Neural Processing Letters (2024) 56:128
https://doi.org/10.1007/s11063-024-11525-z

Central Attention with Multi-Graphs for Image Annotation

Baodi Liu[1] · **Yan Liu**[1] · **Qianqian Shao**[1] · **Weifeng Liu**[1]

Accepted: 8 January 2024 / Published online: 30 March 2024

Abstract

In recent decades, the development of multimedia and computer vision has sparked significant interest among researchers in the field of automatic image annotation. However, much of the research has primarily focused on using a single graph for annotating images in semi-supervised learning. Conversely, numerous approaches have explored the integration of multi-view or image segmentation techniques to create multiple graph structures. Yet, relying solely on a single graph proves to be challenging, as it struggles to capture the complete manifold of structural information. Furthermore, the computational complexity of building multiple graph structures based on multi-view or image segmentation is substantial and time-consuming. To address these issues, we propose a novel method called "Central Attention with Multi-graphs for Image Annotation." Our approach emphasizes the critical role of the central image region in the annotation process. Remarkably, we demonstrate that impressive performance can be achieved by leveraging just two graph structures, composed of central and overall features, in semi-supervised learning. To validate the effectiveness of our proposed method, we conducted a series of experiments on benchmark datasets, including Corel5K, ESPGame, and IAPRTC12. These experiments provide empirical evidence of our method's capabilities.

Keywords Automatic image annotation · Lplacian feature mapping · Multi-graphs · Central attention

1 Introduction

A large number of images appear on social media every day with the development of multimedia technology. How to quickly sift the required image is a crucial problem. The methods

✉ Baodi Liu
thu.liubaodi@gmail.com

Yan Liu
liuyan2021202108@163.com

Qianqian Shao
17353360024@163.com

Weifeng Liu
liuwf@upc.edu.cn

[1] College of Control Science and Engineering, China University of Petroleum (East China), Qingdao, China

for image retrieval are compartmentalized into two categories: content-based image retrieval (CBIR) and tagbased image retrieval (TBIR) techniques. Nevertheless, these methods suffer a sea of difficulties, such as the semantic gap and complexity and arbitrary of manual annotation. Therefore, automatic image annotation (AIA) [3, 10, 12, 19, 27, 35, 38] is proposed to surmount above-mentioned difficulties. AIA is the process that computer automatically lists keywords(tags or labels) that are able to explain the content of images. Consequently, AIA surmounts the semantic gap between visual content and their semantic meanings. Meanwhile, significant time is saved due to computer involvement. Image captioning and image classification are two related but distinct tasks in the field of computer vision. The goal of image captioning is to generate textual descriptions or labels for an image, reflecting the main content contained within the image. On the other hand, image classification aims to categorize images into different classes or predefined categories, with the objective of determining which category an image belongs to. There is some overlap between image captioning and image classification. To some extent, image classification can be considered a special case of image captioning where the labels are predefined categories. The most significant difference lies in the output format: image captioning produces a text description or a set of labels, while image classification outputs one or more category labels. In summary, image captioning and image classification are two related yet distinct tasks, each with its own applications and research directions within the field of computer vision. Image captioning provides richer semantic descriptions for images, while image classification primarily focuses on categorizing images. These two tasks often complement each other and collectively contribute to the advancement of image understanding and applications.

The algorithms of AIA are devided into two categories: model without graph structure and model with graph structure.

(1) Model without graph structure contains the methods [23, 25, 41, 46] that overlook the local geometric structure. Therefore, those algorithms reckon that the relationship between images or between tags is independent. The traditional multi-classification methods [7, 15, 20, 29] are exerted to resolve the problem of AIA. Meanwhile, the meta-learning approach is to tackle the multi-label classification problem in [22]. Then, nearest neighbor based model [1, 21, 32] play a paramount role in image annotation. Further, it is also the common approach applied to AIA that computing the conditional probabilities [18, 40]. On account of the promising performance of deep learning [8, 30, 33, 45], some researchers combine it with image annotation. A quintessential example should be cited that GAN [23] is bestowed to generate the tags of the images. Unfortunately, these methods lost effect during semi-supervised learning due to less training data.

(2) The model with graph structure [16, 17, 26, 28, 44] ponder manifold structure information between images, between labels or between images and tags. A multitude of graph structures, such as Locally Linear Embedding(LLE) [28], Laplacian eigenmaps(LE) [26], [47, 50, 52], Hypergraph [39, 44] and Hessian [14, 31], are embedded into the models of AIA. Meanwhile, GCN [42] is also applied to solve the problem of AIA. An ocean of models employ the single graph to delineate the geometric structure of the entire dataset. However, the single graph is formidable to accurately describe the geometry of the data. Moreover, the multiple graphs, which are computed depending on multi-view [47, 52], [31], comprise the more affluent structure information that comparing with the single graph. More precisely, the methods relying on multi-view necessitate the different features,such as Rgb, RgbV3H1, Lab, LabV3H1, Hsv, and VGG16, to construct multiple graphs. Furthermore, the diverse graphs are computed according to the different parts of an image. Specifically, the image is divided into pieces. And the features of these

pieces respectively are harnessed to compose different graph structures. Nevertheless, it consumes lots of time and has a heavy calculation to obtain the multiple graphs according to the various features or the diverse parts of an image.

Therefore, we propose a method, Central Attention with Multi-graphs for Image Annotation. Our main contributions can be summarized as follows:

(1) The center part of an image is considered as the most crucial part comparing with other parts, such as top left, bottom left, top right, bottom right. In other words, the center part of an image can reflect most of the content of the image. Thus, the performance of the model can be enhanced only relying on the center part of an image without the need for other parts and multiple features of the image.
(2) Multiple graphs, composed by center part and integral image, are calculated to obtain the complete manifold information. And the graph structure of the center part incorporates the more detailed information of an image compared to the single overall graph.
(3) The superiority of our model is demonstrated according to experimenting on three widely employed benchmark datasets. Meanwhile, the computational complexity is reduced because two graph structures are only constructed.

The remaining parts of this paper are organized as follows. Some methods related to AIA are shown in Section II. Then Section III formalizes implementation details of proposed framework. Moreover, the experimental results are particularized in Section IV. Finally Section V draws a conclusion for this paper.

2 Related Work

With the development of deep learning, weakly supervised learning has gained widespread attention in the field of computer vision. Its core idea is to use incomplete or inaccurate supervision information during the training process. This concept has been successfully applied to tasks such as object detection, image classification, and semantic segmentation. Researchers have begun to introduce weakly supervised learning methods into the field of video processing to address the challenges posed by annotating video data. An important application of weakly supervised methods is weakly supervised object detection, where models need to infer the object's location from image-level labels [4–6, 43]. Manifold regularization [2] enhances extremely the performance of the models in semi-supervised learning. Therefore, it appeals increasingly to an army of researchers that the methods conjoining AIA with manifold regularization [49]. The framework combined with manifold regularization is compartmentalized two parts: the model with a single graph and the model with multiple graphs.

The model with a single graph is that the single graph is embedded in the framework. In [26], LE is utilized to build a sample graph in semi-supervised learning. Alternatively, [44] proposes an adaptive Laplacian graph. [34] consider the symbiotic relation between tags to construct tag graphs. And GCN is proposed to resolve the image classification in [24, 37]. An end-to-end model for medical image classification [48] is demonstrated. GCN is also used to achieve multi-label image recognition [9]. Further, it is proposed that a new perspective of image annotation combining GCN with ZeroShot learning in [42]. And [51] combines the graph with attention mechanism during handling the metadata.

The model with multiple graphs annexes more complete manifold information compared to the single graph. [47, 53] harnesses the multi-vision for image annotation to construct the multiple graphs between images and the graph between tags is also computed. Similarly,

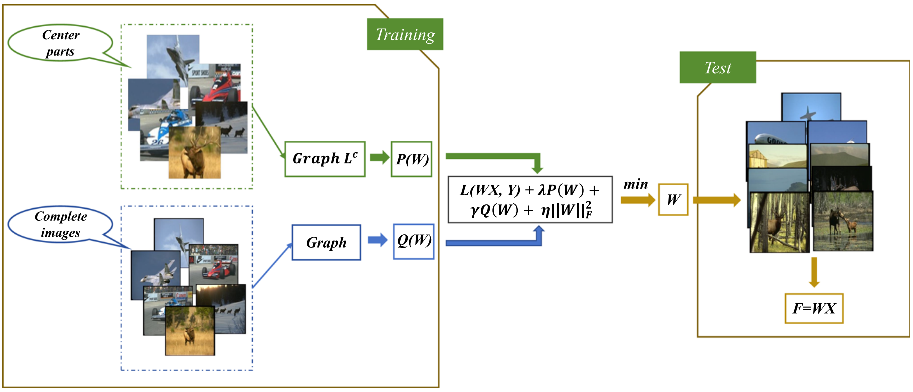


Fig. 1 The framework of proposed method. W is the coefficient matrix. L^C and L are the Laplacian graphs

[17] conjuncts the graph embedding with multi-view nonnegative matrix factorization. Meanwhile, [47] utilizes the multi-view to construct graph structures. This method [50, 52] takes into account the relationship between samples, the relationship between labels and labels, and the relationship between samples and labels. In order to get the utmost out of these three relationships, a set matrix decomposition model is proposed to simultaneously decompose three relationship matrices. And, LLE is exploited to preserve the sample similarities and tag correlations in [28]. More and more work has been based on feature fusion in recent years [11, 13, 36, 54].

However, the single graph is not comprehensive and does not contain detailed information to describe the manifold information of the entire data set. The multiple graphs computed by multi-view or different parts of an image are structurally computationally intensive and consume a slice of time. Even a sea of insignificant parts or views have an unfavorable effect on the performance of the model. A question worth thinking about is whether these graph structures are indispensable. In other words, is there a graph that plays a leading role? We propose central attention with multi-graphs for image annotation. The central attention mechanism reckons that the central part of an image represents the most paramount part of the image. The framework is shown in Fig. 1. The two Laplacian graphs, L^C and L, are respectively computed according to the center part and the entirety of an image. And then the coefficient matrix W is optimized by these graph structures. The predictive labels are computed by the optimal coefficient matrix and the integral feature of the test image.

3 Proposed Method

In this section, we provide a detailed overview of the proposed method. We begin by introducing the relevant notations. Subsequently, in Sect. 3.2, we delve into the introduction of Laplacian eigenmaps. In Sect. 3.3, we elaborate on the central attention mechanism in conjunction with multigraph Laplacian eigenmaps. Finally, Sect. 3.4 presents the objective function.

3.1 Notations

The uppercase letter represents the matric in which the elements are described by two subscripts such as X^{ij}. And the lowercase letter with a subscript expresses the vector, but the scalar without a subscript. Then, X^T is the transpose of matrix X and the trace of a matrix is indicated by Tr. Afterwards, Frobenius norm of a matrix is computed by the trace: $\|A\|_F^2 = \text{Tr}\left(A^T A\right) = \text{Tr}\left(AA^T\right)$.

In our paper, the feature matrix of the integral images is described by $\hat{X} \in \mathbb{R}^{d\times n}$, where d is the dimension and n is the number of images in the training set. In addition, $\hat{X}^C \in \mathbb{R}^{d\times n}$ is the feature matrix in the center of the images. $\hat{Y} \in \mathbb{R}^{m\times n}$ is the groundtruth label matrix and m is the number of labels corresponding to the feature matrix. And $\hat{Y}_{ij} = 1$ manifests the tag i is related to the image j and zero, otherwise. Similarly, $\bar{X}$ is the feature matrix of the integral images and $\bar{X}^C$ is the feature matrix of center parts in the test set. Consequently, $X^T \in \mathbb{R}^{N\times d} = \left[(\hat{X})^T; (\bar{X})^T\right]$ is the feature matrix of both labeled and unlabeled integral images. N is the number of images in the entire dataset containing the training set and test set. And $\left(X^C\right)^T \in \mathbb{R}^{N\times d} = \left[\left(\hat{X}^C\right)^T; \left(\bar{X}^C\right)^T\right]$ is the feature matrix in the center of both labeled and unlabeled. images.

Our goal is to learn an appropriate coefficient matrix $W \in \mathbb{R}^{m\times d}$ in semi-supervised learning. Therefore, the predicted label matrix F is computed by the product of the coefficient matrix and the feature matrix.

$$F = WX. \tag{1}$$

3.2 Laplacian Eigenmap

In semi-supervised learning, the methods of embedding graph structure has achieved pleasurable performance. One of graph structure, Laplacian eigenmaps, exerts a tremendous fascination on a host of researchers due to the handleability. Laplacian eigenmaps assumes that the nearest neighbors of a sample are maintained as soon as possible when those samples are mapped from high-dimensional to low-dimensional. As shown in Fig. 2, the target sample is x_i and its nearest neighbors are circled in high-dimensional space. Corresponding to high-dimensional space, f_i expresses the dimensionality reduction target sample in low-dimensional space. Meanwhile, its nearest neighbors of the target sample are also reduced dimension to f_*.

The adjacency matrix is computed by the Gaussian kernel.

$$S_{ij} = \begin{cases} \exp\left(\frac{-\|x_i - x_j\|_2^2}{\sigma^2}\right), x_i \in N\left(x_j\right) \text{ or } x_j \in N\left(x_i\right) \\ 0, \quad \text{otherwise} \end{cases}, \tag{2}$$

where $N\left(x_*\right)$ is the k nearest neighbors of the sample x_*.

The relationship between the samples is maintained when mapped to a low-dimensional space. Therefore, the local manifold structure information of samples is preserved in label space.

$$\min_W \sum_{i,j=1}^{N} S_{ij} \left\| \frac{Wx_i}{\sqrt{D_{ii}}} - \frac{Wx_j}{\sqrt{D_{jj}}} \right\|_2^2 \Rightarrow \min_W \text{Tr}\left(WXL(WX)^T\right). \tag{3}$$

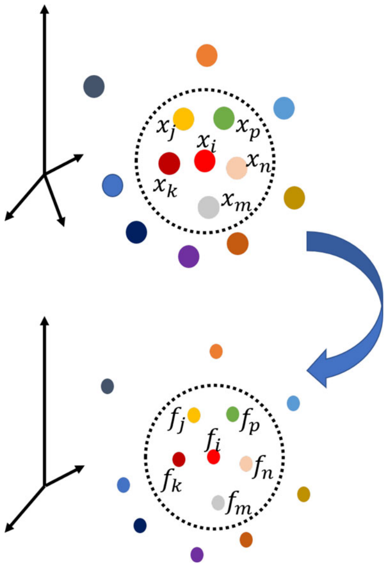


Fig. 2 The above coordinate is a Laplacian eigenmap from higher dimension and the under coordinate is a Laplacian eigenmap from lower dimension. This graph is a Laplacian eigenmap from higher to lower dimensions. Dots represent samples

where $D = \sum_{j=1}^{N} S_{ij}$ is a metric matrix. And L = I - $D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$ is the normalized Laplacian matrix of label a where I is unit matrix.

3.3 Central Attention Mechqanism with Multi-Graphs Laplacian Eigenmap

The single graph structure is arduous to learn the local manifold structure information due to the complexity of sample data. Therefore, the multi-graphs structure is employed. But the traditional methods adapt to the multi-view mechanism to constitute the multifarious graphs. The profuse details in an image are considered according to the multi-view mechanism. The aforementioned graphs are constructed by diverse features such as DenseSift, Rgb, Hsv, and depth feature. Furthermore, the diverse graphs are computed according to the different parts of an image. Specifically, the image is divided into pieces. And the features of these pieces respectively are harnessed to compose different graph structures. However, these graphs are structurally computationally intensive and consume a slice of time. Even a sea of insignificant parts have an unfavorable effect on the performance of the model. Therefore, a question worth thinking about is whether these graph structures are indispensable. In other words, a graph structure is computed just through the main part of an image. Meanwhile, this graph can still accomplish favorable experimental performance.

Therefore, we propose the central attention mechanism with multi-graphs Laplacian eigenmaps. The central attention mechanism assumes that the central part of an image represents the most paramount part of the image. More precisely, an image can be substituted by the central part of this image. As shown in Fig. 3, there are two images which are divided into five parts: Top left, bottom left, top right, bottom right, and center. The center part contains almost all tags comparing with other parts.

Consequently, we only employ the center part of the images to construct the graph considering detailed information instead of five parts. Then, the integral images are also utilized to build the graph structure. Furthermore, the test set images are considered during computing the graphs.

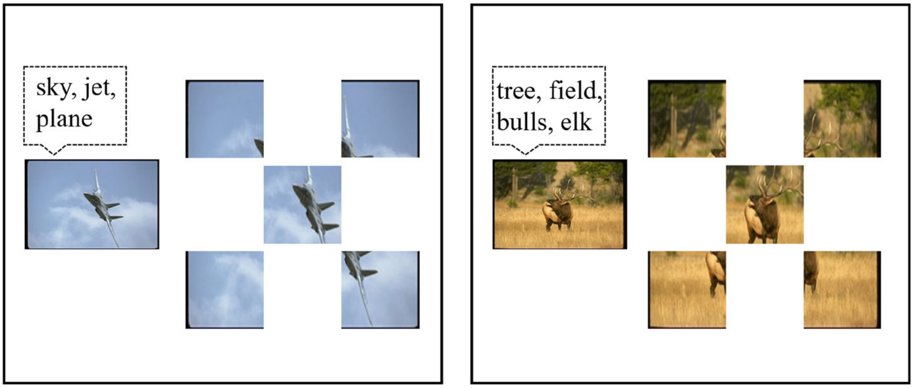


Fig. 3 The center parts of the images

The integral graph is established in the first place:

$$S_{ij} = \begin{cases} \exp\left(\frac{-\|x_i - x_j\|_2^2}{\sigma^2}\right), x_i \in N\left(x_j\right) \text{ or } x_j \in N\left(x_i\right) \\ 0, \quad \text{otherwise} \end{cases}, \tag{4}$$

where $N\left(x_*\right)$ is the k nearest neighbors of the sample $x_* \in X$ including training and test images.

According to Sect. III-B, the integral Laplacian graph can be computed:

$$P(W) = \min_W \operatorname{Tr}\left(WXL(WX)^T\right). \tag{5}$$

There is one more point that the graph structure of the center part is computed.

$$S_{ij}^C = \begin{cases} \exp\left(\frac{-\left\|x_i^C - x_j^C\right\|_2^2}{\sigma^2}\right), x_i^C \in N\left(x_j^C\right) \text{ or } x_j^C \in N\left(x_i^C\right) \\ 0, \quad \text{otherwise} \end{cases}, \tag{6}$$

$$Q(W) = \min_W \operatorname{Tr}\left(WX^C L^C \left(WX^C\right)^T\right), \tag{7}$$

where $L^C = I - \left(D^C\right)^{-\frac{1}{2}} S^C \left(D^C\right)^{-\frac{1}{2}}$ and $D^C = \sum_{j=1}^N S_{ij}^C$.

3.4 The Objective Function

In summary, the overview of our methods is reached.

$$J = \min_W L(WX, Y) + \lambda P(W) + \gamma Q(W) + \eta \|W\|_F^2, \tag{8}$$

$$L(WX, Y) = \|Y - WX\|_F^2, \tag{9}$$

where L(WX, Y) is the loss function. Meanwhile, P(W) and Q(W) are as regularizers, and λ, γ and η are the trade-off parameters.

4 Experimental Results

4.1 Experiment Setting

In this subsection, the experiment datasets are introduced firstly. Moreover, the feature extraction method and evaluation standards are given.

The three image datasets, Corel5K, ESPGame, and IAPRTC12 are exploited to demonstrate the superiority of the presented method. We randomly select 10% of the images as the training set and the remaining 90% of the images are test set in semi-supervised learning. Table 1shows the number of specific images of different data sets. And we delete the samples which have zero labels in this experiment.

Corel5K dataset contains 4992 images. And the vocabulary includes 260 tags. Each image is annotated by less than 5 tags. Meanwhile, the images with zero labels are removed.

ESPGame [18] is composed by 20768 images. And the vocabulary includes 268 labels. However, the images are divided into two parts on account of the device memory issues. Each part contains 10384 images, 10% of which are the training images. And each part is trained separately.

IAPRTC12 [18] consists of 19623 images. As before, this dataset is also divided into two parts. One part incorporates 9811 images and the other part includes 9812. We respectively select 10% of the images from each part. In addition, the vocabulary contains 291 tags.

In this experiment, the pre-trained ImageNet is employed in the feature extraction process. Firstly, the size of the image is modified to 224×224. Then, the pixel values that compose of the center part are from 57 to 168. Similarly, the feature of the center part is also extracted by the pre-trained ImageNet. Further, the feature vector is normalized ensuring that its ℓ_2 norm equals 1.

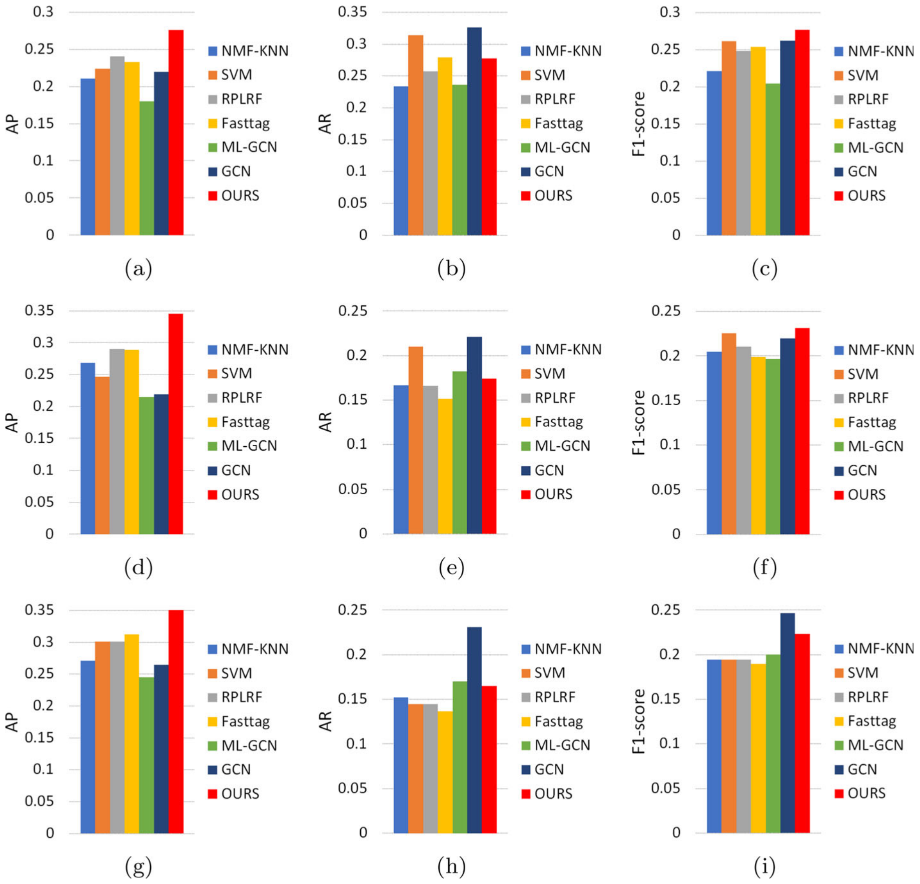
The AP(average precision), AR(average racall) and F1-score are the evaluation indicators in this paper. The top 5 labels of each image are considered as the final prediction labels. Meanwhile, 10% of the images are selected as the training set, and the remaining images are used as the test set in this experiment. The corresponding training and test sets form a data pair. We select randomly three data pairs from the Corel5k dataset to ensure the accuracy of the experiment. However, the three dataset pairs are elected from each part for ESPGame and IAPRTC12. And the final results are the average of all data pairs in a dataset.

4.2 Experimental Performance

In this section, if not most experimental results are provided with respect to the different methods. A host of methods are selected as methods of comparison, “RPLRF” [28], NMF-KNN [21], Fasttag [7], and SVM. Meanwhile, The experiment of the method with $\gamma = 0$ is employed to prove the validity of the center attention mechanism. In addition, we also select the algorithm that contains the graphs composed of the four parts of the images as

Table 1 Statistic of the datasets

	Corel5k	ESP Game	IAPRTC12
No. of images	4992	20768	19623
No. of training images	500	2076	1962
No. of tags	260	268	291



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 4 Annotation results. (**a**–**c**): Corel5k; (**d**–**f**): ESPGame; (**g**–**i**): IAPRTC12

the contrast algorithm to demonstrate the center part dominates in an image. This method is called as "Multi-view" and its objective function is shown as follows:

$$\begin{gathered} J = \min_W L(WX, Y) + \lambda P(W) + \gamma Q(W) + \\ \eta \|W\|_F^2 + \alpha R(W) + \beta T(W) + \mu U(W) + \tau G(W) \end{gathered}, \tag{10}$$

where R (W), T (W), U (W), and G (W) represent the top-left, bottom-right, top-right, and bottom-right graph structures function, respectively. And the sizes of these partial images are the same as center part.

$$* (W) = \min_W \operatorname{Tr}\left(W X^* L^* \left(W X^* \right)^T \right). \tag{11}$$

In Fig. 4, the AP, AR, and F1-score are revealed. As shown in Fig. 4, the superiority of our proposed can be shown comparing with other methods. More exactly, our method improves AP by 3.54%,5.47%, and 3.89% on Corel5k, ESPGame, and IAPRTC12 in comparison with the second-best method. Although it is not the highest values of AR on Corel5K and ESPGame, our method is still extremely effective in semi-supervised learning according to comparing all evaluation indicators.

Table 2 Experimental results

	Corel5k			ESP Game			IAPRTC12		
	AP	AR	F1 Score	AP	AR	F1 Score	AP	AR	F1 Score
$\gamma = 0$	26.59%	27.88%	27.18%	32.35%	17.41%	22.62%	34.43%	16.53%	22.24%
Multi-view	27.31%	27.82%	27.56%	35.01%	18.23%	23.93%	34.15%	16.45%	22.16%
OURS	27.59%	27.74%	27.65%	34.52%	17.41%	23.12%	35.08%	16.51%	22.35%

As Table 2 shown, It proves that the center part of the image is crucial to improve the model performance comparing with "$\gamma = 0$". Simultaneously, the center part of the image is also the main component of an image by comparing AP growth values. More precisely, our AP increases 1 and 2.17% comparing with "$\gamma = 0$" on Corel5k and ESPGame datasets. However, the AP improves 0.72% and 2.66% comparing “Multi-view” with “$\gamma = 0''$. Except for the center part of the image, other parts, the top-left, bottom-right, top-right, and bottom-right, even cause adversely effect on the experimental results for the IAPRTC12 dataset.

4.3 The Predicted Tags

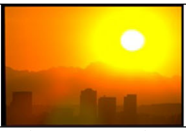
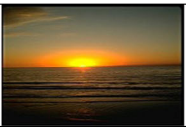
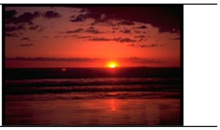
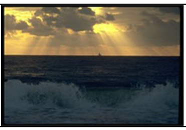
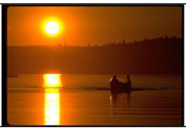
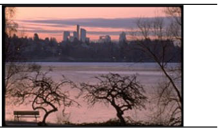
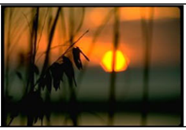
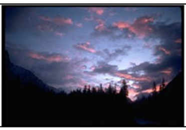
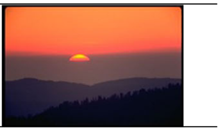
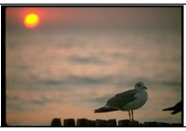
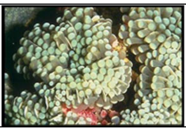
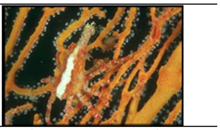
In image captioning tasks, whether the predicted keywords come from a predefined key pool is an important choice. The key pool is a collection containing possible keywords or labels, and these keywords are generated by a computer vision model based on the input image to produce textual descriptions or labels. In our proposed method, the predicted keywords are sourced from the key pool, and the model can only make selections from the known vocabulary. This ensures that the generated labels conform to pre-defined terms, making the approach more controlled. It guarantees label consistency with our task requirements and the definition of the key pool, thereby enhancing label consistency and interpretability.

The specific prediction results are shown in Table 3. The red font indicates no predicted labels and black bold font represents successful prediction labels. The images from the first line to the last line respectively represent 3 tags with successful prediction to 0 tags with successful prediction. There are two reasons for the failure to label: (1) The objects corresponding to a host of labels are not obvious in an image. And thus, their features are susceptible to merge with features of other parts; (2) The similar features of a sea of objects (color and texture, etc.) give rise to the failure to label, such as “water” and “sea”. It can solve these problems considering the relationship between tags.

5 Conclusions

In this paper, we propose central attention with multi-graphs for Image Annotation. The center part of an image is firstly considered as the main part. Further, the center graph comprises detailed information to complement the overall manifold structure. Thus, the superior performance can be available only through the graph structures constructed by central and overall features. Meanwhile, the computational complexity is reduced comparing with multi-view.

Table 3 Label results. The black bold: the correct predicted labels; The red blod: unpredicted labels

			
The groundtruth labels	city, mountain, sky, sun	sky, sun, sea, beach	sky, sun, clouds, sea
The predicted labels	**city**, **sky**, **sun**, clouds, sunset	**sky**, **sun**, clouds, **sea**, boats, hills	**sky**, **sun**, water, **clouds**, sunset
			
The groundtruth labels	sun, clouds, boats	sun, lake, boats, people	city, sun, tree, lake
The predicted labels	mountain, sky, sun, water, **clouds**, land	**sun**, sea, **boats**, land, sunset	mountain, sky, **sun**, water, **tree**
			
The groundtruth labels	sun, branch, leaf	sky, sun, clouds, tree	sky, sun, sea, hills
The predicted labels	sky, **sun**, tree, flowers, frost	mountain, water, **tree**, horizon, sunset	mountain, **sun**, clouds, tree, sunset
			
The groundtruth labels	sun, sea	water, anemone, ocean	close-up, coral, ocean
The predicted labels	sky, water, clouds, birds, street	tree, flowers, rocks, coral, garden	water, tree, snow, rocks, display

6 Future Studies

While our proposed method demonstrates significant performance across three datasets, there are also some limitations that need to be considered. Our proposed method also has certain limitations, as it may encounter challenges in disrupting the original spatial arrangement of images when utilizing central attention and multi-graph structures in deep learning. This means that, after applying these methods, the spatial structure and features of the images may be rearranged or distorted, which may not align with the requirements of the task at hand. To address this challenge, several strategies can be employed:

1. Spatial Transformer Networks (STN): Incorporate Spatial Transformer Networks (STN) to rectify the spatial distortions introduced by central attention and multi-graph structures. STN is a network layer that can learn spatial transformations on images, allowing for

adjustments to specific regions or the entire image, while preserving the spatial layout. This helps maintain the original image layout.

2. Multi-Scale Processing: Employ a multi-scale processing strategy, allowing simultaneous focus on features at different scales. This can be done by introducing multiple graph structures, each concentrating on different scales of image information, thereby preserving a part of the original spatial layout.
3. Image Reconstruction: Conduct image reconstruction on the processed features to restore the original spatial layout of the image. This approach is typically achieved using techniques like autoencoders or Generative Adversarial Networks (GANs) to transform features back into the original image.
4. Interpretability Methods: Employ interpretability methods to visualize the impact of central attention and multi-graph structures to ensure they do not have an unacceptable effect on the spatial arrangement of the original image. This helps in gaining a better understanding of how the model operates.

Future research directions should focus on preserving the original spatial arrangement of images. These methods mentioned above can be used to maintain the integrity of the original image spatial layout as much as possible, aligning with the specific requirements of the task. These strategies can be adjusted and combined based on the specific application context and task demands.

References

1. Bakliwal P and Jawahar CV (2015) Active learning based image annotation. In: 2015 Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp 1–4. IEEE
2. Belkin M, Niyogi P, and Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7:2399–2434
3. Bhagat PK, Choudhary P (2018) Image annotation: then and now. Image Vision Comput 80:1–23
4. Chen C, Li S, Wang Y, Qin H, Hao A (2017) Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans Image Process 26(7):3156–3170
5. Chen C, Wang G, Peng C, Fang Y, Zhang D, Qin H (2021) Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. IEEE Trans Image Process 30:3995–4007
6. Chen C, Wang G, Peng C, Zhang X, Qin H (2019) Improved robust video saliency detection based on long-term spatial-temporal information. IEEE Trans Image Process 29:1090–1100
7. Chen M, Zheng A, and Weinberger K (2013) Fast image tagging. In: International conference on machine learning, pp 1274–1282. PMLR
8. Chen Y, Liu L, Tao J, Chen X, Xia R, Zhang Q, Xiong J, Yang K, Xie J (2021) The image annotation algorithm using convolutional features from intermediate layer of deep learning. Multimed Tools Appl 80:4237–4261
9. Chen Z-M, Wei X-S, Jin X, and Guo Y (2019) Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp 622–627. IEEE
10. Cheng Q, Zhang Q, Peng F, Conghuan T, Li S (2018) A survey and analysis on automatic image annotation. Pattern Recogn 79:242–259

11. Dai Y, Gieseke F, Oehmcke S, Wu Y, and Barnard K (2021) Attentional feature fusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 3560–3569
12. Deane O, Toth E, Yeo S-H (2023) Deep-saga: a deep-learning-based system for automatic gaze annotation from eye-tracking data. Behav Res Methods 55(3):1372–1391
13. Dong H, Pan J, Xiang L, Hu Z, Zhang X, Wang F, and Yang M-H (2020) Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2157–2167
14. Donoho DL, Grimes C (2003) Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci 100(10):5591–5596
15. Fan J, Gao Y, Luo H (2008) Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation. IEEE Trans Image Process 17(3):407–426
16. Feng S, Lang C (2018) Graph regularized low-rank feature mapping for multi-label learning with application to image annotation. Multidim Syst Signal Process 29:1351–1372
17. Ge H, Yan Z, Dou J, Wang Z, and Wang Z (2018) A semisupervised framework for automatic image annotation based on graph embedding and multiview nonnegative matrix factorization. Mathematical Problems in Engineering
18. Guillaumin M, Mensink T, Verbeek J, and Schmid C (2009) Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: 2009 IEEE 12th international conference on computer vision, pp 309–316. IEEE
19. Helmy T, Djatmiko F (2023) Framework for automatic semantic annotation of images based on image's low-level features and surrounding text. Arab J Sci Eng 48(2):1991–2007
20. Huang S-J, Chen J-L, Mu X, and Zhou Z-H (2017) Cost-effective active learning from diverse labelers. In: IJCAI, pp 1879–1885
21. Kalayeh MM, Idrees H, and Shah M (2014) Nmf-knn: image annotation using weighted multi-view non-negative matrix factorization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 184–191
22. Kanda J, Soares C, Hruschka E, and De Carvalho A (2012) A meta-learning approach to select meta-heuristics for the traveling salesman problem using mlp-based label ranking. In: Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, Nov 12-15, Proceedings, Part III 19, pp 488–495. Springer
23. Ke X, Zou J, Niu Y (2019) End-to-end automatic image annotation based on deep cnn and multi-label data augmentation. IEEE Trans Multimedia 21(8):2093–2106
24. Kipf TN and Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907
25. Kong D, Ding C, Huang H, and Zhao H (2012) Multi-label relieff and f-statistic feature selections for image annotation. In: 2012 IEEE conference on computer vision and pattern recognition, pp 2352–2359. IEEE
26. Li J, Feng S, and Lang C (2016) Graph regularized low-rank feature learning for robust multi-label image annotation. In: 2016 IEEE 13th International Conference on Signal Processing (ICSP), pp 102–106. IEEE
27. Li X, Shen B, Liu B-D, Zhang Y-J (2016) A locality sensitive low-rank model for image tag completion. IEEE Trans Multimedia 18(3):474–483
28. Li X, Shen B, Liu B-D, Zhang Y-J (2017) Ranking-preserving low-rank factorization for image annotation with missing labels. IEEE Trans Multimedia 20(5):1169–1178
29. Li Y, Song Y, and Luo J (2017) Improving pairwise ranking for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3617–3625
30. Li Y, Song Y, and Luo J (2017) Improving pairwise ranking for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3617–3625
31. Liu W, Tao D (2013) Multiview hessian regularization for image annotation. IEEE Trans Image Process 22(7):2676–2687
32. Makadia A, Pavlovic V, and Kumar S (2008) A new baseline for image annotation. In: Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, Oct 12-18, Proceedings, Part III 10, pp 316–329. Springer
33. Mamat N, Othman MF, Abdulghafor R, Alwan AA, Gulzar Y (2023) Enhancing image annotation technique of fruit classification using a deep learning approach. Sustainability 15(2):901
34. Ning Z, Zhou G, Chen Z, Li Q (2018) Integration of image feature and word relevance: toward automatic image annotation in cyber-physical-social systems. IEEE Access 6:44190–44198
35. Pulgarín-Ospina CC, del Amor R, Colomera A, Silva-Rodríguez J, and Naranjo V (2023) Histocolai: an open-source web platform for collaborative digital histology image annotation with ai-driven predictive integration. arXiv preprint arXiv:2307.07525

36. Qin X, Wang Z, Bai Y, Xie X, Jia H (2020) Ffa-net: feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence 34:11908–11915
37. Shahraki FF and Prasad S (2018) Graph convolutional neural networks for hyperspectral data classification. In: 2018 IEEE global conference on signal and information processing (GlobalSIP), pp 968–972. IEEE
38. Shi Z, Yang Y, Hospedales TM, Xiang T (2016) Weakly-supervised image annotation and segmentation with objects and attributes. IEEE Trans Pattern Anal Mach Intell 39(12):2525–2538
39. Tang C, Liu X, Wang P, Zhang C, Li M, Wang L (2019) Adaptive hypergraph embedded semi-supervised multi-label image annotation. IEEE Trans Multimedia 21(11):2837–2849
40. Verma Y (2019) Diverse image annotation with missing labels. Pattern Recogn 93:470–484
41. Verma Y, Jawahar CV (2017) Image annotation by propagating labels from semantic neighbourhoods. Int J Comput Vision 121:126–148
42. Wang F, Liu J, Zhang S, Zhang G, Li Y, Yuan F (2019) Inductive zero-shot image annotation via embedding graph. IEEE Access 7:107816–107830
43. Wang G, Chen C, Fan D-P, Hao A, and Qin H (2021) From semantic categories to fixations: a novel weakly-supervised visual-auditory saliency detection approach. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15119–15128
44. Wang L, Ding Z, and Fu Y (2018) Adaptive graph guided embedding for multi-label annotation. In IJCAI
45. Wu B, Chen W, Sun P, Liu W, Ghanem B, and Lyu S (2018) Tagging like humans: diverse and distinct image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7967–7975
46. Xiang Y, Zhou X, Chua T-S, and Ngo C-W (2009) A revisit of generative model for automatic image annotation using markov random fields. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 1153–1160. IEEE
47. Xue Z, Junping D, Zuo M, Li G, Huang Q (2019) Label correlation guided deep multi-view image annotation. IEEE Access 7:134707–134717
48. Zhai Z, Staring M, Zhou X, Xie Q, Xiao X, Els Bakker M, Kroft LJ, Lelieveldt BPF, Boon GJAM, Klok FA et al (2019) Linking convolutional neural networks with graph convolutional networks: application in pulmonary artery-vein separation. In: Graph Learning in Medical Imaging: First International Workshop, GLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 1, pp 36–43. Springer
49. Zhang J, Yang J, Jun Yu, Fan J (2022) Semisupervised image classification by mutual learning of multiple self-supervised models. Int J Intell Syst 37(5):3117–3141
50. Zhang J, He Z, Zhang J, Dai T (2019) Cograph regularized collective nonnegative matrix factorization for multilabel image annotation. IEEE Access 7:88338–88356
51. Zhang J, Wu Q, Zhang J, Shen C, and Lu J (2019) Mind your neighbours: image annotation with metadata neighbourhood graph co-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2956–2964
52. Zhang J, Rao Y, Zhang J, Zhao Y (2019) Trigraph regularized collective matrix tri-factorization framework on multiview features for multilabel image annotation. IEEE Access 7:161805–161821
53. Zhang P, Wei Z, Li Y, Zhao C (2017) Automatic image annotation based on multi-auxiliary information. IEEE Access 5:18402–18411
54. Zhang Z, Zhang X, Peng C, Xue X, and Sun J (2018) Exfuse: enhancing feature fusion for semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 269–284

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.