for future research.

REFERENCES

- [1] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.
- [2] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, Jan. 2020.

- [3] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, Dec. 2020.
- [4] N. Shlezinger and I. V. Bajic, "Collaborative inference for AI-empowered IoT devices," *IEEE Internet Things Mag.*, vol. 5, no. 4, pp. 92–98, Dec. 2022.
- [5] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive feature transmission for split classification at the wireless edge," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3837–3852, Jun. 2023.
- [6] 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, Jan. 2021.
- [7] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.
- [8] X. Huang and S. Zhou, "Dynamic compression ratio selection for edge inference systems with hard deadlines," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8800–8810, Sep. 2020.
- [9] C. Zhang, H. Zou, S. Lasaulce, W. Saad, M. Kountouris, and M. Bennis, "Goal-oriented communications for the IoT and application to data compression," *IEEE Internet Things Mag.*, vol. 5, no. 4, pp. 58–63, Dec. 2022.
- [10] G. Shi, Y. Xiao, Y. Li, and X. Xie, "From semantic communication to Semantic-Aware networking: Model, architecture, and open problems," *IEEE Commun. Mag.*, vol. 59, no. 8, pp. 44–50, Aug. 2021.
- [11] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, Dec. 2021.
- [12] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.
- [13] 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.
- [14] Y. Kim, J. Shin, Y. Cassuto, and L. R. Varshney, "Distributed boosting classification over noisy communication channels," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 141–154, Jan. 2023.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017, pp. 5998–6008.
- [16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Jun. 2021.
- [17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 10 347–10 357.
- [18] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Apr. 2022.
- [19] J. Pan, A. Bulat, F. Tan, X. Zhu, L. Dudziak, H. Li, G. Tzimiropoulos, and B. Martinez, "Edgevits: Competing light-weight CNNs on mobile devices with vision transformers," in *Proc. European Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 294–311.
- [20] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "Efficientformer: Vision transformers at Mobilenet speed," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 12 934–12 949, Dec. 2022.
- [21] G. Xu, Z. Hao, Y. Luo, H. Hu, J. An, and S. Mao, "DeViT: Decomposing vision transformers for collaborative inference in edge devices," *IEEE Trans. Mobile Comput.*, pp. 1–16, 2023.
- [22] T. Liu, P. Li, Y. Gu, and P. Liu, "Efficient transformer inference for extremely weak edge devices using masked autoencoders," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2023, pp. 1718–1723.
- [23] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, Jun. 2022, pp. 16 000–16 009.
- [24] H. Xie, Z. Qin, G. Y. Li, and B. H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, May 2021.
- [25] H. Yoo, L. Dai, S. Kim, and C.-B. Chae, "On the role of ViT and CNN in semantic communications: Analysis and prototype validation," *IEEE Access*, vol. 11, pp. 71 528–71 541, Jul. 2023.
- [26] S. Chang, P. Wang, M. Lin, F. Wang, D. J. Zhang, R. Jin, and M. Z. Shou, "Making vision transformers efficient from a token sparsification view," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, Jun. 2023, pp. 6195–6205.
- [27] S. Son, N. Lee, and J. Lee, "MaskedKD: Efficient distillation of vision transformers with masked images," *arXiv preprint arXiv:2302.10494*, 2023.
- [28] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," *arXiv preprint arXiv:2309.16588*, 2023.
- [29] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.
- [30] H. Chefer, I. Schwartz, and L. Wolf, "Optimizing relevance maps of vision transformers improves robustness," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2022, pp. 33 618–33 632.
- [31] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Jul. 2020, pp. 4190–4197.
- [32] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.
- [33] R. König, R. Renner, and C. Schaffner, "The operational meaning of min- and max-entropy," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4337–4347, Sep. 2009.
- [34] B. Settles, *Active Learning*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.
- [35] D. Liu, G. Zhu, Q. Zeng, J. Zhang, and K. Huang, "Wireless data acquisition for edge learning: Data-importance aware retransmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 406–420, Jan. 2021.