A Multiview Text Imagination Network Based on Latent Alignment for Image-Text Matching

Heng Shang, Guoshuai Zhao, Jing Shi, and Xueming Qian, *Xi'an Jiaotong University, Xi'an, 710049, China*

In image-text matching fields, one of the keys to improving performance is to extract features with more semantic information. Existing works demonstrate that semantic enrichment through knowledge expansion can improve performance. Most of them expand image features, however, the shortage of semantic information in text modality and the unilateral character of the view are often bottlenecks that limit the performance of image-text matching models. To solve the two problems, we aggregate knowledge from multiple views and propose a word imagination graph (WIG). A WIG can be used to expand textual semantic information by imagination based on input images. Then, utilizing WIG, we construct a novel multiview text imagination network (MTIN). A MTIN enables latent alignment of images and texts on tags, which can assist matching on a semantic level. Results from the Flickr30K and MS-COCO datasets demonstrate the effectiveness of our method. The source code has been released on GitHub https://github.com/smileslabsh/Multiview-Text-Imagination-Network.

With the increasing dramatic growth of all kinds of modal information on the Internet, cross-modal retrieval is receiving increasing attention, and the image-text matching task is one of the most crucial branches.

Many effective image-text matching methods have been proposed in recent years. Some of these methods try to achieve better results through knowledge expansion.[1–5] And most knowledge-expansion methods focus only on image modality. Text modality is as important as image modality in bidirectional image-text matching, however, thus far, no work has been done to perform knowledge expansion of text modality. Through our analysis, we found that the shortage of semantic information in text modality made image-text matching tasks extremely difficult. Therefore, it is necessary to research knowledge expansion for text modality in image-text matching.

Specifically, the existing methods face two major challenges. First, the text contains insufficient semantic information. Most of the sentences in the datasets are very short, making it difficult to extract enough semantic information. Take the first sentence in Figure 1 as an example, although this sentence describes the main part of the image, there is still something that is not mentioned, like waves, the surfboard, and the man's action. Second, the caption is unilateral. In both datasets, one image corresponds to five independent, human-generated captions. However, when we look at an image, different people have different views. In Figure 1, the first caption notices the location. The second caption focuses on the waves. Only the third caption mentions the man's movements. So, these captions often describe only part of the image, and it makes the image-text matching task more challenging.

To address the first challenge, we use co-occurrence knowledge to construct the word imagination graph (WIG) and expand the textual information by the WIG. By doing this knowledge expansion, we give our model the power of imagination. For example, when we input the sentence, "A man in blue surfs in the ocean," our model can imagine a *wave* based on the *ocean*, and imagine a *surfboard* based on the *surf*. Then, the feature of *wave* and *surfboard* will be fused into the text feature. Thus, we can obtain a textual representation that contains richer semantic information. In addition, we fuse the detected tag features into the image

1541-1672 © 2023 IEEE
Digital Object Identifier 10.1109/MIS.2023.3265176
Date of publication 10 April 2023; date of current version 5 June 2023.

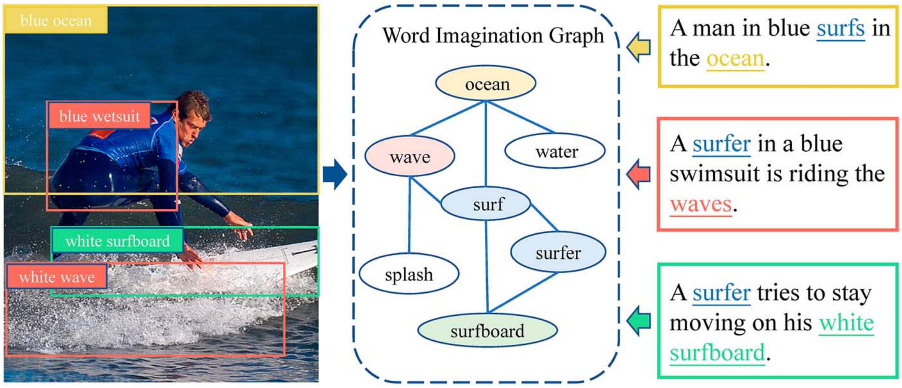

blue ocean
blue wetsuit
white surfboard
white wave
Word Imagination Graph
ocean
wave
water
surf
surfer
splash
surfboard
A man in blue surfs in the ocean.
A surfer in a blue swimsuit is riding the waves.
A surfer tries to stay moving on his white surfboard.

FIGURE 1. Example of a constructed WIG. Two words connected by the same edge form a co-occurrence pair.

feature. Therefore, there is a latent alignment in the matching process as the *wave's* feature is fused into both the image and text features. The latent alignment here refers to the alignment of tags in the image and text features. With latent alignment, we can bridge the semantic gap and achieve better performance. For the second challenge, we aggregate co-occurrence knowledge from different captions that describe the same image and add them into WIG. In this way, we make WIG contain multiview information and alleviate the unilateral character of the view. Furthermore, to make the imagination more reasonable and reduce noise, we use the attention mechanism to weigh the imagination words according to the input image.

The following are the main contributions of this work:

- We propose a WIG, which aggregates knowledge from multiple views. By using a WIG, our model can achieve latent alignment and bridge the semantic gap of images and texts.
- We propose a novel model, a multiview text imagination network (MTIN), which is used to expand semantic information of the text by imagination to obtain a better feature representation.
- We conduct extensive experiments, and the results show the effectiveness of our approach.

RELATED WORK

Knowledge-Expansion-Based Methods

Recently, an increasing number of methods have used external knowledge in deep learning. Du et al.[6] proposed a common-sense knowledge memory module to better incorporate common-sense knowledge into stance classification. Majumder et al.[7] proposed a method based on a novel document modeling technique. Wang et al.[8] proposed a hierarchical reinforcement learning algorithm for multihop knowledge reasoning. For image-text matching, Position Focused Attention Network (PFAN)[2] and PFAN++[3] used object position knowledge to expand region expression. And the following two works are similar to ours. Shi et al.[1] expanded image features by using scene common-sense knowledge. Wang et al.[4] proposed a Consensus-aware Visual-Semantic Embedding (CVSE) model to incorporate common-sense knowledge into image-text matching. Common to the two aforementioned works is that we all hope to apply external knowledge to image and text matching tasks, and both Shi et al.[1] and Wang et al.[4] use the co-occurrence relationship of words.

The following are some main differences between our approach and their two methods:

- We use different ways to extract co-occurrence pairs. Shi et al.[1] use the Visual Genome dataset and extract co-occurrence pairs from human-labeled triplets. Wang et al.[4] extract co-occurrence pairs from image-captioning corpus. In our method, we use the Flickr30K and MS-COCO datasets, and first use the region tags in the image for co-occurrence pair extraction. Moreover, we first use multiview information during co-occurrence pair extraction.
- We have different motivations for using external knowledge. Shi et al.[1] use external knowledge for image feature expansion. CVSE[4] learns the associations and alignments between image and text based on the exploited knowledge. However, we focus on text modality. In view of the existing two shortcomings of image and text matching, we use external knowledge to expand text features.
- Our model design is different. Shi et al.[1] use a multiple-label classification model to extract semantic information from images, and use the Visual

Geometry Group network to extract global representations of images. However, we extract region features and tags of images through bottom-up attention, and use region features as image representations. As for CVSE,[4] our image or text feature extraction methods and the application of external knowledge in the model are all different.

- Finally, we design a special expansion module and an attention module for the imagination mechanism proposed in this article, which have not appeared in previous methods. To the best of our knowledge, we first applied knowledge expansion to text modality in the field of image-text matching.

METHODOLOGY

Image Representation

To take full advantage of the association between images and texts, we extract objects by using a bottom-up attention model faster region-based convolutional neural network[9] pretrained on Visual Genomes[10] by Anderson et al.[11] By using this detector, we get the region feature of each image, represented as $\{v_1, v_2, \cdots, v_n\}$, where n is the number of regions. In addition to outputting the region feature, the detector also outputs tags. We use the Bidirectional Encoder Representation from Transformers (BERT) model[12] to encode these tags. Each region has a tag, so the tag feature can be represented as $\{t_1, t_2, \cdots, t_n\}$. To achieve latent alignment and bridge the semantic gap, we fuse the extracted tag feature with the region features. Then, a fully connected layer is used to transform these fused features into d-dim vectors: $f_i = FC(\text{Concat}(v_i, t_i))$. To input these region features into our model, we convert them to a matrix. So, image features can be represented as $[f_1, f_2, \cdots, f_n]$.

Then, we use the multihead self-attention mechanism to make the inputs interact with each other. Finally, to get the final, one-dim representation $e^v \in \mathbb{R}^{1\times d}$, we use an average pooling layer, followed by the L2 normalization.

Word Imagination Module

WIG

We generate a co-occurrence matrix M, where each row represents a word in texts (original word), and each column represents a tag of a region (expanded word). We initialize the matrix with zero. Then, we traverse all image-text pairs. Each image-text pair $p = \langle I, T\rangle$ contains a tag list $\text{Tag}(I) = \{\text{tag}_1, \text{tag}_2, \cdots, \text{tag}_n\}$, where n is the number of regions, and a word list $\text{Word}(T) = \{\text{word}_1, \text{word}_2, \cdots, \text{word}_{n_w}\}$, where n_w is the number of words in T. During the traversal process, we enumerate all possible word-tag pairs in p: $\langle c_{\text{ori}}, c_{\text{exp}}\rangle \in \text{Word}(T) \times \text{Tag}(I)$, where × is the Cartesian product. After that, we add "1" to the corresponding position of M. Finally, to introduce multiview knowledge into the model, we aggregate word–word co-occurrence pairs from multiple views. For two sentences, T_1 and T_2, which describe the same image, we define their word lists as $\text{Word}(T_1)$ and $\text{Word}(T_2)$, respectively. So, we can extract co-occurrence pairs from these two views: $\langle c_{\text{ori}}, c_{\text{exp}}\rangle \in \text{Word}(T_1) \times \text{Word}(T_2)$. Then, we refine the co-occurrence matrix using the same method.

After the co-occurrence matrix has been constructed, we filter the data in the matrix. The first step is to filter out fewer data than the threshold th_1. The purpose is to avoid generating a graph that is too large, and to avoid introducing noisy data. In the second step, we want the two words making up the co-occurrence pair to be more relevant. The weight of the co-occurrence pair located at the ith row and the jth column is

$$w(p_{ij}) = M[i][j]/\sum_k M[i][k], k = 1, 2, \cdots, n_c, \tag{1}$$

where n_c is the number of columns. So, for each row in M, we only retain elements whose weight is greater than the threshold th_2.

Finally, we construct the WIG from the co-occurrence pairs. In the WIG, each node represents a word. And if two nodes are connected by an edge, they have a relation of co-occurrence.

Word-Expansion Algorithm

We define the input sentence as T and the number of words to be expanded as n_e. For every word word_i in T, we expand n_e concepts for it. Specifically, starting from word_i, we traverse the WIG and record the traversed nodes and their weights. Then, we sort these nodes in descending order by weight and take the top n_e results. In particular, when there are fewer than n_e concepts available for expansion, we fill them with *None*. Afterward, we input these imagination concepts into our model.

Imagination Attention Module

Before the imagination attention module, we transform the expansion feature into a $d-$dim vector $E = [e_1, e_2, \cdots, e_{n_e}]$. The imagination attention module is used to associate the expansion feature with the image embedding e^v and decide how much weight should be paid to each expanded word.

$$\beta_i = \tanh(f(e^v, e_i)),\ i = 1, 2, \cdots, n_e \tag{2}$$

$$f(e^v, e_i) = (e^v)^T M e_i \tag{3}$$

where $M \in \mathbb{R}^{d\times d}$ is the mapping matrix. The weight of each expanded concept is not only related to the image but also to its original weight. So, we fuse the co-occurrence pair's original weight w_i into the attention mechanism

$$\alpha_i = \frac{\alpha_i'}{\sum_i \alpha_i'}, \text{where } \alpha_i' = \frac{\exp(\beta_i)}{\sum_i \exp(\beta_i)} \times w_i. \tag{4}$$

α_i is the final weight of expansion embedding e_i. So, the expansion feature can be represented as

$$E = \sum_i^{n_e} e_i \times \alpha_i. \tag{5}$$

Text Representation

In the previous image representation branch, we use BERT to encode tags. Similarly, as for the text representation branch, we use BERT to encode the imagination words and input sentences. As shown in Figure 2, after the BERT layer, we get both the expansion and caption features.

As for the caption feature, the context convolution module is applied to exploit the context information. This module is the same as the one in SAEM.[13] one-dim convolution neural networks of three window sizes (1, 2, and 3) are used to capture the phrase-level information. Then, we concatenate both the expansion and caption features, and use a fully connected layer to get the final text embedding $e^t \in \mathbb{R}^{1\times d}y$.

Loss Function

In the training process, we use two loss functions. The first loss function is triplet loss, which is a standard ranking objective function. The purpose of the triplet loss in image-text matching is to ensure that similar image-text pairs are closer in the joint space. True matched images and sentences have higher matching scores than mismatched images and words. For an image-text pair $\langle I, T \rangle$, the triplet loss that we use is

$$\begin{aligned} L_{\text{triplet}}(I,T) = \max[0, \beta - s(I,T) + s(I,\hat{T})] \\ + \max[0, \beta - s(I,T) + s(\hat{I},T)] \end{aligned} \tag{6}$$

where β is a margin parameter. For the image I, $\hat{T}$ denotes a negative sentence, and for the text T, $\hat{I}$ denotes a negative image. We employ these hardest negatives in minibatch, and it leads to a better retrieval performance and better computational efficiency.

In addition, we use angular loss in training. By employing a third-order geometric constraint, it can capture a more local structure of the triplet triangles than the triplet loss

$$\begin{aligned} L_{\text{angular}}(I,T) = \log[1 + \exp(f(I,T,\hat{T}))] \\ + \log[1 + \exp(f(T,I,\hat{I}))] \end{aligned} \tag{7}$$

where $\hat{I}$ and $\hat{T}$ are also the hardest negatives in minibatch, and $f(a,b,n) = 4\tan^2\alpha(a+b)n^T - 2(1+\tan^2\alpha)ab^T$, where α is the angular margin parameter.

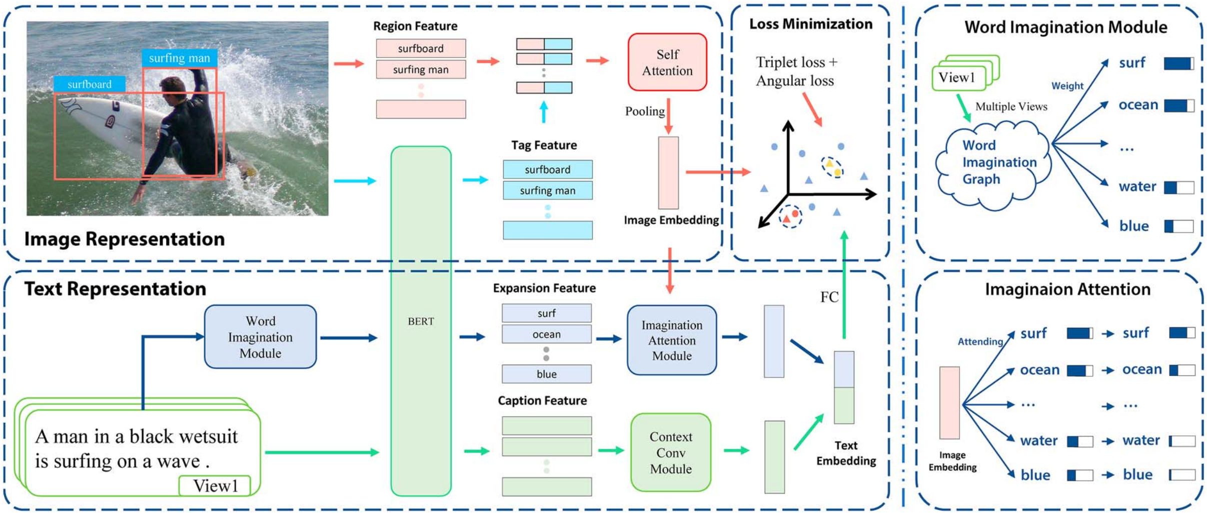


FIGURE 2. Schematic illustration of the MTIN model, which has three components. 1) *Image representation.* To get image embedding, we concatenate region features and tag features. Then, a self-attention mechanism is used to obtain better image representation. 2) *Text representation.* BERT is used to encode captions and expanded words. These two features are enhanced by the Imagination Attention and Context Conv modules, respectively. We fuse these two features as text embedding. 3) *Matching.* Triplet loss and angular loss are used in training. Conv: convolutional; FC: full connection.

EXPERIMENTS

Dataset

Flickr30K[14] is a publicly accessible database of images and their captions. There are 31,783 images and 158,915 English captions in the collection. Each image is accompanied by five captions. We divided the dataset into 1000 images for validation, 1000 images for test, and the remaining images for training.

The MS-COCO dataset[15] is a large-scale dataset for object detection, segmentation, and image captioning. There are 113,682 photos in the collection, with five captions for each image. We follow the work of Faghri et al.[16] to add 30,504 photos to the training set that were previously in the MS-COCO validation set but were not included in this split. The test results are reported by averaging more than five folds of 1000 test images.

Experimental Settings

Evaluation Protocols

We use recall at K, R @ K ($K = 1, 5$, and 10) as the evaluation metrics. In the image-text matching field, we sort images or texts to be matched by a similarity score and select K samples with the highest scores. And R @ K is defined as the proportion of success that can be retrieved from the top-K results.

Implementation Details

The Adam[17] method is employed as a gradient update algorithm, with a learning rate of 0.0001 and a 10% decay for every 10 epochs. The model is trained for 30 epochs. The GPU used in our experiment platform is Nvidia GeForce GTX 2080ti. The batch size is set to 64. For the image representation part, the region feature is 2048 dim, and the number of regions n is 36. The self-attention module has 16 heads. For text representation, the pretrained BERT model has 12 layers and 768 hidden units. For the word imagination part, the number of words to expand n_e is set to five. The $th1$ and $th2$ thresholds are set to 3000 and 0.035, respectively. The final dimension d is 256.

Performance Evaluation

To evaluate effectiveness, we compare it to some strong baselines. Self-Attention Embeddings (SAEM)[13] is our baseline, which also learned the embedding of images and texts directly. CVSE[4] is a way to use external knowledge. We also compare our MTIN model to other influential methods. VSE$++$,[16] SCO,[18] and SCAN[19] are the early classical models of image-text matching. Recently, CAAN[20] has been an influential method. The results are shown in Table 1.

Our method achieves the best result at mean Recall (mR) on both datasets. Compared to our baseline SAEM,[13] the MTIN achieves a 1.3% improvement at mR on Flickr30K, and a 2.3% improvement on MS-COCO. A knowledge-based method like CVSE[4] is the most similar to our approach. Compared with CVSE,[4] our model has better results on both datasets. Our model has significant advantages over some earlier approaches, such as VSE$++$,[16] SCO,[18] and SCAN.[19] Also, the MTIN achieves better results than the recent CAAN method.[20] This supports the efficacy and robustness of our MTIN model.

Ablation Study

To fully verify the validity of our module, we also performed several ablation experiments. As listed in Table 1, 1) "without imagination" means that our model does not use the WIG to expand text features, which can prove the validity of the idea of expanding text features by word imagination; 2) "without attention1" means that we remove the imagination attention module, and all the expanded features will be given equal weights, which can demonstrate effectiveness of the imagination attention mechanism in our method; 3) "without attention2" means that we remove the imagination attention module and directly input the expansion feature and the caption feature into the context convolution module after connecting them; and 4) "without tag" means that we use only region features as inputs for the image representation process. This proves that the fusion of tag features into image features can indeed fit the semantic gap, resulting in better results. It should be noted that the tag information will still be applied to the construction of the co-occurrence matrix M. Because the matrix needs more comprehensive and accurate co-occurrence information, the use of tag information can make the imagination more comprehensive.

As listed in Table 1, word imagination brings major enhancements to our model. mR of the MTIN without imagination on MS-COCO is reduced from 86.4 to 83.9, a reduction of 2.5 compared to the complete model. As for the imagination attention module, the improvement is mainly on image-to-text retrieval. However, on text-to-image retrieval, the result is not significantly improved or even decreased. This is because the model cannot discriminate noise after removing the attention module. Finally, with the latent alignment mechanism, our approach obtains a 0.7 improvement at mR on MS-COCO.

Imagination Parameter Analysis

The impact of various imagination parameters is explained to increase the understanding of our MTIN model. First, when constructing the WIG, we tried using different $th1$

TABLE 1. Comparisons of cross-modal retrieval on Flickr30K and MS-COCO datasets with the competing methods.

	Flickr30K Dataset							MS-COCO Dataset						
	Image to text			Text to image				Image to text			Text to image			
Methods	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**mR**	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**	**mR**
VSE++[15]	52.9	80.5	87.2	39.6	70.1	79.5	68.3	64.6	90	95.7	52	84.3	92	79.8
SCO[17]	55.5	82	89.3	41.1	70.5	80.1	69.8	69.9	92.9	97.5	56.7	87.5	94.8	83.2
SCAN[18]	67.4	90.3	95.8	48.6	77.7	85.2	77.5	72.7	94.8	98.4	58.8	88.4	94.8	84.7
SAEM[12]	69.1	91	95.1	52.4	81.1	88.1	79.5	71.2	94.1	97.7	57.8	88.6	94.9	84.1
CVSE[4]	**73.5**	92.1	95.8	**52.9**	80.4	87.8	80.4	74.8	95.1	98.3	59.9	89.4	95.2	85.5
CAAN[19]	70.1	91.6	**97.2**	52.8	79	87.9	79.8	75.5	95.4	**98.5**	61.3	89.7	95.2	85.9
MTIN (without imagination)	69.1	91	95	52.7	81.1	88.1	79.5	71.8	93.3	97.8	57.7	88.2	94.7	83.9
MTIN (without attention1)	70.7	91	95.6	52.5	81	88.5	79.9	72.5	94.2	98	58.6	88.5	95	84.5
MTIN (without attention2)	69.8	90.7	95.3	51.9	81.4	88	79.5	72.2	93.9	98.2	58.1	88.5	95.3	84.4
MTIN (without tag)	72.1	92	95.9	52.3	81.8	88.7	80.5	74.1	95.3	98.1	61.5	90	95.4	85.7
MTIN (full)	72.2	**92.9**	96.4	52.7	**81.9**	**88.8**	**80.8**	**76**	**95.6**	98.4	**61.9**	**90.6**	**95.8**	**86.4**

and $th2$ thresholds. Because $th1$ and $th2$ are correlated, we first set $th1$ to 3000.

Then, we determine the value of $th2$ based on the results of our model on MS-COCO. When $th2$ takes the value of 0.035, the optimal result on MS-COCO is obtained. After that, we focused on the number of imagination words n_e. As shown in Figure 3(b), when n_e is small, the expansion effect is not obvious, and the advantages of our method cannot be revealed. When there are more imagination words, more noise is introduced, that is, the co-occurrence pairs that are not closely related; this leads to poor results. Finally, n_e is set to five.

WIG-Effectiveness Analysis

To demonstrate the effectiveness of WIG expansion, we calculate the accuracy of expanded words and the attention mechanism on the MS-COCO test set. We want expanded words to be able to associate with the image or other views, so when there are expanded words that come from tags or other views, we think that the expansion is valid on this sentence. But it is still important to note that even if the expanded word does not come from the image or other views, the word may still be meaningful. We calculate the accuracy of the attention by judging whether the most concerned expanded word is valid. The attention accuracy

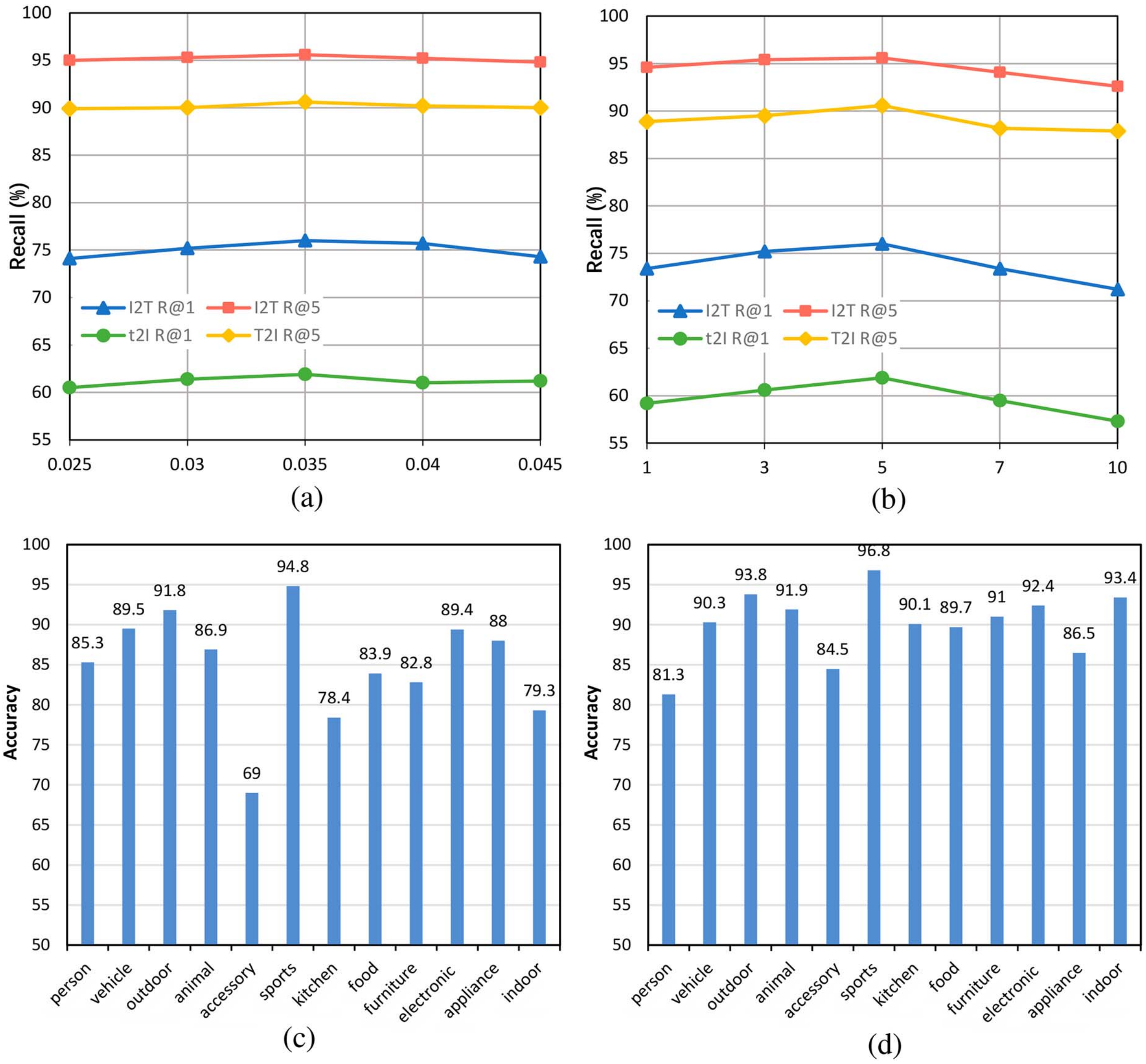


FIGURE 3. Imagination parameter analysis and WIG-effectiveness analysis results. (a) Value of $th2$. (b) Value of n_e. (c) Expansion accuracy on subcategories. (d) Attention accuracy on subcategories.

experiment is only carried out on effectively expanded sentences. The MS-COCO test set contains 1000 images and 5000 captions and has 12 subcategories. On each subcategory, we also calculate accuracy.

As shown in Figure 3(c), expansion accuracy is satisfactory on most of the subcategories. In "Sports," the expansion accuracy is 94.8%. However, in "Accessory," the accuracy is only 69%. This is because in "Accessory," the objects and scenes are not closely related, so the imagination is weakened. There are some subcategories under "Accessory," such as "Hat," "Shoe," and "Eyeglasses." These objects have no special use scenarios and can appear in any scenario. The subcategories of "Sports" include "Surfboard," "Snowboard," and "Tennis Racket." So, the model can image from "Surfboard" to "Ocean," and from "Snowboard" to "Snow."

As shown in Figure 3(d), the highest accuracy rate also appears in "Sports." But the result in "Person" is only 81.3%. This is because the scene under the subcategory "Person" is relatively single, and the connection between expansion words and the person is not close.

Case Study and Analysis

We ran the model on real cases and selected two cases for demonstration in Figure 4. The first case is also in Figure 1. For each caption, our MTIN model gives five expanded words. We can observe that the MTIN expands some concepts that are not mentioned in the caption but appear in the picture. These concepts can be used to enrich semantic information, thereby helping to improve the image-text matching.

CONCLUSION

In this article, we proposed a MTIN to overcome the two major challenges of image-text matching: 1) shortage of the semantic information in text modality and 2) unilateral character of the view. By using external knowledge and multiview information, the MTIN can imagine based on input pictures. Results on the Flickr30K and MS-COCO datasets demonstrate the effectiveness of our proposed MTIN method. Our idea of further mining the semantic components of text and expanding knowledge is valuable for semantic alignment between image and text.

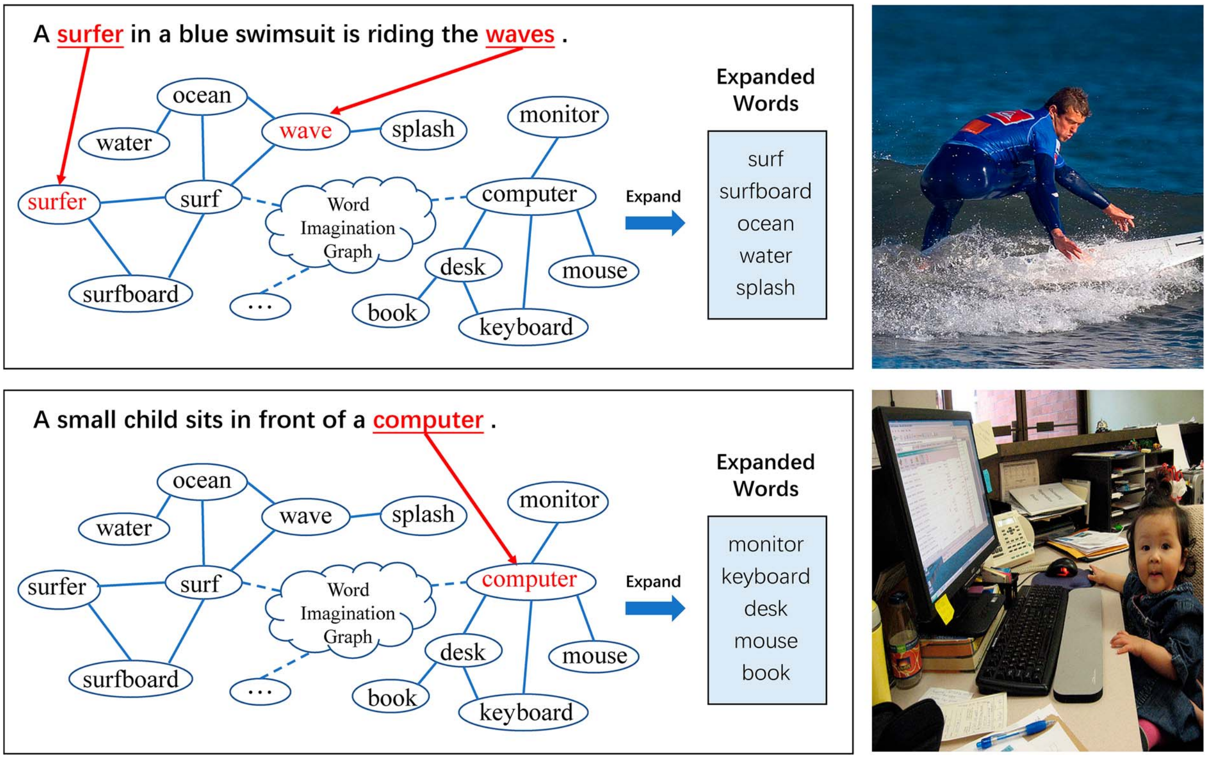


FIGURE 4. A real showcase of intermediate result of our word-expansion module. Both images are from the test set. Words marked in red are nodes in WIG, and we use these words as original words for expansion. We traverse the WIG starting from these original words and expand five top-scoring concepts for the caption. These expanded concepts may come from tags or other views of the image.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 61902309, in part by China Postdoctoral Science Foundation under Grant 2020M683496 and BX20190273, in part by the Fundamental Research Funds for the Central Universities, China (xxj022019003, xzd012022006); in part by the Humanities and Social Sciences Foundation of Ministry of Education, China under Grant 16XJAZH003, and in part by the Science and Technology Program of Xi'an, China under Grant 21RGZN0017. G. Zhao is the corresponding author.

REFERENCES

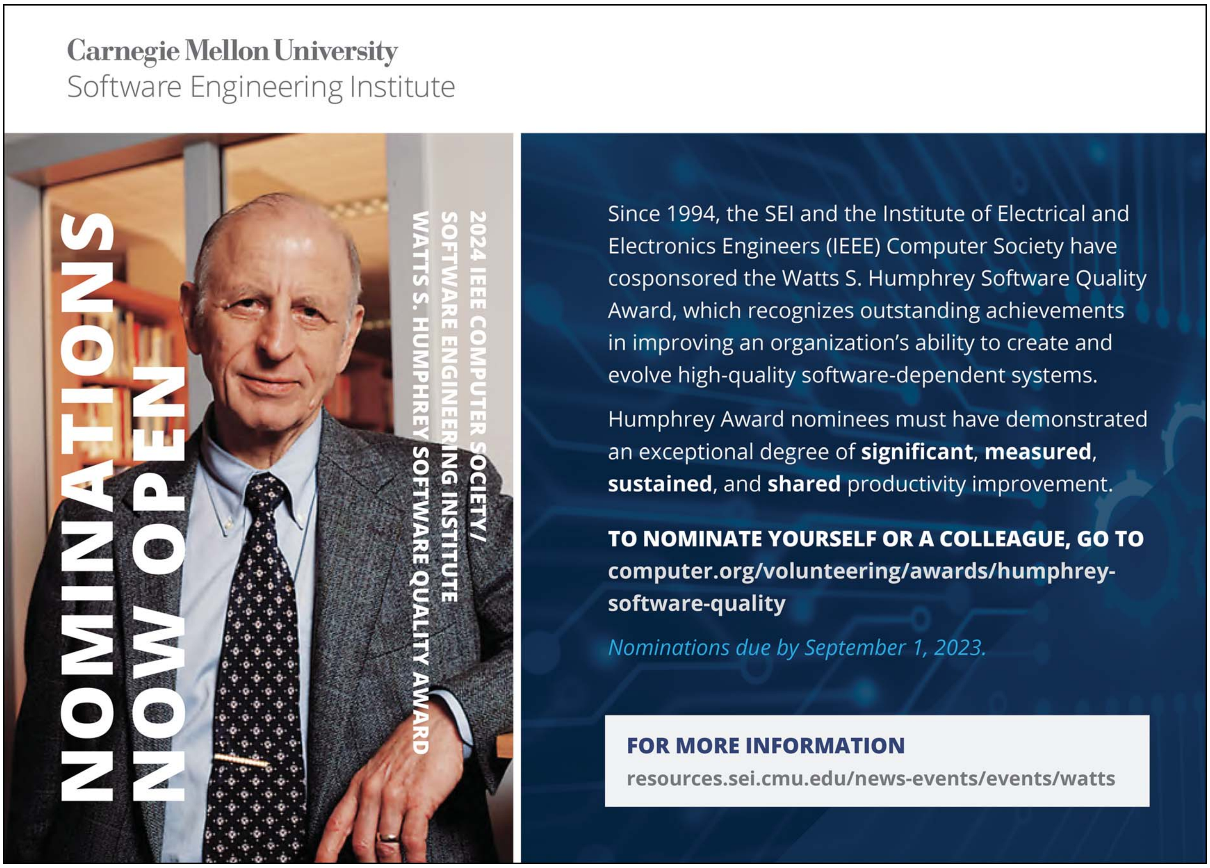
1. B. Shi, L. Ji, P. Lu, Z. Niu, and N. Duan, "Knowledge aware semantic concept expansion for image-text matching," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 5182–5189.
2. Y. Wang et al., "Position focused attention network for image-text matching," in *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2019, pp. 3792–3798.
3. Y. Wang et al., "PFAN++: Bi-directional image-text retrieval with position focused attention network," *IEEE Trans. Multimedia*, vol. 23, pp. 3362–3376, 2021, doi: 10.1109/TMM.2020.3024822.
4. H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2020, pp. 18–34, doi: 10.1007/978-3-030-58586-0_2.
5. G. Zhao, C. Zhang, H. Shang, Y. Wang, L. Zhu, and X. Qian, "Generative label fused network for image-text matching," *Knowl. Based Syst.*, vol. 263, 2023, Art. no. 10280, doi: 10.1016/j.knosys.2023.110280.
6. J. Du, L. Gui, R. Xu, Y. Xia, and X. Wang, "Commonsense knowledge enhanced memory network for stance classification," *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 102–109, Jul./Aug. 2020, doi: 10.1109/MIS.2020.2983497.
7. N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar./Apr. 2017, doi: 10.1109/MIS.2017.23.
8. Z. Wang, L. Li, and D. D. Zeng, "Hierarchical multihop reasoning on knowledge graphs," *IEEE Intell. Syst.*, vol. 37, no. 1, pp. 71–78, Jan./Feb. 2022, doi: 10.1109/MIS.2021.3095055.
9. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
10. R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vision*, vol. 123, no. 1, pp. 32–73, May 2017, doi: 10.1007/s11263-016-0981-7.
11. P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 6077–6086.
12. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol. (NAACL-HLT)*, 2018, pp. 4171–4186.
13. Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2088–2096, doi: 10.1145/3343031.3350940.
14. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014, doi: 10.1162/tacl_a_00166.
15. T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
16. F. Faghri, D. J. Fleet, J. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vision Conf. (BMVC)*, 2018, p. 12.
17. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR)*, 2015.
18. Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2018, pp. 6163–6171.
19. K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, 2018, pp. 212–228.
20. Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2020, pp. 3533–3542.

HENG SHANG is currently working toward his M.S. degree with the School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. His research interests include computer vision, intelligent systems, and multimedia understanding. Shang received his B.S. degree in computer science and technology from Ocean University of China, Qingdao, China. Contact him at shangheng@stu.xjtu.edu.cn.

GUOSHUAI ZHAO is an associate professor with the School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. His research interests include data mining, social user understanding, and recommender systems. Zhao received his Ph.D. degree from Xi'an Jiaotong University, Xi'an, China. Contact him at guoshuai.zhao@xjtu.edu.cn.

JING SHI is currently working toward her M.S. degree with the School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China. Her research interests include computer vision, intelligent systems, and multimedia understanding. Shi received her B.S. degree in computer science and technology from the Shanghai DianJi University, Shanghai, China. Contact her at jincy@stu.xjtu.edu.cn.

XUEMING QIAN is a full professor and director of the SMILES Laboratory at Xi'an Jiaotong University, Xi'an, 710049, China. His research interests include social media, data mining, and computer vision. Qian received his Ph.D. degree from Xi'an Jiaotong University, Xi'an, China. Contact him at qianxm@mail.xjtu.edu.cn.


NOMINATIONS
NOW OPEN
2024 IEEE COMPUTER SOCIETY/
SOFTWARE ENGINEERING INSTITUTE
WATTS S. HUMPHREY SOFTWARE QUALITY AWARD