Benchmarking Multi-Modal Recommendation Models

Ekansh Singh
2020EE10490
ee1200490@iitd.ac.in

Manshi Sagar
2020CS50429
cs5200429@iitd.ac.in

Richa Yadav
2020CS50438
cs5200438.iitd.ac.in

***Abstract*—In today's information-rich digital landscape, personalized recommendations are vital for user engagement and satisfaction. Multimodal recommendation models, which amalgamate various data types such as text, images, and audio, offer a promising avenue to achieve this. This report provides a comprehensive exploration of multimodal recommendation models, delving into their underlying technologies, benchmarking their performance, and shedding light on their potential applications across diverse domains. By offering insights into the efficacy of these models, we aim to contribute to the ongoing evolution of recommendation systems in the digital age.**

***Index Terms*—Multi-Modal Recommendation System, Graph Neural Networks, Collaborative Filtering**

I. Introduction

In today's digital era, the sheer volume of information and media content available to consumers is overwhelming. From movies and music to books and products, the choices seem endless. This abundance of options presents both an opportunity and a challenge. The opportunity lies in tailoring recommendations to individual preferences, enhancing user experience, and boosting engagement. The challenge, on the other hand, is to cut through the noise and deliver relevant content that resonates with users.

To address this challenge, multimodal recommendation models have emerged as a promising solution. These models leverage a diverse array of data types, combining textual descriptions, images, audio, and other multimedia elements to provide recommendations that are not only personalized but also more contextually relevant. Multimodal recommendation models are at the intersection of various cutting-edge technologies, including natural language processing (NLP), computer vision, and audio analysis. This interdisciplinary approach allows them to capture a more comprehensive understanding of users' preferences, thereby improving recommendation accuracy and user satisfaction.

The core idea behind multimodal recommendation models is to integrate different modalities or types of data to provide a holistic understanding of a user's preferences and needs. For example, consider a scenario in which a user is looking for a new book to read. In a traditional recommendation system, only textual descriptions and user behavior data may be considered. However, a multimodal approach would also take into account the book's cover image, audio excerpts, and perhaps even user-generated reviews. This richer data enables the recommendation system to account for nuances that are often missed in unimodal models.

In this report, we delve into the fascinating world of multimodal recommendation models. We explore the underlying technologies, the benefits they offer, and the challenges they face. Our primary objective is to benchmark the performance of various state-of-the-art multimodal recommendation models on our dataset, providing insights into their effectiveness in real-world scenarios.

II. Our Work

Our work focuses on Benchmarking state of the art models on our Enhanced Movielens dataset. First we selected and modified our dataset by scrapping movie posters from the web to make it more relevant followed by that we did a thorough analysis of the six state of the art models that we have benchmarked. Then we used an open source toolbox to compare these models on common metrics (recall, ndcg, precision, mean average precision). At the end we did a comprehensive comparative analysis leading to some conclusions with respect to the time complexity and accuracy. Here is the link to our repository

III. Dataset

The dataset that we have selected is an enhanced version of the Movielens 100k dataset. It contains 100,000 ratings (1-5) from 943 users on 1682 items(movies). It consists of text, audio, video and meta data which has the movie summaries, director, movie cast, etc. where each user has rated atleast 20 movies. The dataset was short of relevant images for movies hence we designed a web scrapper to scrap movie posters from the web.

IV. Sparsity

Here is a comparison for the sparsity of our dataset with respect to the datasets previously used in these models

TABLE I: Sparsity of Our Dataset

Datasets	Users	Items	Interaction	Sparsity
Baby	19445	7050	160792	99.88%
Sports	35598	18357	296337	99.95%
Electronics	192403	63001	1689188	99.99%
Movielens	943	1682	100000	93.69%

V. MMRec ToolBox

MMRec is an open-source toolbox for multimodal recommendation. MMRec simplifies and canonicalizes the process of implementing and comparing multimodal recommendation models. The objective of MMRec is to provide a unified and configurable arena that can minimize the effort in implementing and testing multimodal recommendation models. It enables multimodal models, ranging from traditional matrix factorization to modern graph-based algorithms, capable of fusing information from multiple modalities simultaneously. MMRec supports four popular modalities: Text, Image, Audio, Video.

Compared with the conventional recommender systems that solely leverage user-item interactions for recommendation, multimodal models involves the processes of preprocessing information from multiple modalities. The wide variety of preprocessing methods make the models difficult to reproduce its performance and compare fairly with others. MMRec helps in minimize the effort in implementing and testing multimodal recommendation models and thus, simplifies the research in this field. To ensure model reproducibility and performance consistency, MMRec consumes raw data as input. MMRec provides a full stack toolbox that includes data preprocessing, multimodal recommendation models, multimodal information fusion, performance evaluation to minimize the cost of implementation and comparison of novel models or baselines.

A. Architecture of MMRec

The architecture of MMRec. MMRec consists of 4 modules ranging from raw data preprocessing to model performance evaluation. It takes raw data of multimodal and user-item interaction files as input.

Data Encapsulation: MMRec first preprocesses raw data and encapsulates user interactions and multimodal information into DataLoader of Pytorch. MMRec performs k-core filtering to retain the users and items with at least k interactions, and aligns the multimodal information with the retained items. It then splits the whole interactions into Training/Validation/Test. The raw features of multimodal information are vectorized into numeric values leveraging pre-trained multimodal models, such as transformers.

Trainer: MMRec provides various optimizer to train the models. MMRec unifies the training interface for all models. Customized models are merely required to implement two functions:

calculate_loss : The main part of the model, which defines how loss is generated from the model graph flow.

full_sort_predict: This function predicts the ranking of items for users.

Evaluation: This module features a wide set of commonly used metrics for recommender systems. We have used Recall, NDCG, MAP and Precision for k= 5, 10, 15, and 20.

All modules can be customized and configured by modifying the configuration files. All changes will be reflected and loaded in **config** module. The **config** module also supports grid searching of models on hyperparameters.

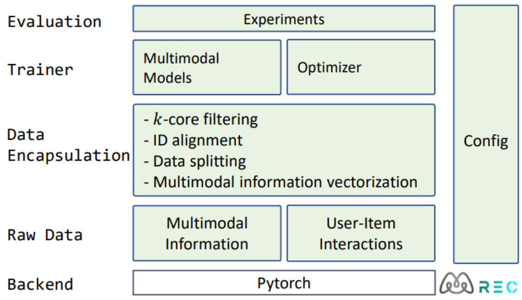

Fig. 1: Architecture of MMRec toolbox

VI. Models

A. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation (DualGNN)

Despite the remarkable performance of prior arts like MMGCN and UVCAN, they are still limited by fusing the user preference derived from different modalities in a unified manner, ignoring the users tend to place different emphasis on different modalities. Furthermore, modality-missing is ubiquity and unavoidable in the micro-video recommendation, some modalities information of micro-videos are lacked in many cases, which negatively affects the multimodal fusion operations.

To remedy the challenges, DualGNN captures the specific multi-modal fusion pattern for each user. It's framework is build upon the user-microvideo bipartite graph and the user co-occurrence graph.

It first simplifies the graph-based model on the multimedia recommendation and devise a new single-modal preference learning module, which performs the graph operations on the user-microvideo graph in each modality to capture single-modal user preferences on different modalities. And then, a multi-modal representation learning module is designed to represent the multi-modal user preference. The process is divided into two parts, the information construction and aggregation operations, in order to explicitly model the user's attentions over different modalities and inductively learn the multi-modal user preference. For information construction, attentive concatenation is used while for information aggregation, it uses the fact that users who have interacted with the same micro-videos are generally close to each other in the multi-modal preference.

Finally, a prediction module is used to rank the potential micro-videos for users by measuring the similarity of each user and micro-video pair.

TABLE II: Metric Values for Different Top-K Values in DualGNN

	Recall	NDCG	Precision	MAP
@5	0.049	0.2178	0.2129	0.1281
@10	0.0868	0.2114	0.2001	0.0997
@20	0.1442	0.2134	0.1775	0.0843
@50	0.2531	0.239	0.1439	0.0819

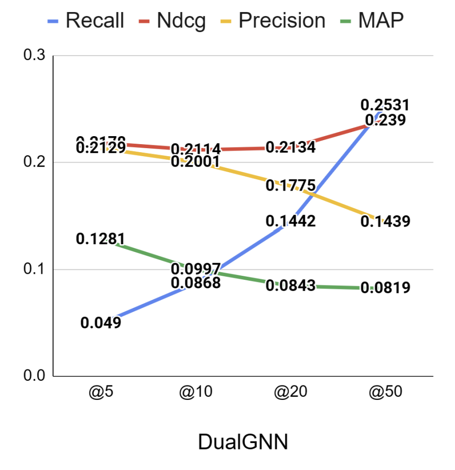

Fig. 2: Metric Values for Different Top-K Values in DualGNN

B. Mining Latent Structures for Multimedia Recommendation (LATTICE)

The majority of previous work follows the traditional CF paradigm and focuses on modeling user- item interactions with multimodal features included as side information. It leads to a gap to the genuine item-item relations that carry *semantic relationships*. LATTICE argues that the latent semantic item-item structures underlying these multimodal contents could be beneficial for learning better item representations and further boosting recommendation. Taking Figure 3 as an example, existing methods will recommend the shirt for u_2 according to collaborative relationships, since shirts, hats, and pants all interacted with u_1. However, previous work may not be able to recommend coats to u_2, which are visually similar to shirts. Firstly, LATTICE learns modality-aware item structures from multimodal features and aggregates modality-aware item graphs to construct *latent multimodal item graphs*. Based on the hypothesis that similar items are more likely to interact than dissimilar items, it first quantifies the semantic relationship between two items by their Cosine similarity and conducts k-NN sparsification on the dense graph to obtain a sparsified, directed graph adjacency matrix.

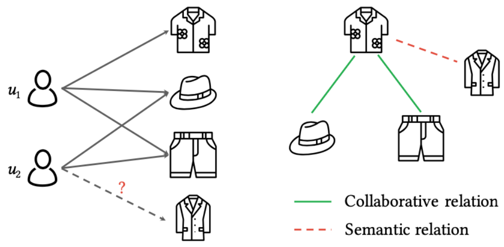

Fig. 3

Then, it transforms raw modality features into high-level featurese and dynamically learns the graph structures by transformed, high-level multimodal features. To keep rich information of initial item graph and stabilize the training process, it adds a skip connection that combines the learned graph with the initial graph to obtain the final graph adjacency matrix representing latent struc-tures for all modalities.

It then introduces learnable weights to assign different importance scores to modality-specific graphs(because users usually focus on different modalities in different scenarios). After this, it performs graph convolution operations to injecting item- item affinities into the embedding process.

Different from previous attempts, LATTICE is flexible and could be served as a play-and-plug module for any CF methods. The play-and-plug paradigm separates the usage of multimodal features with user-item interactions, thus alleviating the cold-start problem, where tailed items are only interacted with few users or even never interacted with users.

It uses Bayesian Personalized Ranking (BPR) loss to compute the pair-wise ranking, which encourages the prediction of an observed entry to be higher than its unobserved counterparts.

TABLE III: Metric Values for Different Top-K Values in lattice

	Recall	NDCG	Precision	MAP
@5	0.0655	0.2956	0.288	0.189
@10	0.12	0.2879	0.2718	0.1527
@20	0.215	0.3023	0.2523	0.1397
@50	0.3946	0.348	0.203	0.1439

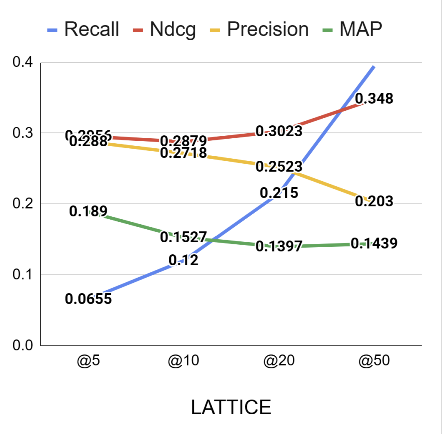

Fig. 4: Metric Values for Different Top-K Values in LATTICE

C. Freezing and Denoising Graph Structures for Multi-Modal Recommendation (FREEDOM)

Generally, prior work fused multimodal features into item ID embeddings to enrich item representations, thus failing to capture the latent semantic item-item structures. In this context, LATTICE proposed to learn the latent structure between items explicitly and achieved state-of-the-art performance for multimodal recommendations. However, we argue the latent graph structure learning of LATTICE is both inefficient and unnecessary. Experimentally, this demonstrated that freezing its item-item structure before training can also achieve competitive performance. Based on this finding, this proposes a simple yet effective model, dubbed as FREEDOM, that FREEzes the item-item graph and denoises the user-item interaction graph simultaneously for Multimodal recommendation. Theoretically, this examine the design of FREEDOM through a graph spectral perspective and demonstrate that it possesses a tighter upper bound on the graph spectrum.

In denoising the user-item interaction graph, it devises a degree-sensitive edge pruning method, which rejects possibly noisy edges with a high probability when sampling the graph. It is evaluated on three real-world datasets and show that FREEDOM can significantly outperform current strongest baselines.

Compared with LATTICE, FREEDOM achieves an average improvement of 19.07% in recommendation accuracy while reducing its memory cost up to 6× on large graphs.

D. Bootstrapped Multi-Modal Model(BM3)

Existing state-of-the-art methods usually use auxiliary graphs (e.g., user-user or item-item relation graph) along

TABLE IV: Metric Values for Different Top-K Values in Freedom

	Recall	NDCG	Precision	MAP
@5	0.0711	0.313	0.3088	0.2011
@10	0.1291	0.3041	0.2889	0.1639
@20	0.2263	0.3151	0.2614	0.1481
@50	0.4119	0.3639	0.2112	0.1542

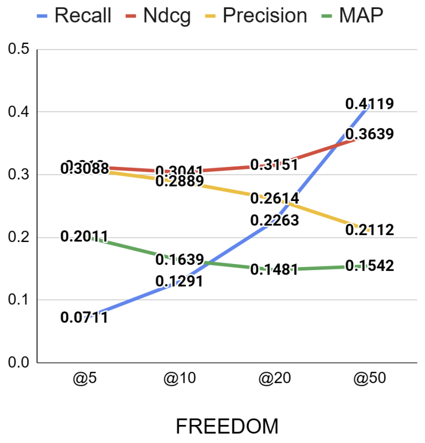

Fig. 5: Metric Values for Different Top-K Values in Freedom

with user-item interaction graph to augment the learned representations of users and/or items. These representations are often propagated and aggregated on auxiliary graphs using graph convolutional networks, which can be prohibitively expensive in computation and memory, especially for large graphs.

Moreover, existing multi-modal recommendation methods usually leverage randomly sampled negative examples in Bayesian Personalized Ranking (BPR) loss to guide the learning of user/item representations, which increases the computational cost on large graphs and may also bring noisy supervision signals into the training process. To tackle the above issues, we propose a novel self-supervised multi-modal recommendation model, dubbed BM3, which requires neither augmentations from auxiliary graphs nor negative samples.

Specifcally, BM3 frst bootstraps latent contrastive views from the representations of users and items with a simple dropout augmentation. It then jointly optimizes three multimodal objectives to learn the representations of users and items by reconstructing the user-item interaction graph and aligning modality features under both interand intra-modality perspectives. BM3 alleviates both the need for contrasting with negative examples and the complex graph augmentation from an additional target network for contrastive view generation.

TABLE V: Metric Values for Different Top-K Values in BM3

	Recall	NDCG	Precision	MAP
@5	0.0561	0.2679	0.2526	0.1721
@10	0.1027	0.2519	0.2291	0.1297
@20	0.1863	0.2619	0.212	0.1148
@50	0.3432	0.3027	0.1742	0.1171

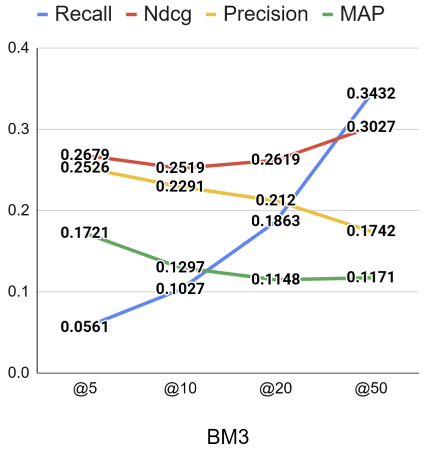

Fig. 6: Metric Values for Different Top-K Values in BM3

E. Multi-View Graph Convolutional Network for Multimedia Recommendation (MGCN)

Current GCN-based methods achieve notable success, they suffer from two limitations: (1) Modality noise contamination to the item representations. (2) Incomplete user preference modeling caused by equal treatment of modality features. MGCN tackles these issues. This model equips three specially designed modules: the Behavior-Guided Purifier, the Multi-View Information Encoder, and the Behavior-Aware Fuser.

Behavior-Aware Fuser: To avoid noise contamination, behavior-guided purifier removes preference-irrelevant modality noise contained in multimodal information, such as the redundant text description, the image background, and the image brightness. Then the preference-relevant modality features are separated from the modality features, with the guidance of behavior features.

Multi-View Information Encoder: Since both the collaborative signals and the semantically correlative signals can significantly influence the efficacy of multimedia recommendation, this module captures collaborative signals from the view of the user-item relationship, and semantically correlative signals from the view of the item-item relationship. For the User-Item View, in particular, to capture high-order collaborative signals, we construct a GCN module to propagate ID embeddings of users and items over the interaction graph. The representations of the l-th layer encode the l-order neighbors' information. For the Item-Item View, it quantifies the item-item affinities based on the Cosine similarity of each raw modality feature. Then, KNN sparsification is conducted on the dense graph to capture the most relevant features from neighbors (for each item, it only preserves edges with the greatest Cosine similarity). Then it constructs a shallow GCN module to propagate modality information. A shallow GCN is used because stacking multiple graph convolution layers not only leads to the node over- smoothing issue, but also easily captures noisy features.

Behavior-Aware Fuser: To accurately capture items' features in different modalities, the behavior-aware fuser does flexible fusion weight allocation based on user modality preferences, which can be distilled from the behavior patterns of users.

During the phase of model training, MGCN adopts the Bayesian Personalized Ranking (BPR) loss as the basic optimization task, which assumes that users prefer historically interacted items over unclicked ones. And it is combined with auxiliary self-supervised tasks to jointly update the representations of users and items.

TABLE VI: Metric Values for Different Top-K Values in MGCN

	Recall	NDCG	Precision	MAP
@5	0.072	0.3099	0.3054	0.1991
@10	0.1272	0.2983	0.2814	0.1586
@20	0.2248	0.3113	0.2578	0.145
@50	0.4113	0.3605	0.2076	0.1517

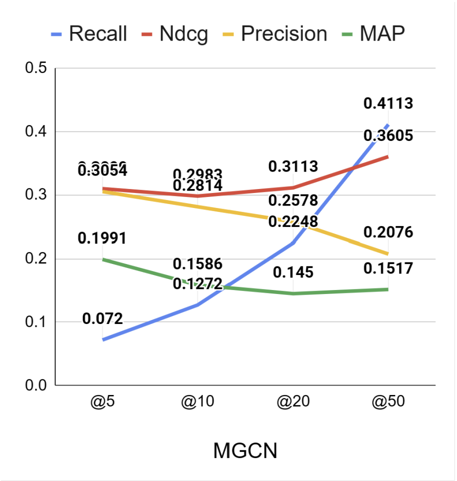

Fig. 7: Metric Values for Different Top-K Values in MGCN

F. Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation (DRAGON)

Earlier state-of-the-art models enhance the dyadic relations between users and items by considering either user-user or item-item relations, leaving the high-order relations of the other side (i.e., users or items) unexplored. DRAGON enhances the dyadic relations by learning **D**ual **R**epresent**A**tions of both users and items via constructing homogeneous **G**raphs for Multim**O**dal Recomme**N**dations.

Previous work explores historical user-item interactions that can be considered as a form of dyadic relation to capturing user preferences. However, these methods show inferior performance due to the sparse nature of interactions between users and items in real-world datasets.

To alleviate the data sparsity problem, recent works focus on modeling user-item interactions as a bipartite graph and integrating multimodal information with graph structure. For example, MMGCN builds a user-item bipartite graph for every modality to obtain modal-specific representation to understand user preference better. DualGNN and LATTICE introduce either user-user or item item relations into the user-item interactions and achieve state-of-the art recommendation performance. Although these models show effective recommendation accuracy, the high-order relations in both sides of the dyadic relations can be explored simultaneously to fully address the data sparsity issue. Inspired by the dual representation learning mechanism, DRAGON enhances the representation learning of users and items by incorporating their dual representations to capture both the interand intra-relations between users and items.

TABLE VII: Metric Values for Different Top-K Values in Dragon

	Recall	NDCG	Precision	MAP
@5	0.0428	0.2163	0.2085	0.1301
@10	0.0775	0.2033	0.1891	0.0965
@20	0.1266	0.2023	0.1701	0.0798
@50	0.2287	0.2203	0.1354	0.0731

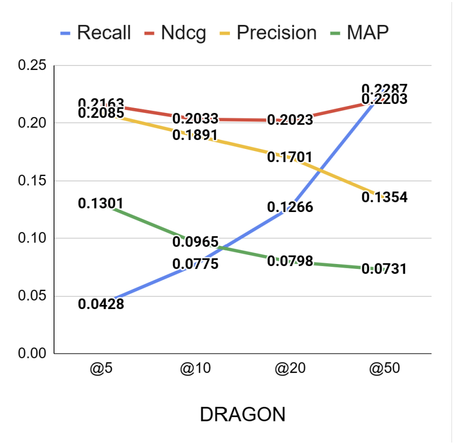

Fig. 8: Metric Values for Different Top-K Values in Dragon

VII. Overall Results

A. Time Analysis

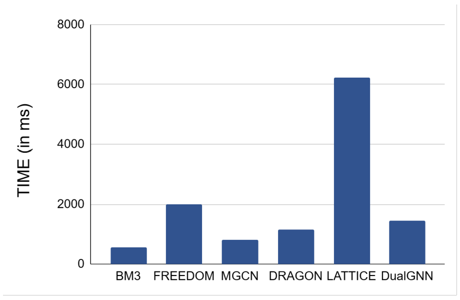

Fig. 9: Time Comparison for the models

We can see that the model with the highest time complexity is LATTICE which is because of its excess graph computations and brute Collaborative Filtering. We can see that BM3 is the fastest model which is mostly due to dropout and bootstrapping the latent contrastive views.

B. Accuracy

We are comparing all the models on our Enhanced Movielens dataset using 4 metrics that are Recall, NDCG, Precision and Mean Average Precision

1) *Recall*

Recall@K is defined as the ratio of the number of relevant items that were recommended in the top-K list to the total number of relevant items in the entire dataset. In other words, it answers the question: ”Out of all the items that the user might find relevant, how many were actually recommended in the top-K list?” A high recall means that the recommendation system is effective at capturing a substantial portion of the relevant items in the top-K list, which is important for ensuring that users don’t miss out on items of interest.

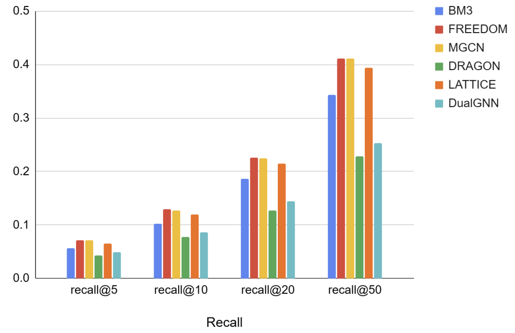

Fig. 10: Performance of models in terms of Recall

2) *Normalized Discounted Cumulative Gain(NDCG)*

NDCG, or Normalized Discounted Cumulative Gain, is a widely used metric in information retrieval and recommendation systems to assess the quality of top-K recommendations. NDCG is particularly useful when you want to consider both the relevance of items and their position in the recommendation list. It quantifies how well the recommendations are ordered and how well the most relevant items are placed at the top of the list. Discounted Cumulative Gain is calculated as :

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

where rel_i is the relevance of the item at rank i in the recommendation list. Typically, relevance is binary (1 for relevant, 0 for not relevant), or it can be a graded score (e.g., 1 to 5, with higher values indicating higher relevance). DCG is normalized to account for varying list lengths and make it a value between 0 and 1. NDCG is calculated by dividing the DCG by the ideal DCG (iDCG), which is the DCG of the list when all the items are perfectly ordered in terms of relevance.

$$NDCG@K = \frac{DCG@K}{iDCG@K}$$

3) *Precision*

Precision@K is defined as the ratio of the number of relevant items recommended in the top-K list to the total number of items in the top-K list. In other words, it answers the question: ”Out of the items recommended in the top-K list, how many are actually relevant to the user?”

It’s worth noting that precision is a valuable metric, but it should be considered in conjunction with other metrics like recall and F1-score to provide a more complete picture of recommendation system performance. Precision measures how many recommended items are truly relevant, but it doesn’t consider whether all relevant items have been included in the recommendations (recall) or the ordering of relevant items in the list.

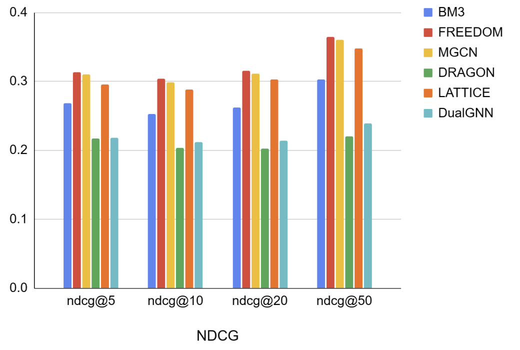

Fig. 11: Performance of models in terms of NDCG

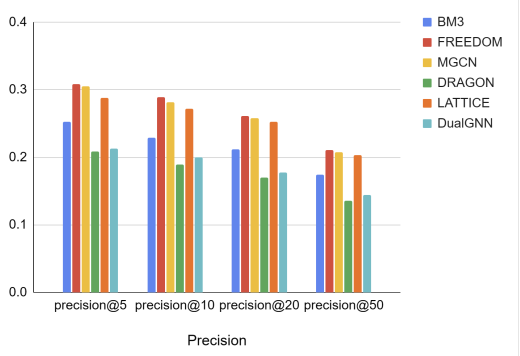

Fig. 12: Performance of models in terms of Precision

4) *Mean Average Precision(MAP)*

Mean Average Precision (MAP) is a widely used metric in the context of measuring the quality of top-K recommendations in recommendation systems, particularly in information retrieval and search engine evaluation. It takes into account both the precision of the recommendations and the order in which relevant items are presented in the top-K list.

MAP measures the quality of recommendations by considering not only whether relevant items are present in the top-K list (precision) but also the order in which they appear. It rewards recommendations that place relevant items at the top of the list, as this is more user-friendly. Higher MAP values indicate better recommendation quality.

VIII. Conclusion

- On benchmarking all these models on our dataset we could see that with respect to time the slowest model is LATTICE whereas the fastest model is BM3, with respect to accuracy, Freedom, LATTICE and MGCN are the best performing whereas Dragon and DualGNN perform the worst.LATTICE can alleviate cold-start problem and outperforms all baselines in the previously tested as well as out dataset. It learns item graphs from multimodal features, along which cold-start items get similar feedbacks from relevant neighbors through neighborhood aggregation. Other Collobarative Filter-

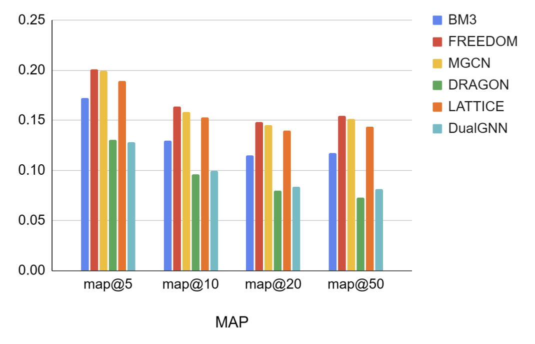

Fig. 13: Performance of models in terms of MAP

ing methods obtain poor performance under cold-start settings in general, primarily because they only leverage users' feedbacks to predict the interactions between users and items. Although these methods work well for items with sufficient feedbacks, they cannot help in cold-start settings, since no user-item interaction is available to update the representations of cold-start items.

- It is very clear that MGCN significantly outperforms both general recommendation models (previously studied datasets) and multimedia recommendation models (our dataset). Specifically, instead of directly incorporating modality information, it first purifies it with the guidance of behavior information which avoids contamination from modality noise. Besides, through the multi-view information encoder, the behavior features and modality features are enriched, by encoding the high-order collaborative signal and semantically correlative signals. It then obtains each user's and item's representations through a behavior-aware fuser, which adaptively fuses modality features according to user modality preference. The indirect injection of modality features mitigates the issue of modality noise contamination.
- We know that LATTICE proposes to learn the latent structure between items explicitly and achieves state-of-the-art performance for multimodal recommendations. However, we argue the latent graph structure learning of LATTICE is both inefficient and unnecessary. This is upon the introduction of Freedom, its results demonstrate that freezing its item-item structure before training can also achieve competitive performance.This model experimentally shows us that the item-item structure learning of LATTICE is not necessary. Specifically, it builds an item-item graph directly from the raw multimodal contents of items and freeze it in training LATTICE. This is the reason LATTICE is very slow as compared to other models. Compared with LATTICE, FREEDOM achieves a very significnat improvement in recommendation accuracy while reducing its memory cost up to 6× on large graph. FREEDOM learns the representations of users and items by integrating the unweighted item-item graph and the sparsified user-item subgraph.
- We can clearly observe that DualGNN does not perform very well in this setting, the reason for that is that our datset does not contain videos as modality. DualGNN consistently outperformed state-of-the-art baselines in all the datasets which involved micro-video recommendations, which could demonstrate its effectiveness for the micro-video recommendation. It's effectiveness in those settings is attributed to the framework to model each user's different preferences for different modalities, and learn the multi-modal user preference in an inductive manner. Thus the DualGNN framework could make multi-modal users preference representation learning more accurately, and increase the accuracy of the top-recommendation in case of video recommendations as well as in the settings where the data is very noisy.
- We can see Dragon does not perform very well as compared to other models, that is because it learns the representations of users and items which is critical for the recommendation system. All representation learning-based techniques assume the existence of a common representation with consistent knowledge of different views of items. Different views of an item contain specific discriminant information in addition to consistent knowledge about this item. It constructs the heterogeneous and homogeneous graphs together to learn the dual representations of both user and item, which could capture both the internal association and the relationship between users and items. The poor performance could be due to the low correlation between the users and items in our dataset.
- BM3 did not achieve good performance in terms of accuracy but it was the fastest among all the models. It removes the requirement of randomly sampled negative examples in modeling the interactions between users and items. To generate a contrastive view in self-supervised learning, It utilizes a simple yet effcient latent embedding dropout mechanism to perturb the original embeddings of users and items. Since our dataset is smaller as compared to other dataset on which it was tested on, the dropout strategy did not work very well.The experimental results of previous studies show that BM3 achieves signifcant accuracy improvements over the state-of-the-art multi-modal recommendation methods as the datsets were large enough to realise the power of dropout, while training 2-9× faster than the baseline methods

References

[1] Xin Zhou. "MMRec: Simplifying Multimodal Recommendation". arXiv preprint arXiv:2302.03497v1, 2023.

[2] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. "Bootstrap latent representations for multi-modal recommendation (BM3)". arXiv preprint arXiv:2207.05969, 2022.

[3] Xin Zhou. "A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation (FREEDOM)". arXiv preprint arXiv:2211.06924, 2022.

[4] Penghang Yu, Zhiyi Tan, Guanming Lu, Bing-Kun Bao*. "Multi-View Graph Convolutional Network for Multimedia Recommendation (MGCN)". arXiv preprint arXiv:2308.03588, 2023.

[5] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu1, Shuhui Wang, Liang Wang. "Mining Latent Structures for Multimedia Recommendation (LATTICE)". arXiv preprint arXiv:2104.09036. 2021.

[6] Hongyu Zhou, Xin Zhou, Lingzi Zhang, Zhiqi Shen. ”Enhancing Dyadic Relations with Homogeneous Graphs for Multimodal Recommendation (DRAGON)”. arXiv preprint arXiv:2301.12097. 2023.

[7] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, Liqiang Nie, ”DualGNN: Dual Graph Neural Network for Multimedia Recommendation”. IEEE Transactions on Multimedia. 2021.