# Model-Driven Deep Learning for MIMO Signal Detection

GuangHua Zhang(✉), Fan Yang, and Sen Li

Northeast Petroleum University, DaQing, China
dqzgh@nepu.edu.cn

**Abstract.** Multiple Input Multiple Output (MIMO) technology is widely applied in various wireless communication systems, significantly improving communication efficiency and reliability. Signal detection is critical for MIMO systems. However, with the increasing integration of deep learning into MIMO signal detection algorithms, challenges such as high complexity and limited interpretability have emerged. To address this, this paper proposes a model driven trainable approximate message passing (AMP) algorithm that combines the iterative process of AMP with deep learning techniques. By introducing trainable parameters and optimizing them through training, and incorporating an attention mechanism to enhance channel feature extraction, the detection accuracy is improved, and the algorithm's generalization capability is enhanced. Simulation results demonstrate that AMP Attention Net achieves lower bit error rates compared to traditional detection algorithms. Furthermore, the proposed algorithm exhibits robust performance under different configurations of transmitting and receiving antennas.

**Keywords:** MIMO · Deep Learning · Signal Detection

## 1 Introduction

MIMO technology holds an important position in modern wireless communications [1,2]. It utilizes multiple transmitting and receiving antennas to achieve higher data transmission rates without increasing the spectral bandwidth, thereby meeting the growing demand for data. Signal detection is crucial for the stable operation and efficient transmission of MIMO communication systems.

Maximum Likelihood (ML) detection is a commonly used algorithm [3]. Its basic principle is to select the most likely transmitted symbol combination from the received signal, which has the highest likelihood, as the final detection result. However, Maximum Likelihood algorithms require enumerating all possible symbol combinations, resulting in exponentially increasing computational complexity as the number of antennas and modulation symbols in the MIMO system increases, making it impractical for real world applications. Zero Forcing (ZF) detection is a common linear detection algorithm that eliminates interference through linear transformations to accurately detect the transmitted symbols [4],


H.-H. Chen and W. Meng (Eds.): WiSATS 2024, LNICST 605, pp. 246–254, 2025.
https://doi.org/10.1007/978-3-031-86196-3_21

While simple and easy to implement, it is highly susceptible to noise. The Minimum Mean Square Error (MMSE) detection algorithm improves the quality and performance of received signals by minimizing the mean square error between the received and estimated signals [5,6]. However, it involves complex matrix inversion, leading to high complexity.

Data driven neural networks are often referred to as black boxes due to their complex structures and highly nonlinear mapping relationships. This characteristic implies that while the model can make predictions based on input data, its internal structure is often inexplicable or difficult to understand [7,8]. Model driven deep learning can reduce training volume and provide stronger interpretability, becoming a research hotspot. In [9], the authors use the Orthogonal Approximate Message Passing (OAMP) algorithm as a detector prototype, replacing matrix inversion with the Conjugate Gradient method to reduce OAMP complexity, further extending the Conjugate Gradient based OAMP algorithm into a network and enhancing detection performance through deep learning. In [10], the authors combine the OAMP algorithm with sparse connected neural networks (ScNet) to form a trainable network structure, proposing the ScNet OAMP network to enhance the detection capability of the MMSE estimator in the OAMP process, improving detection accuracy.

Inspired by the aforementioned work, a new network model, AMP Attention Net, is proposed for signal detection. Specifically, the AMP iterative process is improved by adding several trainable parameters to enhance the accuracy of signal detection, reducing the network complexity and making it easier to implement. Additionally, an attention mechanism is incorporated to extract channel features, thereby enhancing the network's generalization capability.

## 2 System Model

This paper considers a single cell massive MIMO communication system with $N_{\mathrm{t}}$ transmitting antennas and $N_r$ receiving antennas. The received signal vector $\overline{y} \in C^{N_r}$ can be expressed as:

$$\overline{y} = \overline{H}\overline{x} + \overline{n} \tag{1}$$

where $\overline{H} \in C^{N_r \times N_t}$ is the channel gain matrix, $\overline{n} \sim CN\left(0, \sigma^2 I_{N_r}\right)$ is Gaussian white noise, $\sigma^2$ is noise variance, $\overline{x} \in C^{N_t}$ represents the transmitted signal.

Since the vectors and matrices in (1) are complex numbers, they can be converted to real numbers for easier processing in deep learning:

$$y = Hs + n \tag{2}$$

Specifically expressed as:

$$y = \begin{bmatrix} \Re(\overline{y}) \\ \Im(\overline{y}) \end{bmatrix}, H = \begin{bmatrix} \Re\left(\overline{H}\right) & -\Im\left(\overline{H}\right) \\ \Im\left(\overline{H}\right) & \Re\left(\overline{H}\right) \end{bmatrix}, s = \begin{bmatrix} \Re\left(\overline{s}\right) \\ \Im\left(\overline{s}\right) \end{bmatrix}, n = \begin{bmatrix} \Re\left(\overline{n}\right) \\ \Im\left(\overline{n}\right) \end{bmatrix} \tag{3}$$

Here the real and imaginary parts of a complex matrix or vector are tabulated by $\Re(\cdot)$ and $\Im(\cdot)$, respectively.

# 3 Model Driven for Signal Detection

## 3.1 Approximate Message Passing

The core idea of the AMP is to approximate the posterior probability distribution through iterative message updates, with low computational complexity and good convergence. It is widely used in signal detection.

Assume the signal $x = [x_1, x_2, \cdots, x_N]^T \in R^N$,the Gaussian random observation matrix $A \in R^{M \times N}$,then the measurement vector can be expressed as:

$$y = Ax + w \tag{4}$$

where $w \in R^M$ represents noise. The algorithm initializes with the reconstructed vector $x^0 = A^t y$ and residual vector $z^{-1} = y$. The $t\,(t \geq 0)$-th iteration can be expressed as:

$$z^t = y - Ax^t + b_t z^{t-1} \tag{5}$$

$$x^{t+1} = \eta_t(x^t + A^T z^t; \lambda_t) \tag{6}$$

where $x^t$ is the reconstructed vector after the $t$-th iteration, and $b_t z^{t-1}$ is the Onsager correction term.$b_t = \frac{1}{N_r}\|x^t\|_0, \lambda_t = \frac{\alpha}{\sqrt{N_r}}\|z^t\|_2$,$\alpha$ is an adjustable value related to the sparsity and the observation vector. The existence of the Onsager correction term allows the AMP to converge quickly (Table 1).

**Table 1.** Approximate message passing algorithm

| Input: $y, A$, Output: $x^{t+1}$ |
|---|
| Initialize: $x^0 = A^T y, z^{-1} = y, t = 1$ |
| $z^t = y - Ax^t + b_t z^{t-1}$ |
| $x^{t+1} = \eta_t(x^t + A^T z^t; \lambda_t)$ |
| Condition met, exit iteration |

## 3.2 Channel Attention Mechanism

In the signal detection task, the channel information is often complex and changeable. By introducing an attention mechanism, the network can dynamically adjust its attention to different channels, enabling it to pay more attention to the channel information that is critical to the detection task. A channel attention mechanism (CAM) is commonly used in deep learning, and its structure is shown in Fig. 1.

First, apply global max pooling and global average pooling along the spatial dimensions to the input features of size $H \times W \times C$ , resulting in two $1 \times 1 \times C$ feature maps. This step compresses the spatial dimensions, making it easier to learn the channel features in subsequent steps. Next, feed the results of global

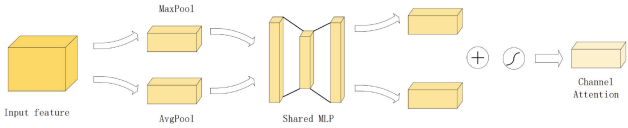


**Fig. 1.** The structure of channel attention mechanism.

max pooling and global average pooling into a shared multi layer perceptron (MLP) to learn, resulting in two $1 \times 1 \times C$ feature maps. Finally, perform an element wise addition on the outputs of the multi-layer perceptron, followed by a Sigmoid activation function to map the results, ultimately obtaining the channel attention weight matrix.The specific calculation process is expressed as:

$$\begin{aligned} M_c &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{\max}^c\right)\right)\right) \end{aligned} \tag{7}$$

### 3.3 Amp Attention Net

Due to the significant influence of parameters on the convergence speed and accuracy of the AMP algorithm, and the complex and variable channel information in communication environments, these issues are addressed by unfolding the AMP detector and adding several trainable parameters. Additionally, an attention mechanism is incorporated to extract channel information features, thereby improving the generalization of the detection algorithm. The specific network structure is shown in Fig. 2. For the overall network, the input is the received signal $y$ and the signal matrix $H$, and the output is the estimated input signal $x$. For the $t+1$-th layer of the network, the input is the estimated signal $x_t$ from the $t$-th layer, and the network at this layer performs an iteration to obtain the output $x_{t+1}$. Compared to the Approximate Message Passing algorithm, AMP NET adds trainable parameters $\omega_t, \psi_t$. Each layer of the network has the same structure but different learnable parameters. The iterative process can be expressed as:

$$z^t = y - Ax^t + \omega_t b_t z^{t-1} \tag{8}$$

$$x^{t+1} = \eta_t(x^t + A^T z^t; \psi_t \lambda_t) \tag{9}$$

Adding a learnable parameter $\omega_t$ is to provide a more suitable step size for the update of the residual $z^t$, while adding parameter $\psi_t$ dynamically selects the threshold. Thus, the MIMO detection problem is transformed into finding the most suitable learnable parameters through deep learning to reduce the error rate.

During the training of the parameters, the gradient descent algorithm is used, and its convergence behavior and performance depend on the appropriate step size for moving to the search point. The optimal step size $\omega_t$ can be learned from the data for updating the residual, and $\psi_t$ follows the same training process.

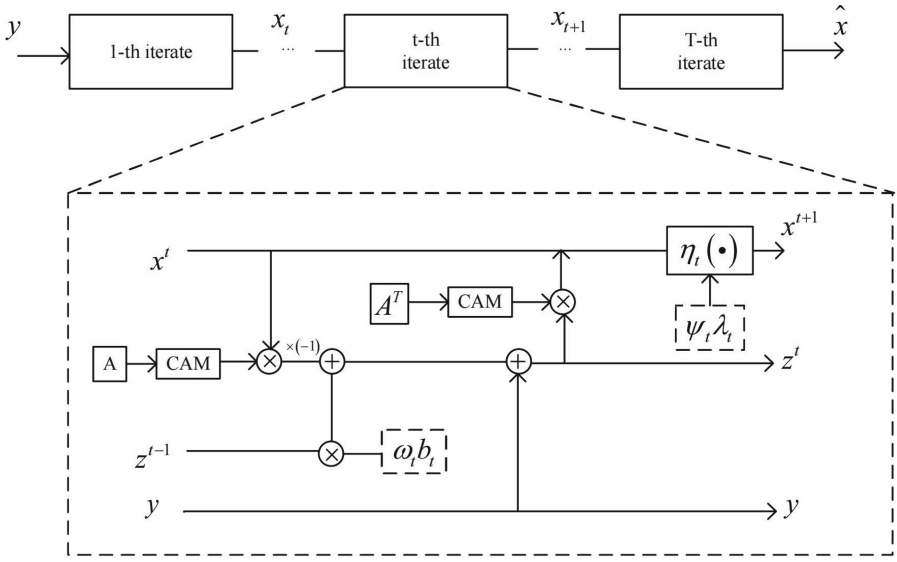


**Fig. 2.** The network structure of the algorithm that combines the AMP algorithm with deep learning.

## 4 Simulation Analysis

### 4.1 Parameter Settings

In practical applications, the dataset used to train the model and the data distribution faced by the model during actual deployment or testing may not match. This mismatch can be caused by various factors, such as the time varying nature of communication channels, changes in noise levels, and differences in hardware devices. To maintain the generalization ability of the model, various experiments are conducted for research.

The wireless communication simulation parameters and network model parameters are shown in Table 2. The number of antennas at the system's transmitter is set to 48, and at the receiver, it is set to 64. The channel model is set to a Rayleigh fading channel, and QPSK modulation is used. A total of 125,000 simulation experiments are conducted. For the network model parameters, 3/5 of the data obtained from the wireless communication simulation is set as the training set, 1/5 as the validation set (used to select the best network during the training phase), and the remaining 1/5 as the test set (used to test the network model). The Adam optimizer and gradient descent method are used for training the network, with the learning rate set to 0.001. The initial values of the variables $\omega_t$ and $\psi_t$ are set to 1, and the batch size is set to 100. The optimizer chosen is the gradient descent method, and the loss function is set to the square of the difference between the predicted values and the true values.

**Table 2.** Simulation Parameters and Network Model Parameters

| Parameter Name | Parameter Value |
|---|---|
| Number of Transmit Antennas | 48 |
| Number of Receive Antennas | 64 |
| Learning Rate | 0.001 |
| Loss Function | MSE |
| Batch Size | 100 |
| Optimizer | Gradient Descent |
| Number of Training Iterations | 500 |
| Training Set Size | 7500 |
| Validation Set Size | 2500 |

### 4.2 Complexity Analysis

For algorithms, the standard measure of complexity generally uses real multiplications. However, the complexity of signal detection algorithms also depends on parameters such as the number of transmit antennas $MT$, receive antennas $MR$, modulation order $M$, and the number of iterations $iter$. ZF and MMSE detection algorithms require matrix inversion operations on the signal matrix. The main complexity of ML algorithm arises from enumerating all possible transmitted symbols and comparing them with the received vector. In cases where the number of transmit and receive antennas is large, the complexity of these three algorithms becomes extremely high, making practical implementation difficult.

The main complexity of the AMP algorithm comes from iterative processing of the channel matrix. In contrast, the complexity of Amp Attention Net mainly depends on the dimensions of matrices and iterations, especially the input matrix and the dimensions of hidden layer variables. Generally, in practical application scenarios, the complexity of Amp Attention Net remains within acceptable limits. Specific complexities are detailed in Table 3.

**Table 3.** Algorithm Complexity

| MIMO Detection Algorithm | Complexity |
|---|---|
| ZF | $O\left(MR^3\right)$ |
| MMSE | $O\left(MR^3 + MT^3\right)$ |
| ML | $O\left(M^{MT} \times MT \times MR\right)$ |
| AMP | $O\left(iter \times MT \times MR\right)$ |
| AMP Attention Net | $O\left(M^2 \times N + iter \times M^2\right)$ |

### 4.3 Performance Analysis

To determine the detection effectiveness of different unfolded layers, the Amp Attention Net with different layers was validated on the validation set, and their bit error rate (BER) were compared. The specific results are shown in Fig. 3. From Fig. 3, it is evident that when the network has 1 layer, the BER on the validation set is significantly higher than that of networks with more layers. When the network has 2 or 3 layers, the BER decreases significantly compared to the 1-layer network but still lags behind networks with more layers. As the number of unfolded layers continues to increase, the overall difference in BER is not substantial. Therefore, considering overfitting and training complexity, the optimal number of unfolded layers is set to 4.

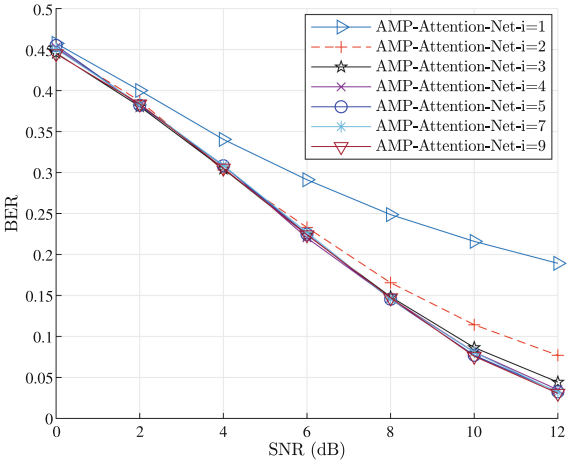


**Fig. 3.** Comparison of BER for Different Layers of AMP Attention Net.

Figure 4 compares the BER performance of the 3-layer Amp Attention Net with the AMP, MMSE, and ZF algorithms. Overall, the ZF algorithm performs the worst. At low SNRs, the performance of MMSE, AMP, OAMP and Amp Attention Net is similar, but as the SNR increases, the BER of Amp Attention Net becomes significantly lower than that of the other algorithms. When the BER is 0.17, Amp Attention Net achieves approximately 1.8dB performance gain compared to AMP, about 0.3dB performance gain compared to OAMP, and about 0.5dB performance gain compared to MMSE. This demonstrates that Amp Attention Net can significantly improve the performance of AMP detection and has strong detection capabilities in MIMO signal detection.

Figure 5 shows the BER of Amp Attention Net for different antenna configurations $64 \times 64$and $128 \times 128$. The results show that although the antenna configuration is different in the simulation, the overall bit error rate is about the same, which shows that the algorithm has strong detection performance and good generalization ability.

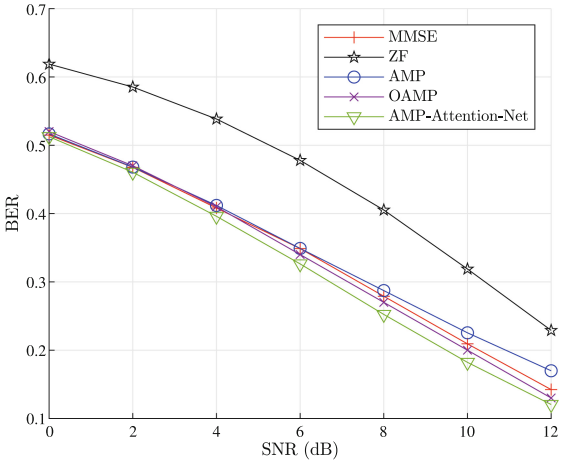


**Fig. 4.** Comparison of BER for MMSE, ZF, AMP, OAMP and AMP Attention Net Detection Algorithms

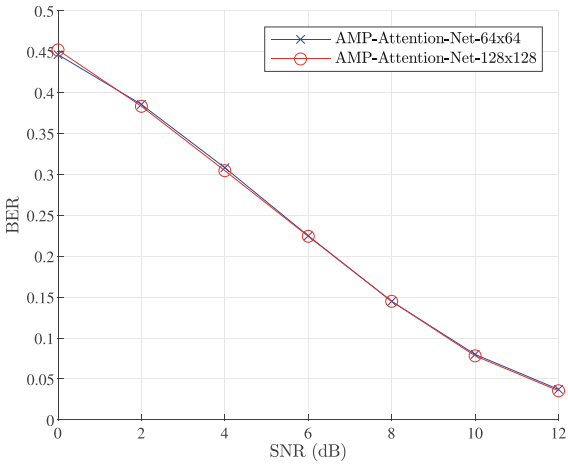


**Fig. 5.** Comparison of BER Under Different Numbers of Antennas

## 5 Analysis

This paper conducts research on signal detection algorithms and proposes a model driven Amp Attention Net signal detection algorithm based on the AMP detection algorithm. This algorithm unfolds the iterative process of the AMP algorithm, adds trainable parameters, and introduces an attention mechanism to enhance the extraction of channel features. By training to obtain the optimal parameters, the accuracy of signal detection is improved. Simulation results indicate that the proposed algorithm has a lower error rate compared to traditional

detection algorithms, and it performs well with different numbers of transmit and receive antennas. Additionally, the algorithm model proposed in this paper is relatively easy to implement and has low cost.

## References

1. Lu, L., Li, G.Y., Swindlehurst, A.L., Ashikhmin, A., Zhang, R.: An overview of massive MIMO: benefits and challenges. IEEE J. Sel. Top. Sig. Process. **8**(5), 742–758 (2014)
2. Larsson, E.G., Edfors, O., Tufvesson, F., Marzetta, T.L.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**(2), 186–195 (2014)
3. He, K., He, L., Fan, L., Deng, Y., Karagiannidis, G.K., Nallanathan, A.: Learning-based signal detection for MIMO systems with unknown noise statistics. IEEE Trans. Commun. **69**(5), 3025–3038 (2021)
4. Liu, Y.: A massive MIMO signal detection method based on ZF method. In: 2021 5th International Conference on Imaging, Signal Processing and Communications (ICISPC), pp. 71–76. IEEE (2021)
5. Khoso, I.A., Kang, C.G.: Extrapolation principle-based low-complexity signal detection in massive MIMO systems. IEEE Wirel. Commun. Lett. **13**(1), 123–130 (2024)
6. Zhao, S., Shen, B., Hua, Q.: A comparative study of low-complexity MMSE signal detection for massive MIMO systems. KSII Trans. Internet Inf. Syst. (TIIS) **12**(4), 1504–1526 (2018)
7. Baek, M.-S., Kwak, S., Jung, J.-Y., Kim, H.M., Choi, D.-J.: Implementation methodologies of deep learning-based signal detection for conventional MIMO transmitters. IEEE Trans. Broadcast. **65**(3), 636–642 (2019)
8. Nguyen, L.V., Nguyen, N.T., Tran, N.H., Juntti, M., Swindlehurst, A.L., Nguyen, D.H.N.: Leveraging deep neural networks for massive MIMO data detection. IEEE Wirel. Commun. **30**(1), 174–180 (2022)
9. Karahan, S.N., Kalaycıoğlu, A., Taşcıoğlu, S.: MIMO detection algorithms–Deep learning based and traditional methods. In: 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), pp. 1–4. IEEE (2022)
10. Zhou, X., Zhang, J., Syu, C.W., Wen, C.K., Zhang, J., Jin, S.: Model-driven deep learning-based MIMO-OFDM detector: design, simulation, and experimental results. IEEE Trans. Commun. **70**(8), 5193–5207 (2022)