2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall) | 979-8-3315-1778-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/VTC2024-Fall63153.2024.10757606

A Radio-Frequency-Based Fully Connected Layer using 1-Bit and 2-Bit Transmissive Intelligent Surfaces

Jingyuan Zhang*, Haige Chen*, Douglas M. Blough
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States
Emails: jingyuan@ece.gatech.edu, hchen425@gatech.edu, doug.blough@ece.gatech.edu

***Abstract*—This paper introduces a novel over-the-air computation method that utilizes low-complexity transmissive intelligent surfaces (TISs) for neural network inference. It is demonstrated that the signal propagation model through TIS closely resembles the fully connected layer of neural networks. And through training, the TIS phase shifts can be determined to perform a specific computation on radio-frequency (RF) signals. Considering the practical constraints of TIS designs with continuous phase shifts in millimeter-wave (mmWave) frequency, we propose a novel discretized complex-valued neural network structure and a training method suitable for low-complexity 1-bit and 2-bit TIS-based neural network layers. It is shown through simulation that the proposed method achieves high accuracy on an image classification task even for 1-bit or 2-bit TISs.**

***Index Terms*—over-the-air, transmissive intelligent surface, fully-connected layer, neural networks, low-complexity**

I. Introduction

In recent years, the demand for machine learning (ML) applications on Internet-of-Things (IoT) devices has been growing. This trend emphasizes the need to conduct ML inference locally on IoT devices, which eliminates the need for data exchange with a centralized cloud. This demand arises from concerns about privacy, real-time processing, and reduced dependence on Internet connectivity. However, IoT devices, typically featuring limited computational capability and memory, encounter challenges in executing local ML inference with constrained resources [1].

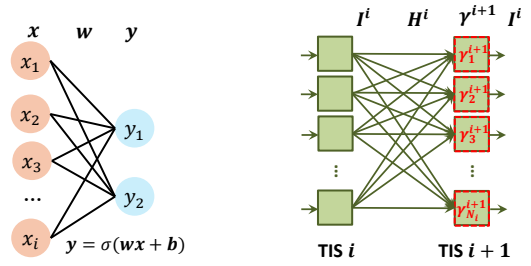
IoT devices commonly depend on wireless signals for communication and sensing purposes. Given the widespread availability of wireless communications, we introduce the concept of over-the-air (OTA) computation to support local ML inference, with the aim to reduce the computational burden on IoT devices. OTA computation is used to modify RF signals in order to simulate mathematical operations, enabling analog computation in the radio-frequency (RF) domain and thereby reducing the computational load on digital processors [2].

Recent studies have proposed transmissive intelligent surfaces (TISs), which are arrays of transmissive electromagnetic elements capable of manipulating transmission coefficients, for wireless communication applications [3], [4]. TISs offer advantages of cost-effectiveness and low power consumption as the electromagnetic elements are passive. Due to the similarity between one TIS and one fully connected (FC) layer in a neural network, we utilize multiple TISs positioned closely in sequence to emulate multiple classic FC layers. The input of the multi-TIS structure contains necessary information for processing, which is then processed by the multi-TIS structure akin to passing through several classic FC layers. In other words, the processing of FC layers can be shifted from digital processors to the RF domain, offering the potential to decrease the computational burden for ML inference on IoT devices.

There has been limited exploration into implementing FC layers using OTA computation. In [5], a novel method is proposed where a trained FC layer can be transformed into the RF domain using a multiple input single output (MISO) system. Here, the input symbol of the FC layer is transmitted through multiple transmit antennas, and a pre-equalization term at the transmitter is utilized to apply the trained weights within the FC layer. At the receiver end, the outputs of different neurons are implemented on different subcarriers in a orthogonal frequency-division multiplexing (OFDM) symbol. Nonetheless, implementation of active antenna arrays poses challenges for IoT devices due to hardware complexity and computational delays, especially when dealing with multiple FC layers and an increased number of neurons. Therefore, there is a need for a low-complexity and real-time solution tailored to ML inference in IoT scenarios. In [6], the authors proposed using TISs, originally designed for wireless communication applications, in the implementation of deep neural networks. That work proposed a diffractive deep neural network using multiple transmissive metasurfaces, where each meta-atom has amplifiers to emulate active neurons and is capable of joint amplitude and phase configuration. In IoT scenarios where low complexity and low power are critical issues, TISs should be kept as simple as possible. Employing TISs featuring 2-bit or 1-bit discretized phases along with passive components would be advantageous in such scenarios.

In this paper, we focus on low-complexity TIS-based neural network layers (TIS-NNLs) for IoT devices. Multiple TISs are cascaded to mimic FC layers of a neural network. Both the design and an efficient training algorithm for such TIS-NNLs are proposed, where 1-bit or 2-bit phase discretization without amplitude adjustment is assumed for the TISs. The

*These authors contributed equally to this work and should be considered co-first authors.



(a) Classic FC layer (b) TIS-FC layer

Fig. 1. Comparison between classic FC layer and TIS-FC layer.

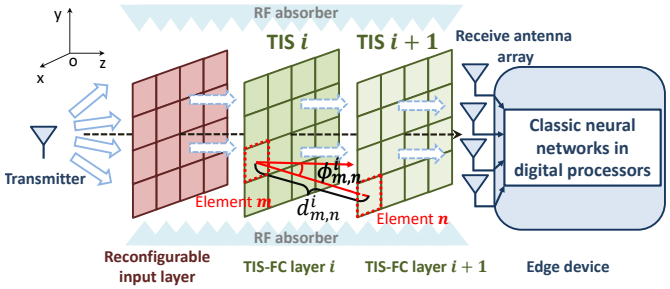


Fig. 2. Illustration of TIS-NNL.

proposed TIS-NNL and training method for 1-bit and 2-bit TISs are applicable to TISs across different frequency bands. In this paper, we focus on TISs at millimeter-wave (mmWave) band due to their relatively smaller array size, which makes them more suitable for edge devices. Simulation results show the effectiveness of 1-bit and 2-bit TIS-NNLs, despite a performance reduction compared to ideal TIS-NNLs where both phase and amplitude are continuously adjustable. This highlights the feasibility of complexity reduction for ML inference on IoT devices through use of 1-bit and 2-bit TISs.

The rest of the paper is organized as follows. The system model of TIS-NNLs is discussed in Sec. II. The training approach proposed for 1-bit and 2-bit TIS-NNLs is introduced in Sec. III. Sec. IV presents the simulation results. Finally, Sec. V concludes the paper.

II. System Design

The transmission model of TISs is introduced in this section, followed by an overview of the proposed TIS-NNL.

A. TIS Transmission Model

The system where several TISs are placed one next to another is considered. Let N_i denote the number of TIS elements on TIS i. Let γ^i_k denote the transmission coefficient of the k^{th} element on TIS i and $\|\gamma^i_k\|_2 \leq 1$. Moreover, let $h^i_{m,n}$ denote the channel between the m^{th} element ($m \in [0, N_i - 1]$) on TIS i and the n^{th} element ($n \in [0, N_{i+1} - 1]$) on TIS $i+1$. Based on Rayleigh-Sommerfeld diffraction, $h^i_{m,n}$ can be modeled as follows [7]

$$h^i_{m,n} = A_t \cos(\phi^i_{m,n})\left(\frac{1}{2\pi d^i_{m,n}} - \frac{j}{\lambda}\right)\frac{e^{\frac{j2\pi d^i_{m,n}}{\lambda}}}{d^i_{m,n}}, \quad (1)$$

where A_t is the aperture size of one TIS element, λ is the wavelength, $d^i_{m,n}$ is the distance between the m^{th} element on TIS i and the n^{th} element on TIS $i+1$, and $\phi^i_{m,n}$ is the angle between the z-axis and signal transmission direction from the m^{th} element on TIS i to the n^{th} element on TIS $i+1$. Note that it is assumed all TISs are enclosed by RF absorbing materials such that all multipaths are eliminated.

Let $\mathbf{I}^i \in \mathbb{C}^{N_i \times 1}$ denote the output signal vector of TIS i with the n^{th} entry being the output signal to the n^{th} element on TIS i. Moreover, let $\mathbf{H}^i$ be the channel matrix between TIS i and TIS $i+1$, where the $(m,n)^{\text{th}}$ entry is $h^i_{m,n}$. Then the signal propagation from TIS i to TIS $i+1$ can be described by

$$\mathbf{I}^{i+1} = \boldsymbol{\gamma}^{i+1} \odot (\mathbf{H}^i \mathbf{I}^i), \quad (2)$$

where $\odot$ represent element-wise multiplication, and $\boldsymbol{\gamma}^{i+1} \in \mathbb{C}^{N_i \times 1}$ represents the transmission coefficient vector of TIS $i+1$ with γ^{i+1}_n being the n^{th} entry.

B. Overview of the Proposed TIS-NNL Structure

A fundamental building block of the neural network structure is the FC layer as illustrated in Fig. 1(a), where the output $\mathbf{y}$ is calculated by the weighted sum of the input vector $\mathbf{x}$ offset by some bias $\mathbf{b}$, then followed by a non-linear activation function $\sigma(\cdot)$:

$$\mathbf{y} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}). \quad (3)$$

Multiple FC layers are often stacked into a deep neural network (DNN) to perform complex tasks such as image classification. The weights $\mathbf{W}$ and the bias $\mathbf{b}$ can be trained in a supervised fashion with gradient descent techniques. However, the DNN typically requires storage of a large number of trained weights in the memory, and the computational and memory complexity can be especially high for inputs with large dimensions such as images, making it prohibitive to implement on resource-constrained IoT devices. The main question we address in this paper is: *Is it possible to offload the memory and computation requirements of DNN inference to a different place while maintaining the locality of the computation?*

Interestingly, if we view $\mathbf{I}^i$ and $\mathbf{I}^{i+1}$ as the input and output of a function, the TIS model in (2) is similar to an FC layer without bias and activation function, in which the $\boldsymbol{\gamma}^{i+1}$ contain the trainable weights. Herein, we define TIS-based fully connected layer, referred to as TIS-FC, as one TIS with trainable transmission coefficients as illustrated in Fig. 1(b). The layer-to-layer operation is defined in (2). Furthermore, multiple TISs can be cascaded to mimic a multi-layer DNN architecture, and by carefully tuning the TIS elements through training, the TISs can also be used to perform computational tasks in inference.

Based on this observation, we propose a system where the low-cost resource-constrained IoT devices offload their inference tasks to a dedicated TIS-NNL structure, in which several TISs are placed between a transmitter and a receiver, and the ML inference takes place in the RF domain directly

as the wireless signals pass through the TISs, as illustrated in Fig. 2. To inject the input into the TIS-NNL structure, a transmitter broadcasts a continuous unmodulated signal, which impinges on a programmable TIS whose elements are tuned according to each entry of the input such that the outgoing signal carries the spatial information of the input. The wireless signals are then modified while passing through the TISs, before reaching the receiver. This way, the computational burden and the storage of the network weights can be offloaded to the TIS-NNL structure from the individual IoT devices.

It should be noted that an important reason why neural networks excel in many tasks is their ability to model complex *non-linear* functions. Since adding non-linearity to passive TIS is still an open problem, in this work, we assume the TIS-NNL only contains linear operations. The non-linear activation functions are applied at the output side of the TIS-NNL, i.e. to the received signal at the receiver. And since the linear layers benefit little from a deep network architecture, we consider a compact TIS-NNL composed of just 2 TIS-FC layers.

III. Quantization-aware training for low-complexity TIS-NNL

There are several further considerations to be made. First, since in the RF domain, the input, the wireless channel, and the TIS transmission coefficients are complex numbers, the proposed TIS-NNL mimics a special type of DNN, namely the complex-valued neural network (CVNN) [8]. In this section, we will first introduce the training of the complex valued TIS-NNL. Second, the reconfigurability of each TIS element is achieved by introducing a tunable phase shift between the input and the output signal. While continuous phase shift designs do exist, they usually incur larger insertion loss, higher costs, and more complicated designs for mmWave [3]. Therefore, we propose a novel method for training a discrete-weight TIS-NNL suitable for the low-complexity TIS designs that use 1-bit or 2-bit phase shift.

A. Complex-Valued Neural Network for TIS-NNLs

In the TIS-FC layer described by (2), the output signal vector $\mathbf{I}^i$, channel matrix $\mathbf{H}^i$, transmission coefficient vector $\boldsymbol{\gamma}^i$ and the output signal vector $\mathbf{I}^{i+1}$ all take on complex values. Each TIS-FC layer has two weight matrices, $\boldsymbol{\gamma}^i_{re}$ and $\boldsymbol{\gamma}^i_{im}$, which represent the real part and the imaginary part of the transmission coefficient $\boldsymbol{\gamma}^i$ of TIS i, respectively. Then the output of TIS-FC layer i is represented by:

$$\begin{aligned}\mathbf{I}^{i+1} = \boldsymbol{\gamma}^{i+1}_{re} \odot \Re(\mathbf{H}^i\mathbf{I}^i) - \boldsymbol{\gamma}^{i+1}_{im} \odot \Im(\mathbf{H}^i\mathbf{I}^i) + \\ j\Big(\boldsymbol{\gamma}^{i+1}_{re} \odot \Im(\mathbf{H}^i\mathbf{I}^i) + \boldsymbol{\gamma}^{i+1}_{im} \odot \Re(\mathbf{H}^i\mathbf{I}^i)\Big).\end{aligned} \tag{4}$$

Equation (4) can be used to combine the results into the final complex output. Consequently, the gradient can also be calculated separately for both parts during back-propagation. This method is often referred to as split-CVNN approach, which is commonly adopted in related works [5], [9].

B. Phase Discretization for TIS-NNLs

In terms of the phase reconfigurability, the existing TIS designs fall under two categories: the continuous phase shift design, or the discrete phase shift design. Continuous shift designs can be achieved with analog components such as varactor diodes. However, they usually introduce a higher insertion loss. Furthermore, for mmWave frequencies, the cost and the design complexity of continuous phase shift can be much higher than discrete phase shift. Therefore for practical reasons, many designs adopt a discrete phase shift with only 1-bit or 2-bit phase reconfigurability. Next, we will discuss the design of TIS-NNLs in the context of low-complexity TISs that have only 1-bit or 2-bit phase reconfigurability.

Considering the low-complexity TIS architecture with discrete phase shift, the transmission coefficient γ of one TIS element can be represented by $e^{j\phi_q}$, where ϕ_q is the phase shift of the TIS element. For an N_q-bit phase shift design,

$$\phi_q \in \{0, \frac{2\pi}{2^{N_q}}, ..., \frac{2\pi}{2^{N_q}} \times (2^{N_q} - 1)\}.$$

In light of the TIS-NNL design, this means $\boldsymbol{\gamma}$ can only take on discrete values from the set $\{1, -1\}$ for a 1-bit TIS, and $\{1, j, -1, -j\}$ for a 2-bit TIS.

Next, we propose a quantization-aware training method for low-complexity TISs. Since in Section III-A, the complex coefficient γ is split into the real part γ_{re} and imaginary pars γ_{im}, we also discretize γ_{re} and γ_{im} individually.

For a 1-bit TIS, since γ is chosen from the discrete set $\{1, -1\}$, only γ_{re} is trainable and γ_{im} is 0. The 1-bit weight discretization function $q_1(\cdot)$ can be simply expressed as the binarization function based on the real part of γ:

$$d_1(\gamma) = bin\left(\gamma_{re}\right) = \begin{cases} 1, & \gamma_{re} \geq 0, \\ -1, & \gamma_{re} < 0. \end{cases} \tag{5}$$

For a 2-bit TIS, both γ_{re} and γ_{im} are trainable. Then γ_{re} and γ_{im} seperately pass through the binarization function as follows

$$\frac{\sqrt{2}}{2}\big(bin\left(\gamma_{re}\right) + j \times bin\left(\gamma_{im}\right)\big), \tag{6}$$

where the factor of $\frac{\sqrt{2}}{2}$ is multiplied to normalize the amplitude to 1. This process discretizes the weight γ to the discrete set $\{e^{j\pi/4}, e^{j3\pi/4}, e^{-j3\pi/4}, e^{-j\pi/4}\}$, corresponding to phase shifts of $\{45^o, 135^o, 225^o, 315^o\}$. However, in many existing 2-bit TIS designs [3], the phase shift is also a function of the operating frequency, causing a common constant phase offset θ across all switching states. Therefore, the actual discrete TIS weights for a particular operating configuration can be $\{e^{j\theta}e^{j\pi/4}, e^{j\theta}e^{j3\pi/4}, e^{j\theta}e^{-j3\pi/4}, e^{j\theta}e^{-j\pi/4}\}$. Therefore, in general, the 2-bit TIS weight discretization function $d_2(\cdot)$ can be expressed as:

$$d_2(\gamma) = \frac{\sqrt{2}}{2}e^{j\theta}\big(bin\left(\gamma_{re}\right) + j \times bin\left(\gamma_{im}\right)\big). \tag{7}$$

Since the constant phase offset θ depends on the specific TIS design and the operating configurations, without loss of generality, in this work we use $\theta = -\frac{\pi}{4}$ so that the 2-bit discrete weights are chosen from the set $\{1, e^{j\pi/2}, e^{j\pi}, e^{j3\pi/2}\}$.

During training, each TIS-FC layer keeps full-precision γ_{re} and γ_{im}, and applies weight discretization function $d_1(\cdot)$ or $d_2(\cdot)$ to get discrete weights during forward propagation. However, since the discretization function is a step function with zero gradient everywhere, the gradient calculation cannot back propagate to the previous layers during training. To solve this problem, we adopt the straight-through estimator (STE) approach, which simply bypasses the zero-gradient layers when computing backpropagation through the discretization function. This is a method commonly used in the hardware-constrained machine learning literature [10], [11].

IV. Evaluation

A basic TIS-NNL is simulated to validate the proposed training method, using a digit classification task for performance evaluation. The simulation setup is introduced first, followed by a comparison of three TIS-NNL structures using continuous phase and amplitude adjustment, 2-bit phase adjustment, and 1-bit phase adjustment, respectively.

A. Simulation Settings

To validate the proposed training approach, the TIS-NNL illustrated in Fig. 2 is trained to conduct a digit classification task using the UCI Digits dataset [12], which includes handwritten digits from 0 to 9. The TIS-NNL structure has one input layer and two TIS-FC layers (referred to as TIS-FC1 and TIS-FC2 in Fig. 3). The RF signal transmitted from TIS-FC2 is received by a uniform planar array of receive antennas at the edge device. To simplify complexity of the edge device, it is assumed that only four receive antennas are used. The received signal is then passed to digital complex-valued fully connected layers with activation function in the edge device to increase dimensionality for digit classification. In what follows, the TIS dimensions in TIS-FC1 and TIS-FC2 are selected as 16×16 and 8×8, respectively. The number of neurons in the two complex-valued FC layers are 6 and 10, respectively. The distance between the transmitter and the center of the input layer, the distance between centers of the two TIS-FC layers, and the distance between the center of TIS-FC2 and the center of the receive antenna array are all set as 0.05 m, which is chosen based on existing stacked TIS designs [7]. It is assumed that the operating frequency is 28 GHz, the noise power is -100 dBm, and the receive antenna spacing and the TIS element spacing for each TIS are both $\frac{\lambda}{2}$.

It is assumed that the reconfigurable input layer in the TIS-NNL has 2-bit phase quantization, where each pixel's value is quantized into 4 discrete phases within the transmission coefficients of TIS elements on the input layer. The 2-bit input layer works well for the selected dataset which will be demonstrated in the following sections. However, for more complex datasets, finer quantization at the input layer may be required to capture more information of the input data. The neural networks used in the simulation are defined as follows.

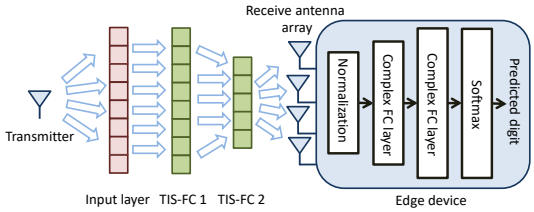


Fig. 3. Illustration of TIS-NNL.

- **WeightC**: This TIS-NNL features continuous amplitude and phase adjustment for the two TIS-FC layers. This TIS-NNL represents an ideal case, with the trainable weights in TIS-FCs offering maximum flexibility.
- **Weight2**: The two TIS-FC layers feature 2-bit phase quantization and no amplitude adjustment.
- **Weight1**: The transmission coefficients in the two TIS-FC layers have 1-bit phase quantization. This case represents the minimal hardware complexity but limited flexibility for trainable weights in TIS-FC layers.
- **NoTisFc**: For the purpose of ablation study on the impact of TIS-FC layers, a case without the two TIS-FC layers is tested. Herein, the input layer maintains 2-bit phase quantization, but no TIS-FC layers are utilized. Additionally, the positions of the transmitter, input layer, and receive antenna array on the edge device remain consistent with other cases.

B. Validation of 1-bit and 2-bit TIS-FC Layers

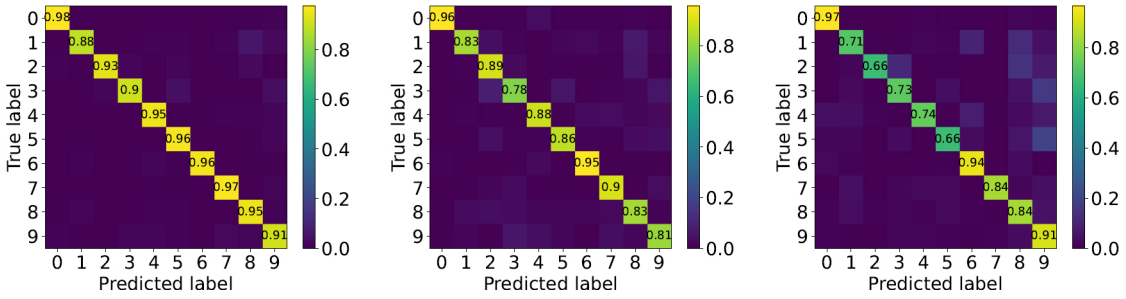
To validate the proposed training method for 1-bit and 2-bit TIS-NNLs, **WeightC**, **Weight2**, and **Weight1** are trained for the digit classification task given that the transmit power is 10 dBm. The confusion matrices for the three cases are shown in Fig. 4. The true positive rates for **WeightC**, **Weight2**, and **Weight1** are 93.96%, 87.10%, and 79.86%, respectively.

According to Fig. 4, **WeightC** shows the highest classification performance, followed by **Weight2** and **Weight1**. This highlights the advantage of continuous adjustment of amplitude and phase when higher complexity is allowed for the TIS. Additionally, despite a roughly 7% decrease in the overall true positive rate compared to **WeightC**, **Weight2** still achieves a rate close over 87%. Furthermore, when utilizing **Weight1**, the reduction in true positive rate is approximately 7% compared to **Weight2**. This validates the proposed training approach for 1-bit and 2-bit TISs, further highlighting the efficacy of employing low-complexity TIS-NNLs while preserving satisfactory performance.

C. Effect of Transmit Power

To provide a more thorough analysis, the effect of transmit power on digit classification accuracy is evaluated across all four types of neural networks discussed in Sec. IV-A. The simulation results are shown in Fig. 5. The insights derived from this figure are summarized below:

- When comparing the performance between the baseline case of **NoTisFc** and all the other cases, it is observed that there is a significant improvement by adding TIS-FC layers. This highlights the importance of the TIS-FC layers in the classification task, and shows that



(a) **WeightC** (b) **Weight2** (c) **Weight1**

Fig. 4. Confusion matrix.

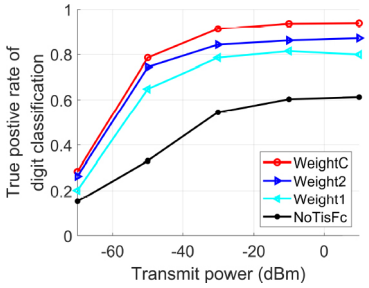


Fig. 5. Transmit power vs. true positive rate for different neural networks.

classification does not solely rely on the classic FC-layers within the edge device.

- The classification accuracy relies on the signal-to-noise ratio (SNR) at the receive antennas of the edge device. Despite mmWave signals suffering from high path loss compared to other frequency bands, closely placing TIS-FC layers together reduces the impact of path loss on the accuracy. Additionally, an appropriately sized TIS array can enhance SNR. As shown in Fig. 5, with the chosen system settings, a robust true positive rate is maintained when transmit power is above -30 dBm.
- When transmit power does not exceed -30 dBm, the true positive rate gap between **WeightC** and **Weight2** remains under 8%, and the gap between **Weight2** and **Weight1** stays within 10%. This demonstrates the robustness and effectiveness of the proposed 2-bit and 1-bit TIS-NNLs for complexity reduction with varied SNR conditions.

V. Conclusion

In this work, we presented an OTA neural network architecture based on low-complexity TISs. We first showed that one of the building blocks of DNNs, the fully connected layer, can be modeled by cascaded TISs, where the TIS phase shifts can be trained to perform specific operations on RF signals. Considering practical constraints of TIS design in mmWave frequencies, e.g. signal attenuation, cost and complexity, we proposed a novel discretized complex-valued neural network structure, the TIS-NNL, and training method that are suitable for 1-bit and 2-bit TIS designs. In simulation, it is demonstrated that a trained TIS-NNL is able to achieve high accuracy on a benchmark image classification task, even when using 1-bit or 2-bit TISs.

References

[1] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge ai: On-demand accelerating deep neural network inference via edge computing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 447–457, 2020.

[2] G. Reus-Muns, K. Alemdar, S. G. Sanchez, D. Roy, and K. R. Chowdhury, "Airfc: Designing fully connected layers for neural networks with wireless signals," in *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, ser. MobiHoc '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 71–80. [Online]. Available: https://doi.org/10.1145/3565287.3610281

[3] J. Tang, M. Cui, S. Xu, L. Dai, F. Yang, and M. Li, "Transmissive ris for b5g communications: Design, prototyping, and experimental demonstrations," *IEEE Transactions on Communications*, vol. 71, no. 11, pp. 6605–6615, 2023.

[4] S. Zhang, H. Zhang, B. Di, Y. Tan, Z. Han, and L. Song, "Beyond intelligent reflecting surfaces: Reflective-transmissive metasurface aided communications for full-dimensional coverage extension," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 905–13 909, 2020.

[5] G. Reus-Muns, K. Alemdar, S. G. Sanchez, D. Roy, and K. R. Chowdhury, "Airfc: Designing fully connected layers for neural networks with wireless signals," in *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, ser. MobiHoc '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 71–80. [Online]. Available: https://doi.org/10.1145/3565287.3610281

[6] C. Liu, Q. Ma, Z. J. Luo, Q. R. Hong, Q. Xiao, H. C. Zhang, L. Miao, W. M. Yu, Q. Cheng, L. Li *et al.*, "A programmable diffractive deep neural network based on a digital-coding metasurface array," *Nature Electronics*, vol. 5, no. 2, pp. 113–122, 2022.

[7] J. An, M. Di Renzo, M. Debbah, and C. Yuen, "Stacked intelligent metasurfaces for multiuser beamforming in the wave domain," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 2834–2839.

[8] C. Y. Lee, H. Hasegawa, and S. Gao, "Complex-valued neural networks: A comprehensive survey," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, pp. 1406–1426, 8 2022.

[9] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1T2hmZAb

[10] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

[11] P. Yin, J. Lyu, S. Zhang, S. Osher, Y. Qi, and J. Xin, "Understanding straight-through estimator in training activation quantized neural nets," 3 2019. [Online]. Available: http://arxiv.org/abs/1903.05662

[12] E. Alpaydin and C. Kaynak, "Optical Recognition of Handwritten Digits," UCI Machine Learning Repository, 1998, DOI: https://doi.org/10.24432/C50P49.