EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Compression and Transmission of Big AI Model Based on Deep Learning

Zhengping Lin[1], Yuzhong Zhou[1], Yuliang Yang[1,*], Jiahao Shi[1], and Jie Lin[1]

[1]Electric Power Research Institute, China Southern Power Grid Company, Guangzhou, China

Abstract

In recent years, big artificial intelligence (AI) models have demonstrated remarkable performance in various AI tasks. However, their widespread use has introduced significant challenges in terms of model transmission and training. To solve this issue, this paper proposes to involve the compression and transmission of large models using deep learning techniques, thereby ensuring the efficiency of model training. To achieve this objective, we leverage deep convolutional networks to design a novel approach for compressing and transmitting large models. Specifically, deep convolutional networks are employed for model compression, providing an effective way to reduce the size of large models without compromising their representational capacity. The proposed framework also includes carefully devised encoding and decoding strategies to guarantee the restoration of model integrity after transmission. In further, a tailored loss function is designed for model training, facilitating the optimization of both the transmission and training performance within the system. Through experimental evaluation, we demonstrate the efficacy of the proposed approach in addressing the challenges associated with large model transmission and training. The results showcase the successful compression and subsequent accurate reconstruction of large models, while maintaining their performance across various AI tasks.

Received on 28 August 2023; accepted on 08 December 2023; published on 11 December 2023

Keywords: Big AI model, compression and transmission, deep learning, convolutional networks

doi:10.4108/eetsis.3803

1. Introduction

The landscape of Information Technology (IT) has undergone a transformative evolution with the advent of wireless communication and transmission technologies [1–3]. Wireless communication has emerged as a pivotal catalyst in reshaping various domains within IT. A foundational work by Shannon laid the groundwork for information theory, serving as a cornerstone for understanding the fundamental limits of data transmission and coding techniques. This theory gained practical relevance through wireless channels, fostering insights into error-correcting codes and channel capacity [4–6]. The inception of cellular networks and their evolution, as chronicled, facilitated ubiquitous connectivity and paved the way for mobile computing paradigms. Wireless mesh networks (WMNs) and their architectural nuances were expounded in [7–9], underscoring their significance in forming Ad Hoc networks for data dissemination. The role of wireless transmission in Internet of Things (IoT) was underscored, exemplifying how wireless connections bind together the fabric of interconnected devices, enabling data exchange. In further, advancements in wireless security mechanisms were explicated in [10, 11], elucidating cryptographic protocols to fortify wireless data integrity and confidentiality. Collectively, the existing works have underscored the indispensable role of wireless communication and transmission as a linchpin in modern IT frameworks, reshaping connectivity, data dissemination, and security paradigms [12, 13].

On the other hand, the evolution of artificial intelligence (AI) has been profoundly shaped by the emergence and advancement of deep learning, with deep convolutional networks playing a pivotal role. Starting from early neural network formulations, deep learning gained traction through seminal works like AlexNet, which harnessed convolutional layers for hierarchical feature learning. This catalyzed the development of subsequent architectures such as GoogLeNet and ResNet, which introduced innovations like inception modules and residual connections. This paradigm

*Corresponding author. Email: yuliangyang@hotmail.com

shift extended to diverse applications including transfer learning, interpretability techniques, reinforcement learning, and generative adversarial networks (GANs), profoundly impacting fields ranging from image classification and natural language processing to medical image analysis and data generation. The trajectory of AI, characterized by the ascendancy of deep convolutional networks, has fundamentally redefined the AI landscape by enabling unprecedented achievements in pattern recognition, perception tasks, and feature extraction.

Recently, researchers have witnessed a surge in the development of massive AI models, accompanied by significant challenges [14–17]. These models, exemplified by GPT-3, have achieved remarkable performance across diverse tasks, yet their sheer size poses multifaceted challenges [18, 19]. Notably, the transmission of these big AI models introduces substantial hurdles, with issues such as transmission outage probability and transmission errors emerging as critical concerns. The intricate interplay between the colossal model size and the limitations of data transmission infrastructure can lead to increased outage probabilities and elevated transmission errors [20–22], impacting the reliability and efficiency of model dissemination. This necessitates innovative approaches to address the intricate trade-offs between model complexity, transmission robustness, and efficient data delivery, thereby ensuring the seamless propagation of these large-scale AI models across networks [23, 24].

Building upon prior literature review, this study addresses challenges by proposing a solution involving deep learning-based compression and transmission of large models to enhance the training efficiency. This entails utilizing deep convolutional networks to compress models effectively, preserving representational capacity. The framework incorporates meticulous encoding/decoding strategies for pre- and post-transmission model integrity and a customized loss function to optimize the transmission and training performance. Empirical validation demonstrates the approach's efficacy in compressing and accurately reconstructing large models across diverse AI tasks, thus contributing to the practical deployment of such models in practical applications.

2. MUSIC based big AI model compression and transmission

With the continuous escalation of computational resources and data availability, there has been an explosive growth in the size of parameters within AI models. Notably, in the realm of machine vision, the evolution is evident as exemplified by the rapid progression from models such as AlexNet with 60 million parameters to the Vision Transformer

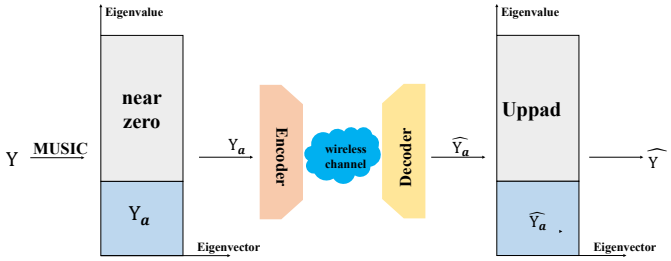

Figure 1. MUSIC based big AI model compression and transmission assisted by deep learning.

(ViT) in 2022, boasting an astounding 1.843 billion parameters. This progression is paralleled by a substantial enhancement in predictive accuracy on the ImageNet dataset, surging from 62% to an impressive 90%.

Let matrix $\mathbf{Y}$ denote the parameters of big AI model, where the associated size is often very large. The multiple signal classification (MUSIC) algorithm is typically used for direction-of-arrival estimation in signal processing, which is then used for data compression. To implement the MUSIC algorithm, the data model is firstly established as,

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{1}$$

where $\mathbf{Y}$ is the received data matrix of size $N \times L$, $\mathbf{X}$ is the original parameter matrix of size $N \times L$, and $\mathbf{N}$ is the noise matrix of size $N \times L$. From this equation, we can calculate the covariance matrix of $\mathbf{Y}$ as,

$$\mathbf{R}_Y = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H \tag{2}$$

where ^{H} denotes the conjugate transpose. From $\mathbf{R}_Y$, we can perform the eigenvalue decomposition as,

$$\mathbf{R}_Y = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^H, \tag{3}$$

where $\mathbf{E}$ is the matrix of eigenvectors, and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. From the eigenvalue decomposition, we can compute the MUSIC spectrum for different angles of arrival θ as,

$$P(\theta) = \frac{1}{\mathbf{a}^H(\theta)\mathbf{E}_n\mathbf{E}_n^H\mathbf{a}(\theta)}, \tag{4}$$

where $\mathbf{a}(\theta)$ is the steering vector for angle θ, and $\mathbf{E}_n$ is the matrix of eigenvectors corresponding to noise eigenvalues. Then, we choose the K angles of arrival that correspond to the lowest values in the MUSIC spectrum as,

$$\mathbf{Y}_a = [\mathbf{y}(\theta_1), \mathbf{y}(\theta_2), \ldots, \mathbf{y}(\theta_K)], \tag{5}$$

where $\mathbf{y}(\theta)$ is the column vector of the source parameter matrix $\mathbf{Y}$ corresponding to angle θ.

Note the complexity of the above MUSIC algorithm mainly lies in the eigenvalue decomposition. When the parameter size of big AI model is increasingly large, the size of the matrix $\mathbf{Y}$ will be also large, causing a huge burden on the implementation of the above MUSIC algorithm. To solve this issue, we provide an efficient eigenvalue decomposition in this paper. An efficient method to perform eigenvalue decomposition is the QR algorithm, which iteratively transforms the matrix into a upper Hessenberg form and then applies QR iterations to converge to the eigenvalues and eigenvectors. This algorithm is often implemented using shifts for improved convergence. Here is the basic outline of the QR algorithm with shifts for eigenvalue decomposition. Specifically, we first perform the Hessenberg reduction by decomposing the matrix $\mathbf{R}_Y$ into an upper Hessenberg matrix $\mathbf{H}$ using Householder transformations,

$$\mathbf{R}_Y = \mathbf{Q}_1 \mathbf{B} \mathbf{Q}_1^H. \tag{6}$$

Then, we perform the shifted QR iteration by iterating the following steps until convergence:

- Apply QR decomposition to $\mathbf{B}$ with a shifted eigenvalue,

$$\mathbf{B}_k - \sigma_k \mathbf{I} = \mathbf{Q}_k \mathbf{R} k \tag{7}$$

- Update $\mathbf{B}$,

$$\mathbf{B}_{k+1} = \mathbf{R}_k \mathbf{Q}_k + \sigma_k \mathbf{I} \tag{8}$$

After the above iteration becomes convergent, we extract the eigenvalues and eigenvectors. In particular, the eigenvalues of $\mathbf{R}_y$ are the diagonal entries of $\mathbf{B}$, and the eigenvectors are the columns of the final $\mathbf{Q}_k$ matrix. The efficiency of the QR algorithm comes from the fact that it reduces the matrix to upper Hessenberg form (which is close to upper triangular) before applying QR iterations. This can significantly reduce the number of iterations required for convergence, hence reducing the computational complexity of eigenvalue decomposition as well as the MUSIC algorithm.

3. CNN based Big Model Compression and Transmission

Drawing inspiration from the observation that the effectiveness of extracting localized correlations within matrices can be achieved by CNN-based autoencoders, we propose the formulation of the deep network. The network introduces a compressed CNN-based autoencoder, as illustrated in Fig. 2. In this scheme, Y_a is input to the encoder, having dimensions of $2 \times N_a \times N_t$. To capture relevant characteristics from this input, a 5×5 header convolution is employed, facilitating the amalgamation of insights from the real and imaginary segments. Subsequent to the 5×5 convolution, an encoder block is invoked, thereby facilitating the extraction of profound, abstract features. Notably, the network adopts a distinct approach compared to the ACRNet. The network optimizes complexity by employing a singular encoder block, which is capable of attaining a RecF of 17×17. The outcome of the encoder reshapes the matrix into a 1-D vector, followed by its transformation through fully connected (FC) layers to accomplish compression, yielding a codeword characterized by a compression ratio $\eta \in (0, 1)$. This codeword is subsequently conveyed wirelessly through wireless channel.

Once the codeword is obtained at the decoder, the decoder initiates the process of reinstating the matrix's original dimensions. This is achieved through a series of operations involving FC layers, along with a reshape operation. Subsequently, we use a head convolution equipped with a kernel measuring 6×6, effectively augmenting the recovery process. This is followed by the implementation of two consecutive compression decoder blocks, which take on the responsibility of rejuvenating the compacted information. It is pertinent to highlight that batch normalization (BN) is thoughtfully integrated into all convolutional stages, where the parametric rectified linear unit (PReLU) activation functions are also utilized. The utilization of PReLU, characterized by a tunable parameter denoted as β, is mathematically represented as,

$$\mathrm{PReLU}(x) = \begin{cases} x, & x \geq 0 \\ \beta x, & x < 0. \end{cases} \tag{9}$$

3.1. Devise on compression encoder and decoder blocks

In pursuit of amplifying RecF to bolster the performance of model matrix reconstruction, without incurring a simultaneous surge in computational overhead, an inventive encoder and decoder block construction is devised. These blocks are comprised of a succession of deep Convolutions (DConvs). Unlike the conventional convolutions, the Dconv employs a specialized interval termed as d, commonly referred to as the compression rate. In a formal context, the operation of the 2D-compression convolution, free of any bias considerations, is articulated as follows,

$$(C \circledast T)[u, v] = \sum_{w} \sum_{q} C[u + dw, v + dq] \cdot T[w, q], \tag{10}$$

In this context, the symbols $\circledast$ and C denote the Depthwise Convolution (DConv), and the 2D input matrix. Notation T represents the convolution kernel. Additionally, the indices w and q pertain to specific positions within the convolution kernel T. The pertinent expression, considering the compression rate

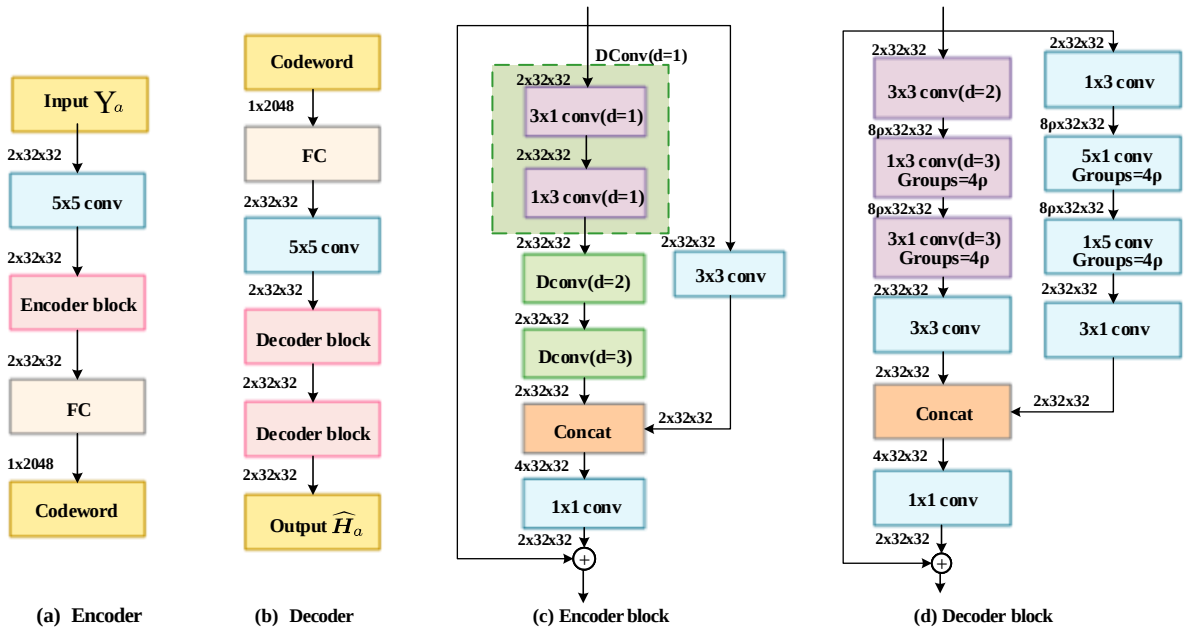

Figure 2. Framework of CNN based big AI model compression and transmission.

p, is given by,

$$t'_u = t_u + (t_u - 1)(p - 1). \tag{11}$$

Here, the values t_u and t'_u represent the actual and effective kernel sizes of convolution, respectively. It's noteworthy that when the compression rate p assumes a value of 1, the compression convolution simplifies to the standard convolution. However, a different scenario unfolds for $p > 1$, where the DConv operation has the capacity to yield an enlarged receptive field in comparison to a standard convolution utilizing an equivalent kernel size. This distinctive characteristic facilitates sparse sampling, particularly advantageous for a block-sparse model matrix strategy.

4. Comparison Results and Discussions

In this part, we perform some simulations on big AI model to validate the proposed studies in this paper. In particular, we employ the Vision Transformer (ViT) proposed in 2022, boasting an astounding 1.843 billion parameters. ViT is a groundbreaking neural network architecture that brought the power of the transformer model from natural language processing into the realm of computer vision. The transformer architecture, initially designed for sequential data like text, revolutionized language understanding by enabling models like generative pre-trained transformer (GPT). ViT extended this idea to image data, leading to remarkable advancements in image classification and other computer vision tasks. Moreover, we consider several compression rates in this work, with p varying in

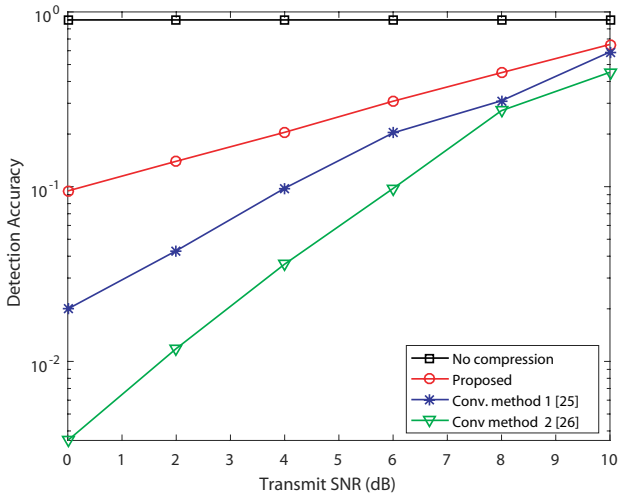

Figure 3. Detection accuracy with $p = 2$.

$\{2, 4, 8, 16\}$. In further, we consider the wireless transmission is through Rayleigh fading, with the average channel gain being unity. For comparison, we compare the proposed method in terms of detection accuracy, with the competitive methods in [25] and [26], in order to show the advantages of the proposed compression and transmission.

Fig. 3 and Table 1 show the detection accuracy of several competitive methods when $p = 2$, where the transmit SNR varies from 0dB to 10dB. From this figure and table, one can see that the proposed method consistently outperforms the other two methods across all SNR levels. It demonstrates the highest accuracy

Table 1 Numerical detection accuracy with $p = 2$.

Method	0dB	2dB	4dB	6dB	8dB	10dB
No compression	0.90	0.90	0.90	0.90	0.90	0.90
Proposed	**0.0945**	**0.1397**	**0.2047**	**0.3084**	**0.4492**	**0.65**
Conv. method 1 [25]	0.0200	0.0430	0.0979	0.2031	0.3103	0.59
Conv. method 2 [26]	0.0035	0.0118	0.0360	0.0977	0.2725	0.45

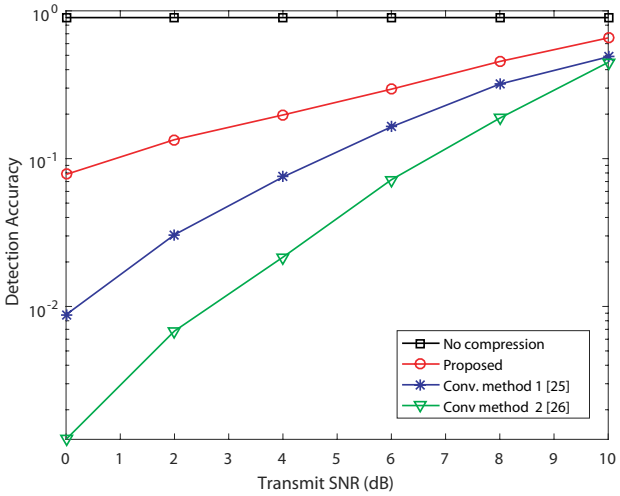

Figure 4. Detection accuracy when $p = 4$.

even under low SNR conditions. This suggests that the proposed method might be more robust to noise and able to extract meaningful information from the received signal. In particular, the detection accuracy of the proposed method is about 0.31 when the transmit SNR is 6dB, while that of the conventional method 1 and 2 is about 0.20 and 0.098. The detection accuracy of the proposed method is about 0.65 when the transmit SNR is 10dB, while that of the conventional method 1 and 2 is about 0.59 and 0.45. Moreover, the gap in accuracy between the proposed method and the other two methods widens as SNR increases. This indicates that while all methods benefit from a higher SNR, the superiority of the proposed method becomes more pronounced as the signal quality improves. Overall, the data in Fig. 3 and Table 1 indicates that the proposed method is the most accurate detection method across various SNR levels, and it is particularly advantageous under low SNR conditions. The convolutional methods also show improvement with increasing SNR, but they fall short compared to the proposed method in terms of accuracy.

Fig. 4 and Table 2 jointly illustrate the detection accuracy performance of multiple competitive methods under the condition of $p = 4$, encompassing a range of transmit SNR spanning from 0dB to 10dB. It is discernible from the information presented in these visual and tabular representations that the proposed method consistently exhibits a superior performance in contrast to the other two methods across the entire spectrum of SNR levels. Notably, even in scenarios characterized by low SNR, the proposed method attains the highest accuracy, a trait indicative of its potential robustness against noise and its proficiency in discerning pertinent information from the received signal. Specifically, the detection accuracy of the proposed method at the transmit SNR of 6dB is around 0.295, while the corresponding results for conventional methods 1 and 2 are approximately 0.164 and 0.072. Similarly, at the transmit SNR of 10dB, the detection accuracy for the proposed method notably ascends to approximately 0.65, while the accuracies for conventional methods 1 and 2 hover around 0.488 and 0.448. In further, the discernible widening of the accuracy gap between the proposed method and the other two methods as SNR levels escalate signifies that while all methods experience benefits from enhanced SNR conditions, the distinct superiority of the proposed method becomes progressively accentuated with improving signal quality. In a holistic perspective, the dataset encapsulated by Figure 4 and Table 2 substantiates the assertion that the proposed method emerges as the most precise detection approach across a spectrum of SNR levels, particularly excelling in conditions characterized by lower SNR levels. Simultaneously, it is evident that the convolutional methods exhibit enhanced performance with the increasing SNR. However, they fall short in terms of accuracy when compared to the proposed method.

Table 2 Numerical detection accuracy with $p = 4$.

Method	0dB	2dB	4dB	6dB	8dB	10dB
No compression	0.90	0.90	0.90	0.90	0.90	0.90
Proposed	**0.0782**	**0.1344**	**0.1974**	**0.2951**	**0.4544**	**0.6558**
Conv. method 1 [25]	0.0088	0.0306	0.0751	0.1641	0.3196	0.4880
Conv. method 2 [26]	0.0013	0.0068	0.0215	0.0720	0.1879	0.4482

Table 3 Numerical detection accuracy with $p = 8$.

Method	0dB	2dB	4dB	6dB	8dB	10dB
No compression	0.90	0.90	0.90	0.90	0.90	0.90
Proposed	**0.0567**	**0.1005**	**0.1891**	**0.2897**	**0.4683**	**0.6355**
Conv. method 1 [25]	0.0056	0.0217	0.0525	0.1456	0.2419	0.4611
Conv. method 2 [26]	0.0006	0.0036	0.0190	0.0642	0.1742	0.3437

Table 4 Numerical detection accuracy with $p = 16$.

Method	0dB	2dB	4dB	6dB	8dB	10dB
No compression	0.90	0.90	0.90	0.90	0.90	0.90
Proposed	**0.0323**	**0.0940**	**0.1711**	**0.2700**	**0.4387**	**0.5600**
Conv. method 1 [25]	0.0022	0.0140	0.0467	0.1022	0.2291	0.4220
Conv. method 2 [26]	0.0001	0.0022	0.0114	0.0457	0.1403	0.3339

Fig. 5 and Table 3 intricately detail the detection accuracy ofthe competitive methods under the context of $p = 8$, encompassing the transmit SNR spanning from 0dB to 10dB. Emanating from the graphical and tabular depictions, it discernibly emerges that the proposed method consistently exhibits a superior performance when juxtaposed against the alternative two methods, prevailing consistently across the entire gamut of SNR levels. This observed supremacy manifests conspicuously, even within scenarios characterized by low SNR, thereby implying an inherent robustness of the proposed method in the presence of perturbing noise, concurrently facilitating the discernment of salient information from the received signal. This characteristic becomes particularly evident as the proposed method attains a detection accuracy of approximately 0.2897 at the transmit SNR of 6dB, whereas its counterparts, conventional methods 1 and 2, approximately 0.1456 and 0.0642, correspondingly. In further, at a higher transmit SNR of 10dB, the detection accuracy of the proposed method escalates to around 0.6355, juxtaposed with the respective accuracies of approximately 0.4611 and 0.3437 attributed to conventional methods 1 and 2. Importantly, the discernible divergence in accuracy between the proposed method and the other two methods perceptibly widens in proportion to the

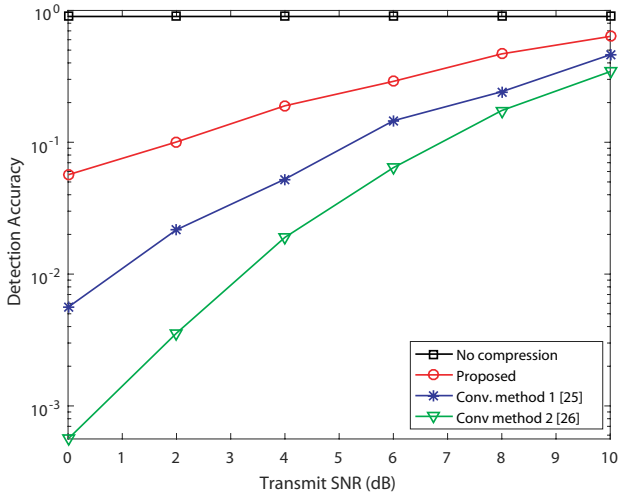

Figure 5. Detection accuracy when $p = 8$.

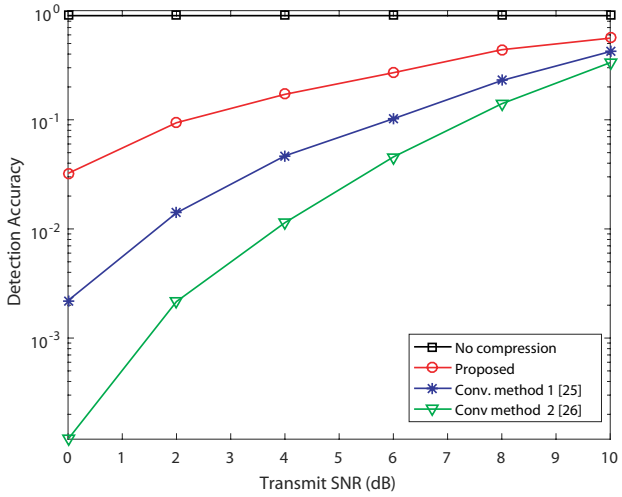

Figure 6. Detection accuracy when $p = 16$.

augmentation in SNR. This progressive widening signifies that while all the methods benefit from escalated SNR conditions, the pronounced supremacy of the proposed method becomes even more accentuated with the enhancement in the signal quality. Overall, the empirical evidence encapsulated within Figure 5 and Table 3 collectively underscores the proposed method as the preeminent modality for accurate detection across diverse SNR levels, particularly showcasing its efficacy within conditions typified by diminished SNR levels. It is notable that the convolutional methods do manifest an augmented efficacy in tandem with the augmentation of SNR, yet their performance remains notably subordinate the precision exhibited by the proposed method.

Figure 6 and Table 4 collectively portray the detection accuracy of the competitive methods under the context of $p = 2$, encompassing the transmit SNR spanning from 0dB to 10dB. The discernments derived from these graphical and tabular representations firmly establish the consistent supremacy of the proposed method over the alternate two approaches across the entire gamut of SNR levels. Evidently, the proposed method maintains a steadfast trajectory of superiority, asserting the highest precision, even within regimes of attenuated SNR. This conspicuous trend conjectures that the proposed method may possess enhanced resilience to noise perturbations, enabling the discernment of substantive insights from the received signal. More concretely, the precision of the proposed method stands at approximately 0.27 at the transmit SNR of 6dB, juxtaposed against the corresponding values of around 0.10 and 0.045 ascribed to conventional methods 1 and 2. Similarly, when the transmit SNR is elevated to 10dB, the detection accuracy of the proposed method escalates to about 0.56, while conventional methods 1 and 2 exhibit accuracies of roughly 0.42 and 0.33, respectively. This widening disparity in accuracy between the proposed method and the other two methods further corroborates the proposition that, although heightened SNR conditions confer benefits upon all methods, the ascendancy of the proposed method becomes increasingly conspicuous as the fidelity of the signal improves. In summation, the comprehensive evidence gleaned from the insights embedded within Figure 6 and Table 4 resolutely advocates the preeminence of the proposed method in terms of detection accuracy across diverse SNR scenarios, particularly underscoring its efficacy in scenarios typified by low SNR levels. Correspondingly, while the convolutional methods exhibit a commendable amelioration with increasing SNR, they remain comparatively outperformed by the proposed method with respect to the detection accuracy.

5. Conclusions

In conclusion, this paper introduced a comprehensive solution to tackle the intricate challenges arising from the widespread use of large AI models. These models have exhibited impressive performance in diverse AI tasks, yet their practical deployment has raised substantial concerns related to the transmission and training efficiency. The proposed approach effectively addresses these concerns by harnessing the power of deep learning techniques. By capitalizing on deep convolutional networks, a novel strategy is devised to compress and transmit large models while preserving their essential representational capabilities. The framework's ingenuity lies in its incorporation of meticulous encoding and decoding techniques, ensuring the models' fidelity post-transmission. Moreover, the introduction of a tailored loss function optimizes not only the model

training process but also the transmission performance, resulting in a harmonious synergy. Empirical validation underscores the efficacy of the approach, demonstrating successful compression and accurate reconstruction of large models without compromising their performance across a spectrum of AI tasks.

5.1. Copyright

References

[1] A. E. Haddad and L. Najafizadeh, "The discriminative discrete basis problem: Definitions, algorithms, benchmarking, and application to brain's functional dynamics," *IEEE Trans. Signal Process.*, vol. 71, pp. 1–16, 2023.

[2] R. Gabrys, S. Pattabiraman, and O. Milenkovic, "Reconstruction of sets of strings from prefix/suffix compositions," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 3–12, 2023.

[3] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical federated learning with quantization: Convergence analysis and system design," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 2–18, 2023.

[4] F. L. Andrade, M. A. T. Figueiredo, and J. Xavier, "Distributed banach-picard iteration: Application to distributed parameter estimation and PCA," *IEEE Trans. Signal Process.*, vol. 71, pp. 17–30, 2023.

[5] Q. Wang, S. Cai, Y. Wang, and X. Ma, "Free-ride feedback and superposition retransmission over LDPC coded links," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 13–25, 2023.

[6] Z. Xie, W. Chen, and H. V. Poor, "A unified framework for pushing in two-tier heterogeneous networks with mmwave hotspots," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 19–31, 2023.

[7] O. Lang, C. Hofbauer, R. Feger, and M. Huemer, "Range-division multiplexing for MIMO OFDM joint radar and communications," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 52–65, 2023.

[8] M. Hellkvist, A. Özçelikkale, and A. Ahlén, "Estimation under model misspecification with fake features," *IEEE Trans. Signal Process.*, vol. 71, pp. 47–60, 2023.

[9] Z. Xuan and K. Narayanan, "Low-delay analog joint source-channel coding with deep learning," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 40–51, 2023.

[10] F. Hu, Y. Deng, and A. H. Aghvami, "Scalable multi-agent reinforcement learning for dynamic coordinated multipoint clustering," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 101–114, 2023.

[11] H. Hui and W. Chen, "Joint scheduling of proactive pushing and on-demand transmission over shared spectrum for profit maximization," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 107–121, 2023.

[12] S. Liu and L. Ji, "Double multilevel constructions for constant dimension codes," *IEEE Trans. Inf. Theory*, vol. 69, no. 1, pp. 157–168, 2023.

[13] Q. Pan, Z. Qiu, Y. Xu, and G. Yao, "Predicting the price of second-hand housing based on lambda architecture and kd tree," *Infocommunications Journal*, vol. 14, no. 1, pp. 2–10, 2022.

[14] Z. Zhang, Z. Shi, and Y. Gu, "Ziv-zakai bound for doas estimation," *IEEE Trans. Signal Process.*, vol. 71, pp. 136–149, 2023.

[15] S. Guo and X. Zhao, "Multi-agent deep reinforcement learning based transmission latency minimization for delay-sensitive cognitive satellite-uav networks," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 131–144, 2023.

[16] X. Fang, W. Feng, Y. Wang, Y. Chen, N. Ge, Z. Ding, and H. Zhu, "Noma-based hybrid satellite-uav-terrestrial networks for 6g maritime coverage," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 138–152, 2023.

[17] R. Gabrys, V. Guruswami, J. L. Ribeiro, and K. Wu, "Beyond single-deletion correcting codes: Substitutions and transpositions," *IEEE Trans. Inf. Theory*, vol. 69, no. 1, pp. 169–186, 2023.

[18] X. Niu and E. Wei, "Fedhybrid: A hybrid federated optimization method for heterogeneous clients," *IEEE Trans. Signal Process.*, vol. 71, pp. 150–163, 2023.

[19] R. Yang, Z. Zhang, X. Zhang, C. Li, Y. Huang, and L. Yang, "Meta-learning for beam prediction in a dual-band communication system," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 145–157, 2023.

[20] X. Chen, W. Wei, Q. Yan, N. Yang, and J. Huang, "Time-delay deep q-network based retarder torque tracking control framework for heavy-duty vehicles," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 149–161, 2023.

[21] Z. Yang, F. Li, and D. Zhang, "A joint model extraction and data detection framework for IRS-NOMA system," *IEEE Trans. Signal Process.*, vol. 71, pp. 164–177, 2023.

[22] T. Zhang, K. Zhu, S. Zheng, D. Niyato, and N. C. Luong, "Trajectory design and power control for joint radar and communication enabled multi-uav cooperative detection systems," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 158–172, 2023.

[23] N. Zhang, M. Tao, J. Wang, and F. Xu, "Fundamental limits of communication efficiency for model aggregation in distributed learning: A rate-distortion approach," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 173–186, 2023.

[24] X. Yue, J. Xie, Y. Liu, Z. Han, R. Liu, and Z. Ding, "Simultaneously transmitting and reflecting reconfigurable intelligent surface assisted NOMA networks," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 1, pp. 189–204, 2023.

[25] M. Zhang, H. Zhang, Y. Fang, and D. Yuan, "Learning-based data transmissions for future 6g enabled industrial iot: A data compression perspective," *IEEE Network*, vol. 36, no. 5, pp. 180–187, 2022.

[26] X. Zhang, X. Zhu, J. Wang, H. Yan, H. Chen, and W. Bao, "Federated learning with adaptive communication compression under dynamic bandwidth and unreliable networks," *Information Sciences*, vol. 540, pp. 242–262, 2020.