Review of Massive MIMO Technology Utilizing Artificial Intelligence

Sanxing Hu, Guo Cheng, Hongyu Chen, Songyi Cao
Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China

Keywords: Massive MIMO; AI, CE-NET; OAMP-NET; model-driven method.

Abstract: In 5G, Massive MIMO has been regarded as an emerging solution to enhance air interface and improve spectrum efficiency. On the other hand, with the rapid development of artificial intelligence applied in our life, Intelligent communication is one of the most attractive directions in wireless communication in the future. Currently the combination of massive MIMO and artificial intelligence is still in preliminary stage. This paper discusses the latest research progress of AI-aided method in massive MIMO including channel estimation, signal detection, hybrid beamforming and end-to-end wireless communication system. Specially, CE-NET and OAMP-NET based on model-driven method is analysed detailly. On this basis, the development trend of AI-aided communication is forecasted in the future, where model-driven aided wireless communication technology will be more competitive after 5G.

1. Introduction

![Figure 1. Outline of wireless communication system](image)

Since 2010, 5G has been arising much attention from academia and industry with its features: ultra-reliability, low-latency, giant channel capacity and dense networks. To achieve this performance, Massive MIMO [1] as well as millimeter wave communication has been regarded as an emerging solution to enhance air interface and improve spectrum efficiency. Though mmWave with high frequency, 30-300GHz, is influenced by the severe propagation pathloss and rain attenuation apparently, it can leverage large antenna array, MIMO, to combat pathloss with beamforming gain. In short mmWave can improve network throughout, spectrum utilization while Massive MIMO allows the space resources to be exploited fully based on the high antenna space dimension, resulting...
in wireless big data and laying foundation for training and testing in Artificial Intelligence-aided wireless communication.

Currently artificial intelligence especially deep learning makes successful achievements in computer vision, NLP etc. Researchers in wireless communication expect to apply it in physical layer to achieve IoE and meet the increasing transmission demand. Hence, Intelligent communication is considered to be one of the mainstream directions in the development of wireless communication after 5G. Though wireless big data could provide a chance for AI applied in communication system, deep learning for wireless communication is still in preliminary stage and it will be faced with many challenges.

Initially, the communication system is just considered as a black-box and an end-to-end DL architecture is used for signal transmission/reception. Encoding, decoding, channel estimation, signal detection and all other functionalities of a communication link are embedded in the DL-block implicitly [2]. However, this kind of data-driven method are faced with many challenges due to large amount of training data and the limitation of hardware impletion.

This paper will discuss the progress of research in deep learning applied in modern communication system, including channel estimation, signal detection, beamforming and end-to-end wireless communication system. To better illustrate it, the block diagram of the wireless communication system in shown below.

2. Deep Learning in Wireless Physical Layer Technology

2.1 Channel Estimation

Channel estimation and signal detection shown in figure 1 is essential in communication system in order to recover the transmitted signal at receiver from noisy channel based on channel state information (CSI). In massive MIMO mmWave system with limited RF chain and dense antenna array scenario, channel estimation and signal detection become more challenge than it before. Large number of antennas in base station and many carriers result in the increasing overhead of spectrum resources and computational complexity. Hence how to reduce the algorithm complexity in Channel estimation and signal detection under the guarantee of normal performance is still a problem need to be solved in future. In this part, we will focus on channel estimation based on deep learning. Later, AI-aided method for signal detection in MIMO will be discussed.

Currently, art-of-state optimization methods regarding estimating channel information in Massive MIMO are OMP, CoSaMP algorithm [3-4] based on compressive sampling. Thanks to the recent advances in deep learning, some study has shown that whether data-driven or model-driven model can reduce the complexity effectively for some specific optimization problems in wireless communication, including channel estimation and signal detection. It even outperforms the conventional design in specific communication environment or under special transmission condition.

Based on this, in [5], the authors propose a channel estimator based on deep learning from basic MMSE algorithm and then use CNN to compensate the error. Specially, channel is modelled as conditional Gaussian distribution. If the random covariance matrix is Toeplitz and have a shifted-invariance structure, computation complexity of channel estimator will decrease significantly. Otherwise, it will be extremely high. Simulation results shows that the NN based channel estimator could reduce the complexity as well as guarantee the accuracy of channel estimation.

On the other hand, some researchers [6] consider the channel under Massive MIMO mmWave communication as a 2D image since it is sparse, and the changes between adjacent elements are subtle which is highly similar to 2D image and then use AI-based image processing method to estimate channel information. Reference [6] exploit LDAMP network which is based on D-AMP algorithm and can learn channel structure and estimate channel from a large amount of training data. LDAMP network. The denoising convolutional neural network, playing an important role in LDAMP network, incorporates the channel matrix into the iterative sparse signal recovery algorithm for channel estimation. The DnCNN denoiser can handle Gaussian denoising with an unknown noise
level, which is more accurate and faster than competing techniques. Instead of learning a mapping directly from a noisy image to a denoised image, learning the residual noise is beneficial which improves both the training times and accuracy of a network. It is shown that this network outperforms state-of-art compressed sensing-based algorithms. Motivated by this, [7] consider the time-frequency response of a mmWave communication as an image similarly and present a DL-based framework in OFDM systems. Specifically, the channel response in pilot positions is considered as a LR image and the estimated channel as a HR image. Following that SRCNN is cascaded with a denoising IR network to estimate channel. The result illustrates that the DL-based algorithm is comparable to the minimum mean square error (MMSE) with full knowledge of the channel statistics, which can be used in channel estimation currently.

2.2 Signal Detection

DL-based method does not only help to estimate channel, but also can be used for detecting the input symbols. In fact, channel estimation and signal detection are usually considered a joint optimization in intelligent communication in future.

In [8], authors use DNN to study the signal detection problem over a known channel model without CSI at the receiver. Different from traditional communication system which estimates CSI firstly and then recover the transmitted signal based on estimated CSI, the FC-DNN decoder regards channel estimation and symbol detection in an end-to-end manner like black box. That is to say, this decoder is trained to output the estimated symbols purely on received signal, which is a data-driven method. Some details in this DNN is that it consists of an input layer, three hidden layers with and output layer with 256, 500-250-120,16 respectively. The OFDM in this article is with 64 subcarriers, a 16-length CP and QPSK modulation. The proposed deep learning-based approach has much better performance than the conventional LS method and is comparable to the MMSE method. Particularly, DNN outperforms traditional MMSE method in non-linear OFDM systems without CP or small peak-to-average SNR. However, this method needs to be improved to model-driven approach since it is not able to explicitly find the channel time-frequency response and BER performs saturation.

To study signal reconstruction in MIMO system, DetNet is proposed in [9]. The authors exploit the structure of a projected gradient solution based on maximum likelihood and design the network architecture accordingly. To test the robust of proposed network DetNet, two scenarios with known CSI, time-invariant channel and time-variant channel respectively is considered. The results show that DetNet with low-complexity and extremely high accuracy achieves better performance than AMP and comparable performance with SDR (running 30 times faster).

Similarly, Reference [10] also consider the signal detection and reconstruction in MIMO and propose a model-driven method based on existed OAMP algorithm. Specifically, Authors combine iterative OAMP algorithm with deep learning network and then exploit OAMP-net which can improve the performance of existing signal detection algorithms by adding several trainable parameters. The OAMP algorithm is proposed in compressed sensing to solve the sparse linear inversion problem and is used in the signal detection in MIMO recently [11], however, the performance will decrease. Since OAMP is an iterative algorithm, the complexity will increase to

![Figure 2. Structure of OAMP-NET](image)
some degree. In order to further reduce the complexity of the algorithm, OAMP-Net shown in figure 2 is exploited [10], containing T cascade layers, which is equivalent to the iterative process of the algorithm.

Each cascade layer achieves the whole process of the OAMP algorithm and adds some trainable parameters to make the OAMP algorithm more flexible. When the parameters change, it can not only adapt to more channel scenarios, but also realize the transformation with other algorithm models. Hence convergence, stability and speed can be improved during this training. Result shows that compared with the previous classical algorithms LMMSE, the complexity of the OAMP algorithm is much lower and the performance is better, which can be used in time-variant channel. What is more, some details about OAMP-net mathematical expression will be presented in section 2.3.

2.3 AI-aided Receiver based on CE-NET & OAMP-NET

Owing to deep learning’s strong ability to address the transceiver’s imperfection (channel estimation and signal detection), wireless communication system can be potentially improved in some specific conditions. An AI-aided receiver for a CP-free OFDM system is developed by Jing Z [12] in 2019 to achieve relatively higher spectrum efficiency than traditional OFDM with enough CP, perform better performance including BER and complexity and has great robustness than traditional competitive algorithms, which is potential in future.

The basic strategy of this model-driven AI receiver is offline training based on generic dataset (channel statistics) and online recovering transmitted data in the current scenario, which is reasonably accurate than data-driven method shown in [8]. Specifically, the AI receiver includes a channel estimation network (CE-NET) and a signal detection neural network based on OAMP, called OAMP-NET. CE-NET is initialized by LS channel estimation algorithm and refined by a linear minimum mean-squared error neural network. The architecture of OAMP-NET is same to [10], improving system performance compared with OAMP algorithm by reducing the training parameters to 2.

Many types of test including BER, computation complexity, capacity based on simulation test and OTA test illustrates the proposed AI receiver has low complexity, great robustness and better BER performance than the existing algorithms. Some details about this AI-receiver will be discussed in the following part.

A. AI-aided CP-Free OFDM Systems

OFDM can effectively deal with delay spread in wireless communication and it has been used widely in wireless system [13]. The basic idea for OFDM can be described as: dividing the given channel with bandwidth B into several orthogonal subcarriers, and then using subcarrier to modulate signal in each subchannel, allowing each subcarrier to transmit concurrently. Modern wireless communication always adds Cyclic Prefix shown in figure 3.

![Figure 3. OFDM symbol with CP](image)

It can be seen easily that the part of the signal at the back of an OFDM symbol is copied and placed at the front of the signal as a prefix. The length of CP should be larger than the channel’s maximum delay spread to guarantee the Orthogonality of each subcarrier and remove ISI since it allows the signal of the previous OFDM symbol to be sufficiently attenuated in the CP of the current
OFDM symbol. Assuming the interval of OFDM symbol and subcarrier is $T_s$ and $\Delta f$ respectively, hence the interval of symbol used for transmission is $T_u = 1/\Delta f$.

Then the smooth interval can be defined:

$$N_{\text{smooth}} = \frac{1}{(T_s-T_u)\Delta f}$$

However, the longer CP with smaller ISI in OFDM will introduce larger overhead and lower available data rate. To increase spectral efficiency of OFDM systems significantly compared with traditional OFDM with enough CP and achieve low BER than classical OFDM signal receiver without CP, AI-aided CP free OFDM system is proposed to address both channel estimation and signal detection. The block diagram of the CP-Free OFDM transmitter is presented in Figure 4 [12].

![Figure 4. Block diagram of a CP-Free OFDM transmitter](image)

It can be seen clearly from the diagram that $b \in R^k$, $u \in R^n$ and $q \in R^n$ represents input bits, transmit symbols and OFDM signals after IFFT while $\hat{b}$ represents the estimated bit. Assuming the length of channel sample space is shorter than that of OFDM blocks.

For a CP-free OFDM system with $N$ subcarriers, the received signal vector $y$ can be expressed as:

$$y = C_q - A_q + A_{q-1} + w$$

$$= CF^H \mu - AF^H \mu + A_{q-1} + w$$

$$= (c - A)F^H \mu + A_{q-1} + w$$

$$= JF^H \mu + A_{q-1} + w$$

$$y = \begin{bmatrix} h_0 & 0 & \cdots & 0 & h_{l-1} & \cdots & h_2 & h_1 \\ h_1 & h_0 & 0 & \cdots & 0 & h_{l-1} & \cdots & h_2 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_{l-1} & \cdots & h_2 & h_1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & \cdots & 0 & h_{l-1} & \cdots & \cdots & h_1 \\ 0 & \cdots & 0 & h_{l-1} & \cdots & h_2 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix}$$

Where $C$ and $A$ is an $N \times N$ cyclic matrix corresponding to current OFDM signal vector and cut-off channel matrix corresponding to previous OFDM signal vector shown above. The first term of (1) denote the received signal concerning ICI in time domain, while IBI is expressed as the second term.

![Figure 5. Block diagram of AI-aided receiver](image)
At the AI-aided Receiver shown in figure 5, including two modules, CE-NET and OAMP-NET, will help to recover the transmitted signal.

B. CE-NET of AI Receiver

The input of channel estimation, $y_s^f$, is first converted into frequency domain by FFT. Following that $\hat{H}_{LS}$ is obtained by LS, which initializes CE-NET to generate more accurate channel estimation $\hat{H}_{out}$. The CE-NET is quite simple with one input layer and output layer. Since $\hat{H}_{LS}$ is always a complex matrix, we must convert it to real matrix as an input. A simple method is combining the real part and imaginary part together as a new matrix, denoted as $\hat{H}_{LS}$:

$$\hat{H}_{LS} = \begin{bmatrix} \text{Re}\{\hat{H}_{LS}\} \\ \text{Im}\{\hat{H}_{LS}\} \end{bmatrix}$$

(3)

If we consider the multiplexing of complex matrix, the principle is generally the same. The CE-NET is trained by minimizing the $l_2$ loss between predictions and knows channel samples by using a specific optimizer. It can be expressed as:

$$\text{Loss} = \arg\min_{\hat{H}_{out}} \| H - \hat{H}_{out} \|_2^2$$

(4)

In addition, we can also use the deep learning method based on LMMSE channel estimator which can improve the accuracy of the channel estimation.

C. OAMP-NET of AI Receiver

The basic idea of OAMP-NET based on OAMP algorithm is discussed in Reference [10]. Similarly, Reference [12] also use OAMP-NET to reduce the complexity by tuning 2 parameters in average. Due to the channel estimator, we get CSI and could eliminate the interference of block from the estimated channel information. After passing the CE-NET, the received signal can be expressed as:

$$\hat{y} = y - A_{q-1} = JF^H \mu + A_{q-1} + w$$

(5)

Where $w'$ represents the noise with a variance $\sigma_{w'}^2$. Then based on OAMP-NET described above in section 2.2, [12] use OAMP-NET work with $L$ layers. The network introduces two scalar trainable parameters, ($\lambda_l, \gamma_l$) instead of 4 trainable parameters: minimum allow variance $\epsilon$, damping procedure $\beta$, number of layers $L$ and initial value $\mu_1$. Thus, the improved AI-aided OAMP algorithm set parameters more flexibly and effectively. Specifically, it can be described as:

$$r_l = \mu_l + \hat{\lambda}_l P_l(y - H\hat{\mu}_l)$$

(6)

$$\nu_l^2 = \frac{\| \hat{y} - H\hat{\mu} \|^2 - M\sigma_{w'}^2}{tr(H^H H)}$$

(7)

$$\tau^2 = \frac{1}{2N} tr(D_l D^H_l) \nu_l^2 + \frac{\gamma^2}{4N} tr(P_l P^H_l) \sigma_{w'}^2$$

(8)

Where $D_l = I - \gamma_l P_l H$. $P_l$ is the optimal matrix discussed in reference [Modern-deep]. Therefore, the total number of trainable variables is equal to $2L$ if there are $L$ layers.

2.4 mmWave Hybrid Beamforming

As discussed in Section 1, mmWave beamforming is important in Massive MIMO technology. However, when the number of antennas reaches hundreds or even thousand, new problems such as expensive hardware cost and high-power consumption of the system appear, which brings the actual deployment of large-scale MIMO technology severe challenges.
One of the effective solutions is Hybrid beamforming technology. The precoding in transmitter and combining in receiver is both divided into two parts: Analog and Digital. To reduce the hardware cost and complexity, limited number of RF chains which is connected to antenna is given, then the object is to find the best solution of the precoder matrix as well as combiner matrix which approximates the full digital precoding. To solve such nonconvex problem, classical algorithm is proposed based on SVD and GMD of channel matrix [14-15].

However, these traditional optimization methods could bring some limitations containing high-computational complexity and less spatial information, [16] propose a deep-learning-based hybrid precoding based on GMD in mmWave Massive MIMO. The diagram of the architecture is shown in Figure 6, where each precoder for obtaining the decoder is regarded as mapping relation in DNN.

In [16], the DNN is simple and considered as a black box, capturing the structural information of hybrid precoding scheme through the training stage and lowering the complexity. The learning strategy is divided the channel matrix into GMD-based analog precoder $R_A$ and GMD-based digital precoder $R_D$ and then the DNN is trained by minimizing the $l_2$ loss between prediction:

$$\text{loss} = \| R_I - R_A R_D \|^2_F$$

$$= \sqrt{tr \left( \left( R_I - R_A R_D \right) \left( R_I - R_A R_D \right)^H \right)}$$

(9)

The result shows that the proposed AI-aided scheme enhance the spectrum efficiency of mmWave Massive MIMO and reduce the computational complexity compared with traditional method.

### 2.5 End-to-End Wireless Communication System

An autoencoder, widely used in machine learning for feature extraction or as generative model for data, is an unsupervised learning technique and try to recover the input at the output. Nowadays autoencoders have been successfully applied to image and video processing, which can be considered as communication over a finite-rate error-free channel. Since the autoencoder structure is correspond to a communication system with an encoder and a decoder, a scheme based on end-to-end learning of autoencoder is proposed in MIMO system over the physical layer [2]. The fundamental strategy is using the autoencoder to optimize the communication system globally, bypassing the separate blocks of system for channel estimation, feedback, coding, decoding, modulation and equalization. The purpose of this autoencoder is maximizing the spectrum efficiency and minimizing the BER. On the other hand, the influence of various uncertain factors in hardware implementation need to be considered when designing the wireless communication.

Based on these factors, in [17], authors develop a wireless communication to discuss the availability of using DNN instead of physical modules. Since the limitation of significant delay in training autoencoders directly, [17] propose a two-phase training. The first phase uses a channel
model to train the encoder and decoder based on the channel model. Following that the DNN is trained again on a real channel based on the training parameters in first phase, which improves the performance of system further by tuning the trainable parameters. In channel module, the delay and phase compensation are considered in DNN. It is generally the same as receiver, another DNN, connected to the channel DNN. This DNN-based method considers the time-varying under real channel, and the result shows system performance is comparable with that of traditional wireless communication systems.

Due to the changing of time and space location during the process of communication, the channel model will have mismatch with the real channel experienced by the transceivers, limiting the overall performance of the system. In [18], authors propose end-to-end GAN based communication system when the channel information is agnostic. Learned GAN is regarded as the channel effects while the encoded signal of the transmitter will serve as the conditioning information. To overcome the fading and time-invariance of channel and get accurate channel information, the pilot is added as a part of conditional information. In this article, the transmitter and receiver are replaced by DNN. Hence the system can be divided into DNN for transmitting, GAN performing as channel and received DNN. The global optimization could be found through the iterative training of 3 neural networks. It is shown that the method of channel estimation using GAN is comparable to the performance of traditional channel estimation, which indicates the availability of end-to-end wireless communication.

3. Discussion and Challenges in Future

Recently, the research in AI-aided wireless communication is developing rapidly in physical layer, which is expected to improve the performance of communication system. There are many different ways to classify these AI-aided methods in fact. Firstly, we can divide these methods based on the functional modules shown in Figure 1. Those are: channel estimation, signal detection, encoding, decoding, equalization etc. Some details regarding channel estimation, signal detection has been introduced in section 2. On the other hand, the AI-aided strategy can be also divided into two types: data-driven method and model-driven method from the perspective of the network topology. The data-driven method regards one or more functional modules as a whole deep learning network, known as a black box, and the network only relies on a large amount of data to complete the training. Some data-driven examples have been introduced in section 2. One of the most typical data-driven models applied in communication is called autoencoder, an end-to-end communication system, making achievement successfully in image processing.

The autoencoder replaces the communication system as an DNN and expects to find the global optimized solution to improve the performance of traditional system. However, it exists some problems when applying it in wireless communication system. In [19], the author compares the autoencoder applied in communication system and image processing deeply and concludes two unique features. First, the objective of communication autoencoder is to find latent coding that can carry information over detrimental wireless channel, generally by increasing the redundancy, while typical autoencoder is intended to find compact representations of structured data in lower-dimensional latent space without concern of pollution of the code. Second, Data in a communication is unstructured and incompressible. Hence communication autoencoder is designed to learn the inherent behaviors of channel rather than the structure of data. It holds promise for further improvement if the sub optimization for each module is replaced by optimizing for end-to-end performance. What is more, data-driven method relies on data heavily, which is always with expensive cost and long time. Besides, data-driven methods have weak robustness, influenced by the changing of network architecture easily.

By contrast, the model-driven method does not change the traditional model, it only uses DNN to enhance or train some tunable parameters to improve the performance of a system or a specific module, which can reduce the cost and save time when training the network. Compared with data-driven method shown in table I, the model-driven method not only has better robustness and accuracy, but also less complexity. In section 2.3, the CP-free OFDM system is aided by
model-driven network, improving the original channel estimation LS algorithm and signal detection OAMP algorithm [12]. In the future, the model-driven AI-aided method based on the classical wireless communication model is more competitive since it could improve the performance of original technology furthermore and need less training data. This kind of method has been used in channel estimation, signal detection and other modules in wireless communication system [5-6,8,10].

Table 1. Comparison between data-driven & model-driven AI-aided method

<table>
<thead>
<tr>
<th>Type</th>
<th>Complexity</th>
<th>Accuracy</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-Driven</td>
<td>High</td>
<td>Low</td>
<td>Weak</td>
</tr>
<tr>
<td>Model-Driven</td>
<td>Low</td>
<td>High</td>
<td>Strong</td>
</tr>
</tbody>
</table>

Moreover, how to process the complex signal and then feed it into the network is also a problem need to be solved. Most articles shown in this report divide the signal into real part and complex part and then combine them together as a single signal [6-10]. This method could lose some information during the processing since the real part of the signal is usually related to the signal’s complex part in wireless communication. In the future, designing a useful and entire complex convolutional neural network will become important and attractive in intelligent.

Apart from the communication system, the resources allocation and the decentralized communication, considering the effect of wireless network, will be also supplemented with Artificial Intelligence in the future. The AI-aided technology is advancing to physical layer. It can be predictable that wireless communication technology plays an increasing important role in future.

4. Conclusion

This report reviews several current major promises of artificial intelligence in wireless communication technology, including channel estimation, signal detection, hybrid beamforming and end-to-end wireless communication system. Intelligent communication is considered as a mainstream of wireless communication after 5G. Introducing AI into wireless communication help researchers solve NP-hard problem effectively. Moreover, the significant integration of wireless communication and artificial intelligent technology, known as model-driven AI-aided method could improve communication system performance and reduce complexity. On the other hand, the performance of deep learning could be improved through the wireless communication, which allows the model-driven AI-aided method to be a potential competitive technology in future.

References


