doi:10.20944/preprints202206.0412.v1

Dictionary Learning for Personalized Multimodal Recommendation

Liu Bo, Henri Millett, Bruno L. Rebola, Plank Svensen

Abstract—In today's Web 2.0 era, online social media has become an integral part of our lives. In the course of the information revolution, the form of information has undergone a radical change, from simple text information to today's integrated video, image, text and audio, and there has also been a great change in the way of dissemination and access, as people nowadays do not just rely on traditional media to passively receive information, but more actively and selectively obtain information from social media. Therefore, it has become a great challenge for us to effectively utilize these massive and integrated multi-modal media information to form an effective system of retrieval, browsing, analysis and usage. Unlike movies and traditional long-form video content, micro-videos are usually short in length, between a few seconds and tens of seconds, which allows users to quickly browse different contents and make full use of the fragmented time in their lives, while users can also share their micro-videos to their friends or the public, forming a unique social way. Video contains rich multimodal information, and fusing information from multiple modalities in a video recommendation task can improve the accuracy of the video recommendation task.According to the micro-video recommendation task, a new combinatorial network model is proposed to combine the discrete features of each modality into the overall features of various modalities through the network, and then fuse the various modal features to obtain the overall video features, which will be used for recommendation. In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

Index Terms—Dictionary learning; Recommender system; Personalized recommendation; Multimodal, Cluster

1 Introduction

With the rapid development of information technology, the form of media data we receive has changed from single text data to multimodal data with more vivid forms and richer contents. At the same time, the popularity of various digital information collection devices and the Internet has made micro-video, a form of data filmed, produced and shared by users, the most emerging and popular one. In today's Web 2.0 era, online social media has become an integral part of our lives. In the course of the information revolution, the form of information has undergone a radical change, from simple text information to today's integrated video, image, text and audio, and there has also been a great change in the way of dissemination and access, as people nowadays do not just rely on traditional media to passively receive information, but more actively and selectively obtain information from social media. Therefore, it has become a great challenge for us to effectively utilize these massive and integrated multi-modal media information to form an effective system of retrieval, browsing, analysis and usage.

Unlike movies and traditional long-form video content, micro-videos are usually short in length, between a few seconds and tens of seconds, which allows users to quickly browse different contents and make full use of the fragmented time in their lives, while users can also share their micro-videos to their friends or the public, forming a unique social way. However, these features of micro-videos also bring some disadvantages. The short length and large quantity make the amount of information too large and complicated, and users find it difficult to find the content they are interested in, and often spend a lot of time on finding quality content. How to let users spend less time to find more interesting content is the core problem that every social media platform needs to solve. For social media platforms, if they rely on traditional search engines for passive push mode, they can only search according to the title of the video, keywords and other information, which does not cover the entire content for micro-videos, and thus usually cannot push the right content to users; for video authors, it also makes them focus more on the naming of the title, which obviously deviates from the original purpose of micro-videos This obviously deviates from the original purpose of micro-videos, which is convenience and sharing. Therefore, platforms are now relying more on personalized recommendation systems to actively push to users, and in this process, how to push quality and interesting content to users while filtering out unpopular content as much as possible is a very important and meaningful task for research.

For the study of micro-videos, Wang et al. [1] studied the characteristics of teaching-related micro-videos and their semantic representation, while focusing on the comparative study of the relevance between different teaching micro-videos. Zhang et al. [2] proposed a micro-video segmentation method based on color histogram and local optimization. Redi et al. [3] conducted a study on the creativity of micro-video data on social media platforms by proposing a definition of creativity, constructing a small

[*]Liu Bo is the corresponding author.

- *Liu Bo, Henri Millett and Plank Svensen are withthe School of Computing, National University of Singapore, Singapore. (e-mail: liubo20222022@gmail.com, HenriMillett@gmail.com, Plank Svensen@gmail.com).*
- *Bruno L. Rebola is with the Indian Institute of Technology, India. (e-mail: brunolRubola@hotmail.com).*

dataset on the creativity of micro-videos, and extracting various features to conduct experiments on creativity prediction. Sano et al. [4] studied the characteristics of looped playback of micro-videos. Nguyen et al. [5] investigated the own characteristics of micro-video data, such as multi-view shooting, multi-clip splicing. The authors argued that micro-videos are inherently ”ready-for-analysis” due to their length limitation and the fact that each frame contains much more important information than traditional long videos. A large open dataset of micro-videos is constructed, classified according to viewpoints and tags, and combined with convolutional neural networks and gesture recognition techniques to study the understanding of micro-videos. As for the topic of popularity prediction, it has attracted the attention of many researchers due to its great potential for commercial applications such as advertising placement [6], and most of the previous research works have been focused on text [7], image [8] and traditional video [9, 10].

With the advent of Convolutional Neural Networks (CNN) [11, 12, 13, 14, 15], people no longer needs to design feature descriptors manually, but automatically learns video semantic features and understands image content, and eventually achieved great success in image detection and retrieval. FaceBook even proposed to use 3D convolutional neural network to extract spatio-temporal information [16].

Researchers have also tried to apply it to video recommender modeling. Such methods usually first extract video frame features using CNNs and then input these features into RNNs [17, 18]. LSTM [19, 20], on the other hand, is a commonly used RNN model for modeling video long-time dependencies. For example, Ng et al. [21]used LSTM to fuse video frames and optical flow f eatures, and experimentally verified t he r obustness o f L STM networks to optical flow n oise a nd t he e ffectiveness o f video sequence feature fusion with LSTM. Video contains rich multimodal information, and fusing information from multiple modalities in a video recommendation task can improve the accuracy of the video recommendation task.The main contributions of this paper are as follows:

- Study the extraction methods of visual features, audio features and title features in videos. In this paper, visual information is mainly extracted by dividing the video into multiple frames, and then each frame is passed through a pre-trained network to extract visual features. The audio information is extracted by dividing the audio into frames, and then each frame is digitally processed to obtain the spectrum, and then the audio features are extracted by VGG network. The title information is mainly obtained by word mapping into word vectors through word cutting technique to get word features.
- According to the micro-video recommendation task, a new combinatorial network model is proposed to combine the discrete features of each modality into the overall features of various modalities through the network, and then fuse the various modal features to obtain the overall video features, which will be used for recommendation.
- In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

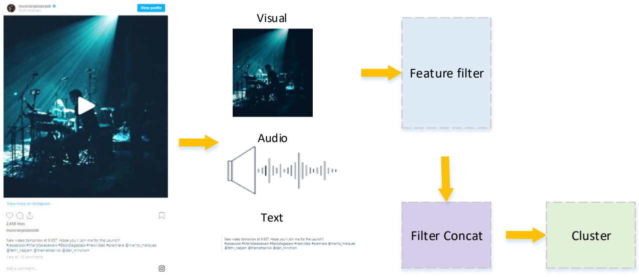


Fig. 1: Schematic diagram of multimodal.

2 Related Work

2.1 Personalized recommendation system

The research on personalized recommendation systems is actually very short-lived. Its starting time would be from the 1990s, but this does not affect the scope of personalized recommendation systems’ impact on human life. Nowadays, it has penetrated into all aspects of our life, such as: e-commerce, news, travel, music, entertainment, etc [22]. With the introduction of personalized recommendation technology, we do not need to consume other resources to select content, but simply select the relevant interest categories, and the recommendation system will ”tailor” the content to the user’s preferences after a few seconds. During the user’s browsing process, the system collects the user’s browsing history and analyzes it to understand the user’s interest distribution, so as to achieve the purpose of ”tailoring” [23]. The content filtering-based recommendation first mines the user’s history, i.e., the user’s browsing history, e.g., when the user reviews or describes an item, then it can be mined, and finally the TF-IDF [24] algorithm is used to determine the importance of words. Collaborative filtering based recommendation Collaborative filtering recommendation systems have been used in many fields nowadays. For example, Amazon recommendation in e-commerce, Instagram and Tiktok in micro-video APP, and google recommendation in search engine. In fact, the key is to use the acquired information data, find out the existence of predicted users with similar preferences to these information data, and calculate the similarity, and then generate the set of neighbors, and finally make recommendations based on the preferences of neighbors with higher similarity, and finally achieve the recommendation effect [25, 26, 27]. The core idea of the user-based collaborative filtering algorithm [28] is to use certain algorithms to calculate and speculate whether there is some connection between users, and this connection can be used by us to find the same interest preferences between two unused users, and then recommend items of interest to the target group of users who have the same connection. If we can obtain sufficient amount of data, then the final recommendation effect will also be more accurate [29]. The core idea of item-based collaborative filtering algorithm is to use the algorithm to calculate whether there is some connection between items and items, and then use this connection to then make effective and reasonable recommendations [30].

2.2 Multimodal recommendation

Co-training is a classical multi-view learning method for dealing with those cases where the dataset contains a large number of unlabeled samples, and belongs to the category of semi-supervised learning in machine learning. However, in real research problems, the requirement of the nature of full redundancy of views is difficult to satisfy, Nigam Ket al. [31] conducted an experimental study on the performance of co-training algorithms on problems that do not have fully redundant views Goldman S et al. [32] proposed a co-training algorithm that does not require fully redundant views. Two different classifiers are trained from the same attribute set, and in the training phase, each classifier still labels the unlabeled example with higher confidence and submits it to the other classifier for learning, while in the testing phase, the labeled confidence of the two classifiers on the test example is first estimated and the one with higher confidence is selected for the final prediction. Thereafter, they also extended the algorithm so that it can use several different kinds of classifiers. To further relax the constraints of co-training, Zhou ZH et al. [33] proposed a tri-training algorithm that neither requires a fully redundant view nor requires the use of different types of classifiers. Subspace learning is an important research area in machine learning that has been widely followed by researchers and the research results have been applied in various specific fields.

In the current study, Liu Meng et al. [34] proposed a Joint Sequential-Sparse approach, where micro-videos are processed in frames, and features are extracted on LSTM according to Visual, Acoustic and Texture respectively, and the obtained features are mapped to the same space. Bo Peng et al. [35] proposed an unsupervised clustering algorithm based on motion scenes, which combines static scene features and dynamic features. A model through contextual constraints was established. First, the complementarity of multi-view subspace representations in each context is explored by single-view and multi-view constraints. Then, by computing the association matrix of contextual constraints and introducing MSIC to mutually regulate the inconsistency of subspace representations in scenes and motions. Finally, an overall objective function is constructed to guarantee the video motion clustering results by jointly constraining the complementarity of multiple views and the consistency of multiple contexts. Jun Yang et al. [36] proposed a method to identify the behavior of construction site employees using semantic information. A nonparametric data-driven scene analysis approach was used to identify the constructed objects. A context-based action recognition model was learned from the training data. Then, action recognition was improved by using the identified constructed objects. Jingyi Hou et al. [37] proposed a decomposed action scene network FASNet. FASNet consists of two parts, one is an Attention-based CANET network, which is able to encode local spatio-temporal features to learn features with good robustness. The other part consists of a fusion network, which mainly fuses spatio-temporal features and contextual The other part consists of a fusion network, which mainly fuses spatio-temporal features and contextual features to learn more descriptive feature information.

3 Methodology

3.1 Single-modal Dictionary Learning

Suppose $X = [x_1, x_2, \ldots, x_N] \in \mathbb{R}^{N\times D}$ is a training data set containing V data samples and the feature dimension of each training sample is D. The unimodal dictionary learning algorithm generally obtains the ideal dictionary $\mathcal{D} \in \mathbb{R}^{D\times K}$ by minimizing the solution formula:

$$argmin_{D\in\mathcal{D}}\frac{1}{2}||X - DA||_F^2 + \lambda||A||, \tag{1}$$

where K denotes the number of atoms in the dictionary, D denotes the j-th column of the dictionary $\mathcal{D}$, A denotes the sparse representation of X with respect to $\mathcal{D}$, and λ denotes the regularization parameter.

3.2 Multi-modal Dictionary Learning

Since real-world objects are usually described by information from multiple modalities in different ways, it is natural to extend single modal dictionary learning to handle multimodal data.

Assume that there are V training samples and each sample is represented by the modal information, i.e., $(x_n^1, x_n^2, \ldots, x_n^m)$, where $x_n^m \in X^m(m = 1, \ldots, M)$ denotes the feature representation information of the m-th modality of sample x_n. $X^m \in \mathbb{R}^{(D\times N)}$ denotes the feature information of the m-th modality of all N samples. D_m denotes the m-th modal feature dimension, and D^m denotes the dictionary of the m-th modality. Thus, the sparse representation $A_n = [A_1, \ldots, A_1^m]$ of the n-th sample and the set of multimodal dictionaries $\mathcal{D} = [D^1, \ldots, D^m]$ can be obtained by solving the reconstruction function with $\updownarrow$ parametrization as follows:

$$min_{A,\mathcal{D}}\frac{1}{2}\sum_{n-1}^{N}\sum_{m=1}^{M}||x_n^m - a_nD^m||_2 + \lambda||A_n||_{2,1}, \tag{2}$$

where a_n^m denotes the sparse representation of the sample x_n^m with respect to the dictionary D. The role of the $\updownarrow$ parametric number $||A_n||_{2,1}$ is to constrain the matrix to have row sparsity, i.e., to force different modalities to reconstruct the input samples using dictionary atoms at the same positions, i.e., the i-th row of the matrix A_v is either all zero or all non-zero values.

3.3 Model Optimization

Update each column of D_t in turn. Here the j-th column of D_t is used as an example to illustrate the process. Define $d_j(t)$ as the j-th column of the dictionary D_t.

$$g(D_t) = \frac{1}{2}\sum_{i=1}||x_iD_ta_i||_2^2, \tag{3}$$

Then we can get:

$$d_j(t) = \frac{\sum_{i=1}^{t} a_{ij}(x_i - Da_i)}{\sum_{i=1}^{t} a_{ij}a_{ij}^T}, \tag{4}$$

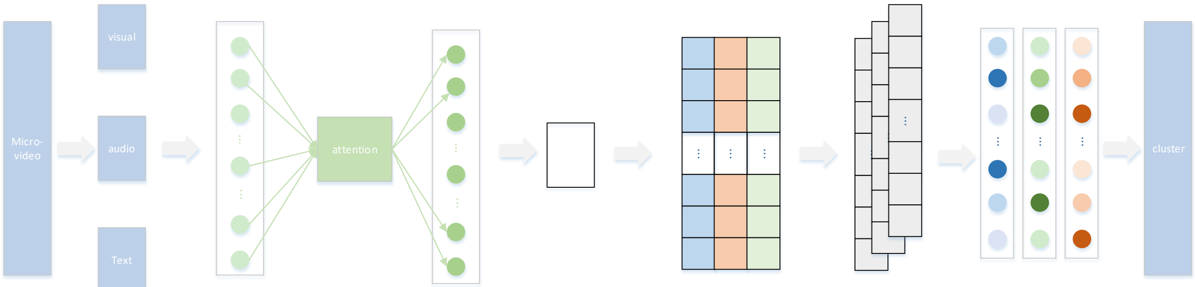


Fig. 2: Overall schematic diagram of our model.

where a_{ij} is the j-th elment of a_i.

Using the additive property of linear solutions, we can get:

$$d_j(t) = \frac{\sum_{i=1}^{t} a_{ij}(x_i - D_{t-1}a_i)}{\sum_{i=1}^{t} a_{ij}a_{ij}^T} + d_j(t-1), \quad (5)$$

As $d_j(t) \leq 1$, the normalization is as following:

$$d_j(t) = \frac{1}{max(||d_j(t)||, 1)} d_j(t), \quad (6)$$

3.4 Multimodal Fusion Network

We build a fusion sub-network to model the interaction information within and between unimodal and multimodal. Specifically, it is assumed that a D-dimensional feature m_c and a D dimensional query feature $\hat{q}$ are obtained from the memory attention network. The goal of mean pooling here is to learn a high-level representation for dimensionality reduction. Representation learning based on mean pooling is equivalent to applying a linear filter of size n to each input embedding, and the value of each element in the output vector is the mean value of the elements in the corresponding convolution window.

$$f_{cq} = [m_c, m_c \otimes \hat{q}, \hat{q}, 1], \quad (7)$$

where $\otimes$ denotes the outer product between vectors. f_{cq} has three distinct subregions and contains all possible combinations of unimodal embeddings, where subregions m_c and $\hat{q}$ form unimodal intra-interactions, and subregion $m_c \otimes \hat{q}$ captures multimodal interactions.

The loss function, formally, is:

$$L = \alpha_1 \sum_{(c,q)\in p} log(1+exp(-s_{cq}))+\alpha_2 \sum_{(c,q)\in N} log(1+exp(s_{cq})), \quad (8)$$

where α_1 and α_2 are hyperparameters of the weights.

4 EXPERIMENTS

4.1 Dataset

The test data used for the experiment is a publicly available dataset from Tiktok. The dataset contains about 1.5 million pieces of data and 2000 users' viewing records of 12,000 micro-videos. The original dataset is randomly divided into a training set and a test set, accounting for 80% and 20%, respectively.

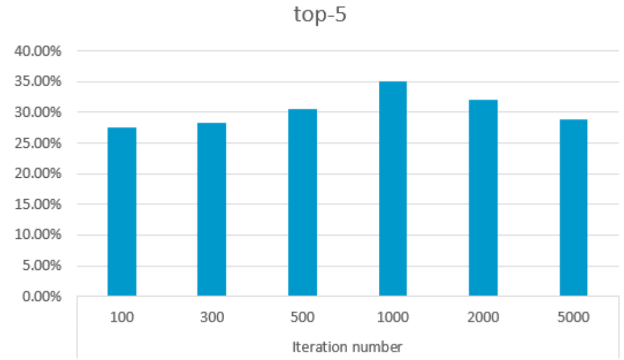


Fig. 3: Test set top-5 result.

4.2 Baselines

FM [38]: Factorization Machine, which simulates first-order feature importance and second-order feature interactions.

DeepFM [39]: DeepFM is an end-to-end model of a joint decomposer and multilayer sensing machine, which uses deep neural networks and factorization machines to model the interactions of higher-order features and lower-order features, respectively.

VBPR [40]:The model integrates visual information into the prediction of people's preferences, and VBPR is a significant improvement over matrix decomposition models that rely only on user hidden vectors and item hidden vectors.

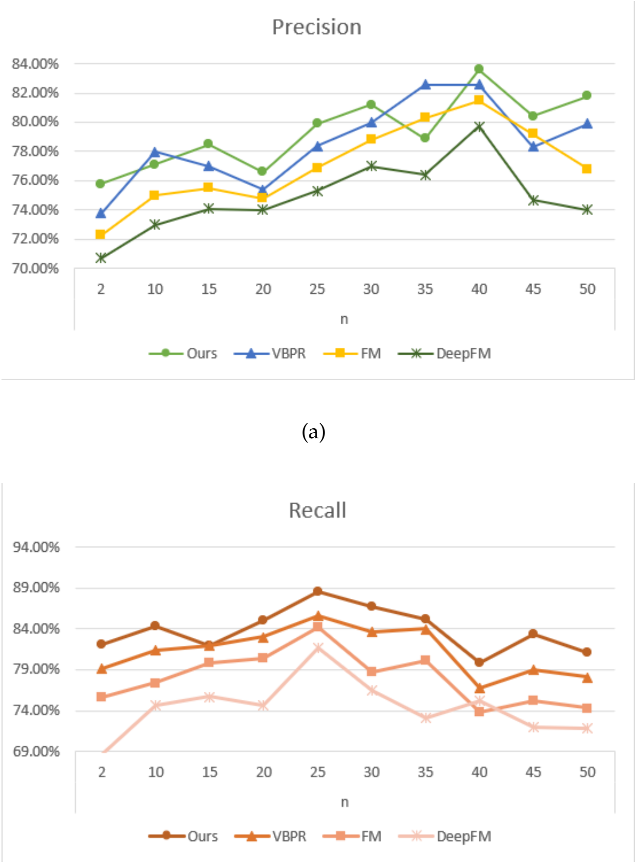
4.3 Metrics

There are many different evaluation systems for the goodness and efficiency of recommendation systems. In this paper, we use Precision and Recall to judge the efficiency and effectiveness of the recommendation function.

$$Precision = \frac{\sum_{u\in U} |R_{(u)} \cap S_{(u)}|}{\sum_{u\in U} |R_{(u)}|}, \quad (9)$$

$$Recall = \frac{\sum_{u\in U} |R_{(u)} \cap S_{(u)}|}{\sum_{u\in U} |S_{(u)}|}, \quad (10)$$

where $R_{(u)}$ denotes the final result recommended to the user, and $S_{(u)}$ denotes the data source generated by all user actions.



(a)

(b)

Fig. 4: Comparison of the experiment results of different models.

4.4 Result Analysis

During the learning process, several experiments were conducted at 100, 200, 500, 1000, 2000 and 5000 iterations, and the average values were obtained. After that, the top-5 precision results with different iterations were compared, and the results are shown in Figure.3. Through this figure, it can be found that the top-5 precision is 0.33 when the number of iterations is 1000, but the setting of the number of iterations is not better, and the accuracy tends to decrease when the number of iterations increases from 1000 to 5000.

As the figure shows clearly the recommendation algorithm (Ours) proposed in this paper is more than FM, DeepFM and VBPR in the field of micro-video recommendation. The recommendation model proposed in this paper is compared with the common item-based collaborative filtering modl FM, where the former focuses on the relationship between users, while the latter focuses on the relationship between micro-videos. For micro-videos, the information itself is less, the quality varies, and the relationship between uploaded videos by uploaders is misleading and not easy to distinguish, which is difficult for finding the relationship between videos. While it is significantly easier to analyze the relationship between users, just find the user groups that are similar to the users. In this paper, the recommendation model is compared with DeepFM, and the uncertain neighborhood collaborative filtering algorithm is one of the star algorithms, which balances the relationship between videos and users by adding an uncertainty factor. However, due to the sparse ratings of micro-videos, it does not fill this data matrix by other means, leading to a decrease in recommendation quality. The model in this paper, by heavily analyzing the explicit and invisible behaviors of users greatly expands the data matrix and ensures the quality of recommendations.

5 Conclusion and Future Work

With the advent of ConvolutionalNeuralNetworks (CNN) , people no longer needs to design feature descriptors manually, but automatically learns video semantic features and understands image content, and eventually achieved great success in image classification, detection and retrieval. The algorithm of micro-video recognition based on deep learning has surpassed the iDT algorithm, making these traditional methods gradually fade out of people's view. FaceBook even proposed to use 3D convolutional neural network to extract spatio-temporal information, which injected new vitality into the research of video recommendation.

Researchers have also tried to apply it to video temporal modeling. Such methods usually first extract video frame features using CNNs and then input these features into RNNs in temporal order for temporal modeling. LSTM, on the other hand, is a commonly used RNN model for modeling video long-time dependencies. For example, Ng et al.used LSTM to fuse video frames and optical flow features, and experimentally verified the robustness of LSTM networks to optical flow noise and the effectiveness of video sequence feature fusion with LSTM. Video contains rich multimodal information, and fusing information from multiple modalities in a video recommendation task can improve the accuracy of the video recommendation task.The main contributions of this paper are as follows: Study the extraction methods of visual features, audio features and title features in videos. In this paper, visual information is mainly extracted by dividing the video into multiple frames, and then each frame is passed through a pre-trained network to extract visual features. The audio information is extracted by dividing the audio into frames, and then each frame is digitally processed to obtain the spectrum, and then the audio features are extracted by VGG network. The title information is mainly obtained by word mapping into word vectors through word cutting technique to get word features. According to the micro-video recommendation task, a new combinatorial network model is proposed to combine the discrete features of each modality into the overall features of various modalities through the network, and then fuse the various modal features to obtain the overall video features, which will be used for recommendation. In order to verify the effectiveness of the algorithm proposed in this paper, experiments are conducted in the public dataset, and it is shown the effectiveness of our model.

6 Conflict of Interest Statement

All authors have no conflict and declare that: (i) no support, financial or otherwise, has been received from any organization that may have an interest in the submitted work ; and (ii) there are no other relationships or activities that could appear to have influenced the submitted work.

References

[1] M. Wang and D. Kang, “Research on semantic representation to promote the correlation of instructional micro video,” in *2015 11th International Conference on Computational Intelligence and Security (CIS)*. IEEE, 2015, pp. 470–473.

[2] B. Zhang and Y. Liu, “Micro-video segmentation based on histogram and local optimal solution method,” in *Chinese conference on image and graphics technologies*. Springer, 2015, pp. 292–299.

[3] M. Redi, N. O’Hare, R. Schifanella, M. Trevisiol, and A. Jaimes, “6 seconds of sound and vision: Creativity in micro-videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 4272–4279.

[4] S. Sano, T. Yamasaki, and K. Aizawa, “Degree of loop assessment in microvideo,” in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5182–5186.

[5] P. X. Nguyen, G. Rogez, C. Fowlkes, and D. Ramanan, “The open world of micro-videos,” *arXiv preprint arXiv:1603.09439*, 2016.

[6] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, “Personalized hashtag recommendation for micro-videos,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1446–1454.

[7] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in twitter,” in *Proceedings of the 20th international conference companion on World wide web*, 2011, pp. 57–58.

[8] P. J. McParlane, Y. Moshfeghi, and J. M. Jose, “” nobody comes here anymore, it’s too crowded”; predicting image popularity on flickr,” in *Proceedings of international conference on multimedia retrieval*, 2014, pp. 385–391.

[9] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu, “On popularity prediction of videos shared in online social networks,” in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 169–178.

[10] S. D. Roy, T. Mei, W. Zeng, and S. Li, “Towards cross-domain learning for social video popularity prediction,” *IEEE Transactions on multimedia*, vol. 15, no. 6, pp. 1255–1267, 2013.

[11] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in neural information processing systems*, vol. 27, 2014.

[12] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, “Sst: Single-stream temporal action proposals,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2911–2920.

[13] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, “Activitynet: A large-scale video benchmark for human activity understanding,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[14] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[15] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, “Towards good practices for very deep two-stream convnets,” *arXiv preprint arXiv:1507.02159*, 2015.

[16] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[17] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[18] H. Jiang, W. Wang, Z. Gao, Y. Wang, and L. Nie, “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*, 2020, pp. 3487–3495.

[19] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[20] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[21] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[22] X. Yu, T. Gan, Z. Cheng, and L. Nie, “Personalized item recommendation for second-hand trading platform,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3478–3486.

[23] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, and T.-S. Chua, “Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video,” in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1437–1445.

[24] G. G. Chowdhury, *Introduction to modern information retrieval*. Facet publishing, 2010.

[25] S. Berkovsky, Y. Eytani, T. Kuflik, and F. Ricci, “Enhancing privacy and preserving accuracy of a distributed collaborative filtering,” in *Proceedings of the 2007 ACM conference on Recommender systems*, 2007, pp. 9–16.

[26] X. Yang, Y. Guo, Y. Liu, and H. Steck, “A survey of collaborative filtering based social recommender systems,” *Computer communications*, vol. 41, pp. 1–10, 2014.

[27] B. Y. Yilmazel and C. Kaleli, “Robustness analysis of arbitrarily distributed data-based recommendation methods,” *Expert Systems with Applications*, vol. 44, pp. 217–229, 2016.

[28] J. Wang, A. P. De Vries, and M. J. Reinders, “Unifying user-based and item-based collaborative filtering approaches by similarity fusion,” in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 501–508.

doi:10.20944/preprints202206.0412.v1

[29] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, and T.-S. Chua, "Contrastive learning for cold-start recommendation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5382–5390.

[30] M. Y. H. Al-Shamri, "Power coefficient as a similarity measure for memory-based collaborative recommender systems," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5680–5688, 2014.

[31] K. Nigam and R. Ghani, "Analyzing the effectiveness and applicability of co-training," in *Proceedings of the ninth international conference on Information and knowledge management*, 2000, pp. 86–93.

[32] S. Goldman and Y. Zhou, "Enhancing supervised learning with unlabeled data," in *ICML*. Citeseer, 2000, pp. 327–334.

[33] Z.-H. Zhou and M. Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 11, pp. 1529–1541, 2005.

[34] M. Liu, L. Nie, M. Wang, and B. Chen, "Towards micro-video understanding by joint sequential-sparse modeling," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 970–978.

[35] B. Peng, J. Lei, H. Fu, C. Zhang, T.-S. Chua, and X. Li, "Unsupervised video action clustering via motion-scene interaction constraint," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 131–144, 2018.

[36] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 327–336, 2016.

[37] J. Hou, X. Wu, Y. Sun, and Y. Jia, "Content-attention representation by factorized action-scene network for action recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1537–1547, 2017.

[38] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*. IEEE, 2010, pp. 995–1000.

[39] H. Guo, R. Tang, Y. Ye, Z. Li, and X. He, "Deepfm: a factorization-machine based neural network for ctr prediction," *arXiv preprint arXiv:1703.04247*, 2017.

[40] R. He and J. McAuley, "Vbpr: visual bayesian personalized ranking from implicit feedback," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.