Enhancing Recommender Systems Through Multimodal Large Language Models and Neural Matrix Factorization

Yizhuo Jiang
School of Management, Shanghai University, Shanghai, China
Jiangyz1002@126.com

2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) | 979-8-3503-8694-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECAI62591.2024.10675129

***Abstract*—This This study proposes a new recommendation system model that integrates multimodal data, large language models (LLM), and neural matrix factorization techniques. By using Vision Transformer (ViT) and the BERT model, we deeply mine the image and text information of users and products, and combine unstructured data with neural network technology for evaluation prediction and recommendation, showing excellent prediction performance. This study also verified the performance advantages of this method compared with traditional recommendation system methods through various ablation experiments. The results highlight the importance of text data in capturing user preferences and improving recommendation accuracy. In addition, the study found that in practical applications, more detailed segmentation processing of visual data is needed to accurately evaluate the specific contribution of visual data to multi-modal recommendation systems. These findings provide important theoretical and practical guidance for further optimizing recommendation systems.**

Keywords-component; recommended system; multimodal; large language model; neural matrix factorization

I. INTRODUCTION

With the development of information and communication technologies and platforms, users are confronted with an overwhelming amount of data and information. In the era of information explosion, the accurate retrieval of content that meets users' needs has been an important research problem. Recommender systems can analyze user characteristics, including past behavior and preferences, to rank items according to certain rules and display items that may be of interest to the user at the top, thus achieving effective information filtering [1]. Recommender systems rely on various types of input. Explicit feedback is the most directly usable data. For example, users give star ratings to products or express their preferences for movies by liking them [2]. In the absence of explicit feedback, many recommender systems infer user characteristics from implicit feedback, such as purchase history, streaming activity, browsing history, search patterns, or time spent watching videos. One of the earliest and most widely used techniques in recommender systems is collaborative filtering, which recommends products or items based on past purchases. Collaborative filtering systems include two main methods: neighborhood-based methods and latent factor models [3]. These two algorithms link the two critical entities: products and users. Neighborhood-based methods focus on the relationships between items or users. For example, when personalizing recommendations for user A, user-based collaborative filtering algorithms find other users with similar interests to user A, and then recommend items that those users are interested in but that user A may not know about. Latent factor models, such as matrix factorization, convert both products and users into the same latent factor, or embedding space, and then describe products and users by automatically inferring embeddings from their interactions.

In addition to collaborative filtering models, content-based filtering is another important type of recommender system. Content-based recommender systems analyses some unstructured data, such as images and text information, and discover hidden patterns in them to make recommendations. The difference between collaborative filtering and content-based recommender systems lies in the data they rely on. Collaborative filtering relies on user and item rating data to make predictions and recommendations, while content-based recommender systems rely on high-level features of users and items to make predictions [4]. Both collaborative filtering and content-based recommender systems have limitations. Collaborative filtering systems can only consider structured data and may ignore feature information from unstructured data, while content-based recommendation systems necessarily ignore the importance of preference similarity between individuals for prediction, which may lead to inaccurate predictions. [5]. Furthermore, collaborative filtering algorithms and content-based recommender systems face issues such as cold start, sparsity, and scalability, which can lead to poor prediction performance [6]. This study develops a comprehensive recommendation model based on both collaborative and content-based recommendations to address performance degradation in scenarios with cold start or small data volumes. This paper employs a large model for deep embedding of textual features and uses neural matrix factorization instead of traditional inner product for collaborative filtering. This approach effectively addresses the current cold start and accuracy issues in recommender systems, making a significant contribution to the improvement of recommender systems and introducing a new paradigm for incorporating multiple modalities to enhance recommendation accuracy.

II. RELATED WORK

The application of artificial intelligence has been widely used in many aspects, including education [7], medical care [8], customer segment [9], etc., and has made outstanding contributions to different fields. In terms of recommended areas, any researchers have turned to neural network models as a result of advances in deep artificial intelligence. Traditionally, matrix factorization is used in collaborative filtering to model the latent

features of users and items, using an inner product to represent interactions. Neural networks can theoretically fit any function, replacing the inner product by learning from the data. This approach is known as neural collaborative filtering [10], which has shown excellent performance in recommendation prediction. A representative method of neural collaborative filtering is neural matrix factorization. As shown in Figure 1, neural matrix factorization combines generalized matrix factorization (GMF) and a multi-layer perceptron (MLP). GMF generates a rating prediction by computing the dot product of user and item embedding vectors to capture linear interactions. At the same time, the user and item embedding vectors are concatenated and fed into the MLP to capture non-linear interactions. Finally, the user and item embedding vectors are shared and the outputs of these two parts are combined using an MLP in the fusion layer. Among them, the formula of collaborative filtering is shown in Formula 1. $\widehat{y_{ui}}$ represents the predicted rating of user u for item i. σ is the final activation function. $\boldsymbol{h}$, $\boldsymbol{W}$, and $\boldsymbol{b}$ are model parameters. $\boldsymbol{p_u}$ and $\boldsymbol{q_i}$ are the embedding vectors for user u and item i, respectively, and are shared between the GMF and MLP parts. a represents an activation function. $\circ$ is the element-wise product.

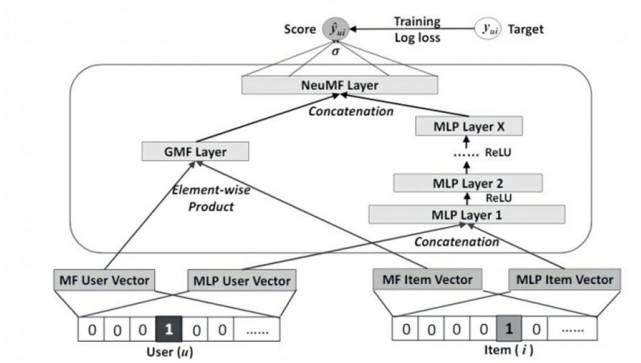

Figure 1. The structure of neural matrix factorization

$$\widehat{y_{ui}} = \sigma\left(\boldsymbol{h}^T a\left(\boldsymbol{p_u} \circ \boldsymbol{q_i}\right) + \boldsymbol{W}\left[\boldsymbol{p_{uq_i}}\right] + \boldsymbol{b}\right) \qquad (1)$$

However, neural collaborative filtering models still face challenges similar to traditional collaborative filtering methods when dealing with cold start or small data scenarios. As large models become increasingly prevalent, they appear to offer solutions to the cold start problem in recommendation systems. By leveraging the deep semantic extraction capabilities of large models on user or item features, it is possible to enhance the retrieval performance of recommendation systems. Features generally include image and text information. For textual information, one effective approach is using large models such as BERT, which can recognize text well [11]. BERT is a pre-trained language representation model based on deep learning and is one of the most important pre-trained models in NLP. As shown in Figure 2, BERT adopts a bidirectional encoder structure, which has been widely proven to capture contextual and semantic relationships in text effectively. The first layer of the BERT model is the word embedding layer. After word embedding, the input proceeds through 12 transformer layers to obtain the final feature representation. BERT's embedding component includes word embeddings, segment embeddings, and position embeddings. Word embeddings map each word or sub-word into a vector in high-dimensional space, while segment embeddings enable BERT to distinguish and process single texts or text pairs. Position embeddings are crucial in the Transformer architecture as they provide sequence information, allowing the model to understand the position of words in a sentence. Additionally, the CLS token, placed at the beginning of the input sequence, has its final hidden state used as the representation of the entire input sequence, effectively representing the sentence-level semantic state. This is highly effective for classifying user-level data [12].

For image information, a popular approach is to use convolutional neural network-based models for recognition, with YOLO series models commonly used for visual recognition and detection [13]. As the powerful capabilities of the Transformer architecture gain attention, visual tasks are also increasingly adopting attention mechanisms and similar methods [14]. An important model in the field of vision is ViT. ViT, like BERT, is a large model based on the Transformer and includes position embeddings and CLS embeddings [15]. The class token in the Transformer encoder output is treated as the encoded feature representation of the input image.

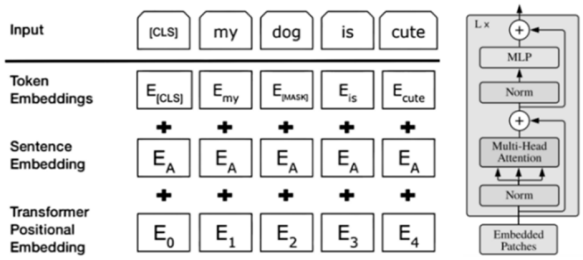

Figure 2. The structure of BERT

III. METHODOLOGY

This study constructs a multimodal dataset based on Taobao's dress data. Taobao is one of the largest online shopping websites in China. The dataset includes both structured data, such as product tags and images, and unstructured data, including brand, fabric, style, material composition, suitable age range, style, combination form, sleeve type, etc. Additionally, user data such as user ID, gender, purchased size and style, preferred color, years of Taobao registration, IP location, as well as unstructured data like user self-introduction and profile picture, were collected. The dataset contains a total of 210 valid entries. Finally, there are star ratings given by different users to the project. For the unstructured data, multi-class classification was performed using one-hot encoding, and some obviously erroneous options were deleted.

To address the current issues of cold start and sparsity in recommender systems, this study proposes a neural matrix factorization model based on large models and multimodal data, as illustrated in Figure 3. The BERT and ViT models were employed to process text and image information of users and items, respectively. To simplify processing, we selected the CLS layer, which represents the entire sentence or image, as the feature embedding. Subsequently, these two CLS layers were concatenated and input into an MLP for fusion. Although ViT

and BERT share embedding layers with the same embedding dimension, we used an MLP to compress the 768-dimensional layer further. Structured data were also concatenated and compressed using an MLP to 768 dimensions. The final embedded vectors were trained and scored through a neural network, ultimately forming the recommended items. It is important to note that since large models decompose user information into high-dimensional embeddings, forming a scoring matrix based on dot product is no longer suitable. Therefore, we directly use neural network models for fitting and prediction.

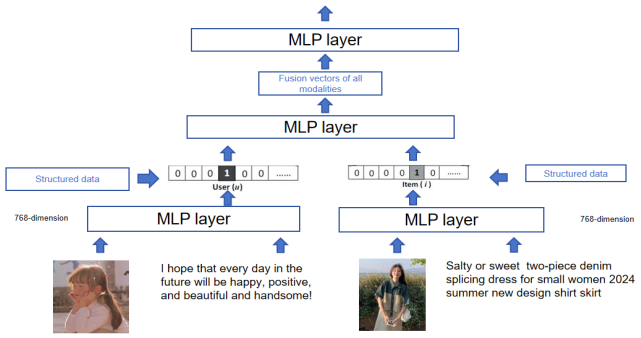

Figure 3. The structure of multimodal recommendation system

IV. Evaluation Criteria

In the paper, there are three methods commonly which are used to evaluate: mean square error, average accuracy, and precision@K, are applied in this paper. Mean squared error (MSE), which shown in Formula 2, is used to evaluate the regression performance of a model and evaluate the average squared difference between predicted and actual values. Such an evaluation method is one of the most widely used benchmarks for measuring overall forecast accuracy. n is the total number of samples, $\hat{r}_i$ is the predicted result of the i-th observation predicted by the model, and r_i is the actual value of the i-th observation. Precision@K is an indicator used to evaluate the performance of recommendation systems, measuring the proportion of the number of relevant items in the top K recommendation results to the total number of recommendations, which shown in Formula 3. U represents the set of users, R_u is the interaction between the user and the item, and $R_u \cap \widehat{R_{u,K}}$ is the results of recommended items that the user is actually interested in. Normalized Discounted Cumulative Gain (NDCG) is also the most common metric for evaluating the relevance and position of recommender system results. DCG@K is the discounted cumulative gain, which is used to calculate the cumulative gain of the first K recommended items for a given user u. IDCG@K is the discounted cumulative gain under ideal conditions, which is the maximum possible value of DCG@K. By weighting the positions, the highest-ranked correlations are given greater weight, and greater weight also means more important rankings. Among them, CG calculates the sum of correlations in recommended results regardless of the position of the results. it shown in Formula 4. DCG introduces the total correlation in the recommendation results of the position loss factor.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_i)^2 \tag{2}$$

$$\text{Precision@K} = \frac{1}{|U|}\sum_{u\in U}\frac{\left|R_u \cap \widehat{R_{u,K}}\right|}{K} \tag{3}$$

$$\text{NDCG@K} = \frac{1}{|U|}\sum_{u\in U}\frac{\text{DCG@K}}{\text{IDCG@K}} \tag{4}$$

V. Experiments and Discussions

In the experiments, all MLPs were configured with a two-layer structure: the first layer consisting of 128 neurons and the second layer consisting of 64 neurons. The third layer served as the output layer. The number of recommended items, KK, was set to 5 for all recommender systems, and all neural networks were trained for 10 epochs. We split the dataset into training and testing sets in a 7:2 ratio and employed cross-validation for training. At the same time, the study removed part of the modal information for comparison, removing images, text, and structural data respectively. To assess the effectiveness of the proposed model, we conducted multiple ablation experiments. These experiments were designed to test the importance of different modalities to the model's performance and the impact of incorporating LLMs on improving semantic vector representations. The experiment compared the traditional matrix factorization model, the neural matrix factorization model, and an advanced CNN-based [16] and DNN-based [17] multi-modal recommendation models.

The final training results of the models are shown in Table 1. The conventional matrix decomposition model as a baseline exhibits the highest MSE and significantly lower Precision@K and NDCG values, which indicates that the conventional approach is less efficient compared to the neural network approach in dealing with data complexity. Moreover, compared with the current advanced neural networks, it shows the superiority of our model based on the transformer structure. In the case where the matrix decomposition is performed without the use of a large model but using a neural network, the model shows higher MSE and lower Precision@K and NDCG compared to the proposed multimodal LLMs model, indicating that high-dimensional semantic representations based on large models are more capable of capturing the complexity of user-item interactions compared to traditional vector embeddings.

Table 1 Ablation experiment results

Models	MSE	Precision	NDCG
Muti-LLMs (mine)	0.37	0.94	0.89
NMD	0.95	0.75	0.71
SVD	6.75	0.75	5.67
Mine-no Figure	0.43	0.93	0.85
CNN-based model	0.42	0.92	0.85
DNN-based model	0.44	0.91	0.85
Mine-no Figure	0.43	0.93	0.85
Mine-no text-	0.54	0.84	0.83
Mine-no structured data	0.44	0.93	0.85

Compared with the complete multi-modal LLMs model, the model that excludes image data exhibits a slight increase in MSE

and a slight decrease in Precision@K and NDCG. This indicates that image data positively contributes to the overall recommendation performance, but its absence does not significantly diminish model performance. The MSE of the model that excludes text data increases significantly, and Precision@K and NDCG decrease significantly, indicating that text information plays a crucial role in capturing user preferences and item characteristics, and has a greater impact on model performance. When structured data is excluded from the large models, the performance is slightly lower than the full model but is comparable to the models that exclude image data. This indicates that, although structured data helps improve performance, its absence does not affect model performance as significantly as text data. Furthermore, some meaningful conclusions can be drawn based on the actual situation and the model's performance. For example, within the product image information, in addition to various dress styles, some images include attractive selling points as text on the image, such as emphasizing "free shipping" or "plus sizes available." Essentially, this information is textual (just printed on the image). Therefore, the semantic information from the images tested by the model may include some textual semantic information. The actual importance of purely image-based semantic information might be lower than the current experimental results suggest.

In future work, we plan to further optimize algorithms related to visual aspects, specifically enhancing ViT's semantic understanding of different regions to better comprehend the contribution of various modalities to the accuracy of the recommendation system from a deeper perspective [18]. On the other hand, we also plan to develop an end-to-end online retrieval system to match more recommendation needs and obtain more usable datasets for better training [19]. Additionally, it is crucial to address data privacy and security issues. This research will continue to explore technical measures to protect user data from being compromised by Trojan programs, leading to data breaches [20].

VI. CONCLUSIONS

This study proposes a recommendation system model that integrates multimodality, large language model (LLM) and neural matrix factorization, using Vit and BERT respectively to collect deep speech information of users and items' pictures and text, while combining unstructured data Information is evaluated, predicted and recommended through simple splicing methods and neural networks, showing good prediction performance. This paper conducts a variety of ablation experiments to prove that this method has higher performance than traditional methods. At the same time, the research results highlight the key role of text data in capturing user preferences and improving recommendation accuracy. At the same time, the research also found that in actual situations, visual data needs to be further segmented to clarify the contribution of visual data to multi-modal recommendation systems and the research will continue to further expand the scope of the dataset to further improve and validate multi-modal recommendation systems.

REFERENCES

[1] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: a survey," Decision Support Systems, vol. 74, pp. 12-32, 2015.

[2] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," Advances in Artificial Intelligence, vol. 2009, 2009.

[3] Z. Yao, J. Wang, and Y. Han, "An improved neighborhood-based recommendation algorithm optimization with clustering analysis and latent factor model," in 2019 Chinese Control Conference (CCC), IEEE, pp. 3744-3748, July 2019.

[4] L. Si and R. Jin, "Flexible mixture model for collaborative filtering," in Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 704-711, 2003.

[5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th International Conference on World Wide Web, pp. 285-295, April 2001.

[6] T. S. Kumari and K. Sagar, "A semantic approach to solve scalability, data sparsity and cold-start problems in movie recommendation systems," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 6s, pp. 825-837, 2023.

[7] Y. Luo, "Identifying factors influencing China junior high students' cognitive ability through educational data mining: Utilizing LASSO, random forest, and XGBoost," in Proceedings of the 4th International Conference on Modern Education and Information Management (ICMEIM 2023), Wuhan, China, pp. 202-207, September 2023.

[8] W. Dai, Y. Jiang, C. Mou, and C. Zhang, "An integrative paradigm for enhanced stroke prediction: Synergizing xgboost and xdeepfm algorithms," in Proceedings of the 2023 6th International Conference on Big Data Technologies, pp. 28-32, September 2023.

[9] Y. Luo, R. Zhang, F. Wang, and T. Wei, "Customer segment classification prediction in the Australian retail based on machine learning algorithms," in Proceedings of the 2023 4th International Conference on Machine Learning and Computer Application, pp. 498-503, October 2023.

[10] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. S. Chua, "Neural collaborative filtering," in Proceedings of the 26th International Conference on World Wide Web, pp. 173-182, April 2017.

[11] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[12] Y. Luo, Z. Ye, and R. Lyu, "Detecting student depression on Weibo based on various multimodal fusion methods," in Fourth International Conference on Signal Processing and Machine Learning (CONF-SPML 2024), vol. 13077, SPIE, pp. 202-207, April 2024.

[13] D. Ma, S. Li, B. Dang, H. Zang, and X. Dong, "Fostc3net: A lightweight YOLOv5 based on the network structure optimization," arXiv preprint arXiv:2403.13703, 2024.

[14] Z. Qi, D. Ma, J. Xu, A. Xiang, and H. Qu, "Improved YOLOv5 based on attention mechanism and FasterNet for foreign object detection on railway and airway tracks," arXiv preprint arXiv:2403.08499, 2024.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... & N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[16] Mu, Y., & Wu, Y. (2023). Multimodal movie recommendation system using deep learning. *Mathematics*, *11*(4), 895.

[17] Choudhury, S. S., Mohanty, S. N., & Jagadev, A. K. (2021). Multimodal trust-based recommender system with machine learning approaches for movie recommendation. *International Journal of Information Technology*, *13*, 475-482.

[18] W. Dai, C. Mou, J. Wu, and X. Ye, "Diabetic retinopathy detection with enhanced vision transformers: The Twins-PCPVT solution," in 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), pp. 403-407, May 2023.

[19] L. Sun, "A new perspective on cybersecurity protection: Research on DNS security detection based on threat intelligence and data statistical analysis," Computer Life, vol. 11, no. 3, pp. 35-39, 2023.

[20] C. Wang, L. Sun, J. Wei, and X. Mo, "A new trojan horse detection method based on negative selection algorithm," in Proceedings of 2012 IEEE International Conference on Oxide Materials for Electronic Engineering (OMEE), pp. 367-36