Performance Analysis of Big Model Transmission under Double Rayleigh Fading

Ying Sun¹, Jiajia Huang¹, and Fusheng Wei¹

¹Guangdong Power Grid Co., Ltd, Guangzhou, China.

Abstract

The recent big model such as GPT-3.5 possesses an extensive understanding of natural language, and it can perform a wide range of tasks, making it a significant advancement in the field of artificial intelligence (AI). A critical challenge in the design and implementation of big model is that it imposes a heavy load on the wireless transmission due to a huge size of the deep network parameters, especially for the distributed implementation. To tackle this challenge, we investigate big model transmission under practical double Rayleigh fading environments, where the big model is simultaneously distributed to multiple training nodes through wireless transmission. To evaluate the system performance, we study the system outage probability (OP) based on the transmission latency, where an analytical expression is derived for the OP, which is used to evaluate the performance of big model transmission under double Rayleigh fading environments. Finally, we present some simulations under double Rayleigh fading environments, in order to show the validity of the proposed big model transmission.

Received on 21 August 2023; accepted on 04 November 2023; published on 09 November 2023

Keywords: Big model, data transmission, double Rayleigh fading, performance analysis.

Copyright © 2023 Ying Sun et al., licensed to EAI. This is an open access article distributed under the terms of the https://creativecommons.org/licenses/by/4.0/Creative Commons Attribution license, which permits unlimited use, distribution and reproduction in any medium so long as the original work is properly cited.

doi:10.4108/eetsis.3776

1. Introduction

The development of big model represents a fascinating journey through the evolution of neural networks, marked by significant milestones in the field of artificial intelligence [1–3]. It began with the inception of neural networks, where early attempts to simulate the human brain's interconnected neurons laid the foundation. However, it was the advent of deep neural networks that truly revolutionized the landscape [4–6]. With the introduction of deep learning, multiple layers of interconnected neurons enabled the extraction of intricate features from complex data, paving the way for more accurate and powerful models [7–9]. This progression led to the birth of big model, exemplified by giants like GPT-3.5. These models, with their massive scale and extensive understanding of natural language, represent the culmination of years of research and innovation in deep learning. As computational power, data availability, and algorithmic advancements grew, big model emerged as pivotal tools, capable of performing a wide range of tasks and driving significant advancements in the field of artificial intelligence [10–12]. Their development reflects the relentless pursuit of pushing the boundaries of AI, enabling applications that were once considered futuristic to become a reality [13–15].

Developing and deploying big model presents a cascade of challenges. These include acquiring the substantial computational resources required for training, curating high-quality and diverse datasets, enhancing generalization to new tasks through techniques like transfer learning, addressing interpretability for transparent decision-making, and ensuring energy efficiency during training and inference. Another challenge is achieving low-latency inference for real-time applications via model compression and hardware acceleration. Additionally, fortifying models against adversarial attacks and distribution shifts and navigating ethical considerations concerning fairness and privacy are important. Maintaining scalability as data and tasks grow is another hurdle. Ultimately, there's the pivotal challenge of devising efficient wireless transmission...
strategies to distribute big model, particularly in distributed settings or resource-constrained environments, where managing the heavy network load imposed by the size of model parameters is a critical factor in realizing the practical utility of big model [16–18].

Efficiently transmitting data, especially for supporting big model, involves a multi-pronged approach. One approach is the leveraging model compression techniques, such as pruning, quantization, and knowledge distillation, which can help reduce the size of the model. One more approach is to utilize efficient compression algorithms, like Huffman coding or arithmetic coding, to minimize the size of the transmitted data. Another approach is to adopt adaptive transmission strategies that prioritize sending critical model components first and utilize error correction codes to ensure robustness against wireless channel errors. In this direction, leveraging distributed computing and edge processing can be used to offload computation and reduce the amount of data needing transmission. In addition, the wireless communication protocol can be optimized by considering factors such as channel conditions, data prioritization, and latency requirements. In further, some advanced techniques have been explored like federated learning to enable collaborative model updates without transmitting raw data, reducing bandwidth needs while maintaining model performance, thus collectively addressing the significant challenge of wireless transmission for big model [19–22].

Reducing transmission outage probability in wireless communication has been a focus of extensive research, with relay, caching, and edge computing techniques playing pivotal roles [23–25]. Cooperative communication, network coding, and relay selection strategies optimize relay-based data transmission, while caching techniques such as content-centric networking and proactive caching reduce content retrieval delays. Edge computing paradigms like Mobile Edge Computing (MEC) and fog computing bring computation closer to the network edge, diminishing latency and enhancing reliability [26–28]. These techniques, coupled with advances in massive MIMO and coding-based caching, are collectively contributing to more robust and efficient wireless networks, addressing the critical challenge of minimizing transmission outage probability in modern communications.

In order to facilitate the advancement of big model, we undertake a comprehensive investigation into the transmission challenges they face within realistic double Rayleigh fading environments, where the big model needs to be efficiently distributed across multiple training nodes. Our focus extends to evaluate the overall system performance, specifically honing in on the critical metric of system outage probability (OP). This performance assessment hinges on the transmission latency, a pivotal factor in real-world deployments of big model. We delve deep into this metric, leveraging our findings to derive a precise analytical expression that quantifies the outage probability. To affirm the robustness and efficacy of our proposed big model transmission strategy, we meticulously execute a series of simulations conducted under the demanding conditions of double Rayleigh fading environments. Through these simulations, we aim to showcase the practical validity and effectiveness of our proposed approach, elucidating its potential to overcome the challenges posed by wireless transmission, especially in distributed settings, which are pivotal for the successful deployment of big model across a spectrum of applications.

2. SYSTEM MODEL of Big model transmission
As shown in Fig. 1, we study the system model of big model transmission with multiple training nodes under double Rayleigh fading channels [29–32], which includes one sender A, and M receivers \{B_{m} | 1 \leq m \leq M\}. Moreover, A has a heavy task of big model parameter that needs to be sent to M training nodes uniformly and simultaneously. Specifically, the wireless channel parameter between A and B_{m}, denoted as \(z_{m}\), could be represented as

\[ z_{m} = g_{1,m} \cdot g_{2,m}, \]  

where \(g_{1,m}\) and \(g_{2,m}\) are the first-hop double Rayleigh fading channel parameter and second-hop double Rayleigh fading channel parameter, respectively. Moreover, the parameters \(g_{1,m}\) and \(g_{2,m}\) can be modeled as

\[ g_{1,m} \sim \mathcal{CN}(0, \xi_{1}), \]  

\[ g_{2,m} \sim \mathcal{CN}(0, \xi_{2}), \]  

where \(\xi_{1}\) and \(\xi_{2}\) are the average channel gain of the first-hop and second-hop for the double Rayleigh fading channels. Thus, the available transmission rate between A and B_{m} is

\[ R_{m} = W \log_{2} \left( 1 + \frac{P}{\sigma^{2}} |z_{m}|^{2} \right), \]  

Figure 1. System model of big model transmission with multiple training nodes under double Rayleigh fading channels.
where $W$ is the allocated bandwidth of the link $A-B_m$, $P$ is the transmit power of $A$, and $\sigma^2$ is the variance of additive white Gaussian noise (AWGN). In this network, the node $A$ has the task of solving for the probability that the transmission latency is between the training node $K$ bits, that needs to be sent to $M$ training nodes uniformly and simultaneously. Hence, the transmission latency from $A$ to the training node $B_m$ is

$$t_m = \frac{k}{R_m},$$

where $k = \frac{K}{M}$ is denoted by the task length allocated to each training node.

3. PERFORMANCE ANALYSIS of big model transmission

In this section, we give a definition of the system outage probability for the big model transmission, and then derive the associated closed-form expression for the outage probability. In particular, the system outage occurs when at least one transmission latency between the sender $A$ to the $M$ training nodes is larger than a given latency threshold $\gamma_{th}$.

$$P_{out} = 1 - Pr(t_1 < \gamma_{th}, t_2 < \gamma_{th}, \cdots, t_M < \gamma_{th}),$$

$$= 1 - Pr(t_1 < \gamma_{th}) \cdot Pr(t_2 < \gamma_{th}) \cdots Pr(t_M < \gamma_{th}),$$

$$= 1 - [Pr(t_1 < \gamma_{th})]^M. \tag{6}$$

$$P_m = 1 - Pr(t_m < \gamma_{th}), \tag{7}$$

$$= 1 - Pr \left( \frac{k}{W \log_2 (1 + \frac{P}{\sigma^2} |Z_m|^2)} < \gamma_{th} \right), \tag{8}$$

$$= Pr \left( |Z_m|^2 > \left( \frac{2^{\gamma_{th}} - 1}{\frac{k}{W \log_2 (1 + \frac{P}{\sigma^2})}} \right) \alpha^2 \right). \tag{9}$$

Secondly, we derive the associated closed-form expression for the outage probability. In the first, because of

$$g_{1m} \sim CN(0, \xi_1), \tag{10}$$

$$g_{2m} \sim CN(0, \xi_2), \tag{11}$$

$$Z_m = g_{1m} \cdot g_{2m}, \tag{12}$$

we can simplify the above notations as,

$$|g_{1m}|^2 = X, \quad |g_{2m}|^2 = Y, \tag{13}$$

and then we have,

$$P_m = 1 - Pr \left[ |Z_m|^2 \leq \left( \frac{2^{\gamma_{th}} - 1}{\frac{k}{W \log_2 (1 + \frac{P}{\sigma^2})}} \right) \alpha^2 \right], \tag{14}$$

$$= 1 - Pr \left[ |g_{1m}|^2 \leq \frac{2^{\gamma_{th}} - 1}{\left|g_{2m}|^2 \right|} \alpha^2 \right], \tag{15}$$

$$= 1 - \sum_{0}^{\infty} \int_{0}^{\infty} f_{X,Y}(x, y) dxdy. \tag{16}$$

$$P_m = 1 - \int_{0}^{\infty} \int_{0}^{\infty} \frac{1}{\xi_1} e^{-\frac{z}{\xi_1}} \frac{1}{\xi_2} e^{-\frac{z}{\xi_2}} dxdy, \tag{17}$$

$$= 1 - \int_{0}^{\infty} \frac{1}{\xi_1} e^{-\frac{z}{\xi_1}} \frac{1}{\xi_2} e^{-\frac{z}{\xi_2}} \left( \frac{2^{\gamma_{th}} - 1}{\frac{k}{W \log_2 (1 + \frac{P}{\sigma^2})}} \right) \alpha^2 dy, \tag{18}$$

By solving the above integral, we can write $P_m$ in further as

$$P_m = 1 - \left[ e^{-\frac{2^{\gamma_{th}} - 1}{\frac{k}{W \log_2 (1 + \frac{P}{\sigma^2})}} \alpha^2} \right] \xi_1 \xi_2 K_1 \left( \frac{4(2^{\gamma_{th}} - 1)\alpha^2}{P \xi_1 \xi_2} \right). \tag{19}$$

$$= 2\alpha \sqrt{\frac{4(2^{\gamma_{th}} - 1)\alpha^2}{P \xi_1 \xi_2}} = 2\alpha \sqrt{\frac{2^{\gamma_{th}} - 1}{P \xi_1 \xi_2}} \xi_1 \xi_2 \tag{20}$$

where $K_1(x)$ denotes the first-order modified Bessel function of the second kind [33].

Finally, the outage probability of the big model transmission is given by

$$P_{out} = 1 - (P_m)^M. \tag{21}$$

$$= 1 - \left[ 2\alpha \sqrt{\frac{2^{\gamma_{th}} - 1}{P \xi_1 \xi_2}} K_1 \left( 2\alpha \sqrt{\frac{2^{\gamma_{th}} - 1}{P \xi_1 \xi_2}} \right) \right]^M. \tag{22}$$

From the above analytical expression of $P_{out}$, we can readily evaluate the failure probability of big model transmission, under double Rayleigh fading environments.
Outage probability

4. SIMULATION RESULTS AND DISCUSSIONS for big model transmission

In this section, we present a series of simulations designed to empirically validate the analytical findings on the big model transmission. If not specified, the parameters of big model transmission are set as follows. Firstly, the task length of the big model is \( K = 1 \text{MBits} \), while the communication bandwidth is \( W = 1 \text{MHz} \) [34–36]. Secondly, the average channel gains \( \xi_1 \) and \( \xi_2 \) are both set to unity [37–39]. The transmit signal-to-noise ratio (SNR) is established at 10 dB, and the variance of AWGN is \( \sigma^2 = 1 \). Finally, the number of training nodes \( M \) is fixed to 5, and the latency threshold is \( \gamma_{th} = 1 \text{s} \).

Fig. 2 and Table 1 show the simulation and analysis results of the big model transmission versus the transmit SNR, where the SNR increases from -5 dB to 30 dB. From Fig. 2 and Table 1, it can be observed that the outage probability decreases with increasing SNR, which is due to a larger SNR leading to a higher transmission rate, which reduces the task transmission latency and eventually the outage probability. In addition, the analytical results overlap with the simulation ones, which validates the closed-form solution. For example, the simulated OP with SNR=25dB and \( \gamma_{th} = 0.1 \) is about 0.19724, while the associated analytical OP is about 0.19722. The simulated OP with SNR=25dB and \( \gamma_{th} = 1 \) is about 0.01752, while the associated analytical OP is about 0.01753. Finally, the outage probability with \( \gamma = 0.1 \) is consistently higher than that with \( \gamma_{th} = 1 \), indicating that a relaxed latency constraint can help reduce the communication outage probability.

Fig. 3 and Table 2 depict the impact of the task length \( K \), on both analytical and simulated outage probabilities of the big model transmission. The value of \( K \) varies from 1 MBits to 9MBits, while \( \xi_1 \) and \( \xi_2 \) take values from the set of \( \{1, 2\} \). In Fig. 3 and Table 2, it becomes evident that as the value of \( K \) increases, the outage probability also increases. This trend indicates that a higher value of \( K \) leads to a worse outage performance. Moreover, the overlap between the analytical and simulation results in Fig. 3 affirms the accuracy and validity of the derived expression for the outage probability. In particular, the simulated OP with \( K = 1 \text{MBits} \) and \( \xi_1 = \xi_2 = 2 \) is about 0.09724, while the associated analytical OP is about 0.09723. The simulated OP with \( K = 1 \text{MBits} \) and \( \xi_1 = \xi_2 = 1 \) is about 0.26963, while the associated analytical OP is about 0.26954. Additionally, the outage performance with \( \xi_1 = 2 \) and \( \xi_2 = 2 \) is notably superior to that with \( \xi_1 = 1 \) and \( \xi_2 = 1 \). This phenomenon appears from the fact that an increased value of \( \xi \) correlates with improved channel quality.

Fig. 4 and Table 3 show the effect of communication bandwidth \( W \) on the analyzed and simulated outage probability of the big model transmission, where \( W \) varies from 1MHz to 9MHz. As shown in Fig. 4 and Table 3, we can see that the outage probability shows a decreasing trend with the increase of \( W \), which indicates that a larger communication bandwidth effectively reduces the outage probability, and the analyzed results overlap with the simulated results, which verifies the closed-form solution. For example, the simulated OP with \( W = 9 \text{MHz} \) and \( \gamma_{th} = 0.1 \) is about 0.290482, while the associated analytical OP is about 0.290627. The simulated OP with \( W = 9 \text{MHz} \) and \( \gamma_{th} = 1 \) is about 0.048095, while the associated analytical OP is about 0.048095. Moreover, the outage probability with \( \gamma_{th} = 0.1 \) is always higher than that with \( \gamma_{th} = 1 \), which is because that a higher
Table 1: Numerical outage probability of big model transmission versus the transmit SNR.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>-5</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim: $\gamma_{th} = 0.1$</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9979</td>
<td>0.9436</td>
<td>0.7217</td>
<td>0.4187</td>
<td>0.1972</td>
<td>0.0821</td>
</tr>
<tr>
<td>ana: $\gamma_{th} = 0.1$</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9979</td>
<td>0.9437</td>
<td>0.7218</td>
<td>0.4185</td>
<td>0.1973</td>
<td>0.0821</td>
</tr>
<tr>
<td>sim: $\gamma_{th} = 1$</td>
<td>0.9796</td>
<td>0.8289</td>
<td>0.5333</td>
<td>0.2695</td>
<td>0.1170</td>
<td>0.0464</td>
<td>0.0175</td>
<td>0.0064</td>
</tr>
<tr>
<td>ana: $\gamma_{th} = 1$</td>
<td>0.9796</td>
<td>0.8289</td>
<td>0.5334</td>
<td>0.2696</td>
<td>0.1169</td>
<td>0.0464</td>
<td>0.0175</td>
<td>0.0064</td>
</tr>
</tbody>
</table>

Table 2: Numerical outage probability of big model transmission versus $K$.

<table>
<thead>
<tr>
<th>$K$ (Mbits)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim: $\xi_{1,2} = 2$</td>
<td>0.0972</td>
<td>0.1741</td>
<td>0.2447</td>
<td>0.3117</td>
<td>0.3756</td>
<td>0.4376</td>
<td>0.4970</td>
<td>0.5542</td>
<td>0.6086</td>
</tr>
<tr>
<td>ana: $\xi_{1,2} = 2$</td>
<td>0.0973</td>
<td>0.1741</td>
<td>0.2447</td>
<td>0.3116</td>
<td>0.3758</td>
<td>0.4376</td>
<td>0.4970</td>
<td>0.5540</td>
<td>0.6084</td>
</tr>
<tr>
<td>sim: $\xi_{1,2} = 1$</td>
<td>0.2696</td>
<td>0.4336</td>
<td>0.5581</td>
<td>0.6566</td>
<td>0.7354</td>
<td>0.7988</td>
<td>0.8491</td>
<td>0.8889</td>
<td>0.9201</td>
</tr>
<tr>
<td>ana: $\xi_{1,2} = 1$</td>
<td>0.2696</td>
<td>0.4339</td>
<td>0.5582</td>
<td>0.6565</td>
<td>0.7353</td>
<td>0.7986</td>
<td>0.8491</td>
<td>0.8890</td>
<td>0.9200</td>
</tr>
</tbody>
</table>

Table 3: Numerical outage probability of big model transmission versus $W$.

<table>
<thead>
<tr>
<th>$W$ (MHz)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim: $\gamma_{th} = 0.1$</td>
<td>0.9436</td>
<td>0.7353</td>
<td>0.5934</td>
<td>0.4999</td>
<td>0.3847</td>
<td>0.2905</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ana: $\gamma_{th} = 0.1$</td>
<td>0.9437</td>
<td>0.7353</td>
<td>0.5934</td>
<td>0.4999</td>
<td>0.3847</td>
<td>0.2907</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sim: $\gamma_{th} = 1$</td>
<td>0.2696</td>
<td>0.1609</td>
<td>0.1173</td>
<td>0.0934</td>
<td>0.0781</td>
<td>0.0531</td>
<td>0.0481</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ana: $\gamma_{th} = 1$</td>
<td>0.2696</td>
<td>0.1609</td>
<td>0.1174</td>
<td>0.0934</td>
<td>0.0780</td>
<td>0.0530</td>
<td>0.0481</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Numerical outage probability of big model transmission versus $M$.

<table>
<thead>
<tr>
<th>$M$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim: $\xi_{1,2} = 2$</td>
<td>0.0901</td>
<td>0.0900</td>
<td>0.0926</td>
<td>0.0951</td>
<td>0.0974</td>
<td>0.0993</td>
<td>0.1012</td>
<td>0.1027</td>
<td>0.1041</td>
<td>0.1056</td>
</tr>
<tr>
<td>ana: $\xi_{1,2} = 2$</td>
<td>0.0899</td>
<td>0.0900</td>
<td>0.0926</td>
<td>0.0951</td>
<td>0.0973</td>
<td>0.0993</td>
<td>0.1011</td>
<td>0.1027</td>
<td>0.1041</td>
<td>0.1054</td>
</tr>
<tr>
<td>sim: $\xi_{1,2} = 1$</td>
<td>0.2334</td>
<td>0.2422</td>
<td>0.2528</td>
<td>0.2618</td>
<td>0.2701</td>
<td>0.2762</td>
<td>0.2818</td>
<td>0.2870</td>
<td>0.2920</td>
<td>0.2955</td>
</tr>
<tr>
<td>ana: $\xi_{1,2} = 1$</td>
<td>0.2334</td>
<td>0.2421</td>
<td>0.2528</td>
<td>0.2619</td>
<td>0.2696</td>
<td>0.2762</td>
<td>0.2820</td>
<td>0.2872</td>
<td>0.2918</td>
<td>0.2960</td>
</tr>
</tbody>
</table>

Latency threshold can effectively improve the outage performance.

Fig. 5 and Table 4 illustrate the influence of the number of training nodes $M$, on both analytical and simulated outage probabilities of the big model transmission, where $M$ varies within the range of 1 to 10, while parameters $\xi_1$ and $\xi_2$ assume values from the set $\{1, 2\}$. From the observation of Fig. 5 and Table 4, it becomes evident that an increase in the value of $M$ leads to a corresponding increase in the outage probability. This observation shows that a higher $M$ is associated with a deteriorating outage performance. Moreover, it is worth noting that the outage performance exhibits noticeable improvement when considering $\xi_1 = 2$ and $\xi_2 = 2$, in contrast to the scenario with $\xi_1 = 1$ and $\xi_2 = 1$. This gap can be attributed to the inherent correlation between the larger values of $\xi$ and enhanced channel quality. It is essential to highlight that the alignment between the analytical and simulation results depicted in Fig. 5 substantiates the accuracy and validity of the derived analytical expression governing the outage probability.

5. conclusion

In conclusion, this paper addressed a critical challenge in the field of artificial intelligence, specifically related to the transmission of large models like GPT-3.5 in practical double Rayleigh fading environments.
The study focused on distributed implementations, which were found to impose heavy loads on wireless transmission due to the massive size of network parameters. The investigation of system outage probability based on transmission latency provided valuable insights, and the derived analytical expression offered a framework for understanding the performance of big model transmission. The presented simulations validated the proposed approach and offered promising strategies for handling the transmission challenges associated with the advanced AI models.

5.1. Copyright

The Copyright licensed to EAI.

References


