MDPI

Article

Transformer-Based Detection for Highly Mobile Coded OFDM Systems

Leijun Wang [1], Wenbo Zhou [1], Zian Tong [1], Xianxian Zeng [1,2,*], Jin Zhan [1], Jiawen Li [1] and Rongjun Chen [1,*]

1   School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China; wangleijun@gpnu.edu.cn (L.W.); a1402057929@gmail.com (W.Z.); t15119803678@gmail.com (Z.T.); gszhanjin@gpnu.edu.cn (J.Z.); lijiawen@gpnu.edu.cn (J.L.)
2   Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen 518000, China
*   Correspondence: zengxianxian@gpnu.edu.cn (X.Z.); chenrongjun@gpnu.edu.cn (R.C.)

Abstract: This paper is concerned with mobile coded orthogonal frequency division multiplexing (OFDM) systems. In the high-speed railway wireless communication system, an equalizer or detector should be used to mitigate the intercarrier interference (ICI) and deliver the soft message to the decoder with the soft demapper. In this paper, a Transformer-based detector/demapper is proposed to improve the error performance of the mobile coded OFDM system. The soft modulated symbol probabilities are computed by the Transformer network, and are then used to calculate the mutual information to allocate the code rate. Then, the network computes the codeword soft bit probabilities, which are delivered to the classical belief propagation (BP) decoder. For comparison, a deep neural network (DNN)-based system is also presented. Numerical results show that the Transformer-based coded OFDM system outperforms both the DNN-based and the conventional system.

Keywords: transformer; deep neural network (DNN); deep learning; orthogonal frequency division multiplexing (OFDM); high-speed railway; wireless communication

1. Introduction

Orthogonal frequency division multiplexing (OFDM) shows great performance in the frequency-selective fading channels. However, in the highly mobile communication system, the Doppler effect destroys the orthogonality among the subcarriers of the OFDM system, resulting in intercarrier interference (ICI), which will degrade the system's performance. In order to mitigate ICI in the mobile OFDM system, an equalizer or detector, e.g., a zero forcing (ZF) detector or minimum mean square error (MMSE) detector, should be implemented. To further combat channel distortion, it is necessary to use a channel coding scheme. Thus, a soft demapper is needed between the detector and the channel decoder. We have already studied the mobile coded OFDM system with a flexible coding scheme called block Markov superposition transmission (BMST) [1,2]. In this paper, we still focus on the mobile coded OFDM system.

In recent years, deep learning (DL) has achieved great success in various fields, such as computer vision [3–6], natural language processing, speech recognition, trajectory prediction [7]. Therefore, researchers in wireless communication are seeking to applying DL to various aspects of communication to achieve good performance. Currently, research in this area has already shown a trend of combining wireless transmission with deep learning in the physical layer, but all studies are still in the initial exploration stage [8–10].

In [11], a deep neural network (DNN) was utilized to solve the problem of channel estimation and signal detection in the OFDM system. In [12], the receiver uses the existing architecture to perform channel estimation and signal detection separately, and the authors combine each module with deep learning to improve the performance. In [13], the proposed

receiver is trained with both offline and real-time online data to capture channel characteristics that are ignored during offline training. In addition, the deep learning approach has been introduced to signal detection in OFDM with index modulation (OFDM-IM) [14]. In particular, a Transformer-based detector was utilized to the OFDM-IM in [15], and a heterogeneous Transformer-based device activity detection method was proposed for the modern machine-type communications [16]. The Transformer [17], which is a network structure based on a self-attention mechanism, was proposed in 2017. Shortly thereafter, it showed remarkable effects in natural language processing [18], and as it demonstrated remarkable effects in computer vision, its application gradually became an important aspect of deep learning.

In this paper, we propose a Transformer-based detection/demapping algorithm for the mobile coded OFDM system. Although there are some studies on deep learning for channel coding [19] and decoding [20–23], the performance is not satisfactory when decoding directly via DL, especially regarding the long capacity-approaching codes. In this paper, we employ the low-density parity check (LDPC) codes, and the soft information given to the conventional belief propagation (BP) decoder is computed by deep learning. The main contributions of this paper are as follows:

- A Transformer-based detection algorithm is proposed for the coded OFDM system. Although DL-based detectors do not outperform the conventional detector in the uncoded OFDM system, they have better performance in the coded OFDM system.
- In our coded OFDM system, the LDPC codes are performed, and the soft information is required by the decoder. Thus, we propose the soft demapping algorithm based on Transformer.
- In the OFDM system, it is difficult to compute the mutual information based on the optimal detector. Thus, we can compute the mutual information with a suboptimal detector, which can be regarded as the soft information quality (SIQ) [24,25]. We compute the SIQ with the assistance of the Transformer network.

The remainder of the paper is organized as follows. Section 2 introduces the overall system model, including some classical detection algorithms. Section 3 proposes the Transformer-based detection and demapping algorithm. Section 4 gives the experimental results. Finally, Section 5 concludes the paper.

2. System Model

2.1. The Coded OFDM System Model

In this section, we introduce a single-antenna coded OFDM system that utilizes N subcarriers over doubly selective channels. Figure 1 depicts the block diagram of the general coded OFDM system. At the transmitter, the information bit stream $\underline{u}$ is first encoded by the encoder to produce a coded sequence $\underline{c}$, which will be mapped to the vector sequence $\underline{x}$. The sequence c, which represents a block of the sequence $\underline{c}$, is then mapped to the vector of N-coded frequency-domain symbols, denoted as $x = [x_0, \cdots, x_{N-1}]^T$. Each x_i in the vector represents the symbol transmitted on the ith subcarrier of one OFDM symbol, chosen from a complex signal constellation $\mathcal{S}$ with $|\mathcal{S}| = 2^{M_c}$, in which M_c represents the bit number carried by one symbol, and we assume that the average symbol energy $\mathbb{E}[|x_i|^2] = 1$.

The time-domain symbols are obtained by using the inverse discrete Fourier transform (IDFT), which can be implemented through the inverse fast Fourier transform (IFFT). A cyclic prefix (CP) or guard interval of length $N_{cp} \geq N_{tap} - 1$ is added to the beginning of the time-domain symbols. The resulting time-domain symbols are then transmitted over the doubly selective channel, which has N_{tap} taps.
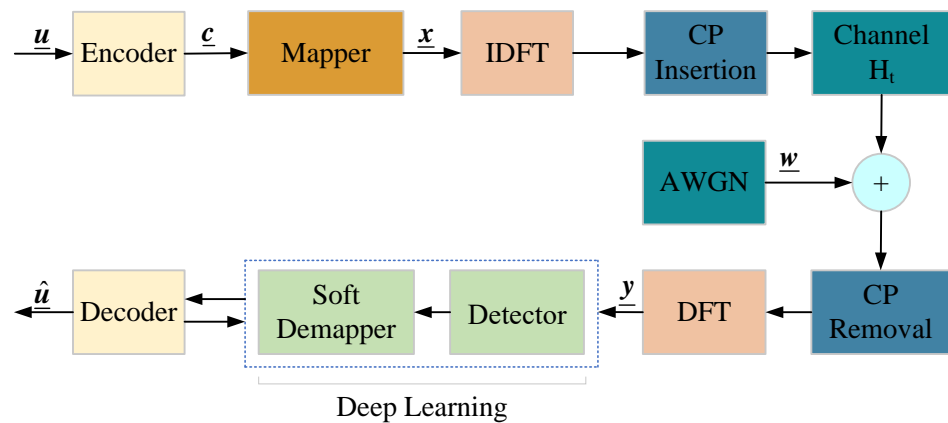
Figure 1. The block diagram of the coded OFDM system.

After the receiver removes the CP and applies the discrete Fourier transform (DFT), the receive vector can be expressed as

$$y = FH_tF^Hx + Fw, \tag{1}$$

where y, x, and w are the sub-blocks of $\underline{y}$, $\underline{x}$, and $\underline{w}$, respectively, corresponding to a single OFDM symbol. F represents the unitary DFT matrix of size N. H_t is the time-domain channel matrix, whose construction will be discussed in detail in the following section. The vector w denotes the time-domain additive white Gaussian noise (AWGN), whose entries are independently and identically distributed (i.i.d.) according to $\mathcal{CN}(0, \sigma^2)$. Then, the signal-to-noise ratio (SNR) observed at the receiver can be defined as SNR $= 1/\sigma^2$.

Denoting the frequency-domain matrix as $H_f = FH_tF^H$, the receive vector can be expressed as

$$y = H_fx + w_f. \tag{2}$$

The frequency-domain noise w_f has the same statistical properties as the time-domain noise w due to the unitary matrix property.

2.2. The Channel Model

As mentioned earlier, doubly selective fading channels are used in high-speed railway communication systems. This study assumes that the receiver has access to channel state information (CSI). In the OFDM system, the discrete-time channel matrix element is represented by the discrete-time impulse response $h_{n,m}$, where n is the subcarrier index, and m is the tap index, with $0 \leq n \leq N - 1$ and $0 \leq m \leq N_{tap} - 1$. The Jakes' Doppler spectrum [26] enables the calculation of $h_{n,m}$ as follows:

$$h_{n,m} = \sqrt{\frac{P_m}{4M}} \left\{ \sum_{i=1}^{M} \sqrt{2}e^{j\psi_i} \left[e^{j(\omega_in+\phi)} + e^{-j(\omega_in+\phi)} \right] \right\} \tag{3}$$

with $\omega_i = \omega_d \cos \alpha_i$, where

$$\omega_d = \frac{2\pi\vartheta_{\max}}{N},$$

$$\alpha_i = \frac{2\pi i - \pi + \theta}{4M}, \quad i = 1, 2, \cdots, M.$$

In this paper, let M be the number of sinusoids ($M = 32$). P_m denotes the power-delay profile (PDP), where ϑ_{max} represents the normalized maximum Doppler shift. The statistical variables θ, ϕ, and ψ are independent and uniformly distributed over the interval

$[-\pi, \pi)$ for all i. The time-domain matrix H_t contains non-zero elements generated using Equation (3). Given the OFDM system structure, H_t can be expressed as follows:

$$H_t = L + U. \tag{4}$$

Equation (4) involves a $N \times N$ lower triangular matrix whose nonzero element $L_{n,m} = h_{n,n-m}, 0 \leq m \leq n \leq N-1$, and U is a $N \times N$ upper triangular matrix whose nonzero element $U_{n,m} = h_{n,N+n-m}, 0 \leq n \leq N_{tap}-1, N-N_{tap}+1 \leq m \leq N-1, n \leq m$.

2.3. The Classical Signal Detection and Demapping Algorithm

When we obtain the receive vector y with optimal maximum likelihood (ML) detection, the transmitted vector x can be obtained as follows:

$$\hat{x} = \arg \max_{x \in \mathcal{S}^N} f(y|x, H_f), \tag{5}$$

where

$$f\left(y\Big|x, H_f\right) = \left(\frac{1}{\pi\sigma^2}\right)^N \exp\left(-\frac{\|y - H_f x\|^2}{\sigma^2}\right). \tag{6}$$

In Equation (5), the size of the symbol set $\mathcal{S}$ and the number of OFDM subcarriers N have a great impact on the complexity, especially when using a large number of subcarriers, as the optimal detection has little practical value. Therefore, it is necessary to perform an equalization operation on the received signal to eliminate the ICI before detection. When the receive vector y—see Equation (2)—is obtained, the process of equalization can be described as follows:

$$Gy = GH_f x + Gw_f, \tag{7}$$

where G is the preprocessing matrix. When the ZF detector is used, Equation (7) can be rewritten as

$$H_f^\dagger y = x + H_f^\dagger w_f, \tag{8}$$

where $H_f^\dagger = (H_f^H H_f)^{-1} H_f^H$ is the pseudoinverse of the frequency-domain matrix H_f.

When MMSE equalization is used, the preprocessing matrix

$$G = \left(H_f^H H_f + \sigma^2 I\right)^{-1} H_f^H. \tag{9}$$

In the uncoded system, the vector $\tilde{y} = Gy$ can be used to estimate the vector x by

$$\hat{x}_i = \arg \min_{x_i \in \mathcal{S}} \|\tilde{y}_i - x_i\|^2, \tag{10}$$

where $\tilde{y}_i$ and x_i are the ith elements of the vectors $\tilde{y}$ and x, respectively. However, in the coded OFDM system, soft information is required by the demapper/decoder. For any symbol x_i transmitted on one subcarrier, the soft information is computed as

$$f(\tilde{y}_i|x_i) \propto \exp\left(-\frac{|\tilde{y}_i - x_i|^2}{\|g_i\|^2 \sigma^2}\right), \tag{11}$$

where g_i represents the ith row of the preprocessing matrix $G = [g_0, g_1, \cdots, g_{N-1}]^T$. In the following section, we will describe the DL (Transformer)-based network, which can realize the functions of a soft demapper and detector; see Figure 1. It is important to note that the long LDPC codes are mainly used in our following experiments, so the codeword set is too large. From the observation of the simulation based on our proposed network, the DL-based decoders do not have good performance. Thus, we still use the conventional BP decoder in this paper.

2.4. Rate Allocation

In the coded modulation system, in order to determine the code rate of the coding scheme, it is necessary to compute or simulate the mutual information. However, as mentioned above, it is difficult to compute the likelihood probability in the OFDM system, resulting in the failure to solve the mutual information. Thus, we can compute SIQ, the mutual information, with the detector. Assume that each subcarrier sends an M-ary QAM symbol, and let X and $\tilde{Y}$ be the corresponding random vectors for the sending symbol x and receiving preprocessed symbol y, respectively. Obviously, X is a discrete random variable and $\tilde{Y}$ is a continuous random variable. Assuming that X follows a uniform distribution and the channel matrix $\boldsymbol{H}$ is known, the SIQ can be computed as

$$I(X; \tilde{Y} | \boldsymbol{H}) = \log_2(M) + \mathbb{E}\left[\sum_X f(x_i | y_i', \boldsymbol{H}) \log_2 f(x_i | y_i', \boldsymbol{H})\right], \tag{12}$$

in which the probability $f(x_i | y_i', \boldsymbol{H})$ can be computed by the classical demapping algorithm described above, or by the output of the DL network.

3. The Detection and Soft Demapping Algorithm with Deep Learning

3.1. Transformer

The traditional Transformer is essentially an encoder–decoder structure, which can be divided into two parts, the encoder and the decoder, as shown in Figure 2.
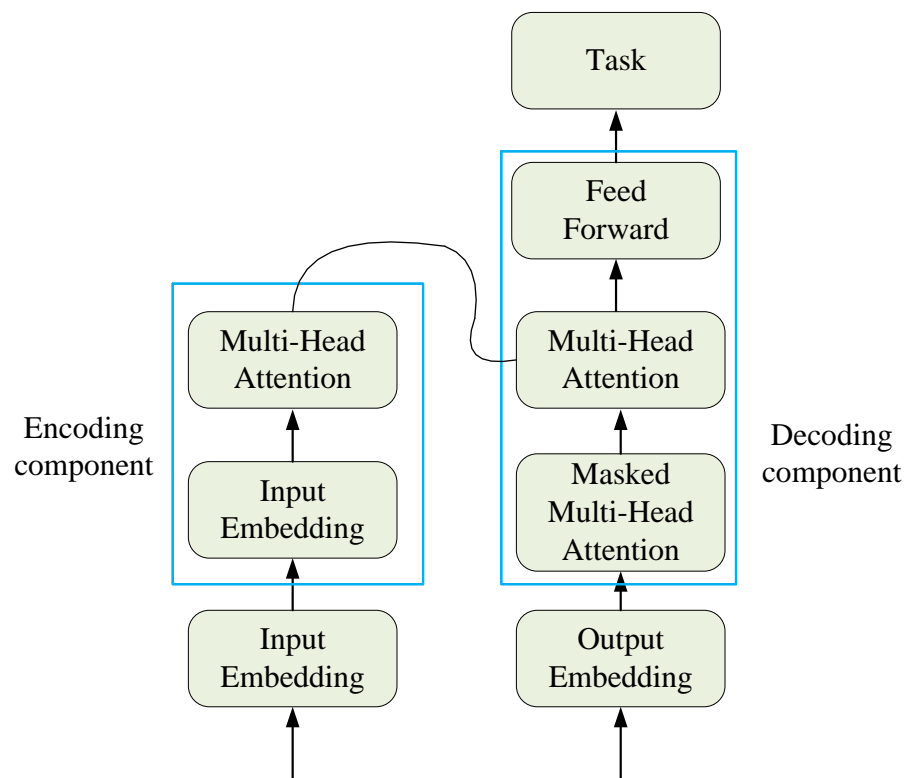

Figure 2. The block diagram of the traditional Transformer structure.

The left part is the encoding component, mainly composed of multi-head attention mechanisms and feed-forward neural networks. Encoding has two sub-layers: one is the multi-head attention mechanism layer, which uses self-attention mechanisms to learn the relationships between different dimensions of the data; the other is the forward propagation layer, which is a simple fully connected layer that performs the same operation on each position vector, including linear transformation and activation functions, and then produces encoding, which is passed to the encoding layer.

The right part is the decoding component, which is mainly composed of masked self-attention mechanisms and feed-forward neural networks. There are three sub-layers in the decoding layer, two of which are multi-head attention mechanism layers. The lower attention mechanism layer learns the relationships within the data using self-attention mechanisms, and then the layer outputs the results together with the results transmitted by encoding to the attention layer above. The attention layer calculates the correlation between the encoding and decoding, and can explore the relationship between the input sequence and the target sequence.

3.2. Experimental Method with Transformer

As shown in Figure 1, we treat the signal detection and demapping processes as a black box and replace them entirely with a Transformer network.

The detailed Transformer structure designed for the OFDM system is depicted in Figure 3. The Transformer network receives a time-domain frame $(\underline{y}, \underline{h})$ as input and recovers the transmitted bit data or probability in an end-to-end manner. The network is trained offline and deployed online. During the training phase, the transmitted bits are randomly generated as labels and modulated with channel information to form frames. A simulated channel is generated using a specific channel model, which varies with each frame. In the deployment phase, no equalization or detection is required, and the trained parameters are directly applied to achieve end-to-end bit recovery.
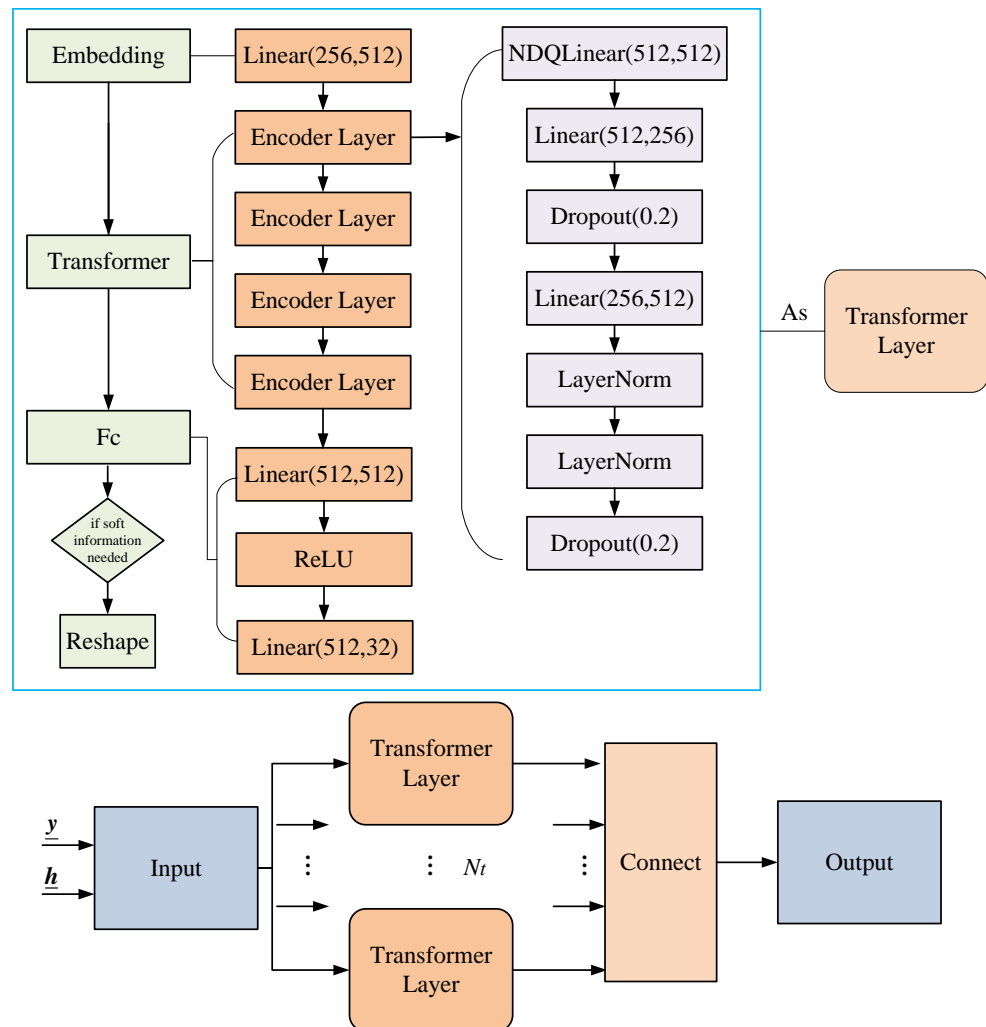

Figure 3. The block diagram of the designed Transformer structure for the OFDM system.

The designed Transformer model is suitable for encoding and non-encoding scenarios. For non-encoding signal detection, the input consists of the received signal $\boldsymbol{y}$ and the CSI $\underline{\boldsymbol{h}}$, which serve as features. The transmitted bits or symbols are used as labels and, after passing through the Transformer, output the hard decision of the transmitted bits or the symbol probabilities using the *softmax* function. The symbol probabilities can be used to compute the SIQ.

In the case of encoding, the labels for the network are the encoded bits, and the network outputs the bit probabilities, which are computed by the *softmax* function. The probabilities are then sent to the BP decoder to recover the original information bits.

3.3. Model Training

The Transformer network structure is related to the input and output dimensions; see Figure 3. The Transformer model used consists of N_t Transformer layers, each of which is composed of an embedding layer, a Transformer layer, and a fully connected layer. The Transformer consists of NDQLinear, Linear, LayerNorm, and Dropout layers, with neuron numbers of 512, 512, and 256, respectively. The number of input layer neurons corresponds to the sum of the real and imaginary parts of two OFDM signals, and the output corresponds to 16 bits. Since the data signal is modulated using QPSK with 64 effective subcarriers and a total of 128 bits, 8 independent networks are required, which are then concatenated for the final output. The hidden layers of the network use ReLU as the activation function.

The parameters in the Transformer network are trained using the training data, with the goal of minimizing the difference between the output of the neural network and the label data. The difference can be described in various ways, and this study uses cross-entropy loss. The cross-entropy between two probability distributions $P(x_i)$ and $Q(x_i)$ for a random variable X can be computed as

$$H(P,Q) = -\sum_{i=1}^{n} P(x_i) \log Q(x_i), \tag{13}$$

which can measure the difference between the two distributions.

Taking encoded data as an example, $\boldsymbol{y}$ represents the received data at the receiving end, $\boldsymbol{h}$ represents the CSI data, and the output encoded data

$$\underline{\hat{\boldsymbol{c}}} = Transformer(\underline{\boldsymbol{y}}, \underline{\boldsymbol{h}}). \tag{14}$$

The goal is to find the best set of weights and biases of the network, denoted by θ^*, that minimize the loss function, which is defined as follows:

$$\theta^* = \arg\min_{\theta} Loss(\underline{\boldsymbol{c}}, \underline{\hat{\boldsymbol{c}}}), \tag{15}$$

in which $\underline{\boldsymbol{c}}$ represents the label.

3.4. DNN Detection Algorithm

DNNs, also known as multi-layer perceptrons, consist of input layers, several hidden layers, and output layers; see Figure 4. Each hidden layer contains several neurons that do not interfere with each other, and they connect to adjacent layers. A single neuron multiplies each input by the corresponding weight and adds the bias parameter, finally reaching the output layer through a non-linear activation function. The core of a DNN is that it can perform self-optimization through back-propagation, but as the number of layers and neurons increases, the training will face problems such as gradient disappearance, slow convergence, and local minimum values. In order to improve the training speed and reduce the computational complexity, the classic gradient descent (GD) method has been replaced by stochastic gradient descent, which randomly selects data to calculate each loss and gradient. However, actual execution may be very slow because the entire data set needs to be traversed. Therefore, a commonly used approach is to randomly extract a subset of

samples for training each time updates are calculated, which is called mini-batch stochastic gradient descent. However, these algorithms may still converge to local optimal solutions.
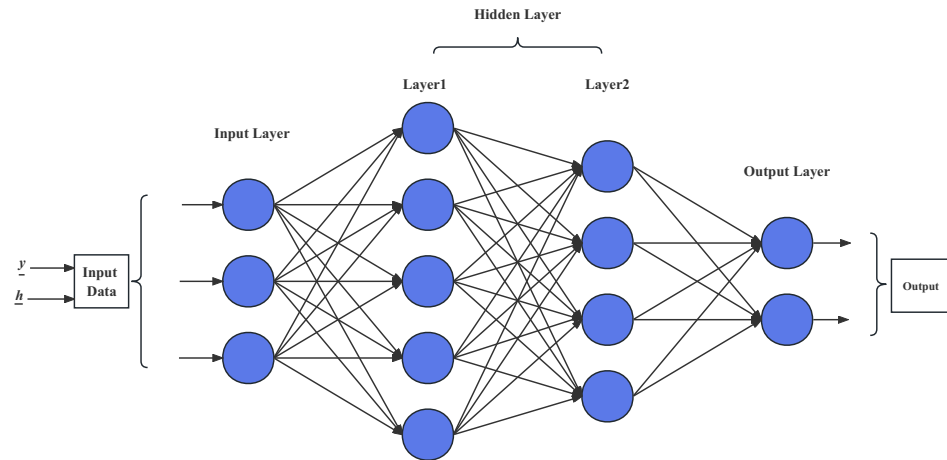

Figure 4. The DNN model for the OFDM system.

The DNN network can be used not only for channel estimation, equalization, and detection in traditional OFDM systems, but also for decoding. In this paper, the DNN network is built for comparison with the Transformer. The signal $\underline{y}$ and the CSI $\underline{h}$ are fed into a DNN network, with the same set of parameters for each network. The hidden layers of the DNN network are set to 256, 512, and 1024, respectively. The output of the two networks are concatenated together and passed through a fully connected layer and then a hidden layer of 256 before outputting 8 values, which are used to calculate the cross-entropy and probability for two dimensions. There are a total of 32 networks, with each network outputting 4 values. The 32 network outputs are merged together to obtain the complete output. The methods for calculating the probabilities and hard decision in the DNN network are identical to those in the Transformer network described above and will not be reiterated here.

4. Experimental Results

In this section, the numerical examples are based on Monte Carlo simulation over time-varying frequency-selective Rayleigh fading channels, and the PDP is $P_i = \alpha e^{-0.6i}$, $0 \le i \le N_{tap} - 1$, where α is a normalization constant. The relation between the normalized maximum Doppler shift $\vartheta_{\max}$ and the relative speed between the transmitter and receiver v is $\vartheta_{\max} = \frac{f_c}{F_c} \frac{v}{c_0}$. The simulation parameters are shown in Table 1.

Table 1. Simulation parameters.

Parameter	Value
Subcarriers N	64
Subcarrier Spacing F_c	15 KHz
Carrier Frequency f_c	2 GHz
Multipaths N_{tap}	9
CP Length N_{cp}	8
Relative Speed	360 km/h
Speed of Light c_0	3×10^8 m/s
Modulation Mapper	QPSK

In the following simulations, we mainly use regular LDPC codes, which are constructed by the progressive-edge-growth (PEG) algorithm. The sum product algorithm (SPA) is employed for the decoding of the LDPC codes, in which the maximum iteration number is 40.

Example 1. *In this example, the bit-error-rate (BER) performance of the uncoded OFDM system with the variety detector is as depicted in Figure 5. The training data are generated when SNR = 20 dB in the DL simulation. From the figure, we have the following observations.*

- *The BER performance of the uncoded OFDM system with the conventional detectors is better than that with the DL detectors in the high SNR region.*
- *The MMSE detector performs better than the ZF detector, while the Transformer-based detector performs better than the DNN-based detector.*
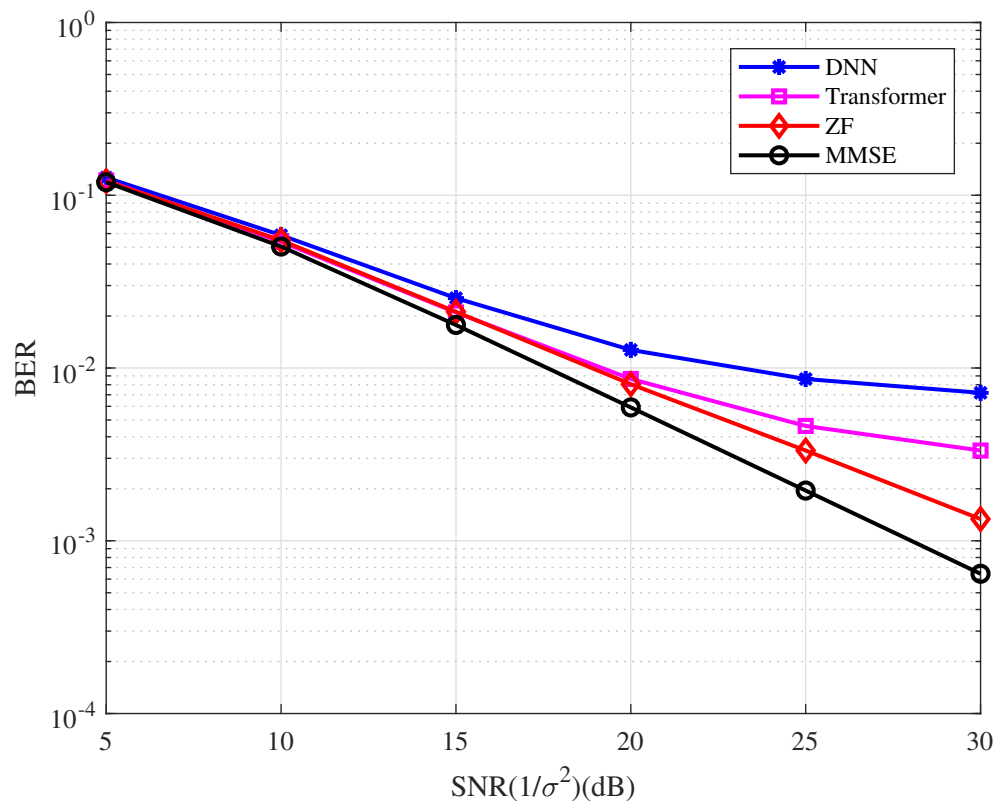
Figure 5. BER performance of the coded OFDM system with the different types of detector.

Example 2. *In this example, the mutual information based on the type of detector, which can be regarded as SIQ, versus the SNR is as depicted in Figure 6. For the ZF detector, we can observe that the Shannon limit for QPSK to achieve the spectral efficiency of 0.75, 1.0, 1.25, 1.5 bits/symbol/subcarrier is approximately −0.1, 2.3, 4.7, 7.5 dB, respectively. With the probability generated by the* softmax *function for the DNN and Transformer, their SIQs are higher than that of the conventional detector in the high SNR region. In the following examples, we take the ZF Shannon limit as a benchmark for comparison.*

Example 3. *In this example, a $[8,3]^{128}$ regular LDPC code is used. The BER performance of the coded OFDM with the type of detector at 0.75 bits/symbol/subcarrier spectral efficiency is depicted in Figure 7. From the figure, we have the following observations.*

- *At the BER of 10^{-4}, the BER performance with the ZF detector is approximately 6.2 dB away from the Shannon limit.*
- *The BER performance corresponding to the Transformer detector has an approximately 2.0 dB gain compared with the DNN detector.*
- *The Transformer-based system performs better than the ZF system in the high SNR region.*
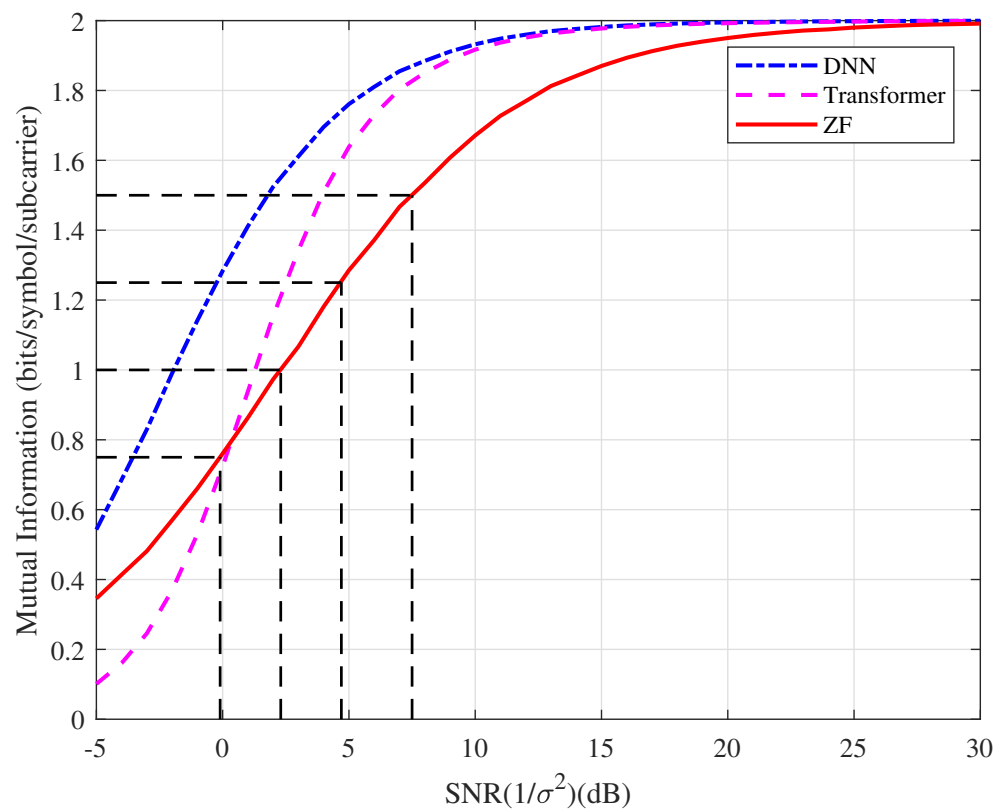- *The ZF detector has better performance than the MMSE detector in the coded OFDM system.*

Figure 6. Mutual information based on the type of detector (SIQ) by Monte Carlo simulation.
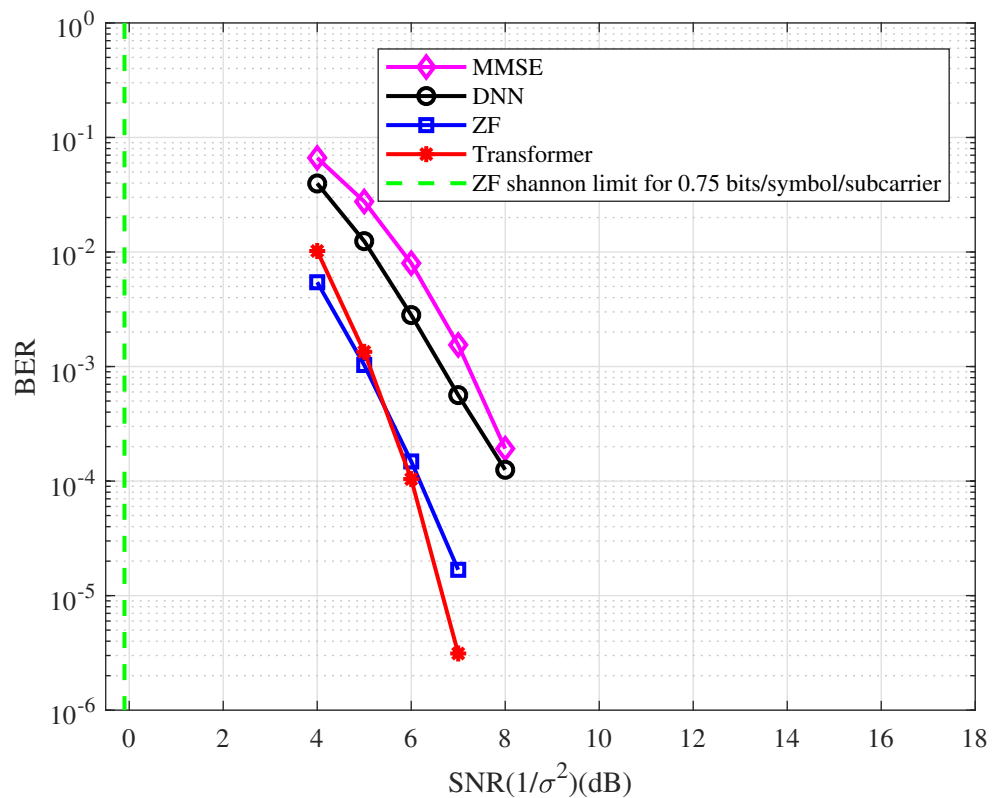

Figure 7. The BER performance of the coded OFDM with the different types of detector at 0.75 bits/symbol/subcarrier spectral efficiency.
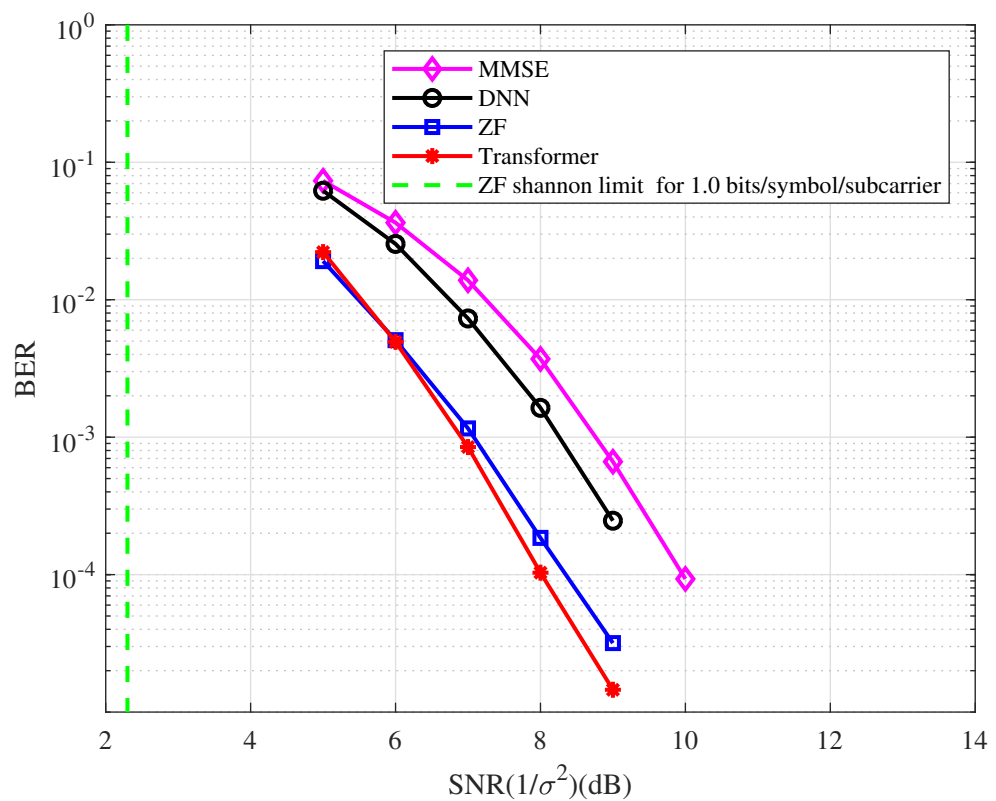
Example 4. *In this example, a* $[2,1]^{512}$ *regular LDPC code is used. The BER performance of the coded OFDM with the type of detector at 1.0 bits/symbol/subcarrier spectral efficiency is depicted in Figure 8. From the figure, we have the following observations.*

- *At the BER of* 10^{-4}, *the BER performance with the ZF detector is approximately 6.0 dB away from the Shannon limit.*
- *At the BER of* 10^{-4}, *the BER performance corresponding to the Transformer detector has an approximately 0.4 dB, 1.4 dB, and 2.0 dB gain compared with the ZF, DNN, and MMSE detectors, respectively.*
- *The BER performance with the ZF detector has an approximately 1.6 dB gain compared with the MMSE detector in the coded OFDM system.*

Figure 8. The BER performance of the coded OFDM with the type of detector at 1.0 bits/symbol/subcarrier spectral efficiency.

Example 5. *In this example, a* $[8,5]^{128}$ *regular LDPC code is used. The BER performance of the coded OFDM with the type of detector at 1.25 bits/symbol/subcarrier spectral efficiency is depicted in Figure 9. From the figure, we have the following observations.*

- *At the BER of* 10^{-4}, *the BER performance with the ZF detector is approximately 6.0 dB away from the Shannon limit.*
- *The Transformer-based system performs better than the ZF system in the high SNR region.*
- *At the BER of* 10^{-4}, *the BER performance corresponding to the Transformer detector has an approximately 1.0 dB and 1.4 dB gain compared with the DNN and MMSE detectors, respectively.*
- *The BER performance with the ZF detector has an approximately 1.1 dB gain compared with the MMSE detector in the coded OFDM system.*

Example 6. *In this example, a* $[4,3]^{256}$ *regular LDPC code is used. The BER performance of the coded OFDM with the type of detector at 1.5 bits/symbol/subcarrier spectral efficiency is depicted in Figure 10. From the figure, we have the following observations.*

- *At the BER of* 10^{-4}, *the BER performance with the ZF detector is approximately 7.0 dB away from the Shannon limit.*

- *The DL-based system performs better than the ZF system in the high SNR region.*
- *At the BER of 10^{-4}, the BER performance corresponding to the Transformer detector has an approximately 0.8 dB gain compared with the MMSE detector.*
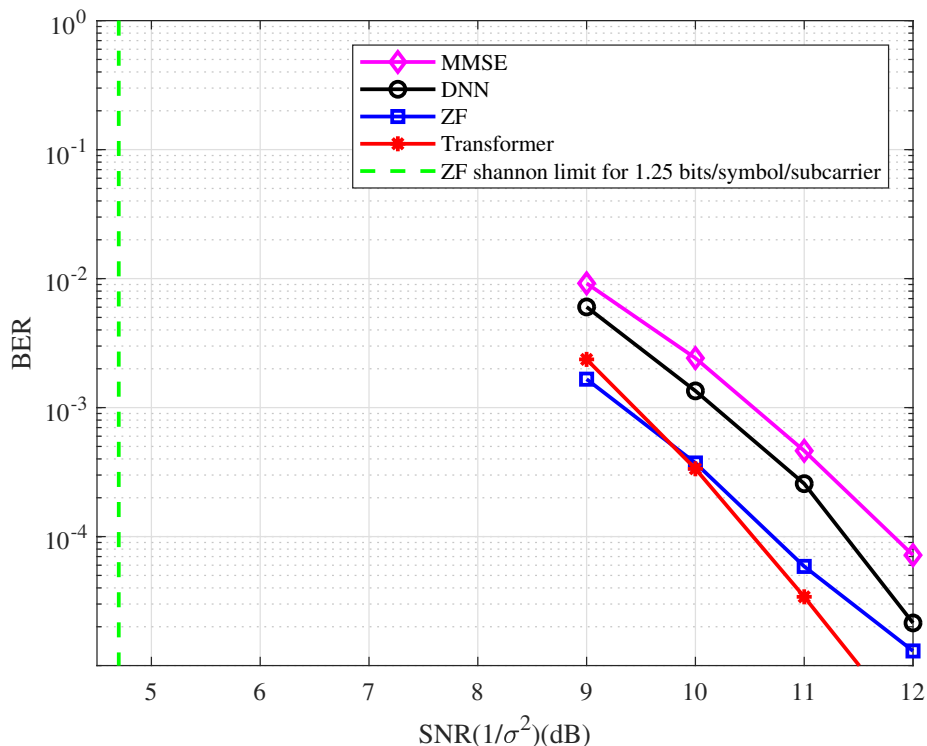

Figure 9. The BER performance of the coded OFDM with the type of detector at 1.25 bits/symbol/subcarrier spectral efficiency.
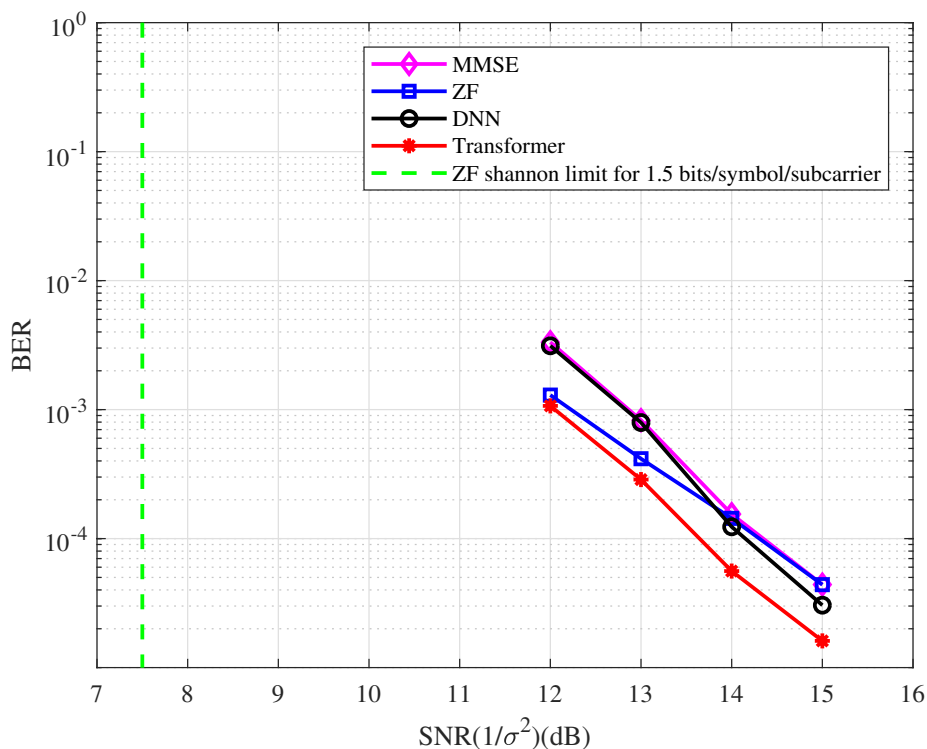

Figure 10. The BER performance of the coded OFDM with the type of detector at 1.5 bits/symbol/subcarrier spectral efficiency.

From the examples described above, we can observe that the conventional detector performs better than the DL-based detector in the uncoded system, while the situation is the opposite in the coded system with the BP decoding algorithm based on the Tanner graph.

An interesting question is whether the system employing our network still has good performance when decoding without BP based on the Tanner graph. We present an example for comparison.

Example 7. *In this example, a* $(3, 2, 2)$ *convolutional code (CC) is used. The CC is defined by the generation matrix* $\begin{pmatrix} 1+D & D & 1+D \\ D & 1 & 1 \end{pmatrix}$. *We use the Bahl–Cocke–Jelinek–Raviv (BCJR) decoding algorithm. The BER performance of the CC-coded OFDM with the type of detector is depicted in Figure 11. From the figure, we have the following observations.*

- *As with the uncoded system, the MMSE detector performs better than the ZF detector.*
- *The DL-based systems perform no better than the ZF and MMSE systems and have an error floor in the high SNR region. This implies that the proposed network is more suitable for the decoding algorithm based on the factor graph.*
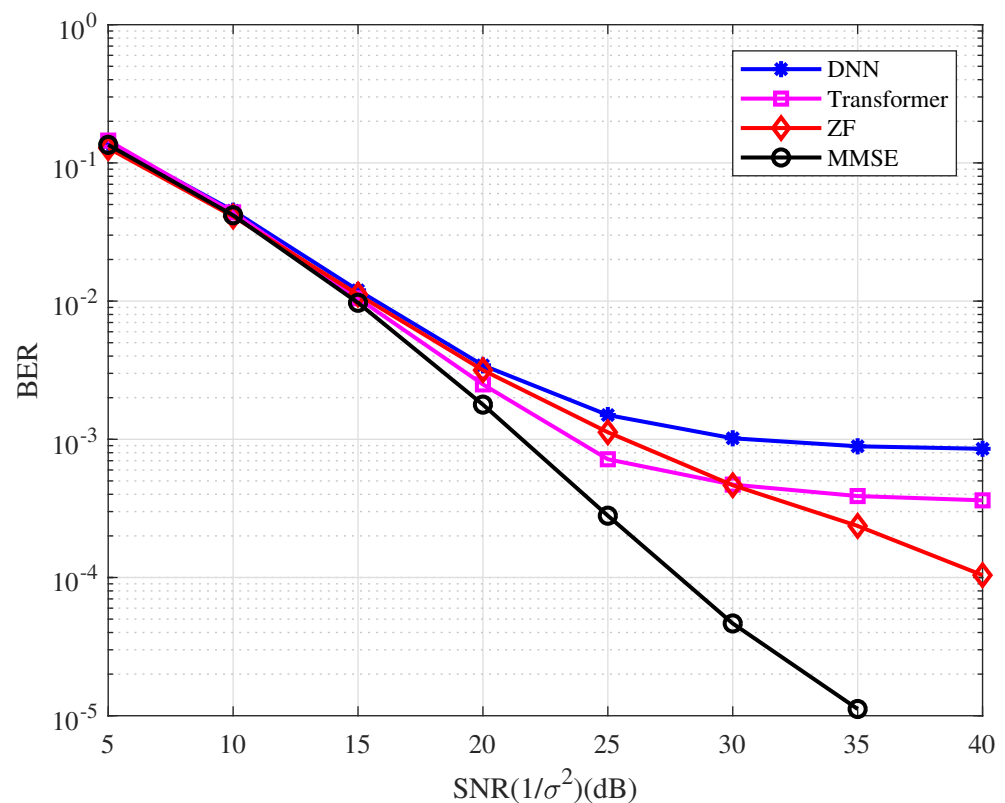

Figure 11. The BER performance of the CC-coded OFDM with the type of detector.

5. Conclusions

In the high-speed railway wireless communication system, an equalizer or detector, e.g., the ZF or MMSE detector, should be used to mitigate the intercarrier interference (ICI). In our schemes, the LDPC codes are employed to improve the error performance. Thus, a soft demapper should be designed for message delivery. In this paper, a Transformer-based detector/demapper is proposed in the mobile coded OFDM system. The proposed network can compute the symbol or bit probabilities, which can be used to calculate the SIQ or deliver it to the decoder. We also designed a DNN-based detector/demapper for comparison. The BER performance of the uncoded system and coded systems with different LDPC codes was presented. Numerical results show that although the Transformer-based uncoded OFDM systems do not outperform the systems utilizing the ZF or MMSE detectors, they

have better performance in the coded OFDM system compared with both the conventional and DNN-based systems.

Abbreviations

The following abbreviations are used in this manuscript:

BCJR	Bahl–Cocke–Jelinek–Raviv
BER	bit error rate
BMST	block Markov superposition transmission
BP	belief propagation
CC	convolutional codes
CSI	channel state information
CP	cyclic prefix
DFT	discrete Fourier transform
DL	deep learning
DNN	deep neural network
GD	gradient descent
ICI	intercarrier interference
IDFT	inverse discrete Fourier transform
IFFT	inverse fast Fourier transform
LDPC	low-density parity check
ML	maximum likelihood
MMSE	minimum mean square error
OFDM	orthogonal frequency division multiplexing
OFDM-IM	OFDM with index modulation
PDP	power-delay profile
PEG	progressive edge growth
SIQ	soft information quality
SNR	signal-to-noise ratio
SPA	sum product algorithm
ZF	zero forcing

References

1. Ma, X.; Liang, C.; Huang, K.; Zhuang, Q. Block Markov superposition transmission: Construction of big convolutional codes from short codes. *IEEE Trans. Inf. Theory* **2015**, *61*, 3150–3163. [CrossRef]
2. Wang, L.; Zhang, Y.; Ma, X. Block Markov superposition transmission for high-speed railway wireless communication systems. In Proceedings of the 2015 International Workshop on High Mobility Wireless Communications (HMWC), Xi'an, China, 21–23 October 2015; pp. 61–65.

3.   Ma, P.; Ren, J.; Sun, G.; Zhao, H.; Jia, X.; Yan, Y.; Zabalza, J. Multiscale Superpixelwise Prophet Model for Noise-Robust Feature Extraction in Hyperspectral Images. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 1–12. [CrossRef]

4.   Chen, R.; Huang, H.; Yu, Y.; Ren, J.; Wang, P.; Zhao, H.; Lu, X. Rapid Detection of Multi-QR Codes Based on Multistage Stepwise Discrimination and A Compressed MobileNet. *IEEE Internet Things J.* **2023**, 1. [CrossRef]

5.   Li, Y.; Ren, J.; Yan, Y.; Liu, Q.; Ma, P.; Petrovski, A.; Sun, H. CBANet: An End-to-end Cross Band 2-D Attention Network for Hyperspectral Change Detection in Remote Sensing. *IEEE Trans. Geosci. Remote. Sens.* **2023**, 1. [CrossRef]

6.   Xie, G.; Ren, J.; Marshall, S.; Zhao, H.; Li, R.; Chen, R. Self-attention enhanced deep residual network for spatial image steganalysis. *Digit. Signal Process.* **2023**, *139*, 104063. [CrossRef]

7.   Mao, Q.; Hu, F.; Hao, Q. Deep learning for intelligent wireless networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2595–2621. [CrossRef]

8.   O'shea, T.; Hoydis, J. An introduction to deep learning for the physical layer. *IEEE Trans. Cogn. Commun. Netw.* **2017**, *3*, 563–575. [CrossRef]

9.   Wang, T.; Wen, C.K.; Wang, H.; Gao, F.; Jiang, T.; Jin, S. Deep learning for wireless physical layer: Opportunities and challenges. *China Commun.* **2017**, *14*, 92–111. [CrossRef]

10.   Qin, Z.; Ye, H.; Li, G.Y.; Juang, B.H.F. Deep learning in physical layer communications. *IEEE Wirel. Commun.* **2019**, *26*, 93–99. [CrossRef]

11.   Ye, H.; Li, G.Y.; Juang, B.H. Power of deep learning for channel estimation and signal detection in OFDM systems. *IEEE Wirel. Commun. Lett.* **2018**, *7*, 114–117. [CrossRef]

12.   Gao, X.; Jin, S.; Wen, C.K.; Li, G.Y. ComNet: Combination of deep learning and expert knowledge in OFDM receivers. *IEEE Commun. Lett.* **2018**, *22*, 2627–2630. [CrossRef]

13.   Jiang, P.; Wang, T.; Han, B.; Gao, X.; Zhang, J.; Wen, C.K.; Jin, S.; Li, G.Y. AI-aided online adaptive OFDM receiver: Design and experimental results. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 7655–7668. [CrossRef]

14.   Van Luong, T.; Ko, Y.; Vien, N.A.; Nguyen, D.H.; Matthaiou, M. Deep learning-based detector for OFDM-IM. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 1159–1162. [CrossRef]

15.   Zhang, D.; Wang, S.; Niu, K.; Dai, J.; Wang, S.; Yuan, Y. Transformer-based detector for OFDM with index modulation. *IEEE Commun. Lett.* **2022**, *26*, 1313–1317. [CrossRef]

16.   Li, Y.; Chen, Z.; Wang, Y.; Yang, C.; Ai, B.; Wu, Y.C. Heterogeneous Transformer: A Scale Adaptable Neural Network Architecture for Device Activity Detection. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 3432–3446. [CrossRef]

17.   Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Red Hook, NY, USA, 4 December 2017; pp. 6000–6010.

18.   Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020, Proceedings of the 16th European Conference, Proceedings, Part I 16, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.

19.   Li, Y.; Chen, Z.; Liu, G.; Wu, Y.C.; Wong, K.K. Learning to Construct Nested Polar Codes: An Attention-Based Set-to-Element Model. *IEEE Commun. Lett.* **2021**, *25*, 3898–3902. [CrossRef]

20.   Gruber, T.; Cammerer, S.; Hoydis, J.; ten Brink, S. On deep learning-based channel decoding. In Proceedings of the 2017 51st Annual Conference on Information Sciences and Systems (CISS), IEEE, Baltimore, MD, USA, 22–24 March 2017; pp. 1–6.

21.   Xu, W.; Wu, Z.; Ueng, Y.L.; You, X.; Zhang, C. Improved polar decoder based on deep learning. In Proceedings of the 2017 IEEE International Workshop on Signal Processing Systems (SiPS), IEEE, Lorient, France, 3–5 October 2017; pp. 1–6.

22.   Liang, F.; Shen, C.; Wu, F. An iterative BP-CNN architecture for channel decoding. *IEEE J. Sel. Top. Signal Process.* **2018**, *12*, 144–159. [CrossRef]

23.   Bi, S.; Wang, Q.; Chen, Z.; Sun, J.; Ma, X. Deep learning-based decoding of block Markov superposition transmission. In Proceedings of the 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi'an, China, 23–25 October 2019; pp. 1–5.

24.   Kavčić, A.; Ma, X.; Mitzenmacher, M. Binary intersymbol interference channels: Gallager codes, density evolution, and code performance bounds. *IEEE Trans. Inf. Theory* **2003**, *49*, 1636–1652. [CrossRef]

25.   Asadi, M.; Huang, X.; Kavcic, A.; Santhanam, N.P. Optimal detector for multilevel NAND flash memory channels with intercell interference. *IEEE J. Sel. Areas Commun.* **2014**, *32*, 825–835. [CrossRef]

26.   Zheng, Y.R.; Xiao, C. Simulation models with correct statistical properties for Rayleigh fading channels. *IEEE Trans. Commun.* **2003**, *51*, 920–928. [CrossRef]