Optimizing Movie Selections: A Multi-Task, Multi-Modal Framework with Strategies for Missing Modality Challenges

Subham Raj[*]
Indian Institute of Technology Patna
Patna, Bihar, India
subham_2221cs25@iitp.ac.in

Pawan Agrawal
Indian Institute of Technology Patna
Patna, Bihar, India
pawan_1901cs40@iitp.ac.in

Sriparna Saha
Indian Institute of Technology Patna
Patna, Bihar, India
sriparna@iitp.ac.in

Brijraj Singh
Sony Research India
Bengaluru, Karnataka, India
brijraj.08@gmail.com

Niranjan Pedanekar
Sony Research India
Bengaluru, Karnataka, India
niranjan.pedanekar@gmail.com

ABSTRACT

Online recommendation systems have become a crucial feature of Over-the-Top (OTT) platforms, which provide streaming media content over the internet. OTT platforms, such as Netflix, Hulu, and Amazon Prime, use recommendation systems to suggest movies, TV shows, and other content to users based on their viewing history and preferences. The accuracy of these recommendations can impact user satisfaction, retention, and revenue for the platform.

In this context, multi-task and multi-modality learning approaches have shown promise in improving the accuracy of movie recommendations. Our proposal is that by simultaneously tackling two correlated tasks, specifically (a) classifying movie genres and (b) identifying user-movie ratings, we can produce high-quality movie embeddings through an end-to-end approach without the use of a rating vector. Additionally, two more correlated tasks, the user's gender and age-group, are also considered for our experiments. We have developed several multitasking models, including fully shared (FS) and shared-private (SP) feature models, to address multiple tasks at once, namely genre classification and user-movie rating prediction. Additionally, we've broadened our method to address situations where certain modalities are missing. Recognizing that real-world situations might not always provide information from every modality, this tackles the challenge of missing modalities. Hence, our proposed model MTM^4F which stands for **M**ulti-**T**ask **M**ulti-**M**odal **M**issing **M**odality **F**ramework includes the integration of a meta-learning framework and a multi-task architecture. In experiments, we utilized a multi-modal format of the MovieLens-100K and MMTF-14K datasets. When tested on these datasets using the root mean square error (RMSE) metric, the suggested multi-task model, combined with its shared private training approach, surpasses state-of-the-art models currently available.

KEYWORDS

Recommender system, multi-modal, missing modality, personalized recommendation,

ACM Reference Format:
Subham Raj, Pawan Agrawal, Sriparna Saha, Brijraj Singh, and Niranjan Pedanekar. 2024. Optimizing Movie Selections: A Multi-Task, Multi-Modal Framework with Strategies for Missing Modality Challenges. In *The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24), April 8–12, 2024, Avila, Spain.* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3605098.3635947

1 INTRODUCTION

Recommender systems (RSs) are a type of artificial intelligence technology that provides personalized recommendations to users. They are widely used in e-commerce, entertainment, social media, and other industries to improve the user experience and increase engagement. Recommender systems use algorithms and user data, such as browsing history, purchase history, and ratings, to predict and suggest relevant items or content that match the user's interests and preferences. These systems can take different approaches, such as collaborative filtering (CF) [11, 12], content-based filtering (CBF) [27], and hybrid models, to provide recommendations. Collaborative filtering (CF) approaches operate under the premise that users with comparable past interaction patterns will likely exhibit similar future behaviors. Typically, these methods employ traditional matrix factorization (MF) methods and their derivatives. In contrast, content-based filtering (CBF) techniques incorporate extra details about items or users, like item characteristics, to suggest items based on a user's prior preferences or direct feedback.

Incorporating multi-modality has proved beneficial for the recommendation task, but the existing works[5, 17, 21] simply concatenated the multi-modal features which might not be capable enough to capture the complex relationships and dependencies between modalitites. Also, plain concatenation can lead to high-dimensional data, making it computationally expensive and prone to issues related to the curse of dimensionality. So, there is a challenge on how we can effectively fuse our multi-modal information, and therefore, we have incorporated CentralNet architecture [30] to fuse the multi-modal data effectively.

Following the fusion of multi-modal data, a multi-task architecture was adopted for the RS task. Multi-task learning is a machine learning approach that aims to learn multiple tasks simultaneously

[*] Primary author

SAC '24, April 8–12, 2024, Avila, Spain

ACM ISBN 979-8-4007-0243-3/24/04…$15.00
https://doi.org/10.1145/3605098.3635947

using shared representations, with the goal of improving the overall performance of the system. In multi-task learning, the model learns to jointly optimize multiple objectives, leveraging the correlations among tasks to improve generalization performance. Multi-task learning has been applied in various fields, including computer vision [8], natural language processing [2], and drug discovery [22]. The authors in [15] proposed a multi-task learning approach to solve recommendation problems as well as to generate explanations.

Our study has also incorporated the real-world scenario where a particular modality might be missing due to security reasons, private contents, etc. This motivates us to include multimodal generative models [18, 25, 26, 34] in our multi-task architecture.

Considering all the above points, we proposed MTM^4F model for movie recommendation having the following contributions:

- Incorporating a new modality to the MMTF-14K dataset, namely the subtitle of the movies.
- Effectively fuse the multi-modal information from audio, video, and text.
- Utilizing the interconnected nature of two main tasks, namely predicting user-item ratings as the primary objective, and classifying movie genres, user genders, and user age groups as auxiliary tasks, to enhance the performance of both objectives by employing efficient multi-modal representations (including audio, video, metadata, and subtitles) for movies.
- Evaluating the performance of multi-task architecture in case of missing modality environment.
- Presenting the model that outperforms existing baselines for recommendation tasks for the MovieLens-100K and MMTF-14K datasets.

The rest of this paper can be outlined as follows. To begin, Section 2 discusses the existing literature on the topic. Subsequently, Section 3 examines the dataset, while Section 4 outlines the problem statement. Section 5 elaborates on the proposed methodology. Lastly, Section 6 presents the experimental findings, and the paper concludes thereafter.

2 RELATED WORKS

2.1 Multimodal Recommender System

Recommendation systems that take into account various types of media content (multimodal recommendation systems) have been effectively employed in various domains such as e-commerce, social media platforms, and instant video platforms, as demonstrated by prior research works [6, 10, 16, 28].

In recent times, Graph Neural Networks (GNNs) have been incorporated in recommendation systems, with a special focus on multimodal recommendation systems. The studies of [31, 35, 37] have introduced GNNs in recommendation systems. Additionally, studies such as [14, 32, 33] have explored the application of GNNs in multimodal recommendation systems. For instance, MMGCN[33] creates a graph specific to each modality and utilizes graph convolutional operations to grasp user preferences concerning individual modalities. This process enables the model to distill item representations concurrently, resulting in learned user representations that reflect their specific interests in items. On the other hand, GRCN[32] follows a similar approach as MMGCN, but it emphasizes adapting and refining the interaction graph's structure to uncover and remove false-positive edges.

2.2 Multimodal Generative models

In the domain of multimodal learning, there are two primary categories of generative models: cross-modal generation and joint-model generation. Authors in [18, 25] proposed cross-modal generation techniques that establish a conditional generative model encompassing all modalities. Meanwhile, authors in [26, 34] proposed joint-model generation techniques like the Multimodal Variational Autoencoder(MVAE) which aims to capture the joint distribution of multimodal data. In this paper, we have tried to modify the existing multimodal generative models algorithm so that it can work very well when a few modalities are not available.

3 DATASET

We have incorporated two publicly available datasets for our experiment. The subsequent sections outline the particulars of these two datasets.

3.1 MovieLens-100K

The original Movielens-100K dataset consists of 1682 movies and 943 users and it consists of 100,000 user-movie pair ratings. It is made multimodal by authors in [19] where they have added audio, video, and textual information for each movie. It is further extended by authors in [1] where they have added subtitles for each movie.

You can find an illustration of the dataset in Figure 1.

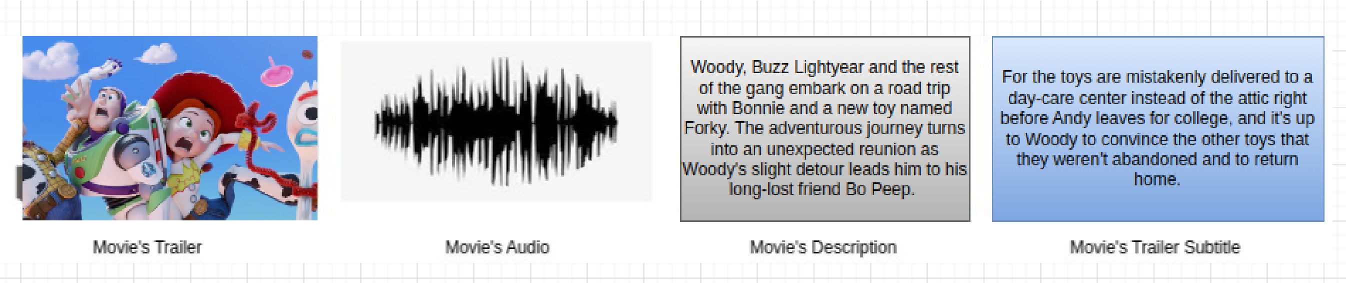
3.2 MMTF-14K

The MMTF-14K dataset [7] has 138,492 users and 13,623 movies with approximately 12 million ratings. It has audio and visual features publicly available.

There are two types of descriptors for audio features: block-level features (BLF)[23] and i-vector features[24]. Similarly, visual features are categorized into two types: Alexnet [13] features and Aesthetic [9] features. Different aggregation methods are used for both types of visual features. The Aesthetic features contain three different types of descriptors: color-based, texture-based, and object-based. On the other hand, Alexnet is a deep neural network model, and the output features are extracted from the fc7 layer. This dataset is extended by authors in [21], where they have introduced textual features such as the movie's summary, its director, and IMDb rating collected from the IMDb website for each movie present in the MMTF-14K dataset. We have now extended this dataset further to incorporate subtitles for each movie.

4 PROBLEM STATEMENT

The objective of our paper is to predict user ratings using a multimodal item representation, which includes Audio, Meta, Subtitle, and Video modalities, and how effectively we fuse this multi-modal data. Following this, different multi-task architectures were adopted to solve the downstream task of recommendation.


Woody, Buzz Lightyear and the rest of the gang embark on a road trip with Bonnie and a new toy named Forky. The adventurous journey turns into an unexpected reunion as Woody's slight detour leads him to his long-lost friend Bo Peep.

For the toys are mistakenly delivered to a day-care center instead of the attic right before Andy leaves for college, and it's up to Woody to convince the other toys that they weren't abandoned and to return home.

Figure 1: A multi-modal instance of a movie from Movielens-100K dataset

The focus is also on handling situations with severe modality gaps, using both modality-complete and modality-incomplete samples to train the model. The dual objectives are flexibility and efficiency, which refer to the ability to handle missing modalities at different stages and enhance the model's efficiency in scenarios with severe modality gaps, respectively. Our goal is to create features for available modalities and use them to estimate the features of the missing modalities, and then generate a user-item representation that will be used in the proposed multi-task model.

To formalize the missing modality task at hand, we define U = u_1, u_2, .., u_n as the complete set of n users in the dataset, and V = v_1, v_2, .., v_m as the complete set of m movies in the dataset. For any given user-movie pair u_x, v_y, we represent the user as $u_x^{user_{embedding}}$ and the movie as $v_y^{[I_{m_1},I_{m_2},..,I_{m_k}]}$, where m_1, m_2, .., m_k are the available modalities for movie v_y. Our model leverages these representations to perform a set of tasks, $T = \{t_1, t_2, t_3, t_4\}$. Task t_1 is our main task of rating prediction, tasks t_2, t_3, and t_4 are auxiliary tasks where we predict the movie's genre, the user's gender, and the user's age group, respectively. The objective is to establish a correlation between these tasks, such that the user-item similarity can be enhanced with the aid of various types of multi-modal information.

5 PROPOSED METHODOLOGY

This section discusses how feature vectors are generated for all modalities and how missing modality features are reconstructed using available modalities. It also introduces an attention-based approach for combining multiple modalities and a multi-task model proposal to perform the desired task. For missing modality experiments, the entire model can be thought of as consisting of three networks: the feature reconstruction network, f_{ϕ_c}, which uses parameter ϕ_c to reconstruct missing modality features; the feature regularization network, f_{ϕ_r}, which uses parameter ϕ_r to regulate latent features; and the main network, f_θ, which uses parameter θ to process the user-item representation and perform the desired tasks.

5.1 Multi-modal Representation via Embedding

Our study comprises four modalities namely, meta, video, audio, and subtitle. Their representation is generated from different methods which is discussed in detail in subsequent sections. These feature vectors form item representation. From these feature vectors, we have also generated user representation. Embedding generation of meta, audio, video, and user is adopted from the method mentioned in [21].

5.1.1 Meta Embedding. The Meta Embedding is a representation that combines a movie's textual description and meta-data, and it is generated using a knowledge graph-based technique. The graph is represented as a list of tuples that include two entity nodes and the relation between them.

The process of creating a knowledge graph for tabulated metadata involves creating directed edges between entities based on their attributes. When dealing with movie descriptions, entities in the knowledge graph are formed using Part-of-speech (POS) information from sentences, with a limitation of three sentences. The network's head, root, and tail are interpreted as the subject, verb, and object, respectively, featuring directed edges from head to root and root to tail. The word2vec representation is applied to transform each word in the graph, and the feature vector is computed by utilizing the word2vec representations of the words at the head, root, and tail nodes. The feature vector is represented as:

$$I_{meta} = ||h + R - t|| \tag{1}$$

Creating a knowledge graph representation of the tabulated meta-information and the initial three sentences of the item description involves applying the equation mentioned above. Subsequently, all generated knowledge graph representations are combined to form a feature vector, denoted as I_{meta}, with a dimensionality of $\mathbb{R}_{1X1220}$.

5.1.2 Video Embedding. Movie trailers were treated as videos, and 5 frames were selected to represent each video. A technique employing K-means clustering was utilized to group the I-frames within the video, identifying the video's most indicative frames. Through experimentation with various K values, it was established that selecting K=5 yielded the most precise representation of the video embedding for the movie. The experimentation here is referred from [21].

The architecture of the model comprises three CNN layers with sizes $(64, 3, 3)$, $(128, 3, 3)$, and $(16, 3, 3)$. Following the CNN layers are batch normalization and max-pooling operations. Next, the resulting output undergoes flattening and is passed through three dense layers of sizes 1024, 512, and 231. After merging the final layer's output with the user vector, it undergoes subsequent processing through extra dense layers to generate the normalized rating value. After completing the comprehensive training, we detach the CNN layers from the two preceding dense layers to function as the

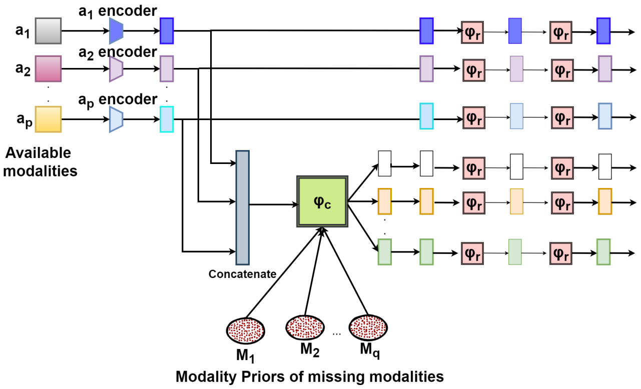


Figure 2: Model proposed for reconstructing missing features and feature regularization

image embedding model. The output of an intermediary layer is regarded as the video embedding, featuring a $\mathbb{R}_{1X231}$ dimensional vector.

5.1.3 Audio Embedding. An end-to-end technique is used to generate an audio embedding from the audio of movie trailers, similar to the technique used to generate the video embedding. The audio data is standardized to a specific length, and then passed through a series of dense layers to generate the audio embedding. The resulting audio embedding is a feature vector from an intermediate layer with dimensions $\mathbb{R}_{1X512}$.

5.1.4 Subtitle Embedding. Movie trailer subtitles were considered for the purpose of creating subtitle embeddings. Sentence transformers have proven their effectiveness in encoding semantic content and producing rich feature vectors. Therefore, we employ the pre-trained SBERT model to produce a feature vector with dimensions of $\mathbb{R}_{1X768}$.

5.1.5 User Embedding. To generate an effective user embedding, the movies rated by the user above a certain threshold value are extracted and their embeddings are averaged together with feature columns. This is concatenated with the user’s meta information to create the user representation. Through experimentation which is mentioned in [21], a threshold value of 3 was found to be optimal. This is necessary as generating user embeddings based solely on metadata such as age, gender, and employment is not sufficient due to limited information. Formally user’s Embedding is calculated as $U = [U_{meta} | \frac{\sum_{i=1}^{m} M_i}{m}]$ where, U_{meta} refers to the metadata information of a user, while m represents the total number of movies that a user has rated above the threshold value of 3.

5.1.6 CLIP Embedding. CLIP (Contrastive Language-Image Pretraining) [20] is a pre-trained model that captures combinations of images and text, training both image and text encoders concurrently to establish a streamlined embedding space. It was trained on 400 million pairs to predict correct matches and create 512-dimensional vectors. The model maximizes the cosine similarity of real pairs. The effectiveness of standard embedding was evaluated by encoding image frames (for video embedding) and subtitles using CLIP embedding and performing ablation studies.

5.2 Feature Reconstruction and Regularization

We have introduced a feature reconstruction network designed to rebuild missing modalities within incomplete data samples. This is achieved by using the existing modalities as a reference for the reconstruction process. Suppose we have a set of available modalities denoted as $a_1, a_2, ..., a_p$, and a set of missing modalities as $m_1, m_2, ..., m_q$ for an incomplete sample. To represent all the available modalities as a single vector, we calculate it by combining individual representations through concatenation. Mathematically, A = Concatenate($[I^{a_1}, I^{a_2}, ..., I^{a_p}]$). In order to reconstruct all missing modalities, the following objective can be optimized:

$$\phi_c^* = \underset{\phi_c}{\operatorname{argmin}} \sum_{i=1}^{q} \mathbf{E}_{p(\hat{I}^{m_i}, A)}(-\log p(\hat{I}^{m_i} | A; \phi_c)) \tag{2}$$

To tackle the difficulty of training the reconstruction network with a small proportion of samples containing all modalities, especially in scenarios with severe modalities missing, we’ve proposed an alternative strategy. This method involves approximating the missing modalities through a weighted sum of modality priors, denoted as $\mathcal{M}_i$. The modality priors, learned independently for each missing modality m_i, are derived from clustering feature vectors. This clustering process utilizes methods such as K-means or PCA, where the centroids of the clusters are employed as the modality priors. Approximatimg missing modality m_i as:

$$\hat{I}^{m_i} \; = < \mathcal{W}_i, \mathcal{M}_i > , \; where \; \mathcal{W}_i \sim \mathcal{N}(F, \sigma) \tag{3}$$

The proposed method employs feature regularization to improve the model’s effectiveness when dealing with severe missing modalities. This regularization is implemented through the feature regularization network at each layer, utilizing the features from the preceding layer as input.

Instead of generating a fixed regularization value for the latent feature h^l at layer l using a deterministic function $f_{\phi_r}(h^{l-1})$, we propose using a multivariate Gaussian distribution $\mathcal{N}(\mu, \sigma)$ to generate the regularization with means and variances obtained from $f_{\phi_r}(h^{l-1})$. The following equation can be used to calculate regularized feature.

$$h^l := h^l \circ r \; , \; where \; r \;\; \sim \;\; \mathcal{N}(\mu, \sigma) \tag{4}$$

The symbol ∘ is a pre-defined operation that is used for feature regularization. It could be either an addition or a dot-product operation. Figure. 2 shows the proposed architecture for reconstructing features for missing modality and feature regularization.

5.3 Multi-Modal Fusion

The proposed approach uses CentralNet [30], which is a hybrid fusion network that combines early and late fusion, to fuse multimodal features. There are k distinct networks for individual modalities, along with a central network that merges features from various modalities by calculating a weighted total of the singular hidden representations and their preceding layer. In CentralNet, $h^i(M_j)$ represents the i^{th} hidden layer of modality M_j. In the central network, the weighted sum of each modality is computed, where weights are trainable scalars. Fusion architecture for 2 modalities is shown in Figure. 3

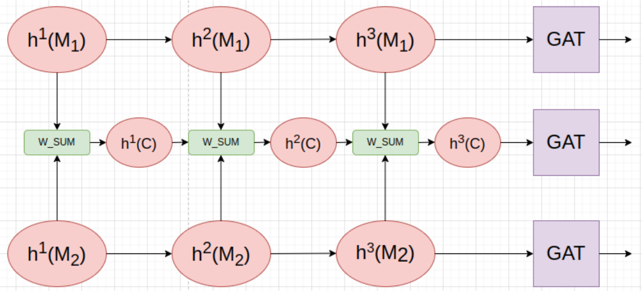


Figure 3: CentralNet Fusion

Outputs at the end of each network are passed through a Graph Attention Network to generate better item representation and are eventually used in our multi-task architecture.

5.3.1 Graph Attention Network. The Graph Attention Network (GAT) was introduced [29] as an approach to handle data structured as graphs through the use of self-attention and masking techniques. In this work, GAT was used to calculate attention weights for movies based on their relationship with other movies. The process involved pairwise concatenation of each movie to create a matrix, which was then multiplied by a learnable layer to produce attention weights. Using the derived attention matrix, a weighted movie vector was generated, with each row being normalized through softmax. Attention weights are calculated as:

$$\alpha_{ij} = \frac{exp(LeakyReLU(a^{-T}[W\vec{h_i}||W\vec{h_j}]))}{\sum_{n\in N_i} exp(LeakyReLU(a^{-T}[W\vec{h_i}||W\vec{h_j}]))} \qquad (5)$$

The movie's representation is then calculated as the weighted sum of all movies using the above attention weights.

5.4 Model Architecture

Fully Shared and Shared Private Multi-Task architectures are discussed in this section. To calculate the loss function for the FS-MT and SP-MT models, mean squared error is used. The formula for the loss function is as : $Loss_{tasks} = \sum_{t=1}^{T} \alpha_t \sum_D (y_{true} - y_{pred})^2$. Here, α_t represents the weight assigned to task t, and y_{true} and y_{pred} represent the true and predicted values, respectively, for each data point D.

5.4.1 Fully Shared Multi-tasking Model (FS-MT). The fully shared multitasking (FS-MT) model aims to handle numerous tasks by utilizing a single parameter set. Its purpose is to acquire shared features and representations that are applicable across various tasks. This model incorporates a fully shared layer that learns both unique and common characteristics for multiple tasks. It accomplishes this by processing inputs through fully connected layers and combining them using a weighted summation method. Nonetheless, when tasks have limited correlation, the shared layer might fail to capture task-specific attributes effectively. You can observe the FS-MT model's architecture in Figure 4 provided alongside.

5.4.2 Shared Private Multi-tasking Model (SP-MT). The SP-MT model is an improvement over the FS-MT model, which separates task-specific features from shared features. It uses a task-specific network for each task and a shared model for all tasks. However, it suffers from the drawback that private features of tasks are mixed with shared features and vice versa, leading to redundancy. Shared-Private Multi-Task architecture is shown in Figure. 5.

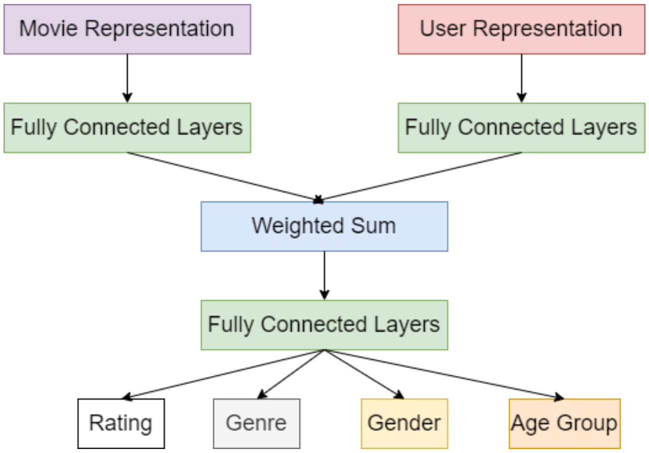


Figure 4: Fully Shared Multi-tasking Model

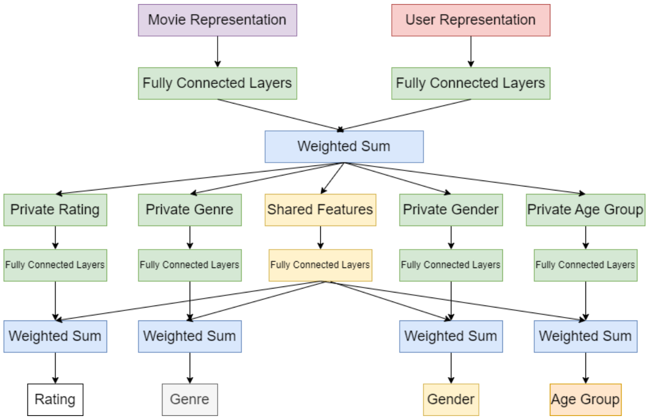


Figure 5: Shared-Private Multi-tasking Model

5.5 Meta-Learning within a Bayesian Framework

In the Bayesian Meta-Learning approach, the main network is trained using feature reconstruction and regularization networks on incomplete modality samples. During the meta-training phase, the loss is calculated as: $\mathcal{L}_{meta-train} = Loss_{tasks}$. In the meta-testing phase, the updated main network is tested using complete modality samples. The network parameters are then updated using the meta-validation stage, which involves calculating the mean squared error between predicted and true feature vectors of missing modalities, as well as the loss from all tasks.

6 EXPERIMENTS AND RESULTS

Within this portion, we carry out experiments on two well-established real-world datasets to respond to the following research questions:

- **RQ1**: How does it affect the proposed model while considering task-specific embedding?
- **RQ2**: How does the model perform under different multi-task settings?
- **RQ3**: How does the proposed model perform in case of missing modality scenario?
- **RQ4**: How the proposed model is performing with respect to existing state-of-the-art models?

6.1 Experimental Setting and Evaluation Metrics

The evaluation was conducted using the U1 split (train-test split) for both the MovieLens-100K and MMTF-14K datasets. We initially removed any entries from the U1 split that were missing any modality type. This left us with 74,815 training instances and 18,751 testing instances for the MovieLens-100K dataset. For the MMTF-14K dataset, outlined in [21], we opted for a random subset of 3,000 users, yielding about 400,000 training instances and 80,000 testing instances. Every instance comprises a user ID, a movie ID, and a rating between 1 and 5. Throughout each trial, a designated set of modalities remained consistent across all training instances, while the rest were incorporated randomly into a specific fraction of the training set.

The efficacy of the proposed design was assessed using the root mean squared error (RMSE) metric [4].

6.2 Results and Analysis(RQ1,RQ2)

We carried out several ablation studies to evaluate the effectiveness of various components in our approach. We firstly tried to explore the different versions of the multi-modal fusion method as mentioned in 5.3 before passing it to the GAT. Once we adopted a particular fusion method, we did thorough ablation studies on different multi-task architectures with two and four tasks.

Table 1: RMSE with different multi-modal embeddings and fusion architecture.

Modalities	$CNet_{fuse}$ *Standard*	$CNet_{fuse}$ *CLIP*	$CNet_{all}$ *Standard*
Audio + Meta + Subtitle + Video	0.936	0.936	**0.930**
Meta + Subtitle + Video	0.939	**0.936**	0.938
Audio + Subtitle + Video	0.933	0.939	**0.932**
Audio + Meta + Video	0.940	**0.937**	0.938
Audio + Meta + Subtitle	0.933	0.935	**0.933**
Subtitle + Video	0.937	0.938	**0.933**
Meta + Video	0.937	0.938	**0.934**
Meta + Subtitle	0.935	0.947	**0.934**
Audio + Video	0.937	0.936	**0.932**
Audio + Subtitle	0.937	0.938	**0.933**
Audio + Meta	0.936	0.937	**0.932**
Audio	0.989	-	**0.984**
Meta	0.940	-	**0.933**
Subtitle	**0.935**	0.938	0.944
Video	0.935	**0.934**	0.936

Table 2: RMSE with different multi-modal embeddings and fusion architecture for MMTF-14K.

Modalities	$CNet_{fuse}$ *Standard*	$CNet_{fuse}$ *CLIP*	$CNet_{all}$ *Standard*
Audio + Meta + Subtitle + Video	0.968	0.969	**0.961**
Meta + Subtitle + Video	0.968	0.969	**0.965**
Audio + Subtitle + Video	0.965	0.967	**0.964**
Audio + Meta + Video	0.971	**0.969**	0.970
Audio + Meta + Subtitle	0.967	0.967	**0.966**
Subtitle + Video	0.970	0.972	**0.969**
Meta + Video	0.969	0.970	**0.967**
Meta + Subtitle	0.970	0.970	**0.969**
Audio + Video	0.974	0.978	**0.971**
Audio + Subtitle	0.977	0.977	**0.970**
Audio + Meta	0.971	0.972	**0.968**
Audio	0.984	-	**0.981**
Meta	0.967	-	**0.966**
Subtitle	0.967	0.968	**0.967**
Video	0.970	**0.969**	0.970

6.2.1 *Study on Multi-modal Fusion.*

We conducted experiments using two versions of CentralNet: $CNet_{all}$, which uses the outputs of all networks for predictions and loss computation, and $CNet_{fuse}$, which only uses the central network for prediction and optimization. Also, we have two kinds of embedding, one is task-specific embedding which we have referred to as standard embedding, and the other is clip-based embedding. Therefore, we have passed both kinds of embedding to the central net architecture and analyzed the results. The results of these experiments are presented in Table. 1,2 for the Movielens-100K and MMTF-14K respectively. The results show that $CNet_{all}$ in which standard embeddings are passed is the best multi-modal fusion architecture, and hence we have adopted this architecture for all our experiments. Also, the multi-task architecture adopted was shared private.

6.2.2 *Effect of Different Tasks.*

We have conducted experiments to evaluate the performance of different multi-task models for various tasks while using all possible combinations of modalities. We have used $CNet_{all}$ architecture for the fusion of different modalities. The outcomes of these experiments are presented in Table 3,4 for the Movielens-100K and MMTF-14K respectively. We can clearly see that shared private architecture with four tasks has performed better in every combination of modalities than other multi-task architectures. The least RMSE was reported 0.930 and 0.961 for the Movielens-100K and MMTF-14K datasets respectively when all modalities were considered. This shows that the multi-modality is having a significant impact while solving the task of recommendation.

6.3 Ablation Studies(RQ3)

We conducted experiments with different combinations of modalities, where the proportion of missing modalities varied. The fusion

Table 3: RMSE with different multi-task models. Column 1: Fully-Shared Model with tasks {t_1, t_2}, Column 2: Fully-Shared Model with tasks {t_1, t_2, t_3, t_4}, Column 3: Shared-Private Model with tasks {t_1, t_2, t_3, t_4}

Modalities	*FS* {t_1, t_2}	*FS* {t_1, t_2, t_3, t_4}	*SP* {t_1, t_2, t_3, t_4}
Audio + Meta + Subtitle + Video	0.946	0.991	**0.930**
Meta + Subtitle + Video	0.948	0.976	0.938
Audio + Subtitle + Video	0.941	0.969	0.932
Audio + Meta + Video	0.948	0.975	0.938
Audio + Meta + Subtitle	1.055	0.973	0.933
Subtitle + Video	0.960	0.981	0.933
Meta + Video	0.936	0.964	0.934
Meta + Subtitle	1.141	0.974	0.934
Audio + Video	0.947	0.979	0.932
Audio + Subtitle	1.151	0.974	0.933
Audio + Meta	1.151	0.971	0.932
Audio	1.151	0.969	0.984
Meta	1.151	0.977	0.933
Subtitle	1.151	0.969	0.944
Video	0.966	0.968	0.936

Table 4: RMSE with different multi-task models for MMTF-14K.

Modalities	*FS* {t_1, t_2}	*FS* {t_1, t_2, t_3, t_4}	*SP* {t_1, t_2, t_3, t_4}
Audio + Meta + Subtitle + Video	0.972	0.971	**0.961**
Meta + Subtitle + Video	0.968	0.968	0.965
Audio + Subtitle + Video	0.977	0.976	0.964
Audio + Meta + Video	0.975	0.975	0.970
Audio + Meta + Subtitle	0.981	0.973	0.966
Subtitle + Video	0.971	0.969	0.969
Meta + Video	0.969	0.970	0.967
Meta + Subtitle	0.989	0.977	0.969
Audio + Video	0.989	0.979	0.971
Audio + Subtitle	1.001	0.974	0.970
Audio + Meta	0.991	0.991	0.968
Audio	1.101	0.999	0.981
Meta	0.989	0.988	0.966
Subtitle	0.996	0.995	0.967
Video	0.985	0.978	0.970

Table 5: Result with different percentages of one missing modality for Movielens-100K

χ% Modalities	20%	50%	75%	90%
A(χ%)+M(100%)+S(100%)+V(100%)	1.016	1.008	0.991	**0.966**
A(100%)+M(χ%)+S(100%)+V(100%)	1.054	**1.052**	1.137	1.051
A(100%)+M(100%)+S(χ%)+V(100%)	1.033	1.071	**1.027**	1.038
A(100%)+M(100%)+S(100%)+V(χ%)	1.146	1.138	**1.136**	1.134

Table 6: Result with different percentages of two missing modalities for Movielens-100K

χ% Modalities	20%	50%	75%	90%
A(χ%)+M(χ%)+S(100%)+V(100%)	1.052	1.052	1.046	**1.046**
A(χ%)+M(100%)+S(χ%)+V(100%)	**1.025**	1.071	1.135	1.138
A(χ%)+M(100%)+S(100%)+V(χ%)	**1.135**	1.145	1.136	1.151
A(100%)+M(χ%)+S(χ%)+V(100%)	1.052	1.146	1.051	**1.048**
A(100%)+M(χ%)+S(100%)+V(χ%)	1.151	1.144	1.136	**1.136**
A(100%)+M(100%)+S(χ%)+V(χ%)	**1.136**	1.146	1.147	1.151

method adopted for this task was $CNet_{all}$ and the multi-task architecture was a shared private network having 4 tasks as this architecture gave the best result when all modalities were completely present. The results are the average of five experiments, and the reported RMSE values are statistically significant. Table. 5,8 shows the results for experiments with only one modality missing. Table.6,9 shows for two modalities missing, and similarly Table.7,10 shows for three modalites missing. Note that Table.5,6,7 is for the Movielens-100K dataset and Table.8,9,10 is for the MMTF-14K dataset. The abbreviations A, M, S, and V represent audio, meta, subtitle, and video embeddings, respectively. The experiments showed that the best RMSE of **0.966** and **0.987**, which was achieved when audio was used in 90% of samples and video, meta, and subtitle were used in all samples for the Movielens-100K and the MMTF-14K datasets respectively. The framework demonstrated its effectiveness even when dealing with severely missing modalities cases, showcasing its adaptable nature.

6.4 State-of-the-art methods(RQ4)

We evaluated the effectiveness of our proposed multimodal movie recommendation approach by comparing it to contemporary state-of-the-art methods employed on the MovieLens-100K and MMTF-14K datasets (results are detailed in Table 11). While numerous previous studies [3, 36] have addressed the MovieLens-100K dataset, they have not incorporated its multimodal variant. It's important to note that the baseline techniques featured in Table 11 are indeed based on the multimodal version of the MovieLens-100K and MMTF-14K datasets. Also, the results mentioned in 11 are the ones where we have considered complete modality and not the missing modality cases because the baselines were executed under the modality complete settings. The baselines used here are described as follows:

- **Siamese Network [19]**: It is a traditional neural network-based RS. Considering multi-modal inputs for an item and concatenating with the user vector to perform the regression task of predicting user-item rating.
- **Graph Attention Network [5]**: State-of-the-art model in which instead of relying on the conventional neural network approach, this system utilizes a graph-based recommendation system. In this approach, the authors have generated

Table 7: Result with different percentages of three missing modalities for Movielens-100K

χ% Modalities	20%	50%	75%	90%
A(100%)+M(χ%)+S(χ%)+V(χ%)	1.138	1.144	1.146	**1.137**
A(χ%)+M(100%)+S(χ%)+V(χ%)	**1.136**	1.138	1.138	1.137
A(χ%)+M(χ%)+S(100%)+V(χ%)	1.146	1.151	1.142	**1.139**
A(χ%)+M(χ%)+S(χ%)+V(100%)	1.047	1.045	1.046	**1.041**

Table 8: Result with different percentages of one missing modality for MMTF-14K

χ% Modalities	20%	50%	75%	90%
A(χ%)+M(100%)+S(100%)+V(100%)	1.034	1.028	1.010	**0.987**
A(100%)+M(χ%)+S(100%)+V(100%)	1.062	**1.055**	1.121	1.093
A(100%)+M(100%)+S(χ%)+V(100%)	1.037	1.069	**1.036**	1.044
A(100%)+M(100%)+S(100%)+V(χ%)	1.141	1.146	**1.139**	1.140

Table 9: Result with different percentages of two missing modalities for MMTF-14K

χ% Modalities	20%	50%	75%	90%
A(χ%)+M(χ%)+S(100%)+V(100%)	**1.056**	1.051	1.050	1.050
A(χ%)+M(100%)+S(χ%)+V(100%)	1.063	**1.061**	1.091	1.098
A(χ%)+M(100%)+S(100%)+V(χ%)	1.098	1.099	**1.094**	1.091
A(100%)+M(χ%)+S(χ%)+V(100%)	1.112	**1.111**	1.004	1.008
A(100%)+M(χ%)+S(100%)+V(χ%)	1.114	1.110	1.109	**1.091**
A(100%)+M(100%)+S(χ%)+V(χ%)	1.101	**1.099**	1.122	1.112

Table 10: Result with different percentages of three missing modalities for MMTF-14K

χ% Modalities	20%	50%	75%	90%
A(100%)+M(χ%)+S(χ%)+V(χ%)	1.166	1.160	1.164	**1.159**
A(χ%)+M(100%)+S(χ%)+V(χ%)	1.163	1.167	**1.161**	1.163
A(χ%)+M(χ%)+S(100%)+V(χ%)	1.159	1.157	1.159	**1.151**
A(χ%)+M(χ%)+S(χ%)+V(100%)	1.158	**1.151**	1.157	1.156

rich item features by evaluating item-to-item relationships through the creation of an attention network. The ultimate goal is to perform regression analysis for rating prediction.

- **Graph Convolution Network(GCN) [17]**: It is a state-of-the-art model. In this approach, an item's representation undergoes a process involving Graph Convolutional Networks (GCN) to produce GCN-based embeddings. This approach outperforms previous approximation methods, as explained in the methodology section.
- **Multitasking based Recommendation[21]**: In this, authors have proposed multi-tasking models where they solved two tasks namely rating prediction which was a primary task and genre prediction which is an auxiliary task.

It is evident that the proposed model has achieved superior performance compared to the prevailing state-of-the-art approach in the context of the multi-modal version of the MovieLens-100K and MMTF-14K datasets.

Table 11: Comparison with SOTA

Model	**RMSE**	
	ML-100K	**MMTF-14K**
Siamese[19]	1.028	1.010
GAN[5]	0.953	0.974
GCN[17]	0.951	0.972
Multitasking[21]	0.940	0.963
MTM^4F	**0.930**	**0.961**

7 CONCLUSION

This paper introduces a new framework for multi-modal multi-task learning. Firstly, we have extended the MMTF-14K dataset to incorporate subtitle modality and then used the CentralNet architecture to merge various modalities in an efficient way to solve the recommendation task. We have outperformed the various multi-modal baselines and achieved state-of-the-art performance for both datasets. The best result obtained was from the combination of all the modalities, which shows the impact of multi-modality. We have explored different multi-task architecture and the results obtained clearly shows that considering the correlated tasks has helped in improving the performance of the recommender system. Additionally, we conducted a thorough analysis of the impact of missing modalities. The study showed that the proposed model outperforms previous multi-modal baselines, and the work represents a significant step towards implementing multi-modal learning in situations where some data modes are missing or hard to obtain.

ACKNOWLEDGMENT

The authors express their appreciation to Sony Research India for the support extended during the execution of this research.

REFERENCES

[1] Pawan Agrawal, Subham Raj, Sriparna Saha, and Naoyuki Onoe. 2023. A Meta-learning Based Generative Model with Graph Attention Network for Multi-Modal Recommender Systems. *Procedia Computer Science* 222 (2023), 581–590.

[2] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 370–379. https://doi.org/10.18653/v1/N19-1034

[3] Julio Barbieri, Leandro GM Alvim, Filipe Braida, and Geraldo Zimbrão. 2017. Autoencoders and recommender systems: COFILS approach. *Expert Systems with Applications* 89 (2017), 81–90.

[4] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions* 7, 1 (2014), 1525–1534.

[5] Daipayan Chakder, Prabir Mondal, Subham Raj, Sriparna Saha, Angshuman Ghosh, and Naoyuki Onoe. 2022. Graph Network based Approaches for Multi-modal Movie Recommendation System. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 409–414.

[6] Zeyu Cui, Feng Yu, Shu Wu, Qiang Liu, and Liang Wang. 2021. Disentangled item representation for recommender systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 2 (2021), 1–20.

[7] Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018. MMTF-14K: a multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference*. 450–455.

[8] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[9] Andreas F Haas, Marine Guibert, Anja Foerschner, Sandi Calhoun, Emma George, Mark Hatay, Elizabeth Dinsdale, Stuart A Sandin, Jennifer E Smith, Mark JA Vermeij, et al. 2015. Can we measure beauty? Computational evaluation of coral reef aesthetics. *PeerJ* 3 (2015), e1390.

[10] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[11] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[12] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[14] Xingchen Li, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua. 2020. Hierarchical fashion graph network for personalized outfit recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 159–168.

[15] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Why I like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 4–12.

[16] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[17] Prabir Mondal, Daipayan Chakder, Subham Raj, Sriparna Saha, and Naoyuki Onoe. 2023. Graph Convolutional Neural Network for Multimodal Movie Recommendation. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. 1633–1640.

[18] Gaurav Pandey and Ambedkar Dukkipati. 2017. Variational methods for conditional multimodal deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 308–315.

[19] Sriram Pingali, Prabir Mondal, Daipayan Chakder, Sriparna Saha, and Angshuman Ghosh. 2022. Towards developing a multi-modal video recommendation system. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[21] Subham Raj, Prabir Mondal, Daipayan Chakder, Sriparna Saha, and Naoyuki Onoe. 2023. A Multi-modal Multi-task based Approach for Movie Recommendation. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[22] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* (2015).

[23] Klaus Seyerlehner, Gerhard Widmer, Markus Schedl, and Peter Knees. 2010. Automatic music tag classification based on block-level. *Proceedings of Sound and Music Computing* (2010).

[24] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass. 2013. Unsupervised methods for speaker diarization: An integrated and iterative approach. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 10 (2013), 2015–2028.

[25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28 (2015).

[26] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891* (2016).

[27] Robin Van Meteren and Maarten Van Someren. 2000. Using content-based filtering for recommendation. In *Proceedings of the machine learning in the new information age: MLnet/ECML2000 workshop*, Vol. 30. 47–56.

[28] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic cooccurrences. In *Proceedings of the IEEE International Conference on Computer Vision*. 4642–4650.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations* (2018).

[30] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. CentralNet: a Multilayer Approach for Multimodal Fusion. arXiv:1808.07275 [cs.AI]

[31] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*. 165–174.

[32] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.

[33] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*. 1437–1445.

[34] Mike Wu and Noah Goodman. 2018. Multimodal generative models for scalable weakly-supervised learning. *Advances in Neural Information Processing Systems* 31 (2018).

[35] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.

[36] Zahra Zamanzadeh Darban and Mohammad Hadi Valipour. 2022. GHRS: Graph-based hybrid recommendation system with application to movie recommendation. *Expert Systems with Applications* 200 (2022), 116850. https://doi.org/10.1016/j.eswa.2022.116850

[37] Mengqi Zhang, Shu Wu, Meng Gao, Xin Jiang, Ke Xu, and Liang Wang. 2020. Personalized graph neural networks with attention mechanism for session-aware recommendation. *IEEE Transactions on Knowledge and Data Engineering* 34, 8 (2020), 3946–3957.