Hindawi
Scientific Programming
Volume 2021, Article ID 5304321, 9 pages
https://doi.org/10.1155/2021/5304321

Research Article

A Multifeature Complementary Attention Mechanism for Image Topic Representation in Social Networks

Lei Shi,[1,2] **Jia Luo**,[3] **Gang Cheng**,[4] **Xia Liu**,[5] **and Gang Xie**[6]

[1]*Institute of Science and Technology Information of China, Beijing 100038, China*
[2]*State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China*
[3]*College of Economics and Management, Beijing University of Technology, Beijing 100124, China*
[4]*School of Computer Science, North China Institute of Science and Technology, Beijing 101601, China*
[5]*School of Opto-Electronic Information Science and Technology, Yantai University, Yantai 264005, China*
[6]*School of Big Data and Computer Science, Guizhou Normal University, Guiyang 550001, China*

Correspondence should be addressed to Jia Luo; 824629061@qq.com

Received 22 April 2021; Revised 12 July 2021; Accepted 15 August 2021; Published 25 August 2021

Academic Editor: Jianping Gou

Image topic representation in social networks is vital for people to get significant and valuable content. However, this task is difficult and challenging due to the complexity of image features. This paper proposes a multifeature complementary attention mechanism for image topic representation named CATR. CATR uses scene-level and instance-level object detection methods to obtain the object information on social networks. Here, the image features are divided into focused features and unfocused features. Focused features are used to learn and express semantic information, while unfocused features are used to filter out noise information in focused feature extraction. The attention mechanism is constructed by combining the object features and the features of the image itself, while the image topic representation in social networks is realized by the complementary attention mechanism. Based on the real image data of Sina Weibo and Mir-Flickr 25K, several groups of comparative experiments are constructed to verify the performance of the proposed CATR by leveraging different evaluation measures. The experimental results demonstrate that the proposed CATR obtains an optimal accuracy and significantly outperforms the other comparison methods in image topic representation.

1. Introduction

With the rapid increase of social networks' content, a large amount of image data is accumulated on social networks. It is very important for short-text modeling, accurate search, and topic clustering to study social networks' image representations. In addition, based on the obtained image information on social networks, we can process and analyze it through relevant methods to get useful image information and mine the topic expressed by its content, which can provide data and underlying support for social networks cross-media search and other applications.

The current mainstream image topic representation methods include the shallow semantic method and deep semantic method. The shallow methods include those based on canonical correlation analysis (CCA) [1] and probabilistic topic model [2]. The method based on deep semantics is mainly based on deep learning [3]. However, although these methods have certain effect on image topic representation, they will encounter severe challenges in the face of the complex image environment of social network. The existing image representation methods, such as the topic model method in the shallow method, have the problem of weak correlation in generating information, and it is difficult to obtain the deep features of the image. Moreover, the deep learning method ignores a central feature of image information and cannot distinguish complex image scenes. In addition, the image features obtained by the existing image topic representation methods are global features, which are not prominent, and encounter bottlenecks in representing

the different features of each image because images on social networks usually contain multiple objects, and the objects in the image have distinct discrimination.

Therefore, to overcome the above problems and improve image topic representation quality, this work proposes a multifeature complementary attention mechanism for image topic representation in social networks (CATR). By distinguishing the focused features and the unfocused features, the object features are mixed into the focused features and the unfocused features to guide and enhance the image feature learning. The complementary attention mechanism based on the focused features and the unfocused features is established to improve the image feature learning effect. The topic representation of social networks' image information is realized by fusing the focused features and unfocused features. Using the object feature can make the focused feature more consistent and make the unfocused feature further away to effectively overcome the noise problem in the process of image topic representation. The attention mechanism can detect the attention region of image information and overcome the problem that the feature-based depth feature extraction method cannot deal with well [4, 5].

Simultaneously, in the research of object features [6], a variety of effective methods are proposed, and multiple applications are achieved. Sin network [7] comprehensively considers the association information between the scene and the surrounding objects and performs object detection through the structural reasoning method. ORN network [8] optimizes object detection through the interaction between appearance features and geometric figures, which obtains better object detection results. To extract high-quality object features, we select Sin network for extracting object features. In our proposed CATR method, first, the focused feature, unfocused feature, and object feature of the image are integrated. Second, the complementary attention mechanism is established to learn the features of the image. Finally, the features of the image itself can be more deeply depicted, and the accurate representation of the image topic in social network can be realized.

The main contribution and innovation of our paper are the following:

(i) An image topic representation method for social networks based on complementary attention mechanisms is proposed, providing ideas for the topic representation and developing multiple social network applications.

(ii) This work innovatively distinguishes the focused feature and unfocused feature of the modeling image to achieve the fine description of image topic semantics.

(iii) The object features are used to guide the learning of image modalities and build complementary attention mechanisms based on the focused and unfocused features of image modalities for obtaining high-quality topic representations.

(iv) Multiple sets of comparative experiments are constructed to verify the performance of the proposed CATR on multiple evaluation indicators in real Sina Weibo image datasets and Mir-Flickr 25K. The experimental results show that the proposed CATR is significantly superior to other comparison methods.

The remainder parts of this paper are organized as follows. In Section 2, we introduce the related works. In Section 3, we describe our CATR method in detail. In Section 4, we give the experimental results on the Sina Weibo image dataset. Finally, conclusions and future work are drawn in Section 5.

2. Related Works

As an essential element of all social networks, the image can effectively supplement the lack of text description. Therefore, the topic representation of an information carried by an image has become a hot research topic [9]. There are two types of image topic representation methods in social networks: image topic representation method based on shallow semantics [10] and image topic representation method based on depth feature [11].

2.1. Image Topic Representation Method Based on Shallow Semantics. In the image topic representation method based on shallow semantics, Tian and Shi [12] using the two-stage topic model with max-bisection for obtaining the image topic semantic representation. Laib et al. [13] used Dirichlet distribution to model the image event topic and estimated the topic based on EM method to realize the classification of topics. Peng et al. [14] propose a weighted constraint-based diagonal matrix for improving the topic representation performance. Chen et al. [15] introduced image modal and emotional modal into the topic model for obtaining the topic distribution of image information in social networks. Qian et al. [2] used the topic model to model the generation of image topic and text topic, respectively, and obtained the distribution of text topic and image topic. Chen et al. [16] proposed a semantic image context mining framework based on image context. The shallow semantic method is fast in training and low in complexity, which is suitable for small data sets. However, the above-mentioned methods cannot obtain more complex and deep semantic information only through simple association or modeling of image information, and there is a certain bottleneck in the accuracy of image topic representation.

2.2. Image Topic Representation Method Based on Depth Feature. In recent years, with the improvement of hardware computing power, image topic representation based on depth feature is becoming more and more popular. The VGG network [3] has gained unprecedented attention and application in image feature extraction and representation, such as gesture recognition, background recognition, and cross-media search [17]. Depth feature has incomparable advantages over other features [18]. Yang et al. [9] constructed a hybrid depth network by combining tensor

decomposition and multiview depth self-encoder. Zhang et al. [19] proposed a CNN method based on different regions to encode the semantic context-aware representation. Yang et al. [20] expressed images by distinguishing image features into two different data streams and combining depth attention mechanism. Chu et al. [21] proposed a deep learning representation method of image features through the deep heterogeneous hypergraph embedding method. Ding et al. [22] used context encoding and multipath decoding to construct segmentation network for enhancing the image expression performance. Liu et al. [23] constructed a residual feature aggregation framework to realize feature extraction. Sun et al. [24] designed a variable-length gene encoding strategy and representation scheme to initialize connection weights for image classification. Wei et al. [25] selectively integrate features and refine multilevel features with feedback mechanisms for accurate object detection. Zhao et al. [26] build the multilevel feature pyramid network using feature pyramids to detect objects for different scales. Inspired by the above methods, the proposed CATR introduces focused features and unfocused features to model the semantic representation of images.

The attention model can generate different attention distribution for various features. It can effectively solve the problem that global features are not highlighted. Gregor et al. [27] realized the generation of complex images by combining the attention model. Bahdanau et al. [28] designed an attention model in the learning task. Jin et al. [29] adopt the attention model to fuse cross-modal features. Zhao et al. [5] explored self-attention fully by using pairwise and patchwise for the effectiveness of image representation models. By introducing attention mechanism, attention region of image information can be detected to overcome the problem that the feature based on depth feature extraction method is not obvious enough. However, most of the features of the image are focused, which only sacrifice the accuracy of the image feature extraction.

3. CATR Method

3.1. Overall Framework of the CATR Method. In this part, we mainly introduce our CATR method. The core of CATR method is to divide image features into focused features and unfocused features and introduce object features to guide feature learning. The complementary attention mechanism is established through the correlation between object and image region to realize the image feature generation under the guidance of object features. The focused features are mainly used to extract information closely related to image semantics. In contrast, unfocused features are used to extract content that may be of visual concern to enhance the focused features. The architecture of the CATR method is shown in Figure 1.

3.2. Description of the CATR Method. For the focused feature of the image, because VGG-19 can obtain more image features and control the number of parameters, this avoids too much calculation and too complex structure. In addition, in the actual experiment, VGG-19 is more robust than other depth features. Thus, we use VGG-19 to get the focused feature and initialize the focused feature as convolution and pooling group [30]. In addition, the output of the last pooling layer is selected as the original image feature. In our method, we formalize the image information as $I_k = \{i_k \mid i_k \in R^{E_i}, k = 1, 2, \ldots, G\}$, where E_i is the dimension of the feature and G is the number of regions in the image modal. There are multiple object features in the image modal; we formalize the image object features as $Q_i = \{q_i \mid q_i \in R^{E_i}, i = 1, 2, \ldots, I\}$, where Q_i represents the object feature of the image modal. We input each object feature and region feature matrix of the image information into the neural network, take the calculated object feature and focused feature of the image as the input, and use the softmax function in the neural network to calculate the attention probability of the focused feature. The calculation method is shown in the following equations:

$$H_i = \tanh\left[W_{i_k} I_k; \left(W_{q_i} q_i + u_{q_i}\right)\right], \tag{1}$$

$$p_i^H = \text{soft}\max\left(W_{p_i^H} H_i + u_{p_i^H}\right), \tag{2}$$

where H_i represents hidden state vector that introduces attention mechanism, W represents the weight parameter, $W_{I_k} \in R^{f_i \times E_i}$, $W_{q_i} \in R^{f_q \times E_q}$, $W_{p_i^H} \in R^{1 \times (f_i + f_q)}$, $p_i^H \in R^E$, u represents the bias, and [;] represents a cascading operation to merge the focused feature and the object feature of the image in a column. p_i^H represents the attention probability set of given object feature q_i and image modal focused feature. Based on equation (3), we can obtain the new feature vector $\hat{i}_i$ after combining the object feature q_i with the focused feature.

$$\hat{i}_i = \sum_{i=1}^{G} p_{i,q}^H \cdot i_k. \tag{3}$$

Compared with the heuristic attention model, DeepFixNet has achieved impressive performance in predicting attention. Using the DeepFixNet method, we can filter out irrelevant features. Thus, DeepFixNet [31] is adopted to initialize unfocused features. After obtaining the unfocused feature, the same calculation method is used as the focused feature to obtain the attention probability of the unfocused feature. In addition, because the DeepFixNet results are not fully applicable to image search requirements, we need to optimize and adjust the image data, and take the focused features and unfocused features after fusing the object features as the output. The fusion method is shown in the following equation:

$$I = \kappa I_{\text{foc}} + (1 - \kappa) I_{\text{unfoc}}, \tag{4}$$

where κ is the weight parameter, which is used to adjust the output of focused features and unfocused features. In this paper, κ is set to 0.6. I_{foc} represents the focused feature of the image, while I_{unfoc} represents the unfocused feature of the image. Through the above fusion, the new features can be obtained including semantic information, object

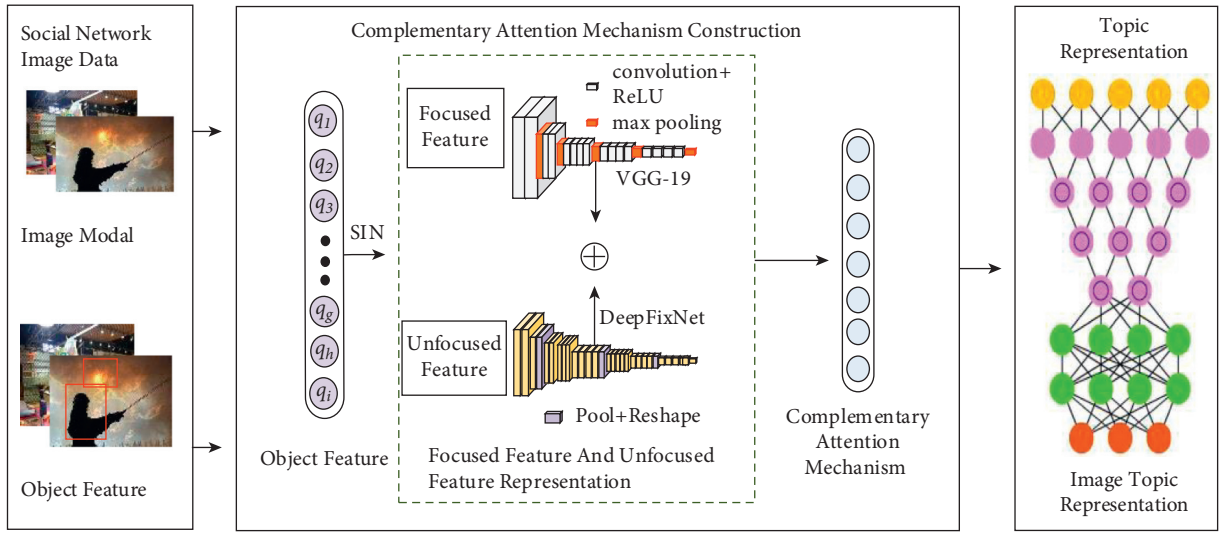

Figure 1: Overall framework of the CATR method.

information, and attention information. Thus, the complementary attention mechanism based on focused feature and unfocused feature can be realized.

The softmax layer is used to adjust the probability information of the output vector and train the classification network of the image retrieval data set. The calculation formula is shown in the following equation:

$$Y = \arg\min_{Y} \sum_{a=1}^{A} \sum_{b=1}^{B} \log\left(c_{ab}, d_{ab}\right) + \beta \|Y\|_2^2, \tag{5}$$

where Y denotes the set of network parameters in CATR. c_{ab} is the bth component of the probability vector generated by softmax layer when processing the ath training image, d_{ab} is a binary switching variable, which is equal to 1 if and only if the ath training image belongs to the bth class of training image, and β is a constant to adjust the parameter norm.

3.3. Implementation Steps of CATR. The implementation steps of CATR based on complementary attention mechanism are shown in Table 1. In CATR, the focused feature and the unfocused feature are distinguished. The object attention mechanism is introduced to guide the learning process of the focused feature and the unfocused feature to realize the topic representation of social network image based on the complementary attention mechanism.

4. Experiment and Result Analysis

4.1. Datasets. Sina Weibo: The text data is removed, and only the image data is retained by using the four social network burst topic datasets of "Kunming railway station terror," "Yutian earthquake," "Tianjin warehouse explosion" and "Hubei flood" crawled from Sina Weibo. To ensure the balance of the number of images in different events, 10,000 images are randomly selected from the image data of each event as the social network image data set used in the experiment, 8,000 of which are used as the training set, and the remaining 2,000 are used as the test set. A total of 40,000 social network images were obtained as experimental data.

Mir-Flickr 25K: a cross-media data set based on social network Flickr platform, which contains 25,000 image contents, each image has several text labels associated with it, and each data point is an image text pair. In the experiment, 20,000 images are randomly selected as the training set, and the remaining 5,000 images are selected as the test set.

4.2. Baseline Methods. The following six methods are used as the comparison methods. In all the comparison methods, mmLDA, M^{3}R, VELDA, and mmETM are the traditional topic representation methods, and VGG-19, ResNet, DenseNet, SkeletonNet, CIN, and our CATR are deep learning methods:

mmLDA [10]: a method for associating image and text, which combines the modeling of text and image by learning image topic distribution and text topic distribution to obtain image distribution. We realize the representation of image information by only modeling the image.

M^{3}R [32]: a joint cross-modal probabilistic graphic model, through the interaction between potential topics of different modal, the mutually enhanced information is transferred to each other, and the consistent cross-media semantic topics are obtained. Through this method, the adaptive image topic representation can be realized.

VELDA [13]: a text and image matching modeling method based on the topic model, which generates text modal information, introduces image information and emotional information for association modeling based on text information, to generate more consistent topic representation, and realizes the topic representation of images in the form of topic distribution.

mmETM [2]: the method can effectively model social network image and text information and learn the

Table 1: The steps of CATR: a multifeature complementary attention mechanism method.

Input: original microblog image, feature dimension
Output: fused image features
(1) The image information obtained by preprocessing unifies the resolution and pixels of the image
(2) Get the object feature of the image
(3) The focused feature of the image is extracted by VGG-19
(4) Cascade object feature Q_i and focused feature of output
(5) According to equations (1) and (2), the attention distribution of the image for different objects is calculated
(6) According to equation (3), the image features under the guidance of the object are calculated
(7) Feature extraction of unfocused image based on DeepFixNet
(8) Using the same method as calculating the attention distribution of focused features, the attention distribution of unfocused features is calculated
(9) According to equation (4), the focused feature and unfocused feature are fused
(10) Use equation (5) to update the weights and parameters in the network

correlation between text and image mode by distinguishing modeling image topics and text topics.

VGG-19 [3]: a network structure based on deep learning. Its deep network structure is often used in image feature extraction and representation and has good image feature extraction and representation performance.

ResNet [33]: it explicitly reformulates the layers as learning residual functions instead of learning unreferenced functions. In the image feature representation and learning, it has achieved wonderful results.

DenseNet [34]: a convolutional network supporting hundreds of layers. It introduces connections between any two layers with the same feature-map size. It has achieved remarkable results in the performance of image topic expression.

SkeletonNet [35]: a hybrid deep learning method. In this method, tensor decomposition mechanism is introduced into the multiview depth automatic encoder to learn the topic representation of images.

CIN [9]: a new deep learning method for fine-grained channel interaction. This method uses self-channel interaction to learn complementary channel information and considers the relationship between channels.

4.3. Evaluation Indicators. To verify the performance of the proposed CATR, we apply the CATR and other methods to the search task and observe the image topic representation performance by the evaluation index of search. We use Normalized Discounted Cumulative Gain (NDCG) [9, 35] and Mean Average Precision (MAP) [36] commonly used in search evaluation as evaluation indicators.

4.3.1. NDCG. NDCG is a standard evaluation index of search performance. The larger the value, the better the search performance of the method. The calculation formula is shown in the following equation:

$$\mathrm{NDCG}(m) = Y_k \sum_{k=1}^{M} \frac{\left(2^{s(k)} - 1\right)}{\log(1+k)}, \tag{6}$$

where Y_k is used for normalization, m represents the top-m ranking of relevance, and $s(k)$ is the relevance of the kth search result.

4.3.2. MAP. The MAP is mean average precision. It is used to evaluate the search results. The AP is average precision. The larger the value, the better the search performance of the method, and the specific calculations are shown in the following equations:

$$\mathrm{MAP} = \frac{1}{C} \sum_{c=1}^{C} \mathrm{AP}(C), \tag{7}$$

$$\mathrm{AP} = \frac{1}{T'} \sum_{t=1}^{T} \mathrm{prec}(t)\delta(t), \tag{8}$$

where C represents the number of queries in the query set, T and T' are the total number of results and the number related to the query, respectively, prec(t) is the accuracy of the tth result, $\delta(t)$ is an indicator function, where $\delta(t) = 0$ means that the returned results are not relevant, and $\delta(r) = 1$ indicates that the returned results are relevant.

4.4. Experimental Results

4.4.1. Experimental Results on Sina Weibo. We use the test set to build the search experiment and choose the MAP@K and NDCG@K as an evaluation indicator, where MAP@K represents the MAP value of the top-k results in the search results, and NDCG@K represents the NDCG value of the top-k results in the search results; K values are 5, 10, 15, and 20, respectively.

(1) The MAP Results of CATR and Other Comparison Methods. The MAP results of the proposed CATR method based on complementary attention mechanism and other comparison methods are shown in Table 2.

It can be seen from the experiments in Table 2 that the proposed CATR method is significantly better than other comparison methods in MAP of social network search, especially when K is 10, the average accuracy of the proposed CATR method is more than 0.8, and the MAP results of the

Table 2: The MAP results of CATR and other comparison methods on Sina Weibo image dataset.

Method	MAP@5	MAP@10	MAP@15	MAP@20
mmLDA	0.496	0.527	0.514	0.482
M^3R	0.561	0.582	0.578	0.562
VELDA	0.624	0.642	0.603	0.591
mmETM	0.703	0.734	0.671	0.627
VGG-19	0.732	0.759	0.735	0.706
ResNet	0.739	0.781	0.738	0.717
DenseNet	0.775	0.783	0.741	0.739
SkeletonNet	0.776	0.786	0.745	0.739
CIN	0.779	0.792	0.747	0.741
CATR	0.781	0.801	0.752	0.743

proposed CATR method are 26%, 19%, 15%, 9%, 4%, 2%, 1.8%, 0.7%, and 0.4% higher than mmLDA, M^3R, VELDA, mmETM and VGG-19, ResNet, DenseNet, SkeletonNet, and CIN, respectively. The above results show that the proposed CATR method based on complementary attention mechanism can effectively express the image topic and improve social network image search performance.

(2) The NGCG Results of CATR and Other Comparison Methods. To further verify the image topic representation effect of the proposed CATR method and other comparison methods, under the same dataset and parameter setting, NDCG is used as the evaluation index to verify the performance of the CATR method and other comparison methods. The experimental results are shown in Figure 2.

From the search experiment results in Figure 2, it can be seen that the NDCG value of the proposed CATR method is significantly better than other comparison methods, and the NDCG value of CIN, SkeletonNet, ResNet, DenseNet, and NetVGG-19 are better than mmLDA, M^3R, VELDA, and mmETM. In all methods, mmLDA method based on the traditional topic model achieves the worst search performance. The experimental results show that the method based on deep learning can improve the quality of image topic representation by describing and expressing image features more precisely. Although the other four methods can also effectively express the image topic, they can only simply associate and model the image and cannot achieve fine-grained image topic representation.

It can be seen from the search results in Table 2 and Figure 2 that the NDCG and MAP values of the proposed CATR are better than those of CIN. The results show that the proposed CATR can obtain more focused features than the global features by extracting focused features and unfocused features and guide them by object features. Through feature complementation, it can obtain more advanced image features.

In addition, some topic representation examples of Sina Weibo datasets are shown in Figure 3. Based on the qualitative judgment of examples, images are often represented not only based on visual semantics, but also based on their high-level object features. For example, in the second row, there are collapsed houses, firefighters, and injured people. In the fourth row, big water, cars, and pedestrians are gathered together, which further reflects the effectiveness of our CATR in theme representation.

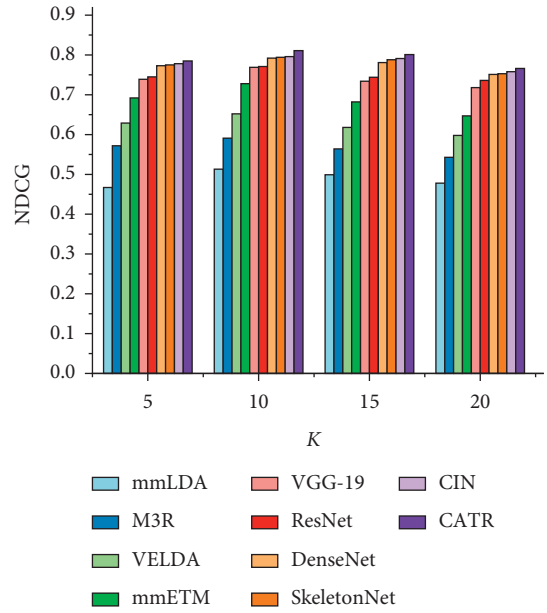

Figure 2: The NDCG results of CATR and other comparison methods on the Sina Weibo image dataset.

4.4.2. Experimental Results on Mir-Flickr 25K. To fully verify the effectiveness of the proposed CATR method, we use the social network cross-media dataset Mir-Flickr 25K as the experimental data and use MAP and NDCG as the evaluation indicator of the experiment.

(1) The MAP Results of CATR and Other Comparison Methods. We use MAP@K as an evaluation indicator to verify the topic representation effect of different algorithms on the public dataset Mir-Flickr 25K. The MAP results of the proposed CATR method and other comparison methods are shown in Table 3.

From the experimental results in Table 3, we can observe that the MAP results of CATR method under different K values are better than other comparison algorithms, which once again show that the superiority of our proposed CATR method in topic representation. At the same time, the experimental results also show that the introduction of object features and complementary attention mechanism can effectively improve the effect of topic representation.

(2) The NGCG Results of CATR and Other Comparison Methods. Similar to the above experiment, the MIR-Flickr 25K dataset is used as the experimental data, and NDCG is used as the evaluation index to verify the image topic representation performance of the proposed CATR method. The experimental results are shown in Figure 4.

Based on the experimental results in Figure 4, we can see that the traditional topic model methods mmLDA, M3R, VELDA, and mmETM perform relatively poorly compared with the deep learning methods VGG-19, ResNet, DenseNet, and SkeletonNet. In the deep learning method, CIN and SkeletonNet have achieved wonderful performances. However, compared with our CATR, the performance of

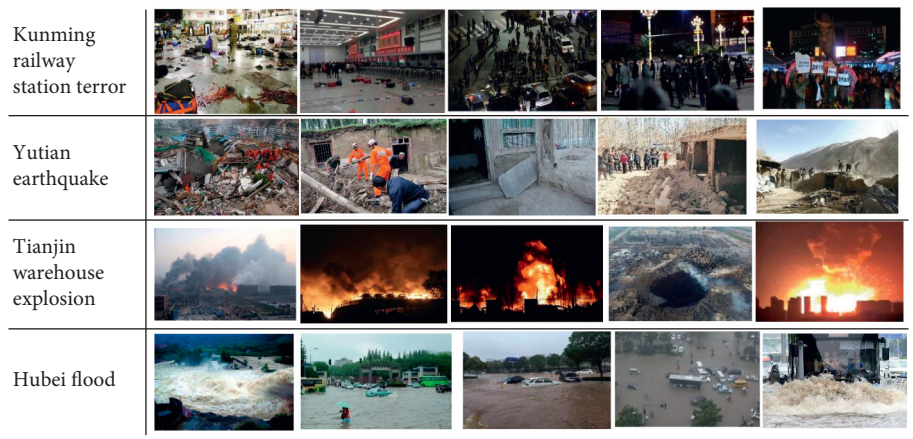

Figure 3: Examples of image topic representation on Sina Weibo. Each row corresponds to the topic representation generated by the CATR.

Table 3: The MAP results of CATR and other comparison methods on the Mir-Flickr 25K dataset.

Method	MAP@5	MAP@10	MAP@15	MAP@20
mmLDA	0.498	0.531	0.522	0.496
M^3R	0.576	0.574	0.583	0.571
VELDA	0.638	0.682	0.624	0.609
mmETM	0.724	0.768	0.735	0.661
VGG-19	0.749	0.773	0.748	0.746
ResNet	0.779	0.794	0.757	0.739
DenseNet	0.782	0.797	0.765	0.741
SkeletonNet	0.786	0.801	0.773	0.762
CIN	0.799	0.802	0.777	0.763
CATR	0.803	0.811	0.784	0.771

CIN and SkeletonNet is relatively poor. This is because our CATR integrates the object features and distinguishes the modeling focused feature and the unfocused feature to comprehensively model the local and global features. Thus, CATR gets the best performance of topic representation in all comparison methods.

4.4.3. Influence of Weight Parameters on the CATR Method for Topic Representation. To thoroughly verify the influence of the weight parameter κ on the image topic representation, we use Sina Weibo image data and MAP as the evaluation indicators to verify the influence of the weight parameter κ on the performance of CATR. The weight parameter κ varies from 0 to 1, and K is 10 and 20, respectively.

The experimental results of the influence of weight parameter κ on CATR for topic representation performance are shown in Figure 5. It can be seen from the experimental results that the focused feature and unfocused feature have a significant influence on the performance of image topic representation. When κ is equal to 0.6, MAP@10 and MAP@20 perform best. The results show that when the focused feature is 0.6 and the unfocused feature is 0.4, the best effect of image topic representation can be achieved.

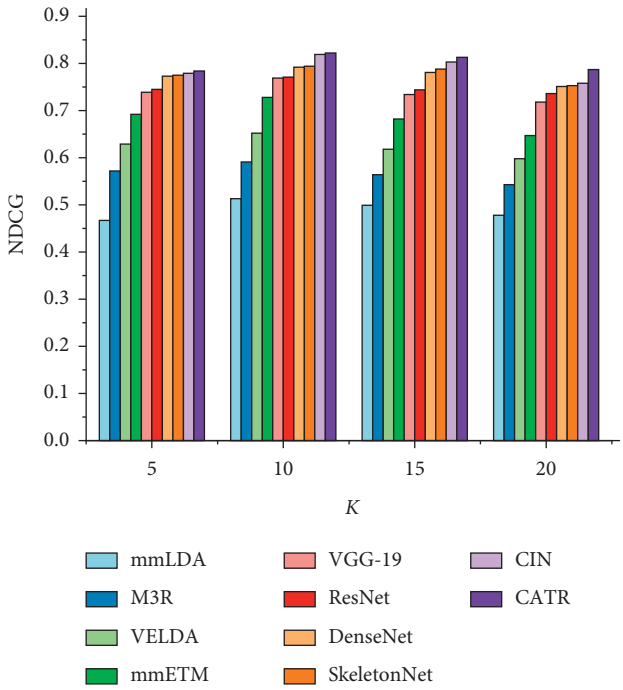

Figure 4: The NDCG results of CATR and other comparison methods on the Mir-Flickr 25K image dataset.

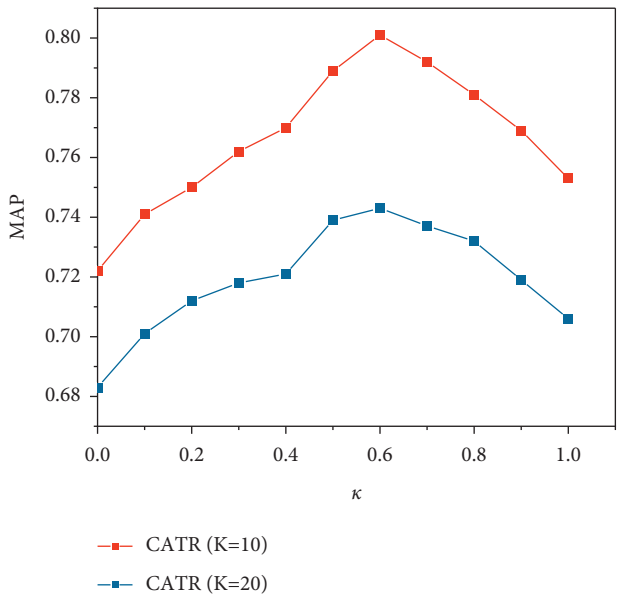

Figure 5: The MAP results comparison of the CATR method under different weight parameter settings.

5. Conclusions

In this paper, the image topic representation method (CATR) based on complementary attention mechanism is proposed. In this method, first, the focused feature and unfocused feature of the modeling image are distinguished to achieve an accurate representation of the feature. Second, the object feature is mixed to construct the complementary attention mechanism to achieve a more focused image feature representation. Finally, the full representation of the image topic of social network is completed. The experimental results show that compared with other comparison methods, the CATR method can obtain an optimal search accuracy, which indicates the superiority of the CATR in image topic representation.

However, there are still social relations, text content, and other information on social networks. In a future work, we will explore and study how to introduce users' social relations and text information inside image topic representation to achieve accurate representation of social network topics based on social relations and multimodal features.

Data Availability

All data used to support the findings of this study are available from the first author upon request.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Beijing Municipal Education Commission (Grant no. SM202110005011), the National Key Research and Development Program of China (no. 2018YFC1505104), the Key Work Project of the Institute of Science and Technology Information of China (no. ZD2020-09), the China Postdoctoral Science Foundation (no. 2019M651786), the Scientific Research Projects of Education Department of Hebei Province, China (no. Z2017043), the opening fund of the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection (Chengdu University of Technology) (no. SKLGP2021K005), and the Hebei IoT Monitoring Engineering Technology Research Center (no. 3142018055).

References

[1] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods," *Neural Computation*, vol. 16, pp. 2639–2664, 2004.

[2] S. Qian, T. Zhang, C. Xu, and J. Shao, "Multi-modal event topic model for social event analysis," *IEEE Transactions on Multimedia*, vol. 18, pp. 233–246, 2016.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the 3rd International Conference on Learning Representations*, pp. 1–14, San Diego, CA, USA, May 2015.

[4] H. Lu, R. Yang, Z. Deng et al., "Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, pp. 1–18, 2021.

[5] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085, Seattle, WA, USA, June 2020.

[6] M. Tamura, H. Ohashi, and T. Yoshinaga, "QPIC: query-based pairwise human-object interaction detection with image-wide contextual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10410–10419, New Orleans, LA, USA, June 2021.

[7] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: object detection using scene-level context and instance-level relationships," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6985–6994, Salt Lake City, UT, USA, June 2018.

[8] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, Salt Lake City, UT, USA, June 2018.

[9] S. Yang, L. Li, S. Wang, W. Zhang, Q. Huang, and Q. Tian, "SkeletonNet: a hybrid network with a skeleton-embedding process for multi-view image representation learning," *IEEE Transactions on Multimedia*, vol. 21, pp. 2916–2929, 2019.

[10] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 127–134, Toronto, Canada, August 2003.

[11] J. Yu, Z. Kuang, B. Zhang, W. Zhang, D. Lin, and J. Fan, "Leveraging content sensitiveness and user trustworthiness to recommend fine-grained privacy settings for social image sharing," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1317–1332, 2018.

[12] D. Tian and Z. Shi, "A two-stage hybrid probabilistic topic model for refining image annotation," *International Journal Machine Learning Cybernetics*, vol. 11, pp. 417–431, 2020.

[13] L. Laib, M. S. Allili, and S. A. Ait-Aoudia, "Probabilistic topic model for event-based image classification and multi-label annotation," *Signal Processing: Image Communication*, vol. 76, pp. 283–294, 2019.

[14] Y. Peng, L. Li, S. Liu, X. Wang, and J. Li, "Weighted constraint based dictionary learning for image classification," *Pattern Recognition Letters*, vol. 130, pp. 99–106, 2020.

[15] T. Chen, H. M. SalahEldeen, X. He, M. Y. Kan, and D. Lu, "VELDA: relating an image tweet's text and images," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 30–36, Austin, TX, USA, January 2015.

[16] T. Chen, X. He, and M. Y. Kan, "Context-aware image tweet modelling and recommendation," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1018–1027, Amsterdam, NL, USA, October 2016.

[17] W. Yang, X. Zhang, Y. Tian, W. Wang, J. H. Xue, and Q. Liao, "Deep learning for single image super-resolution: a brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[18] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," *Applied Soft Computing*, vol. 108, Article ID 107440, 2021.

[19] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 27, pp. 2623–2634, 2018.

[20] F. Yang, J. Li, S. Wei, Q. Zheng, T. Liu, and Y. Zhao, "Two-stream attentive CNNs for image retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1513–1521, Mountain View, CA, USA, October 2017.

[21] Y. Chu, C. Feng, and C. Guo, "Social-guided representation learning for images via deep heterogeneous hypergraph embedding," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, p. 1, San Diego, CA, USA, July 2018.

[22] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Transactions on Image Processing*, vol. 29, pp. 3520–3533, 2020.

[23] J. Liu, W. Zhang, Y. Tang, J. Tang, and G. Wu, "Residual feature aggregation network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2359–2368, Seattle, WA, USA, June 2020.

[24] Y. Sun, B. Xue, and M. Zhang, "Evolving deep convolutional neural networks for image classification," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 394–407, 2019.

[25] J. Wei, S. Wang, and Q. Huang, "F^3Net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12321–12328, New York, NY, USA, February 2020.

[26] Q. Zhao, T. Sheng, Y. Wang et al., "M2det: a single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9259–9266, Honolulu, HI, USA, February 2019.

[27] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: a recurrent neural network for image generation," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, pp. 1462–1471, Lille, France, July 2015.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations*, pp. 1–15, San Diego, CA, USA, May 2015.

[29] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM International Conference on Multimedia*, pp. 795–816, Mountain View, CA, USA, October 2017.

[30] F. Kou, J. Du, W. Cui et al., "Common semantic representation method based on object attention and adversarial learning for cross-modal data in IoV," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 11588–11598, 2019.

[31] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and N. E. Oconnor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 598–606, Las Vegas, NV, USA, June 2016.

[32] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang, "Multi-modal mutual topic reinforce modeling for cross-media retrieval," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 307–316, Orlando, FL, USA, November 2014.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[34] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weiberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, Honolulu, HI, USA, July 2017.

[35] D. Li, Q. Wang, and F. Kong, "Adaptive kernel sparse representation based on multiple feature learning for hyperspectral image classification," *Neurocomputing*, vol. 400, pp. 97–112, 2020.

[36] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, Vancouver, Canada, October 2008.