MFMGC: A Multi-modal Data Fusion Model for Movie Genre Classification

Xiaorui Yang, Qian Zhou, Wei Chen(✉), and Lei Zhao

School of Compute Science and Technology, Soochow University, Suzhou, China
xryang@stu.suda.edu.cn, {qzhou0,robertchen,zhaol}@suda.edu.cn

Abstract. Movie Genre Classification (MGC) is a classic multi-label task that aims to classify movies into different genres. Existing studies have proposed many approaches for this task based on multi-modal data (e.g., synopsis, posters, and trailer). Despite the significant contributions made by them, they usually fuse multi-modal information based on simple operations, e.g., concatenation or weighted sum, failing to effectively capture the interactive information between multi-modal data. In addition, movies with significant overlap in directors and actors tend to own the same genres. This information could potentially improve the performance of MGC, which has been ignored by previous studies. Having observed the shortcomings of existing work, we propose a **M**ulti-modal data **F**usion Model for **MGC** (MFMGC), including two modules: Multi-modal Data Fusion (MDF) and Movie Graph Representation Learning (MGRL). In MDF, we carefully design the fusion layer based on the attention mechanism to effectively capture the modalities' interactive information. In MGRL, we construct a movie graph to extract the structural information between movies. Specifically, the graph is constructed based on the overlap of movies' directors, screenwriters, and actors, and each node in the graph has multi-modal attributes. The experiments conducted on datasets Moviescope and MovieBricks demonstrate the superiority of the proposed model MFMGC over the state-of-the-art approaches.

Keywords: Movie genre classification · Multi-modal movie graph · Movie representation learning

1 Introduction

Over the past decade, the streaming media services have experienced unprecedented growth. Recommending specialized types of content for customers has become an indispensable ability for streaming sites, which is why automatic labeling has attracted increasing attention in recent advances. Especially, the task of Movie Genre Classification (MGC), which is an important branch of automatic labeling and has a wide range of applications (e.g., organizing user videos from social media sites, correcting mislabeled videos, and recommending specific types of films for users), has been paid significant efforts by existing work.

X. Yang et al. (Eds.): ADMA 2023, LNAI 14177, pp. 676–691, 2023.
https://doi.org/10.1007/978-3-031-46664-9_45

Specifically, the task aims to classify movies into different genres and is suffering from new challenges, due to the emerging of more and more movie-related multi-modal information, and the diverse demands of consumers.

Despite the significant contributions on multi-modal data-based MGC made by existing work [1,6,14,25], they usually fuse the features of different modalities via concatenation [4,27] or weighted sum [1,8,25], failing to capture the semantic information contained by multi-modal data effectively. Additionally, the existing studies ignore the movies' metadata (e.g., directors and actors) that is of critical importance to a high performance MGC method. By way of illustration, given a movie and its sequels, they usually share the same directors or main actors and are more likely to have the same genres, compared with other different movies. This information can be effectively exploited by constructing a movie graph and extracting structural features from it. In a nutshell, there remains great scope for further improving the performance of existing MGC approaches due to the following problems: *Problem* 1) the multi-modal fusion strategies of existing studies cannot effectively explore the semantic information of multi-modal data; *Problem* 2) the movies' metadata, which involves abundant structural information, has been ignored by existing work.

Having observed the limitations of above-mentioned studies, we propose a novel model namely MFMGC[1] that is composed of two modules: MDF (Multi-modal Data Fusion) and MGRL (Movie Graph Representation Learning). In detail, the module MDF is designed to address *Problem* 1). Different from most existing studies that rely on late fusion strategies, MDF utilizes the attention mechanism to fuse multi-modal data during the feature extraction process. To be specific, there are two main attention layers in MDF, which are used for exploring the semantic features contained by movie-related multi-modal data. Inspired by VLBert [22], which takes the embeddings of both words in a sentence and region-of-interest (RoI) from images as inputs and utilizes the Transformer encoder to model dependencies among all the input elements, the first attention layer feeds the text and video frames into the Transformer encoder for text-video feature extraction. Then, the second modal attention layer is designed to fuse features of different modalities. In addition, the module MGRL is developed to tackle *Problem* 2). A movie graph is constructed based on the overlap of directors, screenwriters, and actors. Each movie is represented as a node in the graph and has multi-modal representations, which are obtained by fusing the movie-related multi-modal attributes with the module MDF. Next, a Graph Convolutional Network (GCN)-based architecture is applied to capture the structural information between movie nodes. Ultimately, a classification layer is employed to predict the genres of movies.

To fully evaluate the effectiveness of our proposed model MFMGC, the extensive experiments on real-world datasets are very essential. However, most of datasets used in previous studies are either not publicly available or have incomplete data [5,18]. Consequently, apart from the open dataset Moviescope [8], we construct a new multi-modal movie dataset called MovieBricks[1] from

[1] Code and data are available at https://anonymous.4open.science/r/mgc.

Douban, which is the most active online movie database and review platform in China. Specifically, the dataset MovieBrick contains 4063 European and American movies released from 2000 to 2019.

The contributions of this paper are summarized as follows:

- We propose a novel model MFMGC to further improve the performance of existing work on the task MGC, by fully exploring the semantic features involved in movie-related multi-modal data and the structure information between movies.
- Two modules are developed in MFMGC, i.e., MDF and MGRL. The module MDF is designed to tackle *Problem* 1), by capturing the interactive information between different modalities with novel fusion layers. The module MGRL is developed to address *Problem* 2), by extracting structure information from the movie graph that is constructed based on the overlap of movies' directors, screenwriters, and actors.
- We conduct extensive experiments on two real-world datasets, i.e., Moviescope and MovieBricks. Particularly, MovieBricks is the first multi-modal movie dataset in China, comprising over 4000 movies with four different modalities, including synopsis, poster, trailer, and metadata. The results demonstrate the superior performance of the proposed model MFMGC compared with the state-of-the-art methods.

The rest of the paper is organized as follows. The related work is presented in Sect. 2 and the task MGC is formulated in Sect. 3. The proposed model MFMGC is introduced in Sect. 4. We report the experimental results in Sect. 5, which is followed by the conclusion in Sect. 6.

2 Related Work

2.1 Research on Movies

Due to its rich storytelling and high-quality footage, the movies have become a valuable resource for researchers. Current studies on movies can be categorized into three directions: analyzing the content of movies, examining the impact of movies, and studying the characteristics of movies. Researches on movie content mainly use movie trailers as video data, e.g., scene boundary detection [9,20], which aims to divide a video into easily interpretable parts to communicate a storyline effectively, and action recognition [21,26] which utilizes the video scripts that exist for thousands of movies to automatically extract and track faces together with corresponding motion features. Studies on movie influence include movie box office prediction [16,28] and movie review analysis [13,23]. The movie box office prediction before its theatrical release can decrease its financial risk, and movie review analysis is a task of Natural Language Processing, which is able to obtain the emotional or semantic information of the movie's review. In addition, the studies on movie characteristics include understanding the relationships of movie characters [3,15], which aims to weigh the importance of character in defining a story, and movie genre classification [1,4,7,8,18,25].

2.2 Movie Genre Classification

To better contextualize our study, we review existing work focusing on multi-modal data, with particular emphasis on fusion strategies, and introduce them in a chronological order.

Wehrmann et al. [25] propose a novel deep neural architecture called CTT-MMC for multi-label movie-trailer genre classification. The authors utilize both video and audio data, and the fusion strategy involves a Maxout layer before the class prediction, which can be interpreted as a late fusion strategy. John et al. [1] propose a novel model for multi-modal learning based on gated neural networks for MGC. They utilize the plot and poster data for the classification task. The gated mechanism is used to obtain the weights of different modalities and then weight sums them for the final classification. The model is also utilized in other work such as [7]. Cascante et al. [8] compare the effectiveness of visual, audio, text, and metadata-based features in predicting movie genres. They utilize trainable parameters to sum different features of multi-modal data. Behrouzi et al. [4] design a new structure based on Gated Recurrent Unit (GRU) to extract spatial-temporal features from the movie-related data. The authors concatenate the video and audio features to predict the final genres of movies. Mangolin et al. [18] extract features by computing different kinds of descriptors, and then combine classifiers through the calculation of predicted score for each class, and they propose three rules for fusion, i.e., Sum, Prod, and Max.

In summary, current research on MGC with multi-modal data mainly utilizes late fusion strategies, such as concatenation and weighted sum, failing to capture the interaction between different modalities, and they ignore the structural information contained by metadata. To further improve the performance of their designed methods, we propose the novel model MFMGC in this study.

3 Problem Formulation

Given a set of movies $\{M_1, \cdots, M_N\}$, each movie M_i is associated with multi-modal attributes and metadata, i.e., $M_i = \{M_i^t, M_i^p, M_i^v, M_i^a, M_i^m\}$. Detailedly, M_i^t denotes the textual data consisting of the movie's title and synopsis, M_i^p represents the movie's poster, M_i^v denotes the visual data that is a sequence of frame-level patches in the trailer, M_i^a represents the audio fragments extracted from the trailer, and M_i^m is the metadata of the movie. Moreover, $C = \{c_1, \cdots, c_L\}$ is the genre set, where L is the number of movie genres.

Intuitively, movies with a significant overlap of directors, screenwriters, and actors may belong to the same genre, and a corresponding example is presented in Fig. 1. To capture such information effectively, we construct a multi-modal movie graph. Specifically, the graph is denoted as $G = \{V, E\}$, where V is the set of movie nodes, i.e., $V = \{M_1, \cdots, M_N\}$, and E is the set of connections between each pair of movie nodes. Additionally, we design an adjacency matrix A for the edge set E, where A_{ij} represents whether there is an edge between M_i and M_j. Given a threshold $\mathcal{T}$, if the overlap of directors, screenwriters, and actors between M_i and M_j exceeds $\mathcal{T}$, A_{ij} is set to 1. Otherwise, A_{ij} is set to 0.

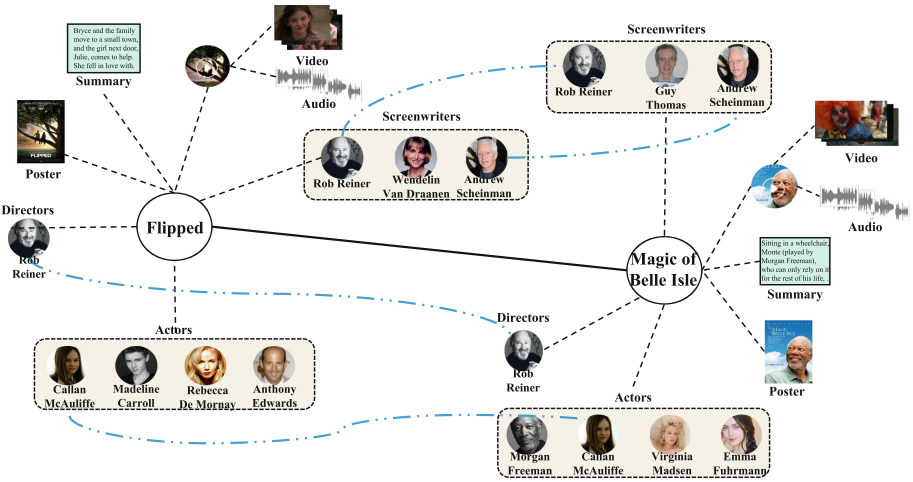

Fig. 1. A multi-modal movie graph, where each node has four multi-modal attributes, i.e., text, poster, video, and audio. Different movie nodes are connected according to the overlap of their directors, screenwriters, and actors.

Definition 1 ***Movie Genre Classification*** **(*MGC*).** *Given a movie M_i from the dataset $\{M_1, \cdots, M_N\}$ and a genre set C, the task of MGC aims to learn a function Φ to predict the genres of movie M_i based on M_i^t, M_i^p, M_i^v, M_i^a, and M_i^m. This process is formulated as follows:*

$$P_i = \Phi(M_i^t, M_i^p, M_i^v, M_i^a, M_i^m), \tag{1}$$

where $P_i = \{c_x, \cdots, c_y\}$ is the set of genres assigned to the movie M_i and each genre in $\{c_x, \cdots, c_y\}$ is from C.

Note that MGC is a multi-label classification task [27] and each movie may belong to multiple genres at the same time. For instance, the movie "X-Men: The Last Stand" has multiple genres, i.e., *Action*, *Horror*, and *Sci-Fic*.

4 Proposed Model

4.1 Overview

To effectively utilize the multi-modal data of movies to conduct MGC, we propose a novel model namely MFMGC. Observed from Fig. 2, the model contains two modules, i.e., Multi-modal Data Fusion (MDF) and Movie Graph Representation Learning (MGRL), and the details of them are as follows.

To feed the movie data into the module MDF, we first segment the audio and frame the video into patches at the frame level, and then use different pre-trained models to embed the text, posters, video frames, and audio segments. Next, these embeddings are fed into MDF, which consists of two stages. Specifically, in the

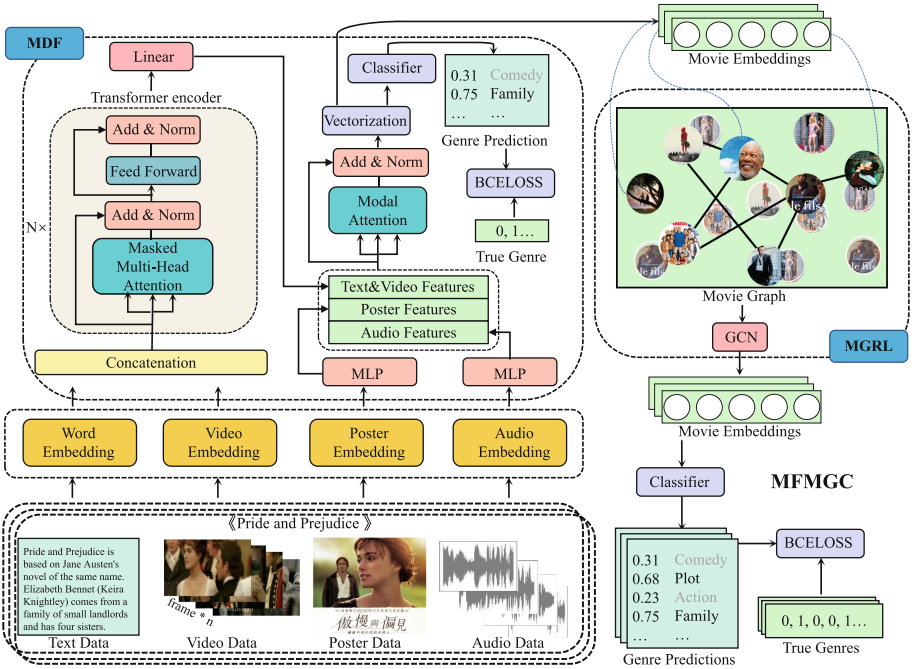

Fig. 2. Overview of the proposed model MFMGC

first stage, feature pre-extraction is performed. For the embeddings of text and video frames, we take them as input and utilize the Transformer encoder as the backbone to fuse text and video modalities, inspired by VLBert [22] that feeds both words in the sentence and region-of-interest (RoI) from the image into the Transformer encoder. For the posters and audio data, we separately design multi-layer perceptron (MLP) layers to perform the feature pre-extraction. In the second stage, we adopt a modal-attention layer to fuse extracted features, ensuring that the multi-modal data could be effectively integrated, resulting in a comprehensive representation of each movie. In MGRL, we deploy a GCN-based architecture to fine-tune the movie representations obtained from MDF and extract structural information between movie nodes.

4.2 Multi-modal Data Embedding

This section details the embedding process of the synopsis, poster, trailer, and audio data. We introduce how to transform these data into a suitable format and then feed the module MDF for feature pre-extraction and fusion.

Text Embedding. We utilize a Transformer Encoder structure to extract text features, where the text data is embedded by the Bert Embedding [12] module. Specifically, given the textual data M_i^t of the movie M_i, M_i^t contains a token

sequence, which is denoted as $\{w_1, w_2, \cdots, w_l\}$ and l is the number of tokens in M_i^t. The pre-trained BertEmbedding module is used to obtain the token sequence's embedding and the process is formally defined as:

$$E_i^t = BertEmbed(M_i^t), \tag{2}$$

where $BertEmbed(\cdot)$ is the Bert Embedding module and $E_i^t \in \mathbb{R}^{l \times h^t}$ is the embedding of M_i^t.

Video Embedding. To obtain valuable information from the video data of the movie, we first extract frames at a rate of one frame per second (FPS). The extracted frames are then processed to obtain high-level dimensional features based on the Swin Transformer. Specifically, the visual data M_i^v of movie M_i consists of p video frames, and we use the following method to embed it:

$$E_i^v = SwinSmall(M_i^v), \tag{3}$$

where $SwinSmall(\cdot)$ is one of Swin Transformer model [17], $E_i^v \in \mathbb{R}^{p \times h^v}$ is the embedding of video frames of the i-th movie, and each frame is embedded to a vector with the dimension of h^v.

Poster Embedding. In addition to video data, posters are also important visual data for movies, containing rich information about the movie's genre to attract audiences with specific preferences. We feed the poster into the Swin Transformer to obtain its embedding and the process can be formally defined as:

$$E_i^p = SwinSmall(M_i^p), \tag{4}$$

where M_i^p is the poster data, and $E_i^p \in \mathbb{R}^{h^v}$ is the poster embedding of the i-th movie.

Audio Embedding. Apart from the above-mentioned information, we also extract features from audio, since different genres of movies usually have different types of soundtracks. For instance, while both *Comedy* and *Action* genres may have visually bright scenes, the background music of *Comedy* movies tends to have a more cheerful instead of intense rhythm. To capture latent features from the audio, we learn corresponding embeddings according to Wav2Vec2 [2]. The audio data is denoted as $M_i^a = \{o_1, o_2, \cdots, o_u\}$, where o_j is the j-th fragment of the given audio, with a sample rate of 16000, and each fragment is a 3-second audio signal. Note that we adopt a mean pooling operation to obtain the audio embedding from the embeddings of fragments, and the process is as follows:

$$E_i^a = MP(Wav2Vec2(M_i^a)), \tag{5}$$

where $Wav2Vec2(\cdot)$ is a Wav2Vec2 layer, $MP(\cdot)$ is the mean pooling operation, and $E_i^a \in \mathbb{R}^{h^a}$ is the audio embedding of the i-th movie.

Ultimately, the embedding of the i-th movie's multi-modal data can be represented as $\mathcal{E}_i = \{E_i^t, E_i^v, E_i^p, E_i^a\}$, which is then fed into the module MDF.

4.3 Multi-modal Data Fusion - MDF

The attention mechanism in Transformer has been proven powerful and flexible to differentially weigh the significance of each part of the input data. In MDF, we utilize this mechanism to fuse multi-modal embeddings, which involve two stages. In the first stage, the Transformer Encoder and MLP are used to extract latent features from different input embeddings. In the second stage, we adopt a modal-attention layer to fuse the features extracted at the first stage.

Feature Extraction of MDF. The Transformer Encoder is particularly effective in extracting sequential features, making it suitable for processing text and video frames. Specifically, in MDF, we first concatenate the embeddings of text and video frames as $\mathcal{E}_i^{tv} = E_i^t \| E_i^v$, where $\|$ denotes the concatenation operation, and $\mathcal{E}_i^{tv} \in \mathbb{R}^{(l+p)\times h^t}$. Then, the concatenated embedding $\mathcal{E}_i^{tv}$ is fed into the fusion module, which consists of a Transformer encoder [24] and a Mean pooling layer. The calculation process is formulated as follows:

$$O_i^{tv} = MP(TransEncoder(E_i^{tv})), \tag{6}$$

where $TransEncoder(\cdot)$ denotes Transformer Encoder.

For poster and audio embeddings, we employ two multi-layer perceptron (MLP) layers to extract their features respectively. The MLP layer consists of two fully connected layers with a ReLU activation function in the middle. The process can be formulated as follows:

$$O_i^p = ReLU(E_i^p W_1^p + b_1^p) W_2^p + b_2^p, \tag{7}$$

$$O_i^a = ReLU(E_i^a W_1^a + b_1^a) W_2^a + b_2^a, \tag{8}$$

where $E_i^{p/a}$ denotes the embedding of posters or audio, $W_1^{p/a}$ and $W_2^{p/a}$ are the weight matrices of the two fully connected layers, $b_1^{p/a}$ and $b_2^{p/a}$ are biases, and ReLU$(\cdot)$ is the Rectified Linear Unit activation function. After the features are extracted, the set of representations $\mathcal{O}_i = \{O_i^{tv}, O_i^p, O_i^a\}$ is obtained.

Modal-Attention Layer of MDF. Following the feature extraction, we apply modal attention to fuse the features of different modalities. Specifically, the feature of text-video O_i^{tv} is first transformed into a new vector $\tilde{O}_i^{tv}$ that has the same dimension with the poster and audio embeddings, through a Linear layer. Then the multi-modal input features are first concatenated to obtain $\hat{\mathcal{O}}_i \in \mathbb{R}^{m\times h}$, where m is the number of features in $\mathcal{O}_i$. Next, the query matrix $Q_i = \hat{\mathcal{O}}_i W_q$ is obtained through the projection matrix W_q, while the key matrix K_i and value matrix V_i are obtained using W_k and W_v, respectively. The scaled dot product function is used as the attention function, and the inter-modal attention matrix P_i is obtained with following method,

$$P_i = softmax(\frac{Q_i K_i^T}{\sqrt{h}}), \tag{9}$$

where $P_i \in \mathbb{R}^{m \times m}$ and each element $P_{i,xy}$ of the matrix represents the inter-modal attention between the x-th and y-th modality of the i-th movie M_i. Then, the multi-modal representation of M_i, which is denoted as F_i, is obtained through attention aggregation and the map function $\mathcal{V}$. Additionally, a residual connection is added to avoid the problem of vanishing gradients during training, and the process can be represented as follows:

$$F_i = \mathcal{V}(P_i V_i + \mathcal{O}_i), \tag{10}$$

where $\mathcal{V}(\cdot)$ denotes the vectorization by row-wise concatenation, and $F_i \in \mathbb{R}^{1 \times mh}$. Finally, we obtain $\mathcal{F} = \{F_1, F_2, \cdots, F_N\}$, which contains the multi-modal representations of all movies in the given dataset.

4.4 Movie Graph Representation Learning - MGRL

To fully explore the structural and semantic information of movies in a unified manner, we construct a multi-modal movie graph based on movies' directors, screenwriters, and actors. Here, the movie nodes have fused representations that are obtained in MDF based on movie-related multi-modal attributes, i.e., synopsis, poster, and trailer. To effectively extract structural information from the graph, we adopt a two-layer GCN to fine-tune the movie representations and the process is as follows:

$$\mathcal{H} = GCN(\mathcal{F}, A) = ReLU(\tilde{A}ReLU(\tilde{A}\mathcal{F}W^0)W^1), \tag{11}$$

where $\mathcal{H} = \{H_1, H_2, \cdots, H_N\}$ denotes the new set of movie representations, $H_i (1 \leq i \leq N)$ is the fine-tuned embedding of movie M_i. A is the adjacency matrix of the movie graph and $\tilde{A} = \tilde{D}^{-\frac{1}{2}}(A + I_N)\tilde{D}^{-\frac{1}{2}}$. I_N is the identity matrix with size $N \times N$, where N denotes the number of movies in the graph. $\tilde{D}$ is the diagonal degree matrix of $\tilde{A}$, which is defined as $\tilde{D}_{ii} = \sum_{j=1}^{N} \tilde{A}_{ij}$. W^0 and W^1 are learnable parameters.

4.5 Classification Layer

Ultimately, to tackle the task of MGC, we use a linear projection followed by a sigmoid function to predict the movie's genre. This can be formally defined as:

$$\mathcal{S}^1/\mathcal{S}^2 = Sigmoid(Linear(\mathcal{F}/\mathcal{H})), \tag{12}$$

where $Sigmoid(\cdot)$ is the activation function that is used to squash the output vector values to range $[0, 1]$, which can be interpreted as the vector of genre probability. Note that, as there has been no work constructed above-mentioned movie graph, to give a more fair comparison, the input of the classification layer can be either $\mathcal{F}$ or $\mathcal{H}$. Consequently, the output can be either $\mathcal{S}^1$ or $\mathcal{S}^2$. Taking $\mathcal{S}^1 = \{S_1, S_2, \cdots, S_N\}$ as an example, $S_i \in \mathcal{S}^1$ is the genre probability vector of the i-th movie M_i, which is denoted as $S_i = \{s_{i1}, \cdots, s_{iL}\}$. Here, $s_{ij} \in S_i$ represents the probability that M_i belongs to the j-th genre, and L is the number of genres.

4.6 Training

The model is optimized by Binary Cross-Entropy Loss (BCELoss). The labels of movies are first embedded and denoted as $\mathcal{C} = \{C_1, C_2, \cdots, C_N\}$. For the i-th movie, the genres set is $C_i = \{c_{i1}, c_{i2}, \cdots, c_{iL}\}$, where $c_{ij} \in \{0, 1\}$, and $c_{ij} = 1$ indicates that the i-th movie belongs to the j-th genre. The formulation for the loss function is as follows:

$$\begin{aligned} \mathcal{L} &= BCELoss(\mathcal{C}, \mathcal{S}) \\ &= -\frac{1}{L}\sum_{i=1}^{N}\sum_{j=1}^{L}(c_{ij}\log(s_{ij}) + (1 - c_{ij})\log(1 - s_{ij})), \end{aligned} \tag{13}$$

where N is the number of movies.

In addition, when adding the module MGRL, we adopt a joint loss function to guide the optimization of both MDF and MGRL:

$$\mathcal{L} = BCELoss(\mathcal{C}, \mathcal{S}^1) + BCELoss(\mathcal{C}, \mathcal{S}^2). \tag{14}$$

5 Experiments

5.1 Dataset

Most of the datasets used in current research are either not open or the access paths have expired, particularly for datasets that contain multiple data sources such as synopses, posters, trailers, and metadata. We start with downloading the dataset Moviescope, which contains movies' synopsis, posters, and URLs of trailers on YouTube, and then develop a Python crawler to obtain the trailers. To enable a more comprehensive evaluation of our model, we create a new dataset from Douban, the most active online movie review and dataset platform in China. The details of the two datasets are as follows.

The dataset Moviescope, all data sources of which are available, contains 4076 movies with 13 different genres. Additionally, the dataset MovieBricks has 4063 movies with 10 different genres, namely *Action*, *Thriller*, *Adventure*, *Story*, *Science-Fiction*, *Love*, *Fantasy*, *Comedy*, *Terror* and *Crime*. Both two datasets are divided into training, validation, and testing sets in a 7:1:2 ratio. Note that a movie may belong to multiple genres at the same time.

5.2 Comparison Method

To validate the effectiveness of MFMGC, we compare its performance with those of several state-of-the-art approaches that are introduced as follows.

- **GMU** [1]. This work develops a model for multi-modal learning based on gated neural networks, which is evaluated on a multi-label scenario for MGC using synopses and posters.

- **Fast-MA** [8]. This work designs a temporal feature aggregator to embed video and text, and compares the effectiveness of visual, textual-based methods on MGC, and it is denoted as Fast Modal Attention (Fast-MA).
- **DL-PO** [19]. This work proposes a simple deep-learning model to predict the genres of a movie with overview and poster. We refer to it as Deep Learning for Posters and Overviews (DL-PO).
- **CMM** [18]. This is a comprehensive study developed in terms of diversity of multimedia sources of information to perform MGC. We refer to it as Comprehensive Multi-modal Model (CMM).
- **MGC-RNN** [4]. This work proposes a new structure based on GRU to derive spatial-temporal features of movie frames and then concatenates them with the audio features to predict the final genres of the movie. We refer to it as MGC-RNN.

5.3 Evaluation Metrics and Parameter Settings

Evaluation Metrics. AUC-ROC [11] is a well-known metric that measures the area under the receiver operating characteristic (ROC) curve. This curve plots the true positive rate against the false positive rate for each possible threshold of the classifier's output. However, relying on a single metric cannot provide a comprehensive evaluation of a multi-label classifier. Therefore, we also calculate the F1 score that has been widely used to evaluate the multi-label classifiers [1,4]. To globally evaluate the performance of different methods, we compute the micro and macro averages of the F1 and AUC metrics. The micro-average calculates the mean of scores without considering genres, while the macro-average computes the score of each genre independently and takes their unweighted means.

Parameter Settings. As mentioned in Sect. 4.2, for the textual modality M^t, a fixed sequence length $l = 256$ is used. For the video modality M^v, we draw $p = 32$ frames from the trailer, and for the audio modality M^a, the number of audio segments is $u = 16$, and the hidden dimension h in our model is set to 256. To reduce the impact of random noise, all experiments are conducted using the 5-fold cross-validation. The results reported are the average of 5 runs using different data partitions. The pre-trained model "Roberta" [10] is utilized to initialize the Transformer Encoder module. To maintain the learned knowledge of pre-trained parameters, we split the learnable parameters into two parts: the learning rate for pre-trained initialized parameters is set to 0.00005, while the learning rate for randomly initialized parameters is 0.0005, and they are denoted as "pre-lr" and "rand-lr" respectively.

5.4 Experiment Results

Experiments are done on a machine with 2 NVIDIA V100 GPUs. The performances are presented in Table 1. As all baselines are designed without considering movie graph, to provide a fair comparison, we present the results of our model's

simplified version MFMGC-P that only utilize partial input data, i.e., synopsis, poster, and trailer of movies. Moreover, MFMGC represents the model that considers the movie's metadata and multi-modal data along with the movie graph. Note that, as the movies from Moviescope only contain few metadata, the movie

Table 1. Experimental results. The used information contains Text (T), Poster (P), Audio (A), Video (V), and Movie Graph (G). Furthermore, "ma" and "mi" are used to represent the macro and micro averages.

Model	Modality	Moviescope				MovieBricks			
		ma-f1	mi-f1	ma-auc	mi-auc	ma-f1	mi-f1	ma-auc	mi-auc
GMU	T	0.5614	0.6158	0.8470	0.8646	0.5563	0.6126	0.8427	0.8657
	P	0.4441	0.5203	0.7425	0.7943	0.4655	0.5371	0.7753	0.8115
	TP	0.5821	0.6328	0.8560	0.8721	0.5947	0.6291	0.8590	0.8748
Fast-MA	T	0.5588	0.6145	0.8459	0.8642	0.5499	0.6063	0.8381	0.8643
	P	0.4102	0.5107	0.7265	0.7727	0.4002	0.5361	0.7339	0.7854
	V	0.4786	0.5492	0.7727	0.8193	0.4832	0.5496	0.7778	0.8160
	TPV	0.6203	0.6497	0.8762	8872	0.5749	0.6300	0.8625	0.8763
DL-PO	T	0.5475	0.5964	0.8415	0.8569	0.5488	0.5945	0.8427	0.8600
	P	0.4201	0.5034	0.7258	0.7809	0.4362	0.524	0.7481	0.7985
	TP	0.5739	0.6251	0.8554	0.8775	0.5818	0.6401	0.8616	0.8857
CMM	T	0.5564	0.6059	0.8416	0.8614	0.5525	0.6041	0.8534	0.8750
	P	0.3548	0.5035	0.6700	0.7478	0.4507	0.5162	0.7113	0.7563
	V	0.4845	0.5817	0.8135	0.8483	0.5219	0.5982	0.8267	0.8535
	A	0.4960	0.5642	0.7985	0.8321	0.4959	0.5693	0.7959	0.8361
	TPVA	0.5588	0.6439	0.8760	0.8916	0.5594	0.6424	0.8754	0.8945
MGC-RNN	V	0.4760	0.5431	0.7666	0.8180	0.4693	0.5424	0.7531	0.8032
	A	0.4957	0.5635	0.8007	0.8306	0.4858	0.5603	0.7897	0.8303
	VA	0.5106	0.5719	0.7886	0.8491	0.5254	0.5981	0.8134	0.8480
MFMGC-P	T	0.5937	0.6364	0.8483	0.8730	0.6531	0.6834	0.8714	0.8836
	P	0.5241	0.5884	0.7895	0.8283	0.5564	0.6187	0.8256	0.8569
	V	0.5412	0.5756	0.8605	0.8799	0.5125	0.5824	0.8157	0.8446
	A	0.4695	0.5247	0.7919	0.8202	0.5072	0.5727	0.7977	0.8363
	TP	0.6347	0.6750	0.8806	0.8976	0.6714	0.7110	0.8966	0.9135
	TV	0.6436	0.6757	0.8801	0.8971	0.6653	0.6981	0.8850	0.8979
	TA	0.6155	0.6659	0.8693	0.8925	0.6578	0.6970	0.8867	0.9015
	PV	0.6054	0.6529	0.8590	0.8875	0.5740	0.6315	0.8355	0.8641
	PA	0.5419	0.6048	0.8159	0.8535	0.5670	0.6345	0.8335	0.8657
	VA	0.5522	0.5860	0.8670	0.8865	0.5889	0.6399	0.8672	0.8880
	TPV	0.6533	0.6859	0.8905	0.9059	0.6843	0.7163	0.9024	0.9155
	TPA	0.6421	0.6878	0.8871	0.9051	0.6765	0.7111	0.9080	0.9201
	TVA	0.6514	0.6933	0.8836	0.9090	0.6702	0.7024	0.8978	0.9050
	PVA	0.6210	0.6662	0.8648	0.8931	0.6038	0.6631	0.8662	0.8907
	TPVA	**0.6600**	**0.6947**	**0.8914**	**0.9065**	**0.6925**	**0.7248**	**0.9046**	**0.9165**
MFMGC	T+G	–	–	–	–	0.6582	0.6935	0.8857	0.9088

graph cannot be constructed, thus MFMGC has no result on this dataset. In addition, we only present the results of MFMGC on MovieBricks when utilizing text data and movie graph, due to the space limitation. More results of the model from P+G, V+G, A+G to TPVA+G are presented on github[1].

Main Results. Observed from Table 1, our proposed model MFMGC-P consistently achieves better performance than baselines. Specifically, the "mi-f1" score of it outperforms GMU by 11.6% on MovieBricks. Even only using poster or video data, MFMGC-P still performs better than other methods, indicating its powerful feature extraction capability. When all multi-modal attributes are taken into account, MFMGC-P achieves higher improvements, which demonstrates the effectiveness of the carefully designed fusion strategy based on attention mechanism. Detailedly, the reasons for the above-mentioned observations are as follows: 1) MFMGC-P utilizes advanced pre-trained models to embed data, the parameters hold abundant knowledge, especially for text and images, which leads to better representations than the traditional models such as Word2vec and VGG. 2) We fuse the multi-modal data via the attention mechanism that differentially weighs the significance of each part of the input data, allowing MFMGC-P to learn a comprehensive representation.

Modality Analysis. To fully investigate the impacts of different modalities on the performance of the proposed model, we compare the results of MFMGC-P when adding single, two, three, and all four modalities. Seen from Table 1, the results of the second part (i.e., TP, TV, TA, PV, PA, and VA) of MFMGC-P outperform those of single-modal data based experiments (i.e., T, P, V, and A). Without surprise, when taking four modalities (i.e., TPVA) into account, MFMGC-P achieves the best performance. These observations demonstrate the effectiveness of the developed module in extracting multi-modal features. Additionally, the higher performance of MFMGC (i.e., T+G) than that of MFMGC-P (i.e., T) demonstrates the significance of the construction of movie graph that can be used to extract structural information between movies.

5.5 Parameter Analysis

To investigate the effect of different learning rates and compare the experimental results with above-mentioned ones more intuitively, the performances of MFMGC-P with varying "pre-lr" and "rand-lr" are reported in Fig. 3(a) and Fig. 3(b). Observed from this, the evaluation metrics present a overall downward trend, where the "ma-f1" score even drops by 21.4%. The reason for the decrease is that setting a lower "pre-lr" can avoid forgetting the knowledge contained by the pre-trained parameters. Additionally, given a too-large "rand-lr", it will lead to faster convergence but makes the model difficult to achieve the best result, as the global optimum may be missed during the iteration.

Furthermore, we analyze the effect of the hidden dimension h, and the results are reported in Fig. 3(c). While varying h from 64 to 512, we first observe the

increase of evaluation metrics, then they present a decreasing tendency, and the best performance is achieved when $h = 256$. Given a too small dimension, the model cannot learn enough information, leading to under-fitting. Conversely, when h is too large, it may introduce unexpected noisy information, resulting in poor performance.

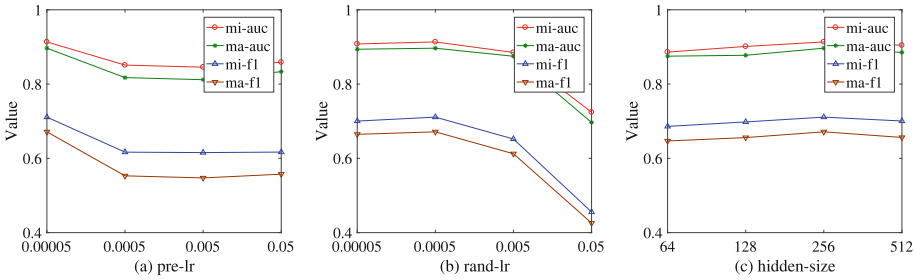

Fig. 3. Parameter analysis for "pre-lr", "rand-lr", and the hidden dimension h.

6 Conclusion

We propose MFMGC, which is a novel model for the task of MGC that utilizes the movie's synopsis, poster, trailer and metadata, and the model comprises two modules: MDF and MGRL. MDF leverages the attention mechanism to capture the modalities' interactive information effectively. In MGRL, we construct a graph to capture the structural relationships between movies based on directors, screenwriters, and actors, where the node in the graph is a movie that has multi-modal attributes and is first represented by MDF. Then a Graph Convolutional Network (GCN)-based architecture is developed to extract structural information between movie nodes. In addition, we also present a new multi-modal movie dataset, i.e., MovieBricks. The experimental results on Moviescope and MovieBricks demonstrate the superior performance of MFMGC.

Acknowledgment. This work is supported by the National Natural Science Foundation of China No. 62272332, the Major Program of the Natural Science Foundation of Jiangsu Higher Education Institutions of China No. 22KJA520006.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv:1702.01992 (2017)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: Wav2Vec 2.0: a framework for self-supervised learning of speech representations. NeurIPS **33**, 12449–12460 (2020)

3. Bamman, D., O'Connor, B., Smith, N.A.: Learning latent personas of film characters. In: ACL, vol. 1, pp. 352–361 (2013)
4. Behrouzi, T., Toosi, R., Akhaee, M.A.: Multimodal movie genre classification using recurrent neural network. Multimedia Tools Appl. **82**, 1–22 (2022)
5. Bi, T., Jarnikov, D., Lukkien, J.: Shot-based hybrid fusion for movie genre classification. In: ICIP, pp. 257–269 (2022)
6. Bi, T., Jarnikov, D., Lukkien, J.: Video representation fusion network for multi-label movie genre classification. In: ICPR, pp. 9386–9391 (2021)
7. Bribiesca, I.R., Monroy, A.P.L., Montes, M.: Multimodal weighted fusion of transformers for movie genre classification. In: Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pp. 1–5 (2021)
8. Cascante-Bonilla, P., Sitaraman, K., Luo, M., Ordonez, V.: Moviescope: large-scale analysis of movies using multiple modalities. arXiv:1908.03180 (2019)
9. Chen, S., Nie, X., Fan, D., Zhang, D., Bhat, V., Hamid, R.: Shot contrastive self-supervised learning for scene boundary detection. In: CVPR, pp. 9796–9805 (2021)
10. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting pre-trained models for Chinese natural language processing. In: EMNLP, pp. 657–668 (2020)
11. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: ICML, pp. 233–240 (2006)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
13. Dridi, A., Recupero, D.R.: MORE SENSE: MOvie REviews SENtiment analysis boosted with SEmantics. In: EMSASW (2017)
14. Huang, Q., Xiong, Yu., Rao, A., Wang, J., Lin, D.: MovieNet: a holistic dataset for movie understanding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12349, pp. 709–727. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_41
15. Kukleva, A., Tapaswi, M., Laptev, I.: Learning interactions and relationships between movie characters. In: CVPR, pp. 9849–9858 (2020)
16. Liao, Y., Peng, Y., Shi, S., Shi, V., Yu, X.: Early box office prediction in China's film market based on a stacking fusion model. Ann. Oper. Res. **308**(1), 321–338 (2020). https://doi.org/10.1007/s10479-020-03804-4
17. Liu, Z., et al.: Swin Transformer: hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
18. Mangolin, R.B., et al.: A multimodal approach for multi-label movie genre classification. Multimedia Tools Appl. **81**(14), 19071–19096 (2022)
19. Nambiar, G., Roy, P., Singh, D.: Multi modal genre classification of movies. In: INOCON, pp. 1–6 (2020)
20. Rao, A., et al.: A local-to-global approach to multi-modal movie scene segmentation. In: CVPR, pp. 10146–10155 (2020)
21. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in Homes: crowdsourcing data collection for activity understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 510–526. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_31
22. Su, W., et al.: VL-BERT: pre-training of generic visual-linguistic representations. arXiv:1908.08530 (2019)
23. Thet, T.T., Na, J.C., Khoo, C.S., Shakthikumar, S.: Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In: CIKM, pp. 81–84 (2009)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30 (2017)

25. Wehrmann, J., Barros, R.C.: Movie genre classification: a multi-label approach based on convolutions through time. Appl. Soft Comput. **61**, 973–982 (2017)
26. Xu, M., et al.: Long short-term transformer for online action detection. NeurIPS **34**, 1086–1099 (2021)
27. Zhang, Z., Gu, Y., Plummer, B.A., Miao, X., Liu, J., Wang, H.: Effectively leveraging multi-modal features for movie genre classification. arXiv:2203.13281 (2022)
28. Zhou, Y., Zhang, L., Yi, Z.: Predicting movie box-office revenues using deep neural networks. Neural Comput. Appl. **31**(6), 1855–1865 (2019)