Improving Micro-video Recommendation via Contrastive Multiple Interests

Beibei Li
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
libeibei16@otcaix.iscas.ac.cn

Beihong Jin*
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
beihong@iscas.ac.cn

Jiageng Song
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
songjiageng20@otcaix.iscas.ac.cn

Yisong Yu
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
yuyisong20@otcaix.iscas.ac.cn

Yiyuan Zheng
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Science
University of Chinese Academy of Sciences
Beijing, China
zhengyiyuan22@otcaix.iscas.ac.cn

Wei Zhuo
MX Media Co., Ltd.
Singapore
zhuowei@mxplayer.in

ABSTRACT

With the rapid increase of micro-video creators and viewers, how to make personalized recommendations from a large number of candidates to viewers begins to attract more and more attention. However, existing micro-video recommendation models rely on expensive multi-modal information and learn an overall interest embedding that cannot reflect the user's multiple interests in micro-videos. Recently, contrastive learning provides a new opportunity for refining the existing recommendation techniques. Therefore, in this paper, we propose to extract contrastive multi-interests and devise a micro-video recommendation model CMI. Specifically, CMI learns multiple interest embeddings for each user from his/her historical interaction sequence, in which the implicit orthogonal micro-video categories are used to decouple multiple user interests. Moreover, it establishes the contrastive multi-interest loss to improve the robustness of interest embeddings and the performance of recommendations. The results of experiments on two micro-video datasets demonstrate that CMI achieves state-of-the-art performance over existing baselines.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Micro-video recommendation, Contrastive learning, Multi-interest learning

ACM Reference Format:
Beibei Li, Beihong Jin, Jiageng Song, Yisong Yu, Yiyuan Zheng, and Wei Zhuo. 2022. Improving Micro-video Recommendation via Contrastive Multiple Interests. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531861

1 INTRODUCTION

In recent years, micro-video apps such as TikTok, Kwai, MX TakaTak, etc. have become increasingly popular. Generally, the micro-video app displays a single video in full-screen mode at a time and automatically plays it in a repetitive way. Usually only after viewing the cover of the micro-video or watching the content of the micro-video for a few seconds, can the user determine whether he or she is interested in this micro-video. With the explosive growth of the number of micro-videos, if the micro-videos exposed to a user do not fall within the scope of his/her interests, then the user might leave the app. Therefore, the efficient micro-video recommendation has become a crucial task.

Existing micro-video recommendation models [3, 7, 9, 17] rely on multi-modal information processing, which is too expensive to deal with large-scale micro-videos. Furthermore, they learn a single interest embedding for a user from his/her interaction sequence. However, most users have multiple interests while watching micro-videos. For example, a user is interested in both tourism and pets, and the user's interactions in the future might involve any one of the user interests. Therefore, a more reasonable approach is to learn multiple disentangled interest embeddings for users, each

[*] Corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain.
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531861

of which represents one aspect of user interests, and then generate recommendations for the users based on the learned multiple disentangled interest embeddings.

On the other hand, contrastive learning has attracted a great deal of attention recently. It augments data to discover the implicit supervision signals in the input data and maximize the agreement between differently augmented views of the same data in a certain latent space. It has obtained the success in computer vision [2], natural language processing [4, 5, 19] and other domains [13]. More recently, contrastive learning also has been introduced to the recommendation, such as sequential recommendation, recommendation based on graph neural network, and etc., which realizes the debiasing [22] and the denoising [14], and resolves the representation degeneration [15] and the cold start problem [16], improving the recommendation accuracy [10, 18, 20, 21]. We note that there exists noise in the positive interactions in the micro-video scenario since micro-videos are automatically played and sometimes users cannot judge whether they like the micro-video or not until the micro-video finishes playing. However, neither existing micro-video recommendation models nor multi-interest recommendation models [1, 8, 11, 12] utilize contrastive learning to reduce the impact of noise in the positive interactions.

In this paper, we propose a new micro-video recommendation model named CMI. Based on the implicit micro-video category information, this model learns multiple disentangled interests for a user from his/her historical interaction sequence, recalls a group of micro-videos by each interest embedding, and then forms the final recommendation result. In particular, contrastive learning is incorporated into CMI to realize the positive interaction denoising, improving the robustness of multi-interest disentanglement. Our contribution is summarized as follows.

(1) We propose CMI, a micro-video recommendation model, to explore the feasibility of combining contrastive learning with the multi-interest recommendation.
(2) We establish a multi-interest encoder based on implicit categories of items, and propose a contrastive multi-interest loss to minimize the difference between interests extracting from two augmented views of the same interaction sequence.
(3) We conduct experiments on two micro-video datasets and the experiment results show the rationality and effectiveness of the model.

2 METHODOLOGY

We denote user and item sets as $\mathcal{U}$ and $\mathcal{V}$, respectively. Further, we denote an interaction between a user and an item as a tritruple. That is, the fact that user u_i interacts with micro-video v_j at timestamp t will be represented by (i, j, t).

Given a specific user $u_i \in \mathcal{U}$, we firstly generate a historical interaction sequence over a period of time, denoted as $s_i = \left[v_{i1}, v_{i2}, \ldots, v_{i|s_i|}\right]$, in which the videos are sorted by the timestamp that user u_i interacts with the video in ascending order, and secondly learn multiple interest embeddings for each user, denoted as $\left[\mathbf{u}_i^1, \mathbf{u}_i^2, \ldots, \mathbf{u}_i^m\right]$. Then, for each interest embedding, we calculate the cosine similarity to each candidate micro-videos and recall K micro-videos with the highest K similarities, that is, a total of mK micro-videos are recalled. Finally, from the recalled micro-videos, we select top-K ones ranked by the cosine similarity as the final recommendations.

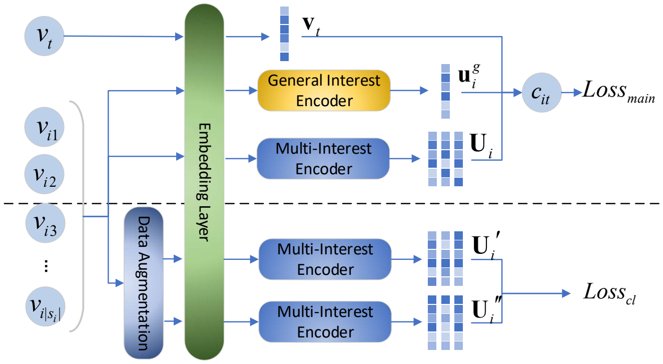

Figure 1: The architecture of CMI

2.1 Multi-interest and General Interest Encoders

We argue that the categories of items are the basis of user interests. User preferences on a certain category of items form an interest of the user. Thus, we assume there are m global categories and set learnable implicit embeddings $[\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m]$ for these m categories. For items in historical interaction sequence s_i of user u_i, we obtain the embedding of each item sequentially through the embedding layer, and form $\mathrm{S}_i = \left[\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{i|s_i|}\right]$. We use the cosine similarity between an item embedding and a category embedding as the score that measures how the item belongs to the category. More specifically, the score of item $v_{ik} \in s_i$ matching category l is calculated by Equation 1.

$$w_{ik}^l = \frac{\mathbf{g}_l^T \mathbf{v}_{ik}}{\|\mathbf{g}_l\|_2 \, \|\mathbf{v}_{ik}\|_2} \tag{1}$$

Next, the probability of item $v_{ik} \in s_i$ assigned to category l is calculated by Equation 2, where ϵ is a hyper-parameter smaller than 1 to avoid over-smoothing of probabilities.

$$p_{ik}^l = \frac{\exp\left(w_{ik}^l/\epsilon\right)}{\sum_{l=1}^m \exp\left(w_{ik}^l/\epsilon\right)} \tag{2}$$

Then, the user interest $\mathbf{u}_i^l$ corresponding to the item category l is calculated by Equation 3.

$$\mathbf{u}_i^l = \Sigma_{k=1}^{|s_i|} p_{ik}^l \mathbf{v}_{ik} \tag{3}$$

While performing the category assignment, we might encounter two degeneration cases. One is that each item has the same or similar probability of belonging to different categories. The reason of causing this degeneration is that learned item category embeddings are quite same with each other. The other is that one item category dominates the entire item embedding space, which means all items belong to that category. In order to avoid degeneration cases, we constrain both category embeddings and item embeddings within a unit hypersphere, that is, $\|\mathbf{g}_i\|_2 = \|\mathbf{v}_*\|_2 = 1$, and constrain every two category embeddings to be orthogonal, thus constructing an orthogonality loss as shown in Equation 4.

$$\mathcal{L}_{orth} = \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} (\mathbf{g}_i^T \mathbf{g}_j)^2 \tag{4}$$

In addition to encoding a user's multiple interests, we use GRU [6] to model the evolution of the general interest of the user, attaining the user's general interest $\mathbf{u}_i^g = GRU\left(\left[\mathbf{v}_{i1}, \mathbf{v}_{i2}, \ldots, \mathbf{v}_{i|s_i|}\right]\right)$.

2.2 Contrastive Regularization

We hold the view that user interests implied in the partial interactions are as same as ones implied in all the interactions. Therefore, we employ random sampling for data augmentation. Specifically, given the historical interaction sequence $s_i = [v_{i1}, \ldots, v_{i|s_i|}]$ of user u_i, we sample $\min(\mu |s_i|, f)$ micro-videos from s_i and form a new sequence s_i' according to their orders in s_i, where μ is the sampling ratio and f is the longest sequence length whose default value is 100. By randomly sampling s_i twice, we get two sequences s_i' and s_i''. Then we feed these two augmented sequences to two multi-interest encoders to learn two groups of user interests, i.e., $\mathbf{U}_i' = \left[\mathbf{u}_i^{1\prime}, \mathbf{u}_i^{2\prime}, \ldots, \mathbf{u}_i^{m\prime}\right]$ and $\mathbf{U}_i'' = \left[\mathbf{u}_i^{1\prime\prime}, \mathbf{u}_i^{2\prime\prime}, \ldots, \mathbf{u}_i^{m\prime\prime}\right]$, as shown in Equation 5, where both $\mathbf{u}_i^{k\prime}$ and $\mathbf{u}_i^{k\prime\prime}$ are interests corresponding to the k-th micro-video category.

$$\begin{aligned} \mathbf{U}_i' &= \text{Multi-Interest-Encoder}\left(s_i'\right) \\ \mathbf{U}_i'' &= \text{Multi-Interest-Encoder}\left(s_i''\right) \end{aligned} \tag{5}$$

Then, we construct a contrastive multi-interest loss as follows. For any interest embedding $\mathbf{u}_i^{k\prime} \in \mathbf{U}_i'$ of user u_i, we construct a positive pair $(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime})$, construct $2m-2$ negative pairs using $\mathbf{u}_i^{k\prime}$ and the other $2m-2$ interest embeddings of user u_i, i.e., $\mathbf{u}_i^{h\prime} \in \mathbf{U}_i'$ and $\mathbf{u}_i^{h\prime\prime} \in \mathbf{U}_i''$, where $h \in [1, m], h \neq k$. Since m is usually not too large, the number of above negative pairs is limited. Therefore, given $\mathbf{u}_i^{k\prime}$, we take the interest embeddings of every other user in the same batch to build extra negative pairs. To sum up, let the training batch be $\mathcal{B}$ and the batch size be $|\mathcal{B}|$, for each positive pair, there are $2m(|\mathcal{B}|-1) + 2m - 2 = 2(m|\mathcal{B}| - 1)$ negative pairs, which forms the negative set $\mathcal{S}^-$. Further, the contrastive multi-interest loss is defined in Equation 6, where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^T \mathbf{b} / (\|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \tau)$ and τ is a temperature parameter [10].

$$\begin{aligned} \mathcal{L}_{cl}\left(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime}\right) = &- \log \frac{e^{\text{sim}\left(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime}\right)}}{e^{\text{sim}\left(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime}\right)} + \sum_{\mathbf{s}^- \in \mathcal{S}^-} e^{\text{sim}\left(\mathbf{u}_i^{k\prime}, \mathbf{s}^-\right)}} \\ &- \log \frac{e^{\text{sim}\left(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime}\right)}}{e^{\text{sim}\left(\mathbf{u}_i^{k\prime}, \mathbf{u}_i^{k\prime\prime}\right)} + \sum_{\mathbf{s}^- \in \mathcal{S}^-} e^{\text{sim}\left(\mathbf{u}_i^{k\prime\prime}, \mathbf{s}^-\right)}} \end{aligned} \tag{6}$$

Through data augmentation and the contrastive multi-interest loss, user interest learning is no longer sensitive to a specific positive interaction, thereby reducing the impact of noisy positive interactions and realizing positive interaction denoising.

2.3 Loss Function

The interaction score between user u_i and candidate item v_t is predicted as $c_{it} = \max_{0<k\leq m}\left(\left\{\mathbf{u}_i^{kT} \mathbf{v}_t / \epsilon\right\}\right) + \mathbf{u}_i^{gT} \mathbf{v}_t$, in which $k \in [1, m]$.

In the training process, for each positive sample v_p^i of user u_i, we need to randomly sample n micro-videos that have never been interacted with from the full micro-videos as negative samples. However, in order to avoid high sampling cost, given a positive sample, we only sample one negative sample, that is, n is 1. Besides, we take the positive sampling items and negative sampling items of other users in the same batch as the negative samples, thus forming a negative sample set $\mathcal{N}$. We then adopt the following cross-entropy loss as the main part of the loss.

$$\mathcal{L}_{\text{main}}\left(u_i, v_p^i\right) = -\ln \frac{\exp\left(c_{ip}\right)}{\sum_{v_* \in \left\{\mathcal{N} \cup v_p^i\right\}} \exp\left(c_{i*}\right)} \tag{7}$$

Finally, our loss function is shown in Equation 8, where λ_* is the regularization coefficient.

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_{cl} \mathcal{L}_{cl} + \lambda_{orth} \mathcal{L}_{orth} \tag{8}$$

3 EXPERIMENTS

3.1 Experiment Setup

3.1.1 *Datasets.*

We conduct experiments on two micro-video datasets.

1. **WeChat**. This is a public dataset released by WeChat Big Data Challenge 2021[1]. This dataset contains user interactions on WeChat Channels, including explicit satisfaction interactions such as likes and favorites and implicit engagement interactions such as playing.
2. **TakaTak**. This dataset is collected from TakaTak, a micro-video app for Indian users. The dataset contains interaction records of 50,000 anonymous users in four weeks.

The statistics of the two datasets are shown in Table 2. For a dataset spanning h day, we construct the training set with the interactions in the first $h-2$ days, the validation set with the interactions in the $(h-1)$-th day, and the test set with the interactions of the h-th day.

3.1.2 *Metrics.*

Here, Recall@K and HitRate@K are used as metrics to evaluate the quality of the recommendations.

3.1.3 *Competitors.*

We choose the following multi-interest recommendation models as competitors.

1. **Ocotopus** [11]: It constructs an elastic archive network to extract diverse interests of users.
2. **MIND** [8]: It adjusts the dynamic routing algorithm in the capsule network to extract multiple interests of users.
3. **ComiRec-DR** [1]: It adopts the original dynamic routing algorithm of the capsule network to learn multiple user interests.
4. **ComiRec-SA** [1]: It uses a multi-head attention mechanism to capture the multiple interests of users.

[1] https://algo.weixin.qq.com/problem-description

Table 1: Recommendation accuracy on two datasets. #I. denotes the number of interests. The number in a bold type is the best performance in each column. The underlined number is the second best in each column.

		WeChat							**TakaTak**					
		Recall			**HitRate**				**Recall**			**HitRate**		
	#I.	**@10**	**@20**	**@50**	**@10**	**@20**	**@50**	**#I.**	**@10**	**@20**	**@50**	**@10**	**@20**	**@50**
Octopus	1	0.0057	0.0125	0.0400	0.0442	0.0917	0.2332	1	0.0076	0.0160	0.0447	0.1457	0.2533	0.4393
MIND	1	0.0296	0.0521	0.1025	0.1774	0.2791	0.4514	1	0.0222	0.0389	0.0773	0.2139	0.3263	0.4977
ComiRec-DR	1	0.0292	0.0525	0.1049	0.1790	0.2893	0.4621	1	0.0226	0.0392	0.0769	0.2345	0.3427	0.5144
ComiRec-SA	1	0.0297	0.0538	0.1079	0.1806	0.2938	0.4684	1	0.0239	0.0409	0.0752	0.2567	0.3665	0.5207
DSSRec	1	0.0327	0.0578	0.1161	0.1971	0.3064	0.4854	8	**0.0244**	0.0408	0.0749	0.2558	0.3704	0.5259
CMI	8	**0.0424**	**0.0717**	**0.1342**	**0.2436**	**0.3612**	**0.5292**	8	0.0210	**0.0415**	**0.0877**	**0.2912**	**0.4172**	**0.5744**
Improv.	/	29.66%	24.05%	15.59%	23.59%	17.89%	9.02%	/	/	1.72%	17.09%	13.84%	12.63%	9.22%

Table 2: The statistics of the two datasets.

Dataset	#Users	#Micro-videos	#Interactions	Density
WeChat	20000	77557	2666296	0.17%
TakaTak	50000	157691	33863980	0.45%

(5) **DSSRec** [12]: It disentangles multiple user intentions through self-supervised learning.

To be fair, we do not compare with models that rely on multi-modal information.

3.1.4 Implementation Details. We implement our model with PyTorch, and initialize the parameters with the uniform distribution $U(-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}})$, where d is the dimension of embeddings. We optimize the model through Adam. Hyper-parameters ϵ, τ and λ_{cl} are searched in [1, 0.1, 0.01, 0.001, 0.0001, 0.00001], and finally we set $\epsilon = 0.1$, $\tau = 0.1$, and $\lambda_{cl} = 0.01$. λ_{orth} is searched in $[15, 10, 5, 1, 0.5]$ and finally we set it to 10. The sampling rate μ is set to 0.5.

For the sake of fairness, in all the experiments, we set the embedding dimension to 64 and the batch size to 1024. We stop training when Recall@50 on the validation set has not been improved in 5 consecutive epochs on the WeChat dataset and 10 consecutive epochs on the Takatak dataset. Besides, for MIND, ComiRec-DR, ComiRec-SA, and DSSRec, we use the open-source code released on Github [2,3].

3.2 Performance Comparison

The experimental results of performance comparison are shown in Table 1, from which we have the following observations.

(1) Multi-interest competitors except for CMI, with few exceptions, reach the best results while the number of interests is 1, indicating that these models cannot effectively capture multiple interests of a user in micro-videos.

(2) The two dynamic routing-based models MIND and ComiRec-DR are not as good as ComiRec-SA and DSSRec. This is probably because both MIND and ComiRec-DR do not fully utilize the sequential relationship between historical interactions, but ComiRec-SA and DSSRec do. In addition, DSSRec adopts a novel seq2seq training strategy that leverages additional supervision signals, thus obtaining better performance.

(3) Octopus performs worst. That is probably because it aggressively routes every item into one interest exclusively at the beginning of training, which makes it easy to trap the parameters in a local optimum.

(4) On two datasets, CMI far outperforms the competitors on most metrics, which demonstrates that CMI generates recommendations with both high accuracy and excellent coverage. The reason is that we avoid model degeneration by setting the category embeddings orthogonal and extract multiple heterogeneous user interests. In addition, we achieve positive interaction denoising via contrastive learning, which improves the robustness of interest embeddings.

3.3 Ablation Study

While setting the number of interests to 8, we observe the performance of two model variants: CMI-CL and CMI-G, where the former is the CMI removing the contrastive multi-interest loss and the latter is the CMI without the general interest encoder. From the Table 3, we find the model variants suffer severe declines in performance. This confirms the feasibility and effectiveness of contrastive learning for multi-interest recommendation, and shows that whether the candidate item matches the general user preferences is also important.

3.4 Effect of the Number of Interests

We set the number of interests to [1, 2, 4, 8, 16] successively and conduct experiments. The experimental results are shown in Table 4. It can be seen that CMI reaches the best performance when the number of interests is 8 rather than 1. This confirms the necessity and effectiveness of extracting multiple interests in the micro-video recommendation scenarios.

[2] https://github.com/THUDM/ComiRec

[3] https://github.com/abinashsinha330/DSSRec

Table 3: Ablation study on WeChat. The values in parentheses are the percentages of decline relative to the original model.

		CMI-CL	CMI-G	CMI
Recall	@10	0.039(-8.02%)	0.0342(-19.34%)	**0.0424**
	@20	0.0665(-7.25%)	0.0589(-17.85%)	**0.0717**
	@50	0.1285(-4.25%)	0.1165(-13.19%)	**0.1342**
HitRate	@10	0.2286(-6.16%)	0.2061(-15.39%)	**0.2436**
	@20	0.3443(-4.68%)	0.3181(-11.93%)	**0.3612**
	@50	0.5188(-1.93%)	0.4935(-6.71%)	**0.5290**

Table 4: The effect of the number of interests on WeChat.

	#I.	1	2	4	8	16
Recall	@10	0.0303	0.0404	0.0409	**0.0428**	0.0412
	@20	0.0530	0.0699	0.0694	**0.0718**	0.0700
	@50	0.1039	0.1343	0.1333	**0.1364**	0.1314
HitRate	@10	0.1969	0.2383	0.2384	**0.2458**	0.2390
	@20	0.3012	0.3547	0.3516	**0.3587**	0.3557
	@50	0.4646	0.5330	0.5271	**0.5322**	0.5238

4 CONCLUSION

This paper proposes a micro-video recommendation model CMI. The CMI model devises a multi-interest encoder and constructs a contrastive multi-interest loss to achieve positive interaction denoising and recommendation performance improvement. The performance of CMI on two micro-video datasets far exceeds other existing multi-interest models. The results of ablation study demonstrate that fusing contrastive learning into multi-interest extracting in micro-video recommendation is feasible and effective.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 62072450 and the 2021 joint project with MX Media.

REFERENCES

[1] Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Controllable Multi-Interest Framework for Recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2942–2951. https://doi.org/10.1145/3394486.3403344

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learnin* 119 (June 2020), 1597–1607. http://proceedings.mlr.press/v119/chen20j.html

[3] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal Hierarchical Attention at Category- and Item-Level for Micro-Video Click-Through Prediction. In *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*. ACM Press, 1146–1153. https://doi.org/10.1145/3240508.3240617

[4] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. *arXiv:2005.12766 [cs, stat]* (June 2020). http://arxiv.org/abs/2005.12766 arXiv: 2005.12766.

[5] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. *arXiv:2006.03659 [cs]* (May 2021). http://arxiv.org/abs/2006.03659 arXiv: 2006.03659.

[6] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 843–852. https://doi.org/10.1145/3269206.3271761

[7] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What Aspect Do You Like: Multi-scale Time-aware User Interest Modeling for Micro-video Recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 3487–3495. https://doi.org/10.1145/3394171.3413653

[8] Chao Li, Zhiyuan Liu, Mengmeng Wu, Yuchi Xu, Huan Zhao, Pipei Huang, Guoliang Kang, Qiwei Chen, Wei Li, and Dik Lun Lee. 2019. Multi-Interest Network with Dynamic Routing for Recommendation at Tmall. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 2615–2623. https://doi.org/10.1145/3357384.3357814

[9] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. 2019. User-Video Co-Attention Network for Personalized Micro-video Recommendation. In *The World Wide Web Conference on - WWW '19*. ACM Press, 3020–3026. https://doi.org/10.1145/3308558.3313513

[10] Zhiwei Liu, Yongjun Chen, Jia Li, Philip S. Yu, Julian McAuley, and Caiming Xiong. 2021. Contrastive Self-supervised Sequential Recommendation with Robust Augmentation. *arXiv:2108.06479 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.06479 arXiv: 2108.06479.

[11] Zheng Liu, Jianxun Lian, Junhan Yang, Defu Lian, and Xing Xie. 2020. Octopus: Comprehensive and Elastic User Representation for the Generation of Recommendation Candidates. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 289–298. https://doi.org/10.1145/3397271.3401088

[12] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled Self-Supervision in Sequential Recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 483–491. https://doi.org/10.1145/3394486.3403091

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]* (Jan. 2019). http://arxiv.org/abs/1807.03748 arXiv: 1807.03748.

[14] Yuqi Qin, Pengfei Wang, and Chenliang Li. 2021. The World is Binary: Contrastive Learning for Denoising Next Basket Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 859–868. https://doi.org/10.1145/3404835.3462836

[15] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2021. Contrastive Learning for Representation Degeneration Problem in Sequential Recommendation. *arXiv:2110.05730 [cs]* (Nov. 2021). https://doi.org/10.1145/3488560.3498433 arXiv: 2110.05730.

[16] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-Start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, 5382–5390. https://doi.org/10.1145/3474085.3475665

[17] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 1437–1445. https://doi.org/10.1145/3343031.3351034

[18] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. *arXiv:2010.10783 [cs]* (May 2021). https://doi.org/10.1145/3404835.3462862 arXiv: 2010.10783.

[19] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. *arXiv:2012.15466 [cs]* (Dec. 2020). http://arxiv.org/abs/2012.15466 arXiv: 2012.15466.

[20] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. 2021. Contrastive Learning for Sequential Recommendation. *arXiv:2010.14395 [cs]* (Feb. 2021). http://arxiv.org/abs/2010.14395 arXiv: 2010.14395.

[21] Junliang Yu, Hongzhi Yin, Xin Xia, Lizhen Cui, and Quoc Viet Hung Nguyen. 2021. Graph Augmentation-Free Contrastive Learning for Recommendation. *arXiv:2112.08679 [cs]* (Dec. 2021). http://arxiv.org/abs/2112.08679 arXiv: 2112.08679.

[22] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive Learning for Debiased Candidate Generation in Large-Scale Recommender Systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, 3985–3995. https://doi.org/10.1145/3447548.3467102