Incorporating Domain Knowledge Graph into Multimodal Movie Genre Classification with Self-Supervised Attention and Contrastive Learning

Jiaqi Li
Southeast University
Nanjing, China
jqli@seu.edu.cn

Guilin Qi*
Southeast University
Nanjing, China
gqi@seu.edu.cn

Chuanyi Zhang
Hohai University
Nanjing, China
20231104@hhu.edu.cn

Yongrui Chen
Southeast University
Nanjing, China
yrchen@seu.edu.cn

Yiming Tan
Southeast University
Nanjing, China
230189757@seu.edu.cn

Chenlong Xia
Southeast University
Nanjing, China
213203677@seu.edu.cn

Ye Tian
Southeast University
Nanjing, China
220224335@seu.edu.cn

ABSTRACT

Multimodal movie genre classification has always been regarded as a demanding multi-label classification task due to the diversity of multimodal data such as posters, plot summaries, trailers and metadata. Although existing works have made great progress in modeling and combining each modality, they still face three issues: 1) unutilized group relations in metadata, 2) unreliable attention allocation, and 3) indiscriminative fused features. Given that the knowledge graph has been proven to contain rich information, we present a novel framework that exploits the knowledge graph from various perspectives to address the above problems. As a preparation, the metadata is processed into a domain knowledge graph. A translate model for knowledge graph embedding is adopted to capture the relations between entities. Firstly we retrieve the relevant embedding from the knowledge graph by utilizing group relations in metadata and then integrate it with other modalities. Next, we introduce an Attention Teacher module for reliable attention allocation based on self-supervised learning. It learns the distribution of the knowledge graph and produces rational attention weights. Finally, a Genre-Centroid Anchored Contrastive Learning module is proposed to strengthen the discriminative ability of fused features. The embedding space of anchors is initialized from the genre entities in the knowledge graph. To verify the effectiveness of our framework, we collect a larger and more challenging dataset named MM-IMDb 2.0 compared with the MM-IMDb dataset. The experimental results on two datasets demonstrate that our model is superior to the state-of-the-art methods. Our code and dataset is available at IDKG.git.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Multimodal, Self-supervised Learning, Contrastive Learning, Knowledge Graph

ACM Reference Format:
Jiaqi Li, Guilin Qi, Chuanyi Zhang, Yongrui Chen, Yiming Tan, Chenlong Xia, and Ye Tian. 2023. Incorporating Domain Knowledge Graph into Multimodal Movie Genre Classification with Self-Supervised Attention and Contrastive Learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612085

1 INTRODUCTION

Movie genre classification is a fundamental task for certain downstream tasks such as movie recommendation [12], understanding [25], editing [9], description [39], etc. Previous studies [16, 53] have achieved unparalleled results in movie genre classification with a single modality such as posters, plot summaries, movie trailers, audio, or metadata. Nowadays more researchers [4, 8, 10, 24, 29, 40, 45, 58] focus on multimodal sources which could be the arbitrary combination of multiple modalities. By taking advantage of multimodal information, existing methods have made great progress in movie genre classification. However, they still leave three issues unsolved:

1) Unutilized group relations in metadata. As illustrated in Figure 1, group relations indicate that entities belonging to the same group usually appear simultaneously. To give two real-scenario examples, if *Nolan* is the director of a movie, it is likely to be a *science fiction*.

*Corresponding authors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612085

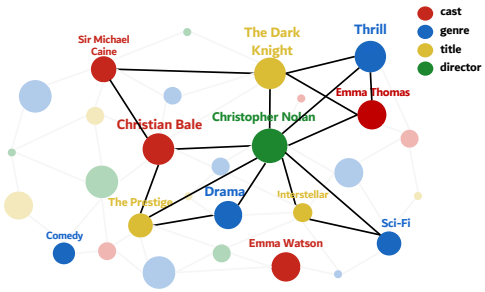


Figure 1: Group relations in metadata. A domain knowledge graph is constructed using titles, casts, directors and genres. An edge between two entities represents their co-appearance in a movie.

If *Emma Watson* starred in a movie, it is probably not a *comedy*. However, recent works typically ignore group relations when modeling metadata. Behrouzi et al. [5] extracts features from metadata and fuses them with other modality features through random forest classifier. Seo et al. [41] trains a graph attention network based on an undirected graph composed of movie nodes. In our opinion, these methods can be further improved if taking group relations into consideration.

2) Unreliable attention allocation. Intuitively, different samples should own varying weights on each modality to boost movie genre prediction with multimodal data. Previous works[1, 5] adopt an attention module to assign different weights to various modalities (e.g., plot summary, poster, trailer, audio). Nevertheless, reliability is not guaranteed due to no supervision in attention module training. Consequently, the produced attention can be irrational.

3) Indiscriminative fused features. Existing methods [1, 8, 40, 45, 52, 55] typically utilize a pre-trained model for each modality to obtain discriminative single-modality features. Then features from each modality are fused for genre classification. However, fused feature space tends to show some distance from the original feature space of each modality, which harms the discriminative ability. As a result, fused features tend to be inefficient in predicting genres.

Inspired by previous methods [7, 49, 54] that leverage knowledge graph, we propose a novel framework named IDKG (**I**ncorporating **D**omain **K**nowledge **G**raph) for movie genre classification. Our motivation is to exploit the knowledge graph from different perspectives to solve the aforementioned issues. To begin with, we propose to construct a domain knowledge graph using metadata which includes directors, casts, titles and genres. Moreover, we adopt a translate model for knowledge graph embedding (such as TransH [50], TransR [31], etc.) to capture the relation among entities in knowledge graph. For the first issue, we leverage group relations present in the metadata to retrieve the pertinent embedding from the knowledge graph. Then the retrieved embedding is integrated with other modalities to improve classification accuracy. To alleviate the unreliable attention allocation problem, we propose an Attention Teacher (AT) module that guides the attention module to produce rational attention scores based on self-supervised learning. Our AT module captures the distribution feature of the knowledge graph to generate pseudo labels for attention scores and utilizes a suitably designed loss function to train the attention module. As to the indiscriminative fused features, we propose a Genre-Centroid Anchored Contrastive Learning (G-CACL) module to strengthen the discriminative ability of features. It can be hard to select positive and negative pairs for samples with multiple genres in contrastive learning. To solve this problem, our G-CACL module defines the centroid of multiple genres embedding as the positive anchor. The enlarged genre space provides feasible optimization directions for fused features to enhance their discriminative ability.

In order to verify the effectiveness of our proposed IDKG, we further create a new dataset, MM-IMDb 2.0, which is more challenging compared with MM-IMDb dataset. It comprises 33,742 movies collected from IMDb website, with genres set the same as MM-IMDb. Notably, the proportion of the number of head and tail genres is enlarged, thereby enhancing the task difficulty.

Our main contributions of can be summarized as follows:

• We propose a novel framework called IDKG that subtly exploits the knowledge graph from different perspectives. To the best of our knowledge, we are the first to incorporate a knowledge graph into the multimodal movie genre classification task.

• We utilize group relations in metadata to obtain relevant embedding from the knowledge graph. With selected embedding as an additional modality source, performance is significantly improved.

• We propose an AT module to alleviate the unreliable attention allocation problem. It obtains pseudo labels from the distribution of the knowledge graph and trains the attention module in a self-supervised manner. Owing to more reliable attention scores, each modality is assigned a more reasonable weight.

• We propose a G-CACL module to alleviate the indiscriminative fused features problem. Centroids of genre embedding from the knowledge graph are regarded as positive anchors. The contrastive learning strategy is applied to enhance fused feature representation.

• We create a new and more challenging dataset MM-IMDb 2.0 to verify the effectiveness of our proposed method. Extensive experiments are conducted to compare our IDKG with current state-of-the-art methods. Experimental results demonstrate that our method outperforms them by a huge margin.

2 RELATED WORK

2.1 Movie Genre Classification

Movie genre classification could be divided into two categories: single-modality-based and multimodal-based. The former predicts genres by one kind of modality data, such as posters, plot summaries, movie trailers, audios, or metadata. [42, 53] extract features from movie trailers, while [51] and [16] focus on using only poster images or plot summaries. As single-modality data is relatively simple to process, these methods have achieved excellent performance. However, for multimodal-based methods [5, 6, 17] which combine two or more modalities , the results have not been as good due to the complexity of data features.

2.2 Self-supervised Learning

Self-supervised learning is a special learning method without direct supervision signal. The supervision signal is generated by using the features of the dataset itself. Nowadays, self-supervised learning is attracting more attention from researchers, such as Next Word Prediction [15, 26], Automated Text Augmentation [18, 33]

in natural language processing and Colorization [46, 56], Context Prediction [35, 36] in computer vision. Moreover, self-supervised learning [2, 11, 23, 34] has great potential to replace fully supervised learning in representation learning domain.

In our paper, we provide a novel insight into the self-supervised learning. We develop a paradigm to the attention module by summarizing the distribution from constructed domain knowledge graph.

2.3 Supervised Contrastive Learning in Multi-label Classification

Contrastive learning is a technique that trains a model to differentiate between similar and dissimilar examples. Such method can be used to learn representations of data. Initially, this approach [20, 37, 48] was explored in the self-supervised setting. The feature embedding is learned without explicit labels by solving a pretext task. Supervised contrastive learning [19, 28] is another form of contrastive learning that employs annotated data to generate positive pairs by selecting samples from various instances of a specific category.

In the contrastive learning paradigm, positive and negative pairs are defined by semantic similarity. Neverthless, it is hard to be applied to multi-label classification because of multiple semantics. Recently, several work attends to bridging combination of multi-label and contrastive learning. [13, 48] compute the contrastive loss by determining a single label that best matches each sample. However, such methods also serve to increase the distance between the sample and other labels. [57] proposes a multi-label contrastive learning framework that utilizes a hierarchical structure to leverage all available labels, but in movie genre classification there is no hierarchical structure of labels. [3, 22] aim to utilize the label embedding space. [22] adopts a center loss but fails to construct negative pairs. [3] focuses on figuring out the similarity between label embedding, which may harm the discriminative ability of fused features. To alleviate the problem of defining positive pairs, our proposed G-CACL module enlarges the genre embedding space and a centroid of genres is generated as the positive anchor for each sample. Negative samples are well-designed for the effectiveness of our module.

3 APPROACH

Problem Definition. In multimodal movie genre classification task, the i-th sample is composed of a text T_i, an image I_i and metadata $Meta_i$. $Meta_i$ includes the directors, casts, title and the genres of the movie. Each sample is annotated with a multi-label vector $y_i \in \{0,1\}^M$, where M is the number of genres of the dataset. The dataset $\mathcal{D}$ is splited into train set $\mathcal{D}_{train}$, test set $\mathcal{D}_{test}$ and validation set $\mathcal{D}_{valid}$. The goal of our task is to train a strong classifier using $\mathcal{D}_{train}$ to predict the multiple genres of each sample in $\mathcal{D}_{test}$.

Model Overview. The overall architecture of our framework is illustrated in Figure 2. Firstly, a Clip model [38] extracts features F_i^T and F_i^I from input source image I and text T. Then, in Section 3.1, we process the metadata from $\mathcal{D}_{train}$ into a knowledge graph. To capture the relations between entities, an embedding matrix of knowledge graph entities is trained by adopting a translate model for knowledge graph embedding. We retrieve all the directors and actors in metadata, and get their embedding from embedding matrix as a new modality K_i. Next, in Section 3.2, we describe an AT module which could ensure the rational allocation of attention module. Multi-modality features are fused according to their attention weights. Finally, in Section 3.3, we introduce a G-CACL module that enhances the discriminability of fused features.

3.1 Knowledge Graph Feature Formation

Knowledge Graph Construction. A domain knowledge graph is constructed full-automatically by using metadata. Our knowledge graph schema is derived from the fields of metadata and we define four types of entities as the name of fields:

$$E = \{d, t, c, g\}, \tag{1}$$

which correspond to directors, titles, casts and genres respectively. Moreover, we define six kinds of relations:

$$R = \{d-t, c-t, g-t, d-c, d-g, c-g\}, \tag{2}$$

which represent the relations of directors and titles, casts and titles, genre and titles, directors and cats, directors and genres, casts and genres. Notably we only use metadata from $\mathcal{D}_{train}$ to avoid data leakage. We visit the fields of metadata and extract each value as an entity which has a unique matching id. Furthermore, we traverse all entity pairs in metadata, and each entity pair forms a triplet with the corresponding R.

Knowledge Graph Embedding. In order to capture the relations between entities, we apply a translate model for knowledge graph embedding, for instance, TransH [50]. Finally we get the embedding matrix of all entities $Mat_e \in \mathbb{R}^{N_k \times D_k}$, where N_k denotes the number of entities in knowledge graph and D_k is the dimension of entity embedding.

Utilization of Group Relations in Metadata. The group relations in metadata can aid the prediction of movie genres as illustrated in Section 1. Since the translate model has enabled knowledge graph capture the relations between entities, for i-th sample we traverse all the director and cast entities in metadata. Finally we obtain the embedding of entities from Mat_e:

$$K_i = \left\{E_i^{d_1}, E_i^{d_2}, ..., E_i^{d_{N_{di}}}, E_i^{c_1}, E_i^{c_2}, ..., E_i^{c_{N_{ci}}}\right\}, \tag{3}$$

where E^d denotes the embedding of an director entity. N_{di} and N_{ci} are the number of directors and casts of the i-th sample respectively. Finally, the feature of knowledge graph $F_i^K \in \mathbb{R}^{D_k}$ is defined as the sum of all embedding in K_i. Notably, in the test phase all E^d and E^c of many samples do not exist in Mat_e. It is because that Mat_e is trained by the entities of $\mathcal{D}_{train}$. F^K becomes the 0 vector at this point and it can be formulated as:

$$F_i^K = \begin{cases} \sum\limits_{E^j \in K_i} E^j, & K_i \cap Mat_e \neq \emptyset, \\ 0, & else. \end{cases} \tag{4}$$

3.2 Attention Teacher Module

We adopt an attention module to balance the weight of each modality for each sample. It is composed of a linear function and a sigmoid function. Notably the parameters of the attention module for three modality features are shared.

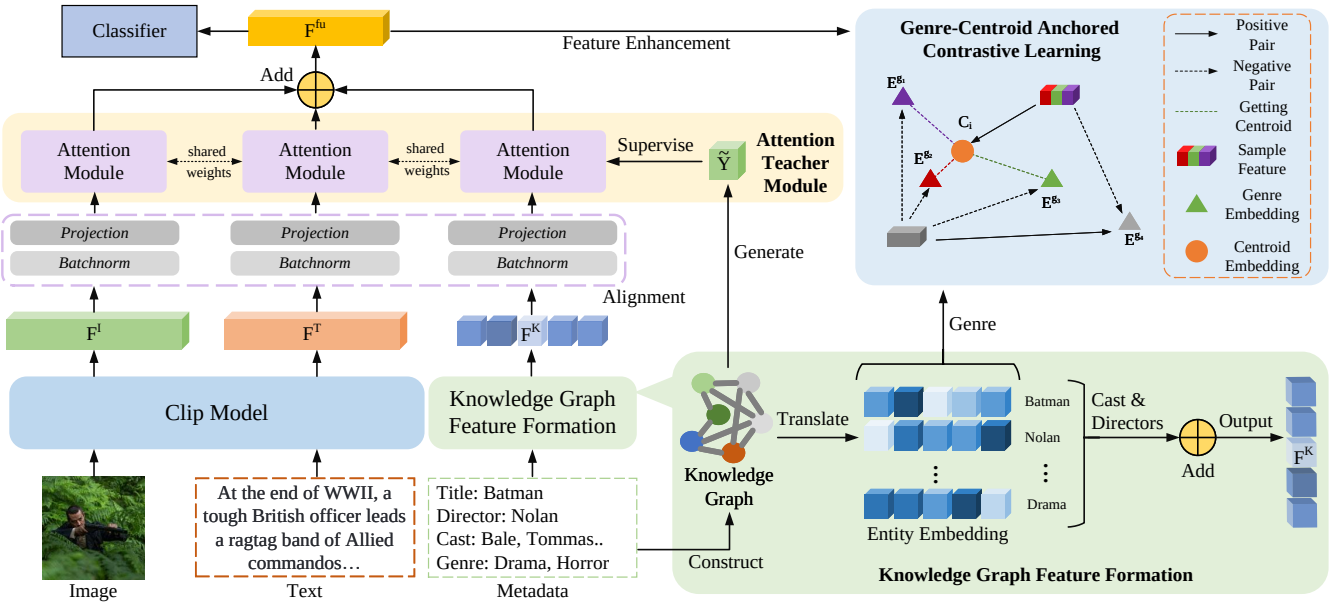


Figure 2: The Overview of our proposed IDKG. We leverage a Clip model to extract visual-textual features. Metadata is utilized to construct a domain knowledge graph, which is then followed by a translate model to obtain entity embedding. The cast and director embedding is regarded as the knowledge graph modality feature. Visual, text and knowledge graph features (F^I, F^T and F^K) are aligned to the same dimension. Next, each modality feature is weighted by the attention module and added to obtain fused feature F^{fu}. The Attention Teacher guides the attention producing with pseudo labels $\widetilde{Y}$ generated from the knowledge graph. After that, the discriminative ability of fused feature is enhanced by a Genre-Centroid Anchored Contrastive Learning module, where the embedding space of anchors C is initialized from the genre embedding E^g in the knowledge graph. Finally, the classifier takes the fused feature F^{fu} as input to produce predictions.

As illustrated in Figure 2, for multi-modality features $F_i^T \in \mathbb{R}^{D_t}$, $F_i^I \in \mathbb{R}^{D_i}$ and $F_i^K \in \mathbb{R}^{D_k}$, where D_t and D_i are the dimension of text feature and image feature extracted from Clip model, we apply batchnorm1d and linear projection function to convert them to the same shape D_p. After alignment, these three feature maps are entered into the attention module to obtain their attention scores A_i^T, A_i^I and A_i^K. Next, these multi-modality features are multiplied by their corresponding attention scores and add up as the ultimate fused feature $F_i^{fu} \in \mathbb{R}^{D_p}$:

$$h(x) = project(batchnorm(x)),$$
$$F_i^{fu} = h(F_i^T) \cdot A_i^T + h(F_i^I) \cdot A_i^I + h(F_i^K) \cdot A_i^K, \tag{5}$$

where $batchnorm$ and $project$ denotes the batchnorm1d and linear projection function respectively.

To ensure the reliable attention allocation of the attention module, we introduce the Attention Teacher (AT) module based on self-supervised learning. Unlike the previous methods which mainly create pretext tasks in vision or language domain, we mine the distribution feature from the knowledge graph. It is observed that different samples contain different numbers of entities when forming F^K. Thus it is natural to consider that varying scores of attention should be assigned to F^K of different samples. Specifically, F^K composed of a small number of entities deserves a lower attention score. Especially in testing phase, if F^K is 0 vector as is presented in Formula (4), the attention scores should be close to 0 at this point. Moreover, we consider the degree of entities in knowledge graph, because an entity is less important if it has very few neighbours in the knowledge graph structure.

Taking above discussions into consideration, we finally utilize the following equations to define the pseudo label $\widetilde{Y}_i$ of attention scores for F_i^K:

$$\overline{N_{d+c}} = \frac{1}{W}\sum_{u=1}^{W}(N_{du} + N_{cu}),$$
$$\overline{V} = \frac{1}{W}\sum_{u=1}^{W}\sum_{E^j \in K_u} V_{E^j},$$
$$\widetilde{Y}_i = \frac{(N_{di} + N_{ci})\sum\limits_{E^j \in K_i} V_{E^j}}{(N_{di} + N_{ci} + \overline{N_{d+c}})(\sum\limits_{E^j \in K_i} V_{E^j} + \overline{V})}, \tag{6}$$

where W denotes the number of samples in $\mathcal{D}_{train}$ and V is the degree of an entity in the knowledge graph. In Formula (6), we introduce the average number of sum of directors and actors $\overline{N_{d+c}}$ and the average degree $\overline{V}$ of all samples in $\mathcal{D}_{train}$. Our intention is that if the sum of directors and actors and the degree of current sample are both beyond the average number in $\mathcal{D}_{train}$, the corresponding pseudo label $\widetilde{Y}_i$ should become higher, otherwise it would be lower. Formula (6) requires the range of $\widetilde{Y}_i$ in $(0, 1)$ to fit the range of produced attention weights which are limited by a sigmoid function.

Then we design an appropriate loss function to make the self-supervised attention scores converge normally. Since A_i^K is a vector while $\widetilde{Y}_i$ is in scalar form according to Formula (6), A_i^K is averaged as scalar form either to compute loss with $\widetilde{Y}_i$. Since the goal of AT module is to guide A_i^K close to $\widetilde{Y}_i$, a regression loss is adopted in this module. Moreover, it is worth noting that A_i^K and $\widetilde{Y}_i$ are both designed to be between 0 and 1. Thus directly applying l1 or l2 regression loss would limit the loss value in range (0,1) and have the risk of underfitting. In order to ensure a large enough gradient, we apply a logarithmic function to guarantee larges gradients instead of directly adopting l1 or l2 regression loss. Considering a batch of input with batchsize B, the self-supervised attention loss is defined as follows:

$$\mathcal{L}_{atten} = -\sum_{i=1}^{B} log(1 - |A_i^K - \widetilde{Y}_i|). \tag{7}$$

In this way the definition domain of loss function is limited to 0 to 1 with large enough gradient and monotonic increasing trend. From Formula (7), it can be observed that only the knowledge graph attention score A_i^K is supervised by its pseudo label $\widetilde{Y}_i$ to train the attention module, while A_i^T and A_i^I do not participate in the training procedure. Nevertheless, in our experiment (Section 4.5) we find that the attention module trained with Formula (7) can produce reasonable scores A_i^T and A_i^I for texts and images.

3.3 Genre-Centroid Anchored Contrastive Learning Module

We propose a Genre-Centroid Anchored Contrastive Learning (G-CACL) module which facilities the genre embedding from knowledge graph to strengthen the discriminative ability of F^{fu} . Considering that in contrastive learning, each sample is typically assigned a single semantic label and positive pairs are defined by whether they belong to the same semantic label. However, in our task each sample is annotated multi-genre. If we regard each genre embedding as a positive anchor, it is hardly achievable to push the feature of sample close to all the positive anchors synchronously. To overcome this limitation, we attempt to represent the semantics of multiple genres in single anchor.

Specifically, for a batch of fused features F^{fu}, the genre embedding set of i-th feature F_i^{fu} is:

$$G_i = \left\{E^{g_1}, E^{g_2}, ..., E^{g_{N_{gi}}}\right\}, \tag{8}$$

where N_{gi} is the number of annotated genres of F_i^{fu}. we enlarge the genre embedding space by computing the centroid C_i of G_i:

$$C_i = \frac{1}{N_{gi}} \sum_{E^{g_k} \in G_i} E^{g_k}, \tag{9}$$

which is regarded as the positive anchor for F_i^{fu}. We could obtain the union of genre embedding of all samples in the current batch:

$$\bigcup_{i=1}^{B} G_i = \left\{E^{g_1}, E^{g_2}, ..., E^{g_{N_{gj}}}\right\}, \tag{10}$$

where N_{gj} is the number of annotated genres of all samples in the current batch. We define the complement set of G_i in $\bigcup_{i=1}^{B} G_i$ as the negative samples of F_i^{fu}:

$$S_i^{Neg} = \bigcup_{i=1}^{B} G_i - G_i. \tag{11}$$

Table 1: The comparison of genres distribution of MM-IMDb and MM-IMDb 2.0. The genres are arranged in descending order of quantity from left-top to right-bottom.

Genre	MM-IMDb	MM-IMDb 2.0	Genre	MM-IMDb	MM-IMDb 2.0
Drama	4188	4773	Fantasy	498	789
Comedy	2609	2861	Music	415	752
Action	1617	1950	History	418	754
Adventure	1588	1673	Western	344	704
Romance	1148	1364	Sci-Fi	280	687
Crime	1081	1298	Musical	292	603
Horror	835	1179	Sport	245	472
Thriller	823	1096	Short	211	428
Biography	584	846	War	164	417
Animation	664	874	Documentary	139	272
Family	646	817	File-Noir	92	73
Mystery	591	800			

Before computing loss, the centroid and genre embedding go through the linear function to be transformed into the same shape as F^{fu}:

$$\begin{aligned} f_{C_i} &= Linear(C_i), \\ f_{E^{g_k}} &= Linear(E^{g_k}). \end{aligned} \tag{12}$$

The loss function of G-CACL module is defined as follows:

$$\begin{aligned} q_i &= \sum_{E^{g_k} \in S_i^{Neg}} exp(F_i^{fu} \cdot f_{E^{g_k}} / \tau), \\ \mathcal{L}_{contra} &= -\sum_{i=1}^{B} log \frac{exp(F_i^{fu} \cdot f_{C_i} / \tau)}{exp(F_i^{fu} \cdot f_{C_i} / \tau) + q_i}, \end{aligned} \tag{13}$$

where τ is the temperature coefficient following [28]. In this loss function, we push the F_i^{fu} close to its corresponding f_{C_i}, and away from the embedding of negative samples S_i^{Neg}. q_i denotes the sum of similarity between F_i^{fu} and each embedding in S_i^{Neg}.

Furthermore, for the multi-label classification, we adopt the binary cross-entropy loss. For a batch of output after the classifier, the multi-label classification loss is:

$$\mathcal{L}_{class} = \sum_{i=1}^{B} \sum_{j=1}^{M} (1 - y_{ij}) \cdot log(1 - p_{ij}) + y_{ij} \cdot log p_{ij}, \tag{14}$$

where p_{ij} denotes the predicted probability that the i-th sample belongs to the j-th genre.

Finally, we compose $\mathcal{L}_{atten}$, $\mathcal{L}_{contra}$ and $\mathcal{L}_{class}$ as the ultimate training loss for our IDKG:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{atten} + \mathcal{L}_{contra}. \tag{15}$$

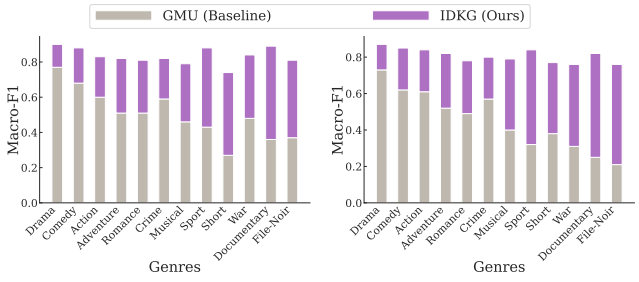
4 EXPERIMENT

4.1 Experiment Setup

Dataset. We evaluate IDKG on two datasets, MM-IMDb and MM-IMDb 2.0. MM-IMDb dataset is a multi-label movie genre classification dataset released by [1] which contains 25959 films. Each sample is composed of a poster, a plot summary and metadata which covers the directors, actors, publication year, etc. Following [1], we

Table 2: Comparison with the state-of-the-art methods. The evaluation metrics are introduced in Section 4.1

Type	Model	MM-IMDb				MM-IMDb 2.0			
		Micro	Macro	Weighted	Samples	Micro	Macro	Weighted	Samples
Multimodal	GMU [1]	0.630	0.541	0.617	0.630	0.617	0.575	0.607	0.588
	CentralNet [45]	0.639	0.561	0.631	0.639	0.622	0.594	0.619	0.606
	MMBT [29]	0.669	0.618	-	-	0.635	0.607	0.652	0.650
	MFM [8]	0.675	0.616	0.675	0.673	0.656	0.608	0.664	0.671
	ReFNet [40]	0.680	0.587	-	-	-	-	-	-
	COCA [55]	0.677	0.626	0.668	0.681	0.659	0.623	0.649	0.670
	BLIP [30]	0.674	0.628	0.663	0.675	0.661	0.618	0.635	0.663
	BridgeTow [52]	0.682	0.633	0.676	0.680	0.668	0.627	0.684	0.679
Graphical	MM-GATBT[41]	0.685	0.645	0.683	0.686	0.674	0.632	0.697	0.685
Graphical+Multimodal	**IDKG**	**0.849**	**0.832**	**0.848**	**0.839**	**0.828**	**0.811**	**0.827**	**0.807**



(a) Macro-F1 on MM-IMDb (b) Macro-F1 on MM-IMDb 2.0

Figure 3: Macro-F1 scores for sampled head and tail classes on two datasets. The GMU is chosen as the compared baseline. Genres are arranged in descending order of quantity from left to right.

randomly split the dataset into train set, test set and validation set at the ratio of 0.6, 0.3 and 0.1.

To further verify our framework, we create a novel and more challenging dataset, MM-IMDb 2.0. We collect 33742 movies with their posters, plots and metadata from the IMDB website. As is illustrated in Table 1, the ratio of the quantity of *Drama* to the quantity of *Film-Noir* is nearly 65:1. Besides, we also constrain several other tail genres. Moreover, we partition our dataset the same proportion as MM-IMDb dataset.

Domain Knowledge Graph. As presented in Section 3.1, we process the metadata of $\mathcal{D}_{train}$ into a domain knowledge graph. For MM-IMDb dataset, there are 2231455 triplets and 264271 entities in its domain knowledge graph. Since MM-IMDb 2.0 dataset has much more movies, the number of triplets and entities are 3512803 and 318299 respectively.

Implementation Details. IDKG is trained on Pytorch with a 2080ti gpu. For the translate model for knowledge graph embedding, we use the toolkit released by [21]. We adopt the SGD optimizer with 0.5 learning rate and the embedding dimension is 200. Moreover, we train the translate model for 500 epochs with 100 batchsize. For the stage two we use the AdamW [32] optimizer and the learning rate is 1e-3. We train 15 epochs with 64 batchsize. To extract the image and text features, we apply the Clip [38] model and the parameters are frozen in our experiments. The unified vector space dimension is 512 which is set the same as [1]. The parameter τ for the G-CACL module is set [0.05, 0.1, 0.3, 0.5, 0.7, 0.9] following [28].

Evaluation Metrics. For the evaluation, we report the Micro-F1, Macro-F1, Weighted-F1 and Samples-F1 scores following [1, 41] as our metrics.

4.2 Experimental Results and Analyses

We compare IDKG with state-of-the-art methods on two datasets and our approach achieves far superior performance. We categorize the existing methods into three types: 1) *Multimodal* type is a straightforward strategy to solve the problem. This type of works [1, 8, 29, 30, 40, 45, 52, 55] mainly focus on the strategy of extracting the well-represented features from each modality respectively by adopting corresponding pre-trained models. It is noted that we also select several recent outstanding *Multimodal* methods [30, 52, 55] as our competitors. 2) *Graphical* method is conducted to explore the potential of structural sources. MM-GATBT [41] leverages graph neural networks to learn the relational semantics of entities by using encoded images as node features. 3) Our method is a comprehensive framework which fully taking advantage of the two types above. Notably the translate model used for Table 2, 3 and 5 is RotateE [43] and the ablation study on translate models is shown in Section 4.3. Moreover, we compare our IDKG with GMU [1] for head and tail genres on two datasets with Macro-F1 scores as the evaluation metric.

Results on MM-IMDb Dataset. The comparison results on MM-IMDb dataset are shown in Table 2. We observe that the Graphical method MM-GATBT [41] outperforms all the Multimodal methods in each metric, which demonstrates that the semantic relations in metadata could serve a great benefits to the capacity of predicting genres. Table 2 shows that IDKG surpasses MM-GATBT by at least 15% on all evaluation metrics. One possible reason may be that in MM-GATBT the graph nodes are composed of the image embedding, where there is no additional knowledge in graph nodes at bottom. IDKG incorporates the knowledge graph embedding with other modalities by using the group relations in metadata which enriches the features for genre prediction. Moreover, two effective modules are designed to address the unreliable attention allocation and indiscriminative fused feature issues, thus boosting the performance of our model.

Table 3: Ablation study on each module. '- *AT*' represents that AT module is removed and the attention module is not trained with pseudo labels. '- *G-CACL*' means that G-CACL module is omitted and fused features are not enhanced through contrastive learning. '- *KG*' denotes that the metadata is not processed into a knowledge graph, thus not being incorporated with visual-textual features. IDKG (GMU) indicates incorporating the domain knowledge graph (KG) and G-CACL module into GMU.

Model	MM-IMDb				MM-IMDb 2.0			
	Micro	Macro	Weighted	Samples	Micro	Macro	Weighted	Samples
IDKG	**0.849**	**0.832**	**0.848**	**0.839**	**0.828**	**0.811**	**0.827**	**0.807**
IDKG - AT	0.842	0.829	0.841	0.831	0.813	0.794	0.812	0.792
IDKG - AT - G-CACL	0.828	0.816	0.825	0.817	0.796	0.779	0.789	0.783
IDKG - AT - G-CACL - KG	0.677	0.625	0.661	0.667	0.668	0.597	0.652	0.631
IDKG (GMU)	0.832	0.816	0.832	0.824	0.797	0.779	0.795	0.783
IDKG (GMU) - G-CACL	0.819	0.804	0.817	0.810	0.782	0.773	0.790	0.772
IDKG (GMU) - G-CACL - KG	0.630	0.541	0.617	0.630	0.617	0.575	0.607	0.588

Table 4: Ablation study on Translate models. We also report the Hit@10 metrics to illustrate the relation between task performance and the effectiveness of each translate model.

Dataset	Trans Model	Micro	Macro	Weighted	Samples	Hit@10
MM-IMDb	TransH [50]	0.833	0.820	0.831	0.827	0.507
	TransR [31]	0.839	0.823	0.838	0.835	0.519
	TransD [27]	0.835	0.827	0.828	0.833	0.508
	ComplEx [44]	0.827	0.813	0.825	0.821	0.485
	ConvE [14]	0.829	0.822	0.836	0.825	0.506
	RotateE [43]	**0.849**	**0.832**	**0.848**	**0.839**	**0.549**
MM-IMDb 2.0	TransH [50]	0.813	0.806	0.815	0.792	0.507
	TransR [31]	0.814	0.809	0.820	0.791	0.519
	TransD [27]	0.817	0.803	0.823	0.796	0.508
	ComplEx [44]	0.806	0.797	0.812	0.784	0.485
	ConvE [14]	0.814	0.805	0.822	0.792	0.506
	RotateE [43]	**0.828**	**0.811**	**0.827**	**0.807**	**0.549**

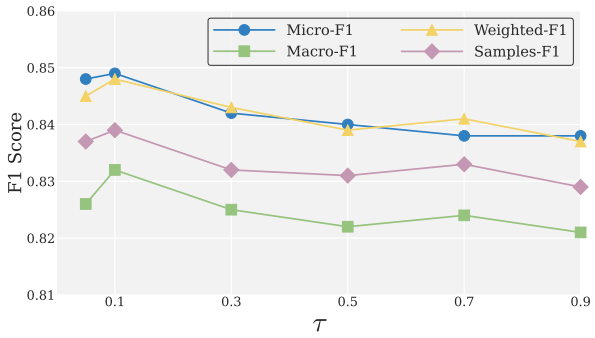


Figure 4: The effect of parameter τ on MM-IMDb. We report F1 scores on different value of τ.

Results on MM-IMDb 2.0 Dataset. The overall results show that the performance of all methods on MM-IMDb 2.0 dataset is inferior to that on MM-IMDb dataset. The reason can be that MM-IMDb2 2.0 dataset is a more challenging dataset. As shown in Table 2, our proposed IDKG also successes in beating all the competitors by at least 12% which is a large margin on all metrics. The experimental result demonstrates that taking advantage of both Multimodal and Graphical methods can remarkably boost the performance.

Macro-F1 score analysis. As shown in Table 1, the distribution of genres is imbalanced in MM-IMDb dataset. Towards severer imbalance problem, we enlarge the proportion of the number of head genres and tail genres when collecting MM-IMDb 2.0 dataset as mentioned in Section 4.1. As Macro-F1 neglects the proportion for each label, it is more sensitive to the imbalance of genres distribution than other evaluation metrics. Thus we compare Macro-F1 between GMU and IDKG for sampling six head genres and six tail genres on two datasets as can be seen in Figure 3 and the genres are arranged in descending order of quantity from left to right. We could observe that for GMU the Macro-F1 of head classes is generally far larger than that of tail classes on two datasets due to the imbalance distribution. However, for IDKG the Macro-F1 of tail genres is close to that of head genres and almost Macro-F1 of all genres is around 80%, which demonstrates distinguished classification ability of IDKG.

4.3 Ablation Study

To evaluate the effectiveness of each component of IDKG, we conduct extensive ablation experiments on two datasets. In particular, we incorporate the domain knowledge graph and the G-CACL module into GMU [1] one by one for more solid proof. After the ensemble of all modules, we name the new model IDKG (GMU). Notably since GMU provides a strategy of feature fusion, AT module is not conducted in its ablation experiment. Moreover, we compare the different translate models for knowledge graph embedding and show the influence on IDKG performance. Finally, we analyze the parameter τ of G-CACL module.

Results on IDKG. Table 3 illustrates that the discarding of domain knowledge graph loses the most performance by at least 10% of all modules which proves that group relations in metadata could contribute to the model performance greatly. With the removal of AT module, the performance of IDKG on all metrics declines to various degrees on two datasets. This is may be that AT module ensure the reliable attention allocation, thus improving the accuracy. The performance goes down if IDKG is not equipped with G-CACL module. The reason is perhaps that G-CACL module could improve the discriminative ability of fused feature which could boost the effectiveness of our model.

Results on IDKG (GMU). As can be seen from Table 3, the trend of performance change on IDKG (GMU) is the same as IDKG. This illustrates the effectiveness of our proposed modules. It is noted

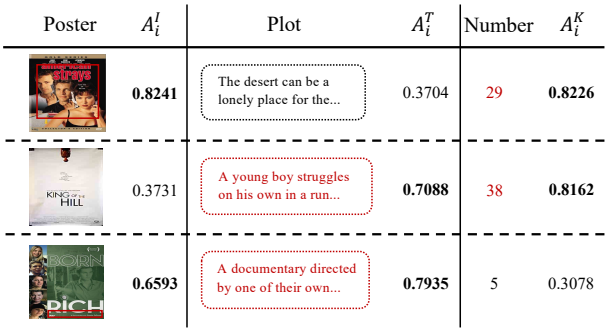


Figure 5: Case study on AT module. The modalities marked red relatively merit higher attention weights. We report the corresponding attention score of each modality (A_i^I, A_i^T and A_i^K) to verify the effectiveness of our AT module.

Table 5: Comparison with other contrastive learning methods on multi-label classification. Notably Ours (random) represents the genre embedding is initialized randomly.

Dataset	Trans Model	Micro	Macro	Weighted	Samples
MM-IMDb	MulCon [13]	0.830	0.818	0.827	0.816
	MCL [22]	0.832	0.813	0.829	0.822
	MLTC [48]	0.835	0.821	0.825	0.826
	C-GMVAE [3]	0.846	0.828	0.846	**0.840**
	Ours (random)	0.841	0.828	0.842	0.835
	Ours	**0.849**	**0.832**	**0.848**	0.839
MM-IMDb 2.0	MulCon [13]	0.806	0.785	0.792	0.784
	MCL [22]	0.811	0.794	0.802	0.787
	MLTC [48]	0.815	0.799	0.812	0.792
	C-GMVAE [3]	0.825	0.807	0.824	0.806
	Ours (random)	0.821	0.808	0.821	0.802
	Ours	**0.828**	**0.811**	**0.827**	**0.807**

that the performance drops significantly when remove G-CACL module for IDKG (GMU). It demonstrates that GMU is weak in the discriminative ability of fused feature and our G-CACL module compensates the shortcoming.

Effect of Translate Models. We compare different translate models including TransH [50], TransR [31], TransD [27], ComplEx [44], ConvE [14] and RotateE [43] to observe the effects on IDKG performance. Moreover, we list the Hit@10 results of predicting missing links on WN18 dataset, which are reported in the toolkit [21]. As show in Table 4, we notice that the RotateE achieves the best performance and ComplEx performs worst. Combining the Hit@10 results, we could draw a conclusion that the performance of IDKG is correspond to the capacity of translate model due to strong embedding representation enhancing the ability of capturing relations between entities.

Analysis on parameter τ. As illustrated in [47], smaller τ is sensitive to difficult negative samples that too small τ would destroy the embedding space due to the unproper negative samples setting. According to Figure 4, when τ is 0.1 the performance of IDKG is the best on MM-IMDb dataset. The reason may be that in our method the negative pairs are rationally constructed, thus small τ would benefit the contrastive loss.

4.4 Comparison with Multi-label Contrastive Learning Methods

To demonstrate the effectiveness of G-CACL module, we not only conduct extensive ablation study as illustrated in Section 4.3, but also replace it with existing multi-label classification methods which adopt contrastive learning. The competitors are MulCon [13], MCL [22], MLTC [48] and C-GMVAE [3] and all of them are introduced in Section 2.3. The results demonstrated in Table 5 verify that our G-CACL module is superior to other contrastive learning methods. We observe that the performance of embedding space initialized randomly is worse than that initialized from the genre embedding of our knowledge graph. We assume that it is because the knowledge graph embedding captures the discrimination between genre semantics, thus improving the effectiveness of our module.

4.5 Analysis on Attention Teacher Module

To test the validity of AT module, we present a case study as shown in Figure 5. We record A_i^I, A_i^T and A_i^K in testing phase and sample 3 cases, where the modalities that contribute more to the prediction are marked in red. It is observed that the attention module outputs the reliable scores for each modality. Notably despite that A_i^I and A_i^T are not directly supervised, they still output the rational scores. For example, the plot summary of the second and third sample should be allocated more attention weights and A_i^T of them are relative high. We infer that A_i^I and A_i^T are soft-supervised because of the shared parameters of the attention module.

5 CONCLUSION

In this paper, we proposed an effective and novel framework named IDKG. To the best of our knowledge, we are the first to apply knowledge graph technology to multimodal movie genre classification. Firstly, IDKG utilized the group relations in the knowledge graph to obtain embedding and incorporated it with other modalities. Furthermore, an Attention Teacher module was proposed to learn the distribution of the knowledge graph and guide the attention module to allocate more reliable weights. Finally, a Genre-Centroid Anchored Contrastive Learning module enhanced the discriminative ability of the fused feature. We also collected a new large-scale dataset named MM-IMDb 2.0 for movie genre classification which faces a severer class-imbalanced problem compared with the MM-IMDb dataset. Finally, we conducted extensive experiments on MM-IMDb and MM-IMDb 2.0 datasets and the experimental results demonstrated that the performance of our model was superior to existing methods. For future work, we plan to construct a multimodal domain knowledge graph of the movie field, as well as applying to more downstream tasks.

ACKNOWLEDGMENTS

We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

REFERENCES

[1] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. In *ICLR*.

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*.

[3] Junwen Bai, Shufeng Kong, and Carla P Gomes. 2022. Gaussian Mixture Variational Autoencoder with Contrastive Learning for Multi-Label Classification. In *PMLR*.

[4] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*.

[5] Tina Behrouzi, Ramin Toosi, and Mohammad Ali Akhaee. 2022. Multimodal movie genre classification using recurrent neural network. *Springer MULTIMED TOOLS APPL* (2022).

[6] Olfa Ben-Ahmed and Benoit Huet. 2018. Deep multimodal features for movie genre and interestingness prediction. In *CBMI*.

[7] Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *ACL*.

[8] Leodécio Braz, Vinícius Teixeira, Helio Pedrini, and Zanoni Dias. 2021. Image-Text Integration Using a Multimodal Fusion Network Module for Movie Genre Classification. In *ICPRS*.

[9] Alexandre Bruckert, Marc Christie, and Olivier Le Meur. 2022. Where to look at the movies: Analyzing visual attention to understand movie editing. *BEHAV RES METHODS* (2022).

[10] Paola Cascante, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180* (2019).

[11] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*.

[12] Sang-Min Choi, Sang-Ki Ko, and Yo-Sub Han. 2012. A movie recommendation algorithm based on genre correlations. *Elsevier Expert Syst. Appl.* (2012).

[13] Son D Dao, Zhao Ethan, Phung Dinh, and Cai Jianfei. 2021. Contrast learning visual attention for multi label classification. *arXiv preprint arXiv:2107.11626* (2021).

[14] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *AAAI*.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[16] Ali Mert Ertugrul and Pinar Karagoz. 2018. Movie genre classification from plot summaries using bidirectional LSTM. In *ICSC*.

[17] Edward Fish, Jon Weinbren, and Andrew Gilbert. 2020. Rethinking movie genre classification with fine-grained semantic clustering. *arXiv preprint arXiv:2012.02639* (2020).

[18] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *ACL*.

[19] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *ICLR*.

[20] Junlin Han, Mehrdad Shoeiby, Lars Petersson, and Mohammad Ali Armin. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *CVPR*.

[21] Xu Han, Shulin Cao, Xin Lv, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. OpenKE: An Open Toolkit for Knowledge Embedding. In *EMNLP*.

[22] Mohammed Hassanin, Ibrahim Radwan, Salman Khan, and Murat Tahtali. 2022. Learning discriminative representations for multi-label image recognition. *JVCIR* (2022).

[23] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. 2019. Using self-supervised learning can improve model robustness and uncertainty. *NIPS* (2019).

[24] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *ECCV*.

[25] Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. *arXiv preprint arXiv:2204.01692* (2022).

[26] Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with Contrastive Sentence Objectives Improves Discourse Performance of Language Models. In *ACL*.

[27] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *IJCNLP*.

[28] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *NIPS* (2020).

[29] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950* (2019).

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

[31] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*.

[32] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

[33] Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *NIPS* (2021).

[34] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *CVPR*.

[35] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*.

[36] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*.

[37] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. 2022. Fair contrastive learning for facial attribute classification. In *CVPR*.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[39] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *IJCV* (2017).

[40] Sethuraman Sankaran, David Yang, and Ser-Nam Lim. 2021. Refining Multimodal Representations using a modality-centric self-supervised module. (2021).

[41] Seung Byum Seo, Hyoungwook Nam, and Payam Delgosha. 2022. MM-GATBT: Enriching Multimodal Representation Using Graph Attention Network. In *ACL(Workshop)*.

[42] Gabriel S Simões, Jônatas Wehrmann, Rodrigo C Barros, and Duncan D Ruiz. 2016. Movie genre classification with convolutional neural networks. In *IJCNN*.

[43] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *ICLR* (2019).

[44] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *PMLR*.

[45] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *ECCV(Workshop)*.

[46] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. 2018. Tracking emerges by colorizing videos. In *ECCV*.

[47] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *CVPR*.

[48] Ran Wang, Xinyu Dai, et al. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *ACL*.

[49] Xiting Wang, Kunpeng Liu, Dongjie Wang, Le Wu, Yanjie Fu, and Xing Xie. 2022. Multi-level recommendation reasoning over knowledge graphs with reinforcement learning. In *WWW*.

[50] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*.

[51] Jeong A Wi, Soojin Jang, and Youngbin Kim. 2020. Poster-based multiple movie genre classification using inter-channel features. *Access* (2020).

[52] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. 2023. Bridge-Tower: Building Bridges Between Encoders in Vision-Language Representation Learning. In *AAAI*.

[53] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. A unified framework of deep networks for genre classification using movie trailer. *APPL SOFT COMPUT* (2020).

[54] Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, and Qinfei Chen. 2017. Incorporating knowledge graph embeddings into topic modeling. In *AAAI*.

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[56] Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *ECCV*. Springer.

[57] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. 2022. Use all the labels: A hierarchical multi-label contrastive learning framework. In *CVPR*.

[58] Zhongping Zhang, Yiwen Gu, Bryan A Plummer, Xin Miao, Jiayi Liu, and Huayan Wang. 2022. Effectively leveraging Multi-modal Features for Movie Genre Classification. *arXiv preprint arXiv:2203.13281* (2022).