Attention-Based Hybrid Precoding for mmWave MIMO Systems

Hao Jiang*, Yu Lu*, Xueru Li†, Bichai Wang†, Yongxing Zhou†, and Linglong Dai*
*Beijing National Research Center for Information Science and Technology (BNRist),
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
†Huawei Technologies Company, Ltd., Shenzhen, Guangdong 310051, China
Email: jiang-h18@mails.tsinghua.edu.cn

2021 IEEE Information Theory Workshop (ITW) | 978-1-6654-0312-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITW48936.2021.9611432

***Abstract*—Hybrid precoding design is a high-complexity problem due to the coupling of analog and digital precoders as well as the constant modulus constraint for the analog precoder. Fortunately, the deep learning based hybrid precoding methods can significantly reduce the complexity, but the performance remains limited. In this paper, inspired by the attention mechanism recently developed for machine learning, we propose an attention-based hybrid precoding scheme for millimeter-wave (mmWave) MIMO systems with improved performance and low complexity. The key idea is to design each user's beam pattern according to its attention weights to other users'. Specifically, the proposed attention-based hybrid precoding scheme consists of two parts, i.e., the attention layer and the convolutional neural network (CNN) layer. The attention layer is used to identify the features of inter-user interferences. Then, these features are processed by the CNN layer for the analog precoder design to maximize the achievable sum-rate. Simulation results demonstrate that the attention layer could mitigate the inter-user interferences, and the proposed attention-based hybrid precoding with low complexity can achieve higher achievable sum-rate than the existing deep learning based method.**

I. Introduction

Millimeter-Wave (mmWave) multiple-input multiple-output (MIMO) has been regarded as one of the key techniques for 5G wireless communications [1], [2]. For the classical fully-digital precoding, the hardware cost and power consumption are unaffordable due to the use of a very large number of expensive radio-frequency (RF) chains, which have high energy consumption (about 250 mW per RF chain [3]). To address this issue, hybrid precoding has been proposed by using fewer RF chains to design a low-dimensional digital precoder [4]. At the same time, the analog phase shifters (PSs) with low cost and power consumption are introduced to design the high-dimensional analog precoder to achieve high array gains. However, the joint design of analog and digital precoders is difficult due to the constant modulus constraint of the analog PSs [5].

Existing dominant hybrid precoding algorithms could be generally divided into two categories, i.e., the codebook-based beamforming and the non-codebook beamforming. For the codebook-based beamforming, the analog beam to be used is acquired by searching from the predefined codebook, aiming to maximize the desired received signal power for each user. An intuitive design of the codebook is the discrete Fourier transform (DFT) codebook [6], in which the beams are orthogonal to each other, and all beams cover the entire beam domain. Unfortunately, this method suffers from a low degrees of freedom due to the finite codebook space, and consequently the performance is limited. By contrast, the non-codebook beamforming could obtain the near-optimal performance by optimizing the hybrid precoders to approach the performance of the fully digital precoding. In [7]–[12], the manifold optimization [7], Barzilai-Borwein gradient [8], matrix decomposition [9], gradient projection [10], geometric mean decomposition [11], and complex oblique manifold [12] are utilized to minimize the Euclidean distance between the hybrid precoder and the fully digital precoder. However, the optimization algorithms mentioned above have high computational complexity.

To reduce the complexity of the optimal hybrid precoding designs, several deep learning (DL) based hybrid precoding methods have been recently proposed by using low-complexity neural networks [13]–[17]. The DL-based methods could be generally divided into two categories according to the learning style, i.e., the supervised learning based method and the unsupervised learning based method. For supervised learning based method [14], the hybrid precoders are labeled according to the classical singular value decomposition (SVD), which requires a lot of computation resources to calculate the SVD. Moreover, it is difficult to have a performance gain based on supervised learning, since the results are bounded by classical optimization algorithms [15]. Unlike supervised learning, the unsupervised learning based method [16] learns the precoder without labeled samples. In this case, the unsupervised learning methods directly take sum-rate as loss function to train the DL model. However, the existing works just use simple fully-connected (FC) neural networks, which is difficult to learn the constrained analog precoder. As a result, the performance of unsupervised learning based method remains limited.

To improve the performance of the unsupervised learning based method without using labeled samples, in this paper, inspired by the recently developed attention mechanism [18], we propose an attention-based hybrid precoding scheme for mmWave MIMO systems. To be specific, the proposed attention-based hybrid precoding mainly consists of the following two parts. First, based on the attention mech-

978-1-6654-0312-2/21/$31.00 ©2021 IEEE

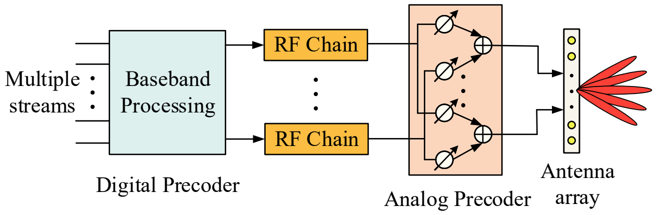

Fig. 1. The architecture of hybrid precoding.

anism, we design the attention layer to identify the features of inter-user interferences. Then, we apply a low-complexity convolutional neural network (CNN) layer to learn the analog precoder according to the features of inter-user interferences. Simulation results show that the proposed low-complexity attention-based hybrid precoding could achieve better sum-rate performance than existing unsupervised learning based scheme.

Notation: We denote the column vectors and matrices by boldface lower-case and upper-case letters, respectively. $(\cdot)^T, (\cdot)^H, (\cdot)^{-1}$ denote the transpose, the conjugate transpose, and the inverse of the matrix, respectively. $\mathbf{X}_{i,j}$ denote the element of the matrix $\mathbf{X}$ in row i and column j. $|\cdot|$ and $\|\cdot\|_2$ denote the absolute value and l_2-norm, respectively. $\mathcal{CN}(0,1)$ is the standard complex Gaussian distribution with mean 0 and variance 1.

II. System Model and Problem Formulation

We consider a downlink hybrid precoding mmWave massive MU-MIMO system, as shown in Fig. 1, in which the base station (BS) are deployed with N_t antennas and N_t^{RF} RF chains to transmit N_s data streams in parallel. In the practical hybrid precoding system, the number of RF chains is far less than the number of antennas, i.e., $N_t^{\text{RF}} < N_t$. Assuming there are K single antenna users to be severed by a BS, the downlink system model can be given by

$$\mathbf{y} = \sqrt{\rho}\mathbf{HADs} + \mathbf{n}, \tag{1}$$

where ρ is the transmitted power, $\mathbf{y} = [y_1, \cdots, y_k]^T \in \mathbb{C}^{K\times 1}$ is the received signal vector for N_s single-antenna users; $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_k]^T \in \mathbb{C}^{K\times N_t}$ is the channel matrix; $\mathbf{A} \in \mathbb{C}^{N_t \times N_t^{\text{RF}}}$ and $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_{N_s}] \in \mathbb{C}^{N_t^{\text{RF}} \times N_s}$ are analog precoder and digital precoder matrices, respectively; $\mathbf{s} \in \mathbb{C}^{N_s \times 1}$ is the transmitted signal; and $\mathbf{n} \in \mathbb{C}^{K\times 1}$ is the received zero mean additive white Gaussian noise (AWGN).

At the BS, the N_s independent data streams in baseband are processed by the well-designed low-dimensional digital precoder $\mathbf{D}$. Then, the RF chains convert the digital signals into analog signals. Finally, the high-dimensional analog precoder $\mathbf{A}$ shapes the transmitted beam to users by a PS network, which has low hardware cost and energy consumption [4], so each element of $\mathbf{A}$ should satisfy $|\mathbf{A}_{i,j}|^2 = 1/N_t$. For the convenience of our discussion, we suppose $K = N_s = N_t^{\text{RF}}$, which means that each data stream serves one user.

It is well known that mmWave channel is sparse due to the limited number of scatters in mmWave propagation environment [19]. So, in this paper, we adopt the geometric Saleh-Valenzuela channel model [20] to describe the mmWave channel, which is represented as

$$\mathbf{h}_k = \sqrt{\frac{N_t}{L}} \sum_{l=1}^{L} \alpha_l \mathbf{f}_t(\phi_l^k), \tag{2}$$

where L is the number of the effective signal propagation paths, and we usually have $L \leq N_t$; α_l is the complex gain of the lth path; ϕ_l^k is the angles of departure (AoDs) of the lth path and kth user; and $\mathbf{f}_t(\phi_l^k)$ is the antenna array steering vector, which depends on the BS array geometry. When we consider the widely used uniform linear arrays (ULAs), we have

$$\mathbf{f}_t(\phi_l^k) = \frac{1}{\sqrt{N_t}}[1, e^{j\frac{2\pi}{\lambda} d\sin(\phi_l^k)}, \cdots, e^{j(N_t-1)\frac{2\pi}{\lambda} d\sin(\phi_l^k)}]^T, \tag{3}$$

where λ is the wavelength of the transmitted signal, and d is the antenna spacing.

Then, the received signal-to-interference-plus-noise ratio (SINR) at the kth user can be denoted by

$$\rho_k = \frac{\rho|\mathbf{h}_k^T \mathbf{A} \mathbf{d}_k|^2}{\sigma^2 + \rho \sum_{j\neq k}^{K} |\mathbf{h}_k^T \mathbf{A} \mathbf{d}_j|^2}, \tag{4}$$

where σ^2 is the variance of the AWGN. In this paper, we aim to maximize the total achievable rate expressed as $R = \sum_{k=1}^{K} R_k$, where $R_k = \log_2(1+\rho_k)$ is the data rate of the kth user. The optimal digital precoder $\mathbf{D}_{\text{opt}}$ and analog precoder $\mathbf{A}_{\text{opt}}$ could be found by solving the optimization problem formulated as

$$\begin{aligned} \max_{\mathbf{A}_{\text{opt}}, \mathbf{D}_{\text{opt}}} \quad & R(\mathbf{H}, \mathbf{A}, \mathbf{D}) \\ s.t. \; & |\mathbf{A}_{i,j}|^2 = 1/N_t \\ & ||\mathbf{d}_k||_2^2 = 1/K, \; \forall k = 1, 2, \cdots, K. \end{aligned} \tag{5}$$

Unfortunately, it can be seen that (5) is a non-convex optimization problem due to the non-convex objective function and the constant modulus constraints. The optimal solution could be found by the exhaustive search method. However, the extremely high complexity makes exhaustive search impossible in the limited channel coherence time. Instead, the optimization techniques [7]–[10], [12] can be applied to maximize the achievable sum-rate, nevertheless, these techniques still suffer from high complexity.

III. Attention-Based Hybrid Precoding

In this section, to reduce the complexity of hybrid precoding for MIMO, we propose an attention-based hybrid precoding scheme, which contains the attention layer and CNN layer. First, inspired by the attention mechanism [18], we design the attention layer to identify the features of inter-user interferences. Then, we apply a low-complexity CNN layer to learn the analog precoder according to the learned features of inter-user interferences.

A. Overview of attention mechanism

The attention mechanism could be regarded as a resource allocation mechanism [21]. Unlike conventional deep learning techniques, which distribute the same resources for each component of the signal, the attention mechanism distributes more resources to more important components. Intuitively, the attention mechanism can be explained by the human visual mechanism. Specifically, the human visual systems obtain the key areas after scanning the visual field, and then the visual systems pay more attention to the stimuli information from these more important areas, while restraining the information from less important areas. By distributing different resources to different areas, the human visual system is able to greatly reduce the complexity and improves the accuracy of visual signal processing [21].

To realize resource allocation based on importance, the attention mechanism introduces the normalized attention matrix to measure the importance degree between the output and each part of the input. The important parts are given larger weights, and then the normalized attention matrix is multiplied by the input as a mask. Especially, taking advantage of the attention matrix, we can also measure the importance between the different components of the input, which is called the self-attention mechanism. Next, we will introduce the self-attention mechanism, which is mainly considered in this paper.

For the input $\mathbf{X} \in \mathbb{R}^{n\times m}$, which has n independent components and the embedding dimension of each component is m, there are several basic matrices that are used to characterize the self-attention mechanism: the query matrix, the key matrix, the value matrix, and attention matrix.

(1) *Query matrix:* Aim to match or query the importance of other components. $\mathbf{Q} = W^Q(\mathbf{X}) \in \mathbb{R}^{n\times d}$, where W^Q denotes the linear transformation of the input $\mathbf{X}$, d denotes the query dimension of the each component.

(2) *Key matrix:* Aim to be matched or queried the importance by other components. $\mathbf{K} = W^K(\mathbf{X}) \in \mathbb{R}^{n\times d}$, where W^K denotes the linear transformation of the input $\mathbf{X}$, d denotes the key dimension of the each component.

(3) *Value matrix:* Aim to extract or keep the feature of the input. $\mathbf{V} = W^V(\mathbf{X}) \in \mathbb{R}^{n\times m}$, where W^V denotes the linear transformation of the input $\mathbf{X}$. In addition, the value matrix could be obtained without linear transformation, i.e., $\mathbf{V} = \mathbf{X}$.

(4) *Attention matrix:* Aim to measure the importance between the different components of the input, which could be denoted by $\mathrm{Softmax}(\frac{\mathbf{QK}^T}{\sqrt{d}}) \in \mathbb{R}^{n\times n}$, where the Softmax normalization is adopted to make each row add up to 1. The element in row i and column j denotes the attention weight of component i to component j.

Then, we get the weighted output of the attention mechanism, which could be represented as the product of normalized attention matrix and value matrix as follows:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}(\frac{\mathbf{QK}^T}{\sqrt{d}})\mathbf{V}, \qquad (6)$$

where the attention matrix is adjusted by $\sqrt{d}$, since the variance of the product of $\mathbf{Q}$ and $\mathbf{K}^T$ increases as d increases. It is useful to balance the increased variance by dividing $\sqrt{d}$ [18].

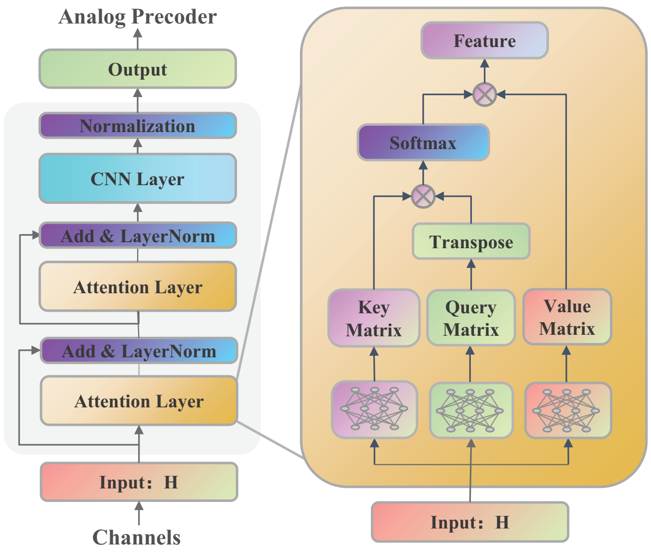

Fig. 2. The structure of the proposed attention-based hybrid precoding scheme.

B. Attention layer

In this subsection, we present the proposed attention-based scheme for hybrid precoding in mmWave massive MU-MIMO system, utilizing the attention mechanism to process multi-user raw channel state information (CSI) and identify the interferences between users. The CSI could also be obtained by the DL-based channel estimation method [22]. The structure of the attention-based neural network model is shown in Fig. 2, which is comprised of the attention layer and CNN layer.

First of all, the input channel $\mathbf{H}$ is processed by the attention layer. In the attention layer, we firstly transform the complex value input $\mathbf{H} \in \mathbb{C}^{K\times N_t}$ into real value $\mathbf{H}^r \in \mathbb{R}^{K\times 2N_t}$ by moving the imaginary part to the embedding dimension. Next, according to the introduction of the attention mechanism in subsection III-A, there are three independent linear transformations $W^Q(\mathbf{H}^r), W^K(\mathbf{H}^r)$, and $W^V(\mathbf{H}^r)$ to simultaneously calculate the corresponding query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$, and value matrix $\mathbf{V}$, respectively. In general, the linear transformation could be implemented by a linear FC network without an activation function. Then, the attention matrix $\mathrm{Softmax}(\frac{\mathbf{QK}^T}{\sqrt{d}}) \in \mathbb{R}^{K\times K}$ is calculated, which denotes the level of attention among users. At last, by adding residual connection and layer normalization, we can get the output of the attention layer, which could be represented as follows:

$$\mathbf{H}^r_{\mathrm{Att}} = \mathrm{Layernorm}(\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) + \mathbf{H}^r), \qquad (7)$$

where $\mathrm{Layernorm}(\mathbf{X}) = \frac{\mathbf{X}_{i,j}-\mu_i}{\sqrt{\sigma_i^2+\epsilon}}$, and μ_i and σ_i^2 denote the expectation and variance of the ith row of $\mathbf{X}$, respectively. The normalization operation in this layer is used to speed up the training by normalizing the data into the standard normal distribution. To avoid the denominator to be zero, a small

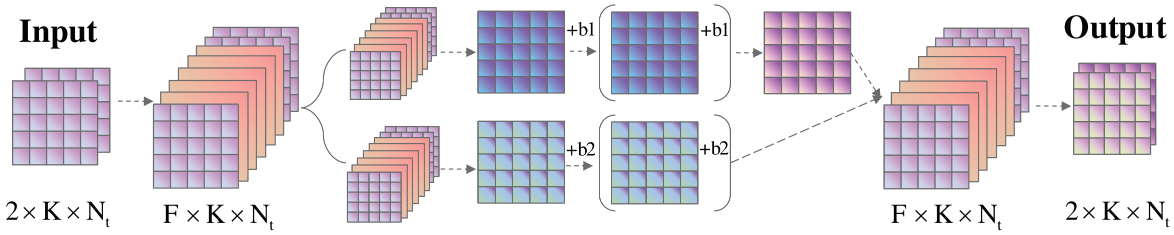

Fig. 3. The structure of the CNN layer.

quantity ϵ is added in the denominator. Note that we can cascade multiple attention layers to improve the accuracy of the self-attention algorithm.

C. CNN layer

Further, we apply the low complexity CNN layer as shown in Fig. 3 to learn the analog precoder matrix according to the features of inter-user interferences. In the CNN layer, we firstly reshape the input of the CNN layer $\mathbf{H}^r_{\text{Att}} \in \mathbb{R}^{K\times 2N_t}$ to $\mathbb{R}^{2\times K\times N_t}$, which could be considered to have real and imaginary two-channels. Next, the channel number of the reshaped input is extended to $F \gg 2$, to extract more features. Then, we divide the F-channels features into two streams, and we apply convolution kernels in different sizes to extract features at different scales for two streams. At last, the features at different scales are integrated into real and imaginary two-channels matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{2\times K\times N_t}$, which is waiting to be normalized.

Finally, to satisfy the constant modulus constraint of the analog precoder, the real and imaginary two-channel $\widetilde{\mathbf{A}}$ should be normalized in channel dimension as follow:

$$\mathbf{A}_{j,i} = \frac{1}{\sqrt{N_t}} \frac{\widetilde{\mathbf{A}}_{1,i,j} + j\widetilde{\mathbf{A}}_{2,i,j}}{\sqrt{\widetilde{\mathbf{A}}^2_{1,i,j} + \widetilde{\mathbf{A}}^2_{2,i,j}}}, \tag{8}$$

where $\mathbf{A} \in \mathbb{C}^{N_t\times K}$ is the complex value analog precoder. Then, the digital precoder is designed by the classical zero-forcing (ZF) algorithm [23] according to the effective channel $\mathbf{H}_{eq} = \mathbf{HA}$. That is to say, the digital precoder matrix $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_K]$ could ve computed as:

$$\begin{aligned} \widetilde{\mathbf{D}} &= [\widetilde{\mathbf{d}}_1, \cdots, \widetilde{\mathbf{d}}_K] = \mathbf{H}^H_{eq}(\mathbf{H}_{eq}\mathbf{H}^H_{eq})^{-1}, \\ \mathbf{d}_k &= \frac{1}{\sqrt{K}} \frac{\widetilde{\mathbf{d}}_k}{||\widetilde{\mathbf{d}}_k||_2}, \ k = 1, \cdots, K. \end{aligned} \tag{9}$$

Note that in this paper we distribute the same power to different users, and the power distribution problem could be considered in future research.

D. Unsupervised learning

To realize the training weights of the attention layer and CNN layer, we utilize unsupervised learning to maximize the achievable sum-rate in (5). Unlike supervised learning which needs lots of data annotation based on conventional algorithms, unsupervised learning could automatically learn to minimize the loss function without guiding by conventional algorithms. Moreover, the performance of supervised learning method will be bounded by the performance of conventional algorithms, so it is difficult to have a performance improvement based on supervised learning.

We define the negative number of the objective function in (5) as the loss function, which could be represented as

$$\text{loss}(W^Q, W^K, W^V, W^{CNN}, \mathcal{H}) = -\frac{1}{B}\sum_{i=1}^{B} R(\mathbf{H}^{(i)}, \mathbf{A}^{(i)}, \mathbf{D}^{(i)}), \tag{10}$$

where $\mathcal{H} = \{\mathbf{H}^{(1)}, \cdots, \mathbf{H}^{(B)}\}$ is the channel set, and the B is the batch size (the number of training samples to estimate the loss function). By using the back propagation (BP) algorithm [24] to train the weights in the attention layer and CNN layer, the minimum loss function, i.e., the maximum achievable sum-rate, could be ultimately acquired.

TABLE I
COMPLEXITY COMPARISON WITH OTHER HYBRID PRECODING ALGORITHMS.

Scheme	Complexity
Proposed attention-based	$\mathcal{O}(KN_t)$
MO-AltMin [7]	$\mathcal{O}(N_t^{\text{RF}}K^2N_t^3)$
GP-AltMin [10]	$\mathcal{O}(N_t^{\text{RF}}K^2N_t^3)$
Fast optimization [12]	$\mathcal{O}(N_t^{\text{RF}}(N_t^{\text{RF}}+K)N_t)$
CE-based [17]	$\mathcal{O}(I_{\text{iter}}K^2N_t)$
Two-stage [25]	$\mathcal{O}(KN_t)$

Table I shows the computation complexity of the attention-based hybrid precoding and other conventional algorithms. We observe that the complexity of the attention-based hybrid precoding increase linearly, while the complexity of conventional algorithms is proportional to the power of the number of users and antennas, which shows the complexity advantage of the proposed attention-based hybrid precoding scheme.

IV. SIMULATION RESULTS

In this section, we provide simulation results to evaluate the performance of the proposed attention-based hybrid precoding for mmWave massive MU-MIMO systems. The simulation parameters are described as follows. We consider the mmWave massive MU-MIMO system with hybrid precoding, where $N_t = 64$, $K = N_t^{\text{RF}} = N_s = 16$, $d = \lambda/2$. The channel is generated according to the channel model [20] described in

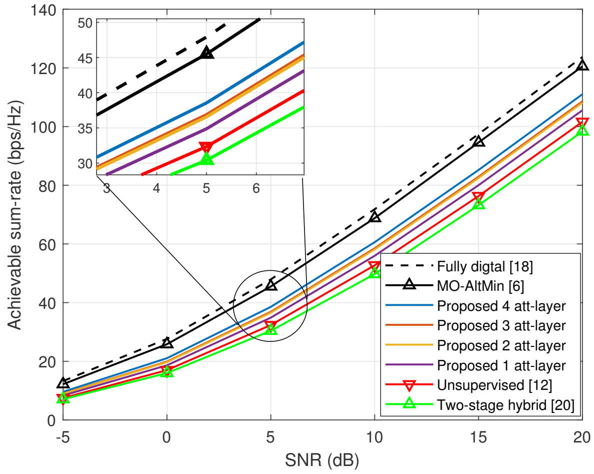

Fig. 4. Achievable rate comparison against the SNR when $K = 16$.

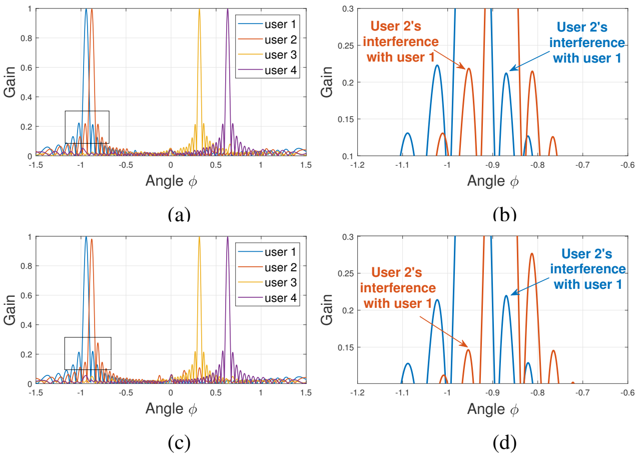

Fig. 5. (a) beam pattern realized by unsupervised learning based algorithm [16]; (b) enlarged version of Fig. 5 (a); (c) beam patterm realized by attention-based algorithm with one attention layer; (d) enlarged version of Fig. 5 (c).

Section II, in which the AoDs are assumed to follow the uniform distribution $\mathcal{U}[-\pi/2, \pi/2]$. The number of channel path $L = 3$, and the complex path gain α_l is assumed to be Gaussian, i.e., $\alpha_l \sim \mathcal{CN}(0, 1)$. The signal-to-noise ratio (SNR) is defined as $\text{SNR} = \rho/\sigma^2$.

Fig. 4 compares the achievable sum-rate against SNR of the proposed attention-based hybrid precoding scheme with that of unsupervised learning based hybrid precoding [16], conventional two-stage hybrid precoding algorithm [25], MO-AltMin optimization based hybrid precoding method [7], and fully digital precoding [23]. We can observe that the proposed attention-based hybrid precoding scheme outperforms the unsupervised learning based hybrid precoding, and it has a higher achievable sum rate with cascading more attention layers. Meanwhile, Fig. 4 also verifies the attention-based hybrid precoding scheme could achieve about 20% improvement in sum-rate compared with the classical two-stage algorithm [25].

To explain the performance improvement of the attention-based hybrid precoding, we show the beam pattern and array gain of the proposed attention-based hybrid precoding and unsupervised learning based hybrid precoding in Fig. 5. In the simulation, we assume $K = 4$, $L = 1$, and $\alpha_1 = 1$ for each user. The AoDs for four users are $\phi_1^1 = -0.3\pi, \phi_1^2 = -0.28\pi, \phi_1^3 = 0.1\pi$, and $\phi_1^4 = 0.2\pi$, respectively. After the calculation of attention layer, the normalized attention weights of user 2 to other user are [0.02, 0, 0.87, 0.11]. We observe that the attention weight for user 2 to user 3 is significantly larger than user 1, which shows that user 2 pays less attention to user 1 and pays more attention to the user 3. Thereby, the user 2's interference with user 1 could be mitigated by shifting the beam to user 3. From Fig. 5 (b), we can observe that the sidelobe jamming is significant between user 1 and user 2, which causes the reduced achievable sum-rate. By contrast, in Fig. 5 (d), user 2's sidelobe is suppressed, weakening the interference with user 1 when we use one attention layer.

V. Conclusions

In this paper, we have proposed an attention-based hybrid precoding scheme, to improve the performance of the deep learning based hybrid precoding scheme for the mmWave MIMO systems. Specifically, inspired by the attention mechanism, we designed the attention layer to identify the interferences among users, so that the interference features can be obtained. Then, we designed a CNN layer to learn the analog precoder according to interference features. Furthermore, we applied unsupervised learning to train the weights of the attention layer and the CNN layer to maximize the achievable sum-rate. It is showed that the proposed attention-based hybrid precoding scheme could realize the lower complexity compared with conventional algorithms. Simulation results verified the 20% sum-rate performance improvement of the proposed attention-based hybrid precoding scheme compared with the classical algorithm. For future research of hybrid precoding for mmWave MIMO systems, we will focus on user pairing and scheduling policies based on deep learning techniques.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (Grant No. 2020YFB1805005), in part by the National Natural Science Foundation of China (Grant No. 62031019), and in part by the European Commission through the H2020-MSCA-ITN META WIRELESS Research Project under Grant 956256.

References

[1] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, "Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communications using lens antenna array," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2370–2382, Oct. 2017.

[2] L. Dai, B. Wang, M. Peng, and S. Chen, "Hybrid precoding-based millimeter-wave massive MIMO-NOMA with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 131–141, Jan. 2019.

[3] P. V. Amadori and C. Masouros, "Low RF-complexity millimeter-wave beamspace-MIMO systems by beam selection," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2212–2223, Jun. 2015.

[4] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[5] X. Gao, L. Dai, and A. M. Sayeed, "Low RF-complexity technologies to enable millimeter-wave MIMO with large antenna array for 5G wireless communications," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 211–217, Apr. 2018.

[6] D. Love and R. Heath, "Limited feedback unitary precoding for spatial multiplexing systems," *IEEE Trans. Inf. Theory*, vol. 51, no. 8, pp. 2967–2976, Aug. 2005.

[7] X. Yu, J.-C. Shen, J. Zhang, and K. B. Letaief, "Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 485–500, Apr. 2016.

[8] M. Mulla, A. H. Ulusoy, A. Rizaner, and H. Amca, "Barzilai-borwein gradient algorithm based alternating minimization for single user millimeter wave systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 508–512, Apr. 2020.

[9] W. Ni, X. Dong, and W.-S. Lu, "Near-optimal hybrid processing for massive MIMO systems via matrix decomposition," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 3922–3933, Aug. 2017.

[10] J.-C. Chen, "Gradient projection-based alternating minimization algorithm for designing hybrid beamforming in millimeter-wave MIMO systems," *IEEE Commun. Lett.*, vol. 23, no. 1, pp. 112–115, Jan. 2019.

[11] T. Xie, L. Dai, X. Gao, M. Z. Shakir, and J. Li, "Geometric mean decomposition based hybrid precoding for mmwave massive MIMO systems," *China Commun.*, vol. 15, no. 5, pp. 229–238, May. 2018.

[12] H. Kasai, "Fast optimization algorithm on complex oblique manifold for hybrid precoding in millimeter wave MIMO systems," in *Proc. IEEE Global Conference on Signal and Inf. Process. (IEEE GlobalSIP'18)*, Nov. 2018, pp. 1266–1270.

[13] L. Dai, R. Jiao, F. Adachi, H. V. Poor, and L. Hanzo, "Deep learning for beamspace channel estimation in millimeter-wave massive MIMO systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 133–139, Aug. 2020.

[14] T. Peken, S. Adiga, R. Tandon, and T. Bose, "Deep learning for SVD and hybrid beamforming," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6621–6642, Oct. 2020.

[15] Y. Lu and L. Dai, "Reconfigurable intelligent surface based hybrid precoding for THz communications," *arXiv preprint arXiv:2012.06261*, Dec. 2020.

[16] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink MIMO," *IEEE Access*, vol. 7, pp. 7599–7605, Dec. 2019.

[17] Y. Zhang, X. Dong, and Z. Zhang, "Machine learning-based hybrid precoding with low-resolution analog phase shifters," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 186–190, Jan. 2021.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Information Processing Systems (NIPS'17)*, Jun. 2017, pp. 6000–6010.

[19] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[20] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[21] S. Treue and S. Katzner, "Visual attention," *Encyclopedia of Neuroscience*, pp. 243–250, 2009.

[22] X. Wei, C. Hu, and L. Dai, "Deep learning for beamspace channel estimation in millimeter-wave massive MIMO systems," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 182–193, Jan. 2021.

[23] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[25] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.