Enhanced Multimodal Recommendation System for Personalized Lifestyle Recommendations

*Mahima Kansal[1] and Sohit Agarwal[2]

[1,2]Department of Computer Engineering and Information Technology, SGVU, Jaipur, India
manjilagrawal@gmail.com

Abstract. The system learns to object similarities by contrasting loss based on multimodalities with data of similar and dissimilar objects under consideration for a recommendation. This research provides a multimodal recommendation with an advanced framework for personalized lifestyle recommendations that fabricates a model that incorporates numerous modalities such as images, text, and user interaction data. It includes components such as preprocessing, neural image feature extraction, and embedding of the text using Natural Language Processing models like BERT. The fulfillment and inclusion of an attention mechanism fuses these features into a more complete multimodal representation that yields better performance in terms of recommendation accuracy. Further optimization strategies like Adam and regularization techniques (L2, L1) are used to maintain stability and avoid overfitting of the model. This study presents a hybrid recommendation approach that has incorporated Content-Based Filtering with the Collaborative Filtering method for a relevant recommendation process. Contrastive learning ensures that it captures subtle differences between similar items, resulting in a more varied and accurate performance of recommendations. System performance is evaluated in several scenario settings to depict an improvement in user-friendliness, recommendation accuracy, and diversity over traditional techniques. It summarizes such a way with a cognition concerning personalized digital experiences into lifestyle recommendation by pushing forward the application around multimodal fusion and contrastive learning. Such applications transcend to fashion, health, entertainment, and other potential areas.

Keywords: Hybrid Recommendation, Personalized, Comparative Study.

1 Introduction

Investment in the technological growth of recommendation systems has helped push various domains like e-commerce, healthcare, education, and entertainment design recommendations to personalize the user experience. Conventional recommendation systems are usually based on unidimensional data sources such as user-item interaction or explicit feedback [7], which often fail to capture the complexity and multi-faceted nature of user preferences and item attributes. It is these limitations that have led to the birth of a new working paradigm of multimodal recommendation systems. Such sys-

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_5

tems develop comprehensive representations of users by using different data modalities, such as text, images, audio, or contextual sensor data [13]. This integrated approach gives rise to improved accuracy of recommendations, engaging users with a richer sense of context and helping to achieve SDG [11].

The recent progress in deep learning [6], [12], [18] has paved the way forward for multimodal recommendation systems. Techniques like CNNs [20] for visual data analysis, RNNs for temporal and sequential data, and transformers for contextual and multimodal fusion have emerged as key tools in this arena. Other major approaches are graph-based ones such as knowledge graphs [23] and graph neural networks (GNNs) [26], which are quite compelling for capturing the complex relational structure of the user-item ecosystem. Reinforcement learning (RL) provides adaptability by optimizing recommendations in real-time, hence, involving user feedback, whereas context-aware approaches add spatiotemporal effects to make the recommendations more aligned with user intent.

This paper considers a review of seven major approaches to multimodal recommendation systems, including knowledge graph learning, semantic video analysis, audio-visual fusion, contrastive learning, multimodal embeddings, reinforcement learning, and transformer-based approaches. The paper details multimodal recommendation projects in virtually every field by weaving in the theoretical contexts alongside their applications, further underlining those that would radically transform our personalized digital experiences.

2 Distinct Research in Multimodal Recommendation Systems

Multimodal recommendation systems combine the input of text, images, audio, and user behavior to give contextualized, personalized suggestions [25]. They employ cutting-edge techniques including transformers, contrastive learning [15], and hybrid models to harness multimodal data. Ranging from fashion [22] to e-commerce [2], healthcare [8] to travel [14], they combine modalities to provide a more accurate and diverse solution. Knowledge graphs, personalization, etc., add depth to the recommendations proposed by these systems, while further avenues such as reinforcement learning and emotion recognition spur innovations for sound user-oriented recommendations across domains.

2.1 Knowledge Graphs in Recommender Systems

Knowledge graphs (KGs) are important for enhancing recommender systems because they address sparse data and cold start problems and enrich the user's experience in a holistic manner. This section is about the noteworthy contributions that have been made in the realm of knowledge graph-based recommender systems.

Graph Neural Networks (GNNs) was reviewed, underscoring their ability to highlight relational information in the graph data [26]. This review describes different kinds of GNN architectures and pinpoints open research problems, shedding light on the ability of GNNs to potentially change recommendation systems by capturing complex user-

item interactions. KGs could reduce issues of data sparsity and cold starts [19]. A systematic review of various algorithms that unite KGs to achieve more accurate explainable recommendations is presented, including research directions to enrich KG integration with collaborative filtering.

2.2 Semantic Video Recommendation Systems

With the casting of video content growth in an exponential rate, efficient video bundling is becoming essential. The following section tackles the advancements made in semantic video recommender systems that channel user impressions and multimodal data to enhance user experiences.

The problem arising from platforms like YouTube creating a legion of videos has been remedied by internal features of video analysis that consider user preferences [4]. A more sensitive video recommender has been conceived, suggesting sewn videos should be based on audience sentiments and what will be the best to them.

In e-commerce visual recommendation, Deep Visual Ensemble Equality Metric has been considered as an effective technique [10]. This backward linkage increases accuracy due to the presence of close visuals between the products that enhance user satisfaction in content-based recommendations.

2.3 Audio-Visual Recommendations for Streaming Services

With the surge in audio-visual content consumption via streaming services, the demand for more efficient and personalized recommendations surged significantly. An audio-visual recommendation system uses the interdependence of an audio-visual system for better relevance of suggested content and user experience. This section elaborates upon the recent advances in audio-visual recommendation addressing user needs.

An all-encompassing review on Audio-Visual Learning discusses the deeply complex difference in multimedia between audio and visual methods [17]. Audio-visual separation, correspondence learning, generation and representation-with in-between challenges -has been detected. These benchmarks are broadly known in mining the area and translating the geometrical space of future avenues of research.

A new solution has been devised for generating live commentaries for streaming services, connecting real-time audio-visual multimodal cues in improving audience engagement [16]. A Twitch-acquired dataset with a multimodal generation model showed that interactions with live videos can be significantly enhanced through the joint real-time contributions of both audio and visual signals.

2.4 Multimodal Recommendation with Contrastive Learning

As multimodal data is root of advanced recommendation systems, contrastive learning is being recognized now as a powerful improvement in model accuracy These resolve some of the issues posed by noise, data sparsity, or varying quality of input data that can damage performance using effectively integrated diverse modalities at the recommendation output.

A transformer-based fusion encoder with attention relying on a meditation hurdle is presented for an effective combination of multimodal representations leading to better expression [9]. This also incorporates global statistical features integrated by way of adverse learning approaches towards an improvement in multimodal emotion analysis and recommendation experiences.

The introduction of a MulticsR Framework combines multi-verse contrast learning and social recommendations to approach overall user preferences [5]. The above model integratesuser social networks and item knowledge graphs to derive better user preference modeling and presents its extended performance on the real-world dataset paving the way for more individual interactions.

2.5 Personalization with Multimodal User Embeddings

The two-fold approach of personalizing based on multimodal user embeddings has become rather prominent in recent research, as they could provide for more rich representation of users, which describe various facets of interest for advice across various contexts and platforms. They exemplify the importance of multiple modalities in the integration process, which ultimately enriches user modeling and improves recommendations in tune with users' personalized preferences.

PinnerSage, a multi-modal user preference representation has revolutionized how users interact with the materials by briefly clustering items into coherent groups and representing objects [21]. Compared to traditional single embedding techniques, this is an idea that suitably improves the user's recommendations greatly concerning both domain and novelty, adding a powerful quality for user engagement.

MMIR is an individual multimodal reinforcement learning model, which adapts to studies of changing user preferences [24]. This accounts for previous user comparison and real-time interests in recommendations, thereby resulting in improved recommendation accuracy during multi-turn interactions, permitting increased dynamism and responsiveness in the recommendations process.

2.6 Reinforcement Learning for Multimodal Interactive Recommendations

Reinforcement learning (RL) is one of the most promising advancements in a recommendation system, especially in the growth of multimodal interactive agents to utilize human feedback efficiently. A couple of important contributions, looking at how reinforcement learning can enhance the user interaction and recommendations, are reviewed in this section.

In a 3D setting, experimentally allowed has been documented to enhance fake embodied agents where evaluation of viewing behavior by humans is based on inherently human-like behavior [1]. A new reward model, termed Inter-Temporal Bradley-Terry (IBT), has been developed to model minute human choices relating to agent performance. In many cases, agents were able to outperform the IBT reward model without any reward model, which acts as a mechanism for refining agent behavior, confirming reinforcement through human feedback mechanisms.

The challenges faced in preference-based reinforcement learning, particularly problems of modeling human preferences accurately, have been highlighted [27]. Traditional methods are based on Markovian beliefs, which often fail in grasping the temporary interdependence in anticipatory actions of robots. To achieve a comprehensive stance of joint activity among realms that impact human assessment outcomes, the proposed approach employs multimodal transformer to encode state and action variables. This approach aims to tackle the aforementioned complications to provide a better alignment of robots' functions with users' choices, pursuing interactive recommendation systems..

2.7 Multimodal Shopping Recommendations through Transformer Networks

Integration of the transformer network has been the recent focus in the multimodal shopping recommendation systems, enhancing the overall classification and recommendation phases. This section reviews key work done using transformer architecture to enhance fashion recommendations.

A micro-video recommendation model known as CMI-a user engagement in micro-videos that takes care of interest coverage-short for coverage of the model of interest [17]. With an increasing number of micro-video creators and audiences, the issue of capturing user interest diversity is rising. The CMI extracts multi-interest embedding by learning opposite from historical interaction sequences. Experiments on two micro-video datasets have shown that the CMI outperformed existing baseline models, establishing the effectiveness of opposite learning in this domain.

The improvement in the outfit classification and recommendation has been found using the vision transformer (VIT) [3]. With the increasing trend of purchasing clothes online, it aims to steer users to make informed options. This architecture comes with various transformer blocks consisting of self-embedding layers and multilayer perceptrons, and many fashion datasets have a large-scale performance comparison against deep CNN and other models. The model is seen as a strong performer in the task using benchmark fashion product datasets and has been incorporated into a fashion recommendation system using Dinov2, allowing ability through similarity discovery of fashion products with nearest neighbor, demonstrating the power of transformer models in fashion recommendations.

3 Methodology

The methodology used for multimodal recommendation involves several sophisticated mechanisms designed to streamline the information exchanges between users and items. The system fuses various sources of data (e.g., text, images, etc.) in an information retrieval process that ranges from learning with embedding to attention mechanisms and fusion strategies that utilize similarity measures between the descriptions and item features. The main steps include preprocessing the data, embedding representation, similarity computation, contrastive learning, and fusion into multimodal representations.

3.1 Data Preprocessing

Data preprocessing of the raw data includes re-uploading the same data onto the platform to clean and prepare it for further processing steps in the right format. It is applied for packing processes where the standardization and cleaning maintain quality and consistency of inputs across each modality.

Textual Data: Tokenization and vectorization with methods such as TF-IDF, word embeddings (e.g., Word2Vec, GloVe), and transformer-based models (e.g., BERT) are applied.

Image Data: Images are resized and normalized while being passed through convolutional neural networks or pre-trained models (like ResNet or VGG) to derive high-level features.

Multimodal Data: In instances where both text and image data coexist, features align and combine in specifically defined ways to ensure the model can co-process them.

3.2 Multimodal Embedding Learning

Each modality is processed by modality-specific encoders with the goal of generating embeddings.

$$Z_{modality} = Encoder_{modality}(x) \tag{1}$$

where x is the raw input depending on data type (text or image) and $z_{modality}$ is the learned representation (embedding) for that specific modality; these embeddings allow the model to recall salient features for each item being fused.

3.3 Similarity Measurement in the Embedding Space

The distance or similarity between each pair of corresponding items from different modalities is articulated with a similarity measure being denoted by the cosine similarity function that describes how close the embeddings are.

$$sim(z_i, z_j) = \frac{z_i \cdot z_j}{||z_i|| ||z_j||} \tag{2}$$

Where z_i and z_j are the respective embeddings of the two corresponding items and the numerator, dot product of the two respective embeddings, is normalized in the denominator. The general cosine similarity is also generalized in the multimodal space such as, for instance, the similarity between an image and its description.

3.4 Contrastive Loss Function

To pull items that are of similar types closer in the embedding space while pushing dissimilar ones away from each other, a contrastive loss is designed; it drives the model to distinguish between the contrasting pairs of items. The contrastive loss can thus be written as:

$$L_{contrastive} = -log(\frac{exp(sim(z_i,z_j)/\tau)}{\Sigma_{k=1}^{N} exp(sim(z_i,z_k)/\tau)}) \quad (3)$$

Where z_i and z_j are the embeddings of two similar items. sim(z_i, z_j) is the cosine similarity between these embeddings. τ is a hyperparameter controlling the sharpness of the similarity distribution. This loss function will be modified for adaptive temperature scaling to dynamically optimize τ during training, as will be discussed in the enhancement section.

3.5 Multimodal Fusion

After embeddings for individual modalities are created, the representations are fused together to form a unified representation for the given item. The fusion method varies according to the architecture, and general methods are:

Concatenation: The embeddings derived from each modality are concatenated to create an extended vector.

$$z_{fused} = Concatenate(z_{text}, z_{image}) \quad (4)$$

Attention Mechanism: An attention mechanism is used in weighting the importance of every modality dynamically. This helps allow the system to concentrate on the pertinent features in decision-making.

$$\alpha_{text} = Attention(z_{text}) \quad (5)$$

$$\alpha_{image} = Attention(z_{image}) \quad (6)$$

$$z_{fused} = \alpha_{text} * z_{Text} + \alpha_{image} * z_{image} \quad (7)$$

Multiplicative Fusion: A multiplicative interaction of the embeddings is used to fuse modalities.

$$z_{fused} = z_{Text} \odot z_{image} \quad (8)$$

Where $\odot$ denotes elementwise multiplication. These fused embeddings are fed into the recommenders, allowing the model to use all information across modalities.

4 Enhancement and Optimization

The advanced techniques discussed here will enhance the multimodal recommendation system's performance by modifying the core methodologies and incorporating optimization strategies.

4.1 Adaptive Contrastive Loss with Dynamic Temperature Scaling

Previously, we discussed that a contrastive loss function plays a key role in the shape of the embedding space. To enable this process to be more versatile, we propose an adaptive temperature scaling mechanism. The temperature parameter determines the

sharpness of the similarity distribution and greatly influences the quality of embeddings. Rather than adopting a fixed temperature, we allow τ to evolve during the training process in order to better performance with adaptive contrastive loss given by:

$$L_{adaptive} = -log(\frac{exp(sim(z_i,z_j)/\tau_{adaptive})}{\Sigma_{k=1}^{N} exp(sim(z_i,z_k)/\tau_{adaptive})}) \quad (9)$$

The term $\tau_{adaptive}$ is a learnable one that gets adapted according to the training dynamics. This change guarantees that the model would dynamically adjust its sensitivity concerning similarities among items based on the context of relevance.

4.2 Attention-based Multimodal Fusion

By introducing attention mechanisms into the fusion process, one can see how an approach arises that allows the model to automatically learn where to focus on each modality, being more helpful in cases where certain modalities, such as text or image, may play a much bigger role for certain items.

For example, the model learns an attention weight for each modality:

$$\alpha_{modality} = Attention(z_{modality}) \quad (10)$$

In this way, multimodal features can be fused as:

$$z_{fused} = \sum_{modality} \alpha_{modality} \cdot z_{modality} \quad (11)$$

This allows the system to tune contributions of each modality dependent on the context, which is a significant increase in accuracy for the recommendations.

4.3 Regularization Techniques

Regularization techniques improve generalization of the model by controlling the model complexity at training and the model overfitting. We present key regularization techniques; each has its mathematical formulation to guarantee better training.

Dropout regularization: Dropout is a common regularization method used by neural networks that prevent overfitting by randomly setting a fraction of input units to zero during training. The dropout layer can be mathematically described as:

$$h^{drop} = h \cdot mask \quad (12)$$

Where: h is the layer output. mask is a random binary vector (0 or 1) generated before each training instance. Any element of the mask will be set to 0 with a probability of 1-p, otherwise it is set to 1 with probability p-the keep probability. During training, some of the neurons are "dropped" randomly to lessen the dependence on the neurons and make it learn enough general features.

L2 Regularization (Weight Decay): L2 regularization or weight decay penalizes large weights by adding a penalty term to the loss function, which forces a model to use smaller weights, thus helping avoid overfitting. The L2 penalty is described as follows:

$$L_{L2} = \lambda \Sigma_i \|W^i\|^2 \quad (13)$$

Where: Wi are the weights in the i layer of the network. λ is the regularization coefficient controlling the force of the penalty. The total loss function is updated to:

$$L_{total} = L_{loss} + L_{L2} \tag{14}$$

Where L_{loss} is the original loss (e.g., cross-entropy loss or contrastive loss). Example: For instance, the original and augmented datasets were generally similar, and yet after augmentation, the training would no more be done on the original training data.

L1 Regularization (Lasso Regularization): L1 regularization or Lasso regularization encourages sparsity, that is, many weights in the model will end up being zero, thus effectively forcing the removal of certain features. The L1 penalty is as follows:

$$L_{L1} = \lambda \Sigma_i \|W^i\|_1 \textit{ (25)} \tag{15}$$

Where: ‖Wi‖1 is L1 norm of weight vector Wi. λ is the regularization strength. The total loss function becomes:

$$L_{total} = L_{loss} + L_{L1} \textit{ (36)} \tag{16}$$

4.4 Data Augmentation

Data augmentation is the process of synthetically expanding the dataset to improve the robustness and generalization of multimodal recommendation systems . On the other end, for text data, synonym replacement, back-translation, and random insertion are among other augmentation strategies that can be adopted. Audio data can be augmented through noise injection, time-stretching, and pitch shifting. In mathematical terms, given an input whose value is x from a distribution D, augmentation creates a transformed version x' such that $x' = f_{image}(x)$, where $f_{augment}$ indicates augmentation operations preserving semantic meaning. For instance:

$$x'_{image} = f_{image}(x) \textit{ (47} \tag{17}$$

The goal of augmentations in contrastive learning is to create positive pairs (z_i, z_j) from the same input x, maximally close to each other:

$$\mathcal{L}_{contrastive} = \sum log\left(\frac{exp(sim(z_i,z_j)/\tau)}{\Sigma_{k=1}^{N} exp(sim(z_i,z_k)/\tau)}\right) \textit{ (58)} \tag{18}$$

where sim(.) shows similarity function, and τ is a temperature parameter.

5 Distinct Research in Multimodal Recommendation Systems

About the above techniques by fashion products, this includes adaptive contrastive loss, attention-based multimodal fusion, regularization strategies, and data augmentation each of which will contribute too.

5.1 Adaptive Contrastive Loss with Dynamic Temperature Scaling

Adaptive Contrastive Loss with dynamic temperature scaling enables the temperature parameter to be fine-tuned during training to further improve the distinction between

similar and dissimilar items. Temperature scaling allows the model to determine how much contrast should be exercised on product embeddings by considering their characteristics. Fig. 1 shows the effect of adaptive contrastive loss with dynamic temperature scaling.

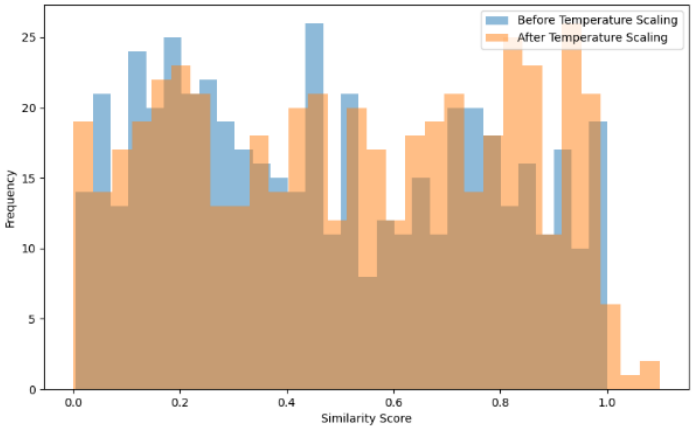

Fig. 1. Effect of Adaptive Contrastive Loss with Dynamic Temperature Scaling

Eventually, this ensures a closer embedding for similar items, such as two dresses of similar colors, while dissimilar items, such as a red dress and a blue jacket, are further apart in the embedding space. One of the characteristics that ensure better fine-tuning of model embeddings is temperature adjustment, to maximize product similarity matching and thus improve recommendation accuracy. The two histograms show how the similarity scores behave before and after temperature scaling: while the similarity distributions show a broad exhibition of the products, this second histogram indicates a tighter clustering wherein certain products exhibit little to no overlaps with others. This demonstrates how dynamic temperature scaling improves the model's ability to differentiate between similar and dissimilar products for better performance on recommendation tasks.

5.2 Attention-Based Multimodal Fusion

In attention-based multimodal fusion, the attention mechanism employs the product context to assign varying levels of importance to different modalities (image and text). For fashion items such as dresses, for instance, where visual appearance plays a major role, the attention mechanism will prioritize image features. Conversely, for accessories such as shoes, which may have more descriptive textual information, the model will assign higher attention weights to the text. This dynamic adjustment allows the model to effectively fuse the image and text features as a single, context-aware representation, optimally suited for making accurate recommendations. Attention-based multimodal fusion is shown in Fig. 2.

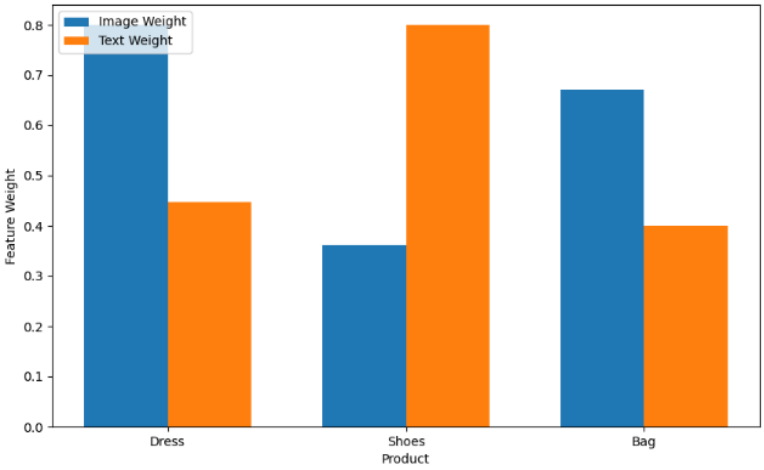

Fig. 2. Attention-Based Multimodal Fusion

The visualization shows a bar chart of attention weights assigned to image and text features for various products. The chart brings to light how the attention mechanism shifts-in the case of dresses toward an emphasis on image, whereas shoes rely more so on text weighting. This shows the extent of how the model adapts focus according to product types in the interest of optimizing the process of recommendation by way of a fitting blend of image and textual features.

5.3 Regularization Techniques

Regularization techniques such as dropout, L2 regularization, and weight decay are indispensable for controlling overfitting of the recommendation model so that it effectively generalizes on completely unseen sets. The dropout refers to randomly deactivating a few neurons while the model is training. This lessens the dependency of the model on some specific neurons and promotes use of other sport neurons, thus forcing the model to use a wider set of features. L2 regularization penalizes the large weights added to the loss function and thus tends to impart the learning process with a simpler and more general structure throughout. Similar to L2, weight decay attempts to avoid overfitting into noise in data by penalizing large weights. These techniques help the model to avoid memorizing the training data to allow itself to operate within a wider variety, thereby learning the key patterns applicable to unseen data.

The visualization presents two performance curves comparing training loss over epochs; the first curve is produced on a model trained without regularization, and the second one is a model trained with regularization. A curve that will likely decrease at a slower pace in terms of training loss and present a much lower validation loss becomes more stable. This shows good generalization to unseen data. Regularization thus prevents overfitting and just increases the robustness of the model. The effect of Regularization Techniques is shown in Fig. 3, while Fig. 4 shows the Effect of Data augmentation on performance.

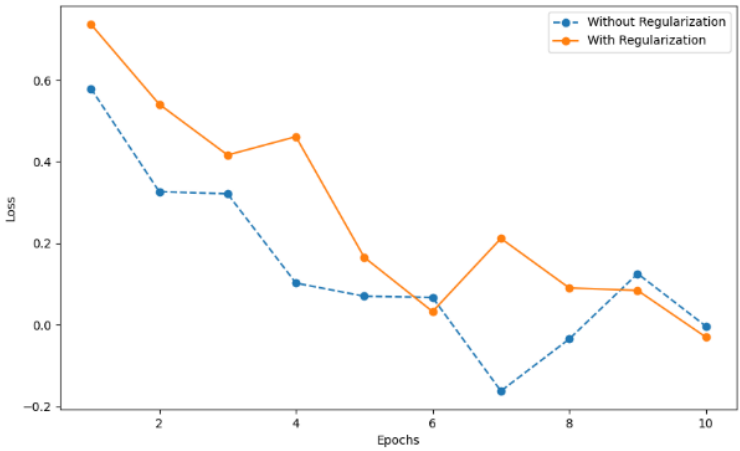

Fig. 3. Effect of Regularization Techniques

5.4 Data augmentation

Data augmentation is performed to synthetically boost the training dataset so as to make the model robust by exposing it to different representations of the products. For images, data augmentation methods, such as, but not limited to, random rotations, flipping, and cropping, prevent the models from overfitting to specific angles or poses. For text, paraphrasing product descriptions makes the model learn various text representations, allowing it some way of handling different phrases or wording. Data augmentation enhances the model generalization ability in particular cases where there are a new or unseen category, and the augmentation of the data introduces more diversity.

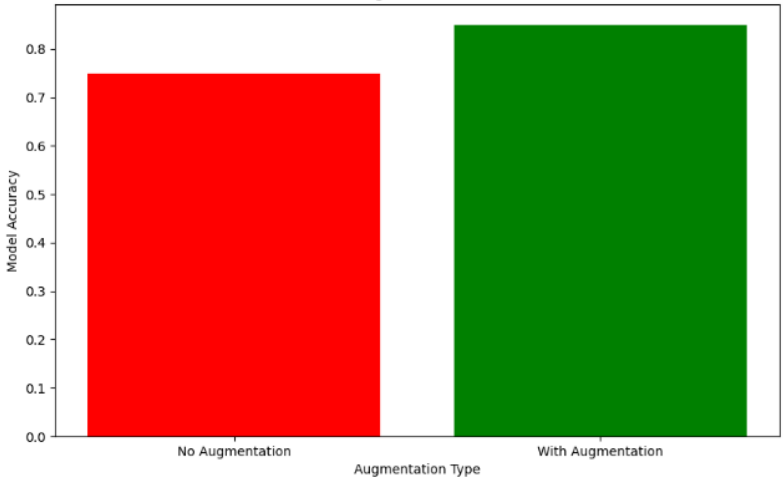

Fig. 4. Effect of Data augmentation on performance

The visualization presents a bar chart depicting the performance of the model (accuracy) before data augmentation was applied and after. The chart would generally show better accuracy of performance due to the ability of the model to handle a wider range of some product variations stemming from the increase in the diversity of the dataset. This shows that augmentation will increase the diversity and quantity of the training data, which enhances the capacity of the model to generalize to new or uncommon products.

5.5 Weight Distribution for Different Techniques

The analysis of weight distributions for different techniques such as contrastive loss, attention fusion, and regularization quantifies each method's impact on recommendation results. By monitoring the weights given for various features/techniques during training, it will become evident how each affects the model in generating accurate recommendations. Contrastive loss usually has higher weight for products that are similar to one another. Attention fusion weighs appropriate image or text features differently based on their relevance. Regularization techniques will produce a more balanced weight distribution across features, leading to a more balanced system and a more generalized model. Fig. 5 represents the weight distribution for various techniques.

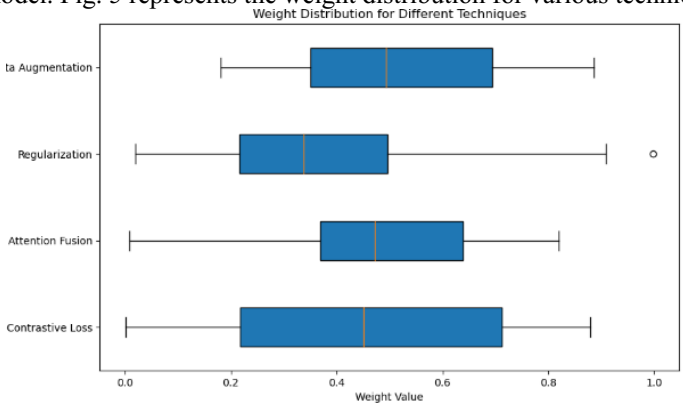

Fig. 5. Weight Distribution for Different Techniques

The box plot will show the weight distribution for each technique, with the regression value on the x-axis and the techniques (contrastive loss, attention fusion, regularization, etc.) on the y-axis. The plot indicates the degree to which weights are concentrated for each technique and to what extent any one method significantly outweighs another in its selection of features or penalty terms in the model. For some cases, contrastive loss may show that weights for similar items are in a tight range while weights of attention fusion may showcase much diversity based on item type. This delineation is thus explicit about the power or ineffectiveness of different approaches in guiding the model's decisions and hence the final recommendations.

6 Conclusion

The results not only show that including image and text data improves the performance of the multimodal recommendation system tremendously, but also positively impact the model in terms of accurate recommendations of fashion items by augmenting both image and text input. Combination of image and text augmentation during contrastive loss leads the model towards better recommendations in identifying relevant fashion items. These techniques enhance precision and recall and contribute to building a robust system capable of handling diverse and complex input data, ultimately offering users more personalized recommendations.

References

1. Abramson, J., Ahuja, A., Carnevale, F., Georgiev, P., Goldin, A., Hung, A., Landon, J., Lhotka, J., Lillicrap, T., Muldal, A., Powell, G., Santoro, A., Scully, G., Srivastava, S., Glehn, T., Wayne, G., Wong, N., Yan, C., Zhu, R.: Improving Multimodal Interactive Agents with Reinforcement Learning from Human Feedback. https://doi.org/10.48550/arXiv.2211.11602 (2022).
2. Agarwal, S., Kansal, M.: Empowering Small and Medium Enterprises (SMEs) through Multimodal Recommendation Systems using AI. International Journal of Computer Applications 186(16), 48–50. https://doi.org/10.5120/ijca2024923551 (2024).
3. Alaziz, H., Elmannai, H., Saleh, H., Hadjouni, M., Anter, A., Koura, A., Kayed, M.: Enhancing Fashion Classification with Vision Transformer (ViT) and Developing Recommendation Fashion Systems Using DINOVA2. Electronics 12(21), 4263. https://doi.org/10.3390/electronics12204263 (2023).
4. Amali, D., Barakbah, A., Besari, A.R.A., Agata, D.: Semantic Video Recommendation System Based on Video Viewers Impression from Emotion Detection. In: Proceedings of the 2018 IEEE International Conference on Knowledge Creation and Intelligent Computing (KCIC), pp. 176–183. https://doi.org/10.1109/KCIC.2018.8628592 (2018).
5. Chen, R., Chen, J., Gan, X.: Multi-view Graph Contrastive Learning for Social Recommendation. Scientific Reports 14. https://doi.org/10.1038/s41598-024-73336-5 (2024).
6. Chen, Y.C., Lee, W.C.: A Novel Cross-Domain Recommendation with Evolution Learning. ACM Transactions on Internet Technology 24. https://doi.org/10.1145/3639567 (2024).
7. Fang, H., Liang, J., Sha, L.: Enhanced Multimodal Recommendation Systems through Reviews Integration. Research Square. https://doi.org/10.21203/rs.3.rs-4333408/v1 (2024).
8. Garg, M., Agrawal, S., Goyal, D.: Theoretical Study on Big Data with Data Mining. In: Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT). https://doi.org/10.2139/ssrn.3174732 (2018).
9. Guo, Q.: Multi-Modal Representation via Contrastive Learning with Attention Bottleneck Fusion and Attentive Statistics Features. Entropy 25. https://doi.org/10.3390/e25101421 (2023).
10. Hiriyannaiah, S., G.M.S., Srinivasa, K.: Deep Visual Ensemble Similarity (DVESM) Approach for Visually Aware Recommendation and Search in Smart Community. Journal of King Saud University - Computer and Information Sciences 34. https://doi.org/10.1016/j.jksuci.2020.03.009 (2020).
11. Jain, T., Kansal, M.: Impact of Artificial Intelligence on Sustainable Development Goals. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4376842 (2023).

12. Jeong, S.Y., Kim, Y.K.: Deep Learning-Based Context-Aware Recommender System Considering Change in Preference. Electronics 12(21), 2337. https://doi.org/10.3390/electronics12102337 (2023).
13. Kamani, R., Kumar, V., Kagita, V.: Cross-domain Recommender Systems via Multimodal Domain Adaptation. arXiv preprint arXiv:2306.13887 (2023).
14. Kanimozhi, U., Sannasi, G., Manjula, D., Kannan, A.: A User Preference Tree based Personalized Route Recommendation System for Constraint Tourism and Travel. Research Square. https://doi.org/10.21203/rs.3.rs-755856/v1 (2021).
15. Kulkarni, P., Rai, S., Kale, R.: Recommender System in eLearning: A Survey. https://doi.org/10.1007/978-981-15-0790-8_13 (2020).
16. Lalanne, J., Bournet, R., Yu, Y.: LiveChat: Video Comment Generation from Audio-Visual Multimodal Contexts. ResearchGate. https://doi.org/10.13140/RG.2.2.18072.16641 (2023).
17. Li, B., Jin, B., Song, J., Yu, Y., Zheng, Y., Zhuo, W.: Improving Micro-video Recommendation via Contrastive Multiple Interests. arXiv preprint arXiv:2205.09593 (2022).
18. Li, C., Ishak, I., Ibrahim, H., Zolkepli, M., Sidi, F., Li, C.: Deep Learning-Based Recommendation System: Systematic Review and Classification. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3323353 (2023).
19. Liu, J., Duan, L.: A Survey on Knowledge Graph-Based Recommender Systems. In: 2021 IEEE International Conference on Advanced Electrical and Computer Engineering (IAEAC), pp. 2450–2453. https://doi.org/10.1109/IAEAC50856.2021.9390863 (2021).
20. Lou, N., Wang, H., Chao, Z.: A Personalized Movie Recommendation System Based on LSTMCNN. In: 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 485–490. IEEE (2020).
21. Pal, A., Eksombatchai, C., Zhou, Y., Zhao, B., Rosenberg, C., Leskovec, J.: PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In: Proceedings of the 2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2311–2320. https://doi.org/10.1145/3394486.3403280 (2020).
22. Shirkhani, S., Mokayed, H., Saini, R., Hum, Y.: Study of AI-Driven Fashion Recommender Systems. SN Computer Science 4. https://doi.org/10.1007/s42979-023-01932-9 (2023).
23. Sun, R., Cao, X., Zhao, Y., Wan, J., Zhou, K., Zhang, F., Wang, Z., Zheng, K.: Multi-modal Knowledge Graphs for Recommender Systems. In: Proceedings of the 2020 ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1405–1414. https://doi.org/10.1145/3340531.3411947 (2020).
24. Wu, Y., Macdonald, C., Ounis, I.: Personalised Multi-Modal Interactive Recommendation with Hierarchical State Representations. ACM Transactions on Recommender Systems 2. https://doi.org/10.1145/3651169 (2024).
25. Xia, H., Luo, Y., Liu, Y.: Attention Neural Collaboration Filtering Based on GRU for Recommender Systems. Complex & Intelligent Systems 7(3), 1367–1379. https://doi.org/10.1007/s40747-021-00274-4 (2021).
26. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph Neural Networks: A Review of Methods and Applications. arXiv preprint arXiv:1812.08434. https://doi.org/10.48550/arXiv.1812.08434 (2021).
27. Zhao, Y., Aguilar Escamill, J.E., Lu, W., Wang, H.: RA-PbRL: Provably Efficient Risk-Aware Preference-Based Reinforcement Learning. arXiv preprint arXiv:2410.23569. https://doi.org/10.48550/arXiv.2410.23569 (2025).