Decentralized and Versatile Edge Encoding Methods for Task-Oriented Communication Systems

Hoon Lee, *Member, IEEE*, and Seung-Wook Kim, *Member, IEEE*

***Abstract*—In this paper, we develop task-oriented edge networks in which separate edge nodes perform decentralized inference processes with the aid of a cloud. Individual ENs compress their local observations into uplink messages using task-oriented encoder neural networks (NNs). Then, the cloud carries out a remote inference task by leveraging received signals. We develop fronthaul-cooperative DNN architecture along with proper uplink coordination protocols. Inspired by the nomographic function, an efficient cloud inference model becomes an integration of a number of shallow DNNs. This modulized architecture brings versatile calculations that are independent of the number of ENs. Numerical results demonstrate the viability of the proposed method for optimizing task-oriented edge networks.**

I. Introduction

Recent advantages in deep learning techniques have got great attention in implementing intelligent edge networks by means of powerful deep neural network (DNN) models installed at network clouds. This triggers recent studies on task-oriented edge networks that provide deep learning inference services to edge nodes (ENs) [1], [2]. To achieve this goal, ENs are requested to send their data samples to the cloud through fronthaul links that are subject to wireless fading and resource constraints.

Thus, an essential optimization challenge for task-oriented networks involves a joint design task of neural edge encoders and cloud inference models over resource-constrained wireless fronthaul channels. Cooperative edge-cloud DNN architectures were proposed for the task-oriented edge networks to execute remote inference tasks, such as network management [3]–[5] and image classification [6]–[10]. Neural edge quantization techniques were presented [3], [4], [6], [7] for noiseless fronthaul links. The additive Gaussian noise channels have been recently incorporated [4], [8]. However, the impact of wireless fading has not been studied adequately.

The massive connectivity requirements from ENs need versatile DNN architectures at the cloud whose calculations are independent of the number of ENs. Existing DNN models [6] accept concatenated edge messages, and thus they fail to achieve the scalability with respect to the EN population. The sum-pooling-assisted models in [8] aggregations received signals for the scalable design, but this model cannot adjust the compression rate.

This paper develops decentralized and scalable learning architecture for task-oriented edge networks. We interpret an oracle edge-cloud inference rule as the nomographic function. Our investigation reveals that by exploiting the Kolmogorov-Arnold (KA) representation [11], the oracle inference model can be decomposed into a set of decentralized edge encoding functions followed by a sum-pooling layer together with a cloud inference function. This offers a scalable architecture whose computations are irrelevant to the number of ENs. The performance of the proposed approach is examined over classification tasks of the Tiny ImageNet dataset [12]. Numerical results validate the effectiveness of the proposed approach over conventional methods.

II. Task-Oriented Edge Network

In task-oriented edge networks, a cloud and N ENs cooperatively carry out an inference of DNN models for a global information sample $\mathbf{a}$ via resource-constrained fronthaul links. Let $\mathcal{N} \triangleq [1, N]$ be the set of ENs. Due to the distributed nature of practical EN deployment scenarios, each EN i can only know its own local observation $\mathbf{a}_i \in \mathbb{R}^A$, which is given by a subset of the global information $\mathbf{a}$. The cloud collects these local observations and infers a desired output variable $\mathbf{x} \in \mathbb{R}^X$ through a DNN g_φ with trainable parameter φ.

Each fronthaul link contains $S \leq A$ time-frequency resource blocks (RBs). To accomodate this fronthaul resource constraint, each EN i compresses its local observation $\mathbf{a}_i$ into a fronthaul message $\mathbf{s}_i \in \mathbb{R}^S$ of length S. An encoder of EN i is denoted by a learnable function $f_{\psi_i} : \mathbb{R}^A \to \mathbb{R}^S$ with ψ_i being a trainable parameter. The edge encoding process of EN i is given by

$$\mathbf{s}_i = f_{\psi_i}(\mathbf{a}_i). \tag{1}$$

The transmit power budget p_E at the ENs is imposed to limit the magnitude of each element of $\mathbf{s}_i$.

The ENs send the fronthaul message $\mathbf{s}_i$, $\forall i \in \mathcal{N}$, to the cloud through orthogonal fronthaul RBs. Each fronthaul channel is corrupted by the wireless fading $\mathbf{h}_i \in \mathbb{R}^S$ and the the additive Gaussian noise $\mathbf{n}_i \in \mathbb{R}^S$. The corresponding received signal $\mathbf{y}_i \in \mathbb{R}^S$ is written by

$$\mathbf{y}_i = \mathbf{h}_i \odot \mathbf{s}_i + \mathbf{n}_i \triangleq h_i(\mathbf{s}_i), \tag{2}$$

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant 2021R1I1A3054575 and Grant 2022R1F1A1074009, and in part by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) grant funded by the Korea government (DAPA(Defense Acquisition Program Administration)) (21-106-A00-007, Space-Layer Intelligent Communication Network Laboratory, 2022).

H. Lee is with the Department of Electrical Engineering and the Artificial Intelligence Graduate School, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 44919, South Korea (e-mail: hoonlee@unist.ac.kr).

S.-W. Kim is with the Division of Electrical and Communication Engineering, Pukyong National University, Busan 48513, South Korea (e-mail: swkim@pknu.ac.kr).

where $\odot$ is the element-wise multiplication and $h_i : \mathbb{R}^S \to \mathbb{R}^S$ is the channel transfer function.

The inference result $\mathbf{x}$ at the cloud is then obtained as

$$\mathbf{x} = g_\varphi(\mathbf{y}_1, \cdots, \mathbf{y}_N) \tag{3a}$$

$$= g_\varphi(h_1 \circ f_{\psi_1}(\mathbf{a}_1), \cdots, h_N \circ f_{\psi_N}(\mathbf{a}_N)), \tag{3b}$$

where $\circ$ is the composition operator. For supervised learning tasks, the inference performance is evaluated by a loss function $l(\mathbf{x}, \mathbf{t})$ where $\mathbf{t} \in \mathbb{R}^X$ denotes the desired label. Then, we can formulate the training problem as

$$\min_\theta L(\theta) \triangleq \mathbb{E}[l(\mathbf{x}, \mathbf{t})], \tag{4}$$

where $\theta \triangleq \varphi \bigcup \{\psi_i : \forall i \in \mathcal{N}\}$ accounts for the trainable parameter. The above problem can be addressed by state-of-the-art stochastic gradient descent (SGD) algorithms. At each training epoch, θ is updated as

$$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta), \tag{5}$$

where η indicates the learning rate.

III. Inference Design

We present an effective inference model for the task-oriented edge network by leveraging the properties of the nomographic function. Without loss of the generality, we consider scalar label $t \in \mathbb{R}$ and scalar local information $a_i \in \mathbb{R}$, $\forall i \in \mathcal{N}$. It is assumed that the domain $\mathbb{A}$ of a_i is a compact set. Notice that our goal is to design g_φ similar to an oracle mapping $c : \mathbb{A}^N \to \mathbb{R}$ which estimates the ground-truth label t expressed by

$$t = c(a_1, \cdots, a_N). \tag{6}$$

According to the analysis in [13], every function can be categorized into the nomographic function. Thus, the oracle mapping c can be rewritten as

$$c(a_1, \cdots, a_N) = u\left(\sum_{i \in \mathcal{N}} v_i(a_i)\right), \tag{7}$$

for some mappings $u : \mathbb{R} \to \mathbb{R}$ and $v_i : \mathbb{R} \to \mathbb{R}$. Here, the inner mapping v_i can be regarded as the composition of the fronthaul channel and encoding DNN, i.e., $v_i = h_i \circ f_{\psi_i}$. Therefore, it suffices for the cloud DNN to accept the aggregated received signal $\sum_{i \in \mathcal{N}} \mathbf{y}_i$ instead of the concatenation $[\mathbf{y}_1^T, \cdots, \mathbf{y}_N^T]^T$. By doing so, we can achieve the scalability to the EN population N.

However, (7) is generally not viable for continuous mappings u and v_i [14]. Such a restriction is not suitable for the continuous-valued wireless fronthaul channels h_i. This difficulty can be addressed via the KA representation theorem [11], which factorizes c as follows:

$$c(a_1, \cdots, a_N) = \sum_{m \in \mathcal{M}} u_m \left(\sum_{i \in \mathcal{N}} v_{mi}(a_i)\right), \tag{8}$$

where $\mathcal{M} = [1, M]$ and u_m, $\forall m \in \mathcal{M}$, and v_{mi}, $\forall (m, i) \in \mathcal{M} \times \mathcal{N}$ are continuous.

To identify mappings u_m and v_{mi}, we employ learnable functions u_{λ_m} and $v_{\mu_{mi}}$ with λ_m and μ_{mi} being the trainable parameters. With these DNNs at hand, the cloud DNN g_φ can be designed as

$$g_\varphi(\mathbf{y}_1, \cdots, \mathbf{y}_N) = \sum_{m \in \mathcal{M}} u_{\lambda_m} \left(\sum_{i \in \mathcal{N}} z_{\zeta_m} \circ h_i \circ f_{\psi_i}(\mathbf{a}_i)\right), \tag{9}$$

where the cloud DNN parameter ψ is given as $\varphi = \{(\lambda_m, \zeta_m) : \forall m \in \mathcal{M}\}$. Here, additional trainable functions z_{ζ_m}, $\forall m \in \mathcal{M}$, each having parameter set ζ_m, establishes the inner mapping function $v_{\mu_{mi}}$ as $v_{\mu_{mi}} = z_{\zeta_m} \circ h_i \circ f_{\psi_i}$ with $\mu_{mi} \triangleq (\zeta_m, \psi_i)$. Consequently, the proposed cloud DNN can be built by using M component DNNs u_{λ_m} and z_{ζ_m}, $\forall m \in \mathcal{M}$, which is independent of N.

IV. Numerical Results

We assess the proposed method for the image classification task of the Tiny ImageNet dataset [12]. It consists of $120,000$ color images of 200 classes, each with 500 training images, 50 validation images, and 50 test images. We first resize the image size of all samples into $3 \times 64 \times 64$. Only a subset of this full-size image is available at ENs. For the simulation, we use randomly cropped images with window size 48×48, i.e., $\mathbf{a}_i \in \mathbb{R}^{3 \times 48 \times 48}$. The encoder DNN f_{ψ_i} has one convolutional layer, four residual blocks (ResBlocks) [15], and one fully-connected layer. The convolutional layer has 64 kernels of size 7×7 with stride 2. The ResBlock comprises two convolutional layers with kernel size 3×3 and a skip connection link. The rectified linear unit (ReLU) activation is employed. The output of the fourth ResBlock of the encoder DNN is flattened into the vector processed by the fully-connected producing the fronthaul message $\mathbf{s}_i$. For the cloud DNN, we use $M = 17$ component modules z_{ζ_m} and u_{λ_m} each realized with two-layer MLP having 128 neurons. The Adam optimizer is applied with the learning rate 10^{-4} and the batch size 256.

The transmit power constraint is set to $p_E = 1$. Then, the fronthaul signal-to-noise ratio (SNR) is defined as SNR $= 1/\sigma^2$. For each training sample, the SNR values are uniformly generated within $[0$ dB, 30 dB$]$. The Rayleigh fading is considered. We train the proposed model at $N_{train} = 8$ ENs and directly test the trained model over a wide range of test EN population $N_{test} \in [4, 8]$. We consider the following three baseline cloud DNN models.

- *BaseNet*: The cloud is assumed to have the perfect access to the full-size image input. The DNN is constructed with the encoder DNN followed by a three-layer MLP with a hidden dimension of 2048.
- *CatNet*: A three-layer MLP accepts the concatenated received signal as an input feature.

The depth and width of these baseline models are set to have a similar level of model complexity to the proposed cloud DNN.

Fig. 1 presents the test accuracy performance of various methods with respect to the fronthaul SNR for $S = 16$. For all simulated N_{test}, the proposed approach is superior to CatNet. The proposed scheme approaches the ideal performance of BaseNet as the SNR grows. Increasing N_{test} improves the accuracy of all schemes. CatNet exhibits a good accuracy performance for small N_{test}, but its performance severely

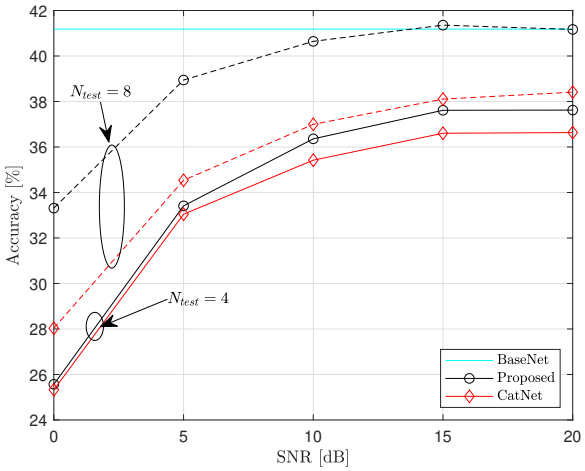

Fig. 1. Accuracy performance with respect to SNR for $S = 16$.

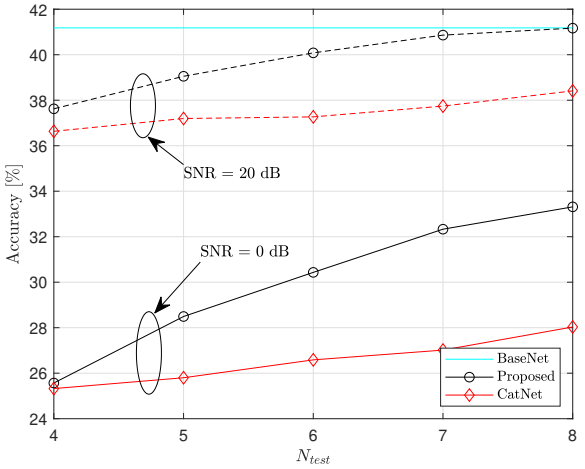

Fig. 2. Accuracy performance with respect to N_{test} for $S = 16$.

degrades for $N_{test} = 8$. The proposed scheme optimized for fixed EN population $N_{train} = 8$ clearly outperforms CatNet, demonstrating its scalability.

We depict the accuracy performance in Fig. 2 by changing N_{test} for $S = 16$. The proposed framework outperforms other baseline methods regardless of N_{test}. The performance of CatNet does not improve with the test EN population since it fails to extract useful features from edge-encoded signals.

V. Conclusions

In this paper, we have proposed decentralized and versatile inference models for multiedge task-oriented communication networks. We have exploited the notion of the nomographic function to establish an efficient cloud DNN model along with decentralized edge encoding DNNs over wireless fronthaul links. The effectiveness of the proposed approach has been demonstrated through simulation results.

Acknowledgement

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) grant funded by the Korea government (DAPA(Defense Acquisition Program Administration)) (21-106-A00-007, Space-Layer Intelligent Communication Network Laboratory, 2022) and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1F1A1074009).

References

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

[2] 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.

[3] H. Lee, S. H. Lee, and T. Q. S. Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.

[4] H. Lee, J. Kim, and S.-H. Park, "Learning optimal fronthauling and decentralized edge computatin in fog radio aacess networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5599–5612, Sep. 2021.

[5] H. Lee, S. H. Lee, and T. Q. S. Quek, "Artificial intelligence meets autonomy in wireless networks: A distributed learning approach," *IEEE Netw.*, vol. 22, no. 1, pp. 100–107, Nov./Dec. 2022.

[6] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multi-device cooperative edge inference," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 73–87, Jan. 2023.

[7] S. Xie, S. Ma, M. Ding, Y. Shi, M. Tang, and Y. Wu, "Robust information bottleneck for task-oriented communication with digital modulation," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 8, pp. 2577–2591, Aug. 2023.

[8] 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.

[9] X. Xu, B. Xu, S. Han, C. Dong, H. Xiong, R. Meng, and P. Zhang, "Task-oriented and semantic-aware heterogeneous networks for artificial intelligence of things: Performance analysis and optimization," *IEEE Internet Things J.*, to be published.

[10] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Sep. 2022.

[11] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," *Doklady Akademii Nauk*, vol. 114, no. 5, p. 953–956, 1957.

[12] Y. Le and X. Yang, "Tiny ImageNet visual recognition challenge," 2015, [Online] Available: http://tiny-imagenet.herokuapp.com.

[13] R. Buck, "Approximate complexity and functional representation," *J. Math. Anal. Appl.*, vol. 70, no. 1, pp. 280–298, 1979.

[14] R. C. Buck, "Nomographic functions are nowhere dense," *in Proc. Amer. Math. Soc.*, vol. 85, no. 2, pp. 195–199, Jun. 1982.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *in Proc. IEEE Conf. Comput. Vis. Comput. Recognit. (CVPR).* S, 2016, pp. 770–778.