Multimedia Tools and Applications (2024) 83:72043–72062
https://doi.org/10.1007/s11042-024-18431-5

Bridging the gap: dual perception attention and local-global similarity fusion for cross-modal image-text matching

Xiangyu Shui[1] · Zhenfang Zhu[1] · Yun Liu[1] · Hongli Pei[1] · Kefeng Li[1] · Huaxiang Zhang[2]

Received: 14 November 2023 / Revised: 8 January 2024 / Accepted: 21 January 2024 /
Published online: 5 February 2024

Abstract

Current image-text matching methods implicitly align visual-semantic segments within images, and employ cross-modal attention mechanisms to discover fine-grained cross-modal semantic correspondences. Although region-word pairs constitute local matches across modalities, they may lead to inaccurate measurements of relevance when viewed from a global perspective of image-text relationships. Additionally, cross-modal attention mechanisms may introduce redundant or irrelevant region-word alignments, which can reduce retrieval accuracy and limit efficiency. To address these challenges, we propose a **D**ual perception **A**ttention and local-global **S**imilarity **F**usion framework(DASF). Specifically, We combine two types of similarity matching, global and local, to establish a more accurate correspondence between images and text by simultaneously considering global semantics and local details during the matching process. Simultaneously, we integrate dual-perception attention mechanisms to learn the relationship between images and text, utilizing attention polarity to determine the degree of matching and better consider contextual and semantic information, thereby reducing interference from irrelevant regions. Extensive experiments on two benchmark datasets, Flickr30K and MSCOCO, demonstrate the superior effectiveness of our DASF, achieving state-of-the-art performance.

Keywords Image-text Matching · Dual perception attention · Cross-modal

1 Introduction

Image-text matching is a foundational task with widespread applications in various domains, such as image search in search engine, tagging and filtering on social media, product recognition in e-commerce, and medical image analysis. Image-text matching aligns images with textual descriptions to discover semantic similarity and establish meaningful connections between the two. Common practices involve semantically aligning vision and language,

✉ Zhenfang Zhu
zhuzf_sdjtu@126.com

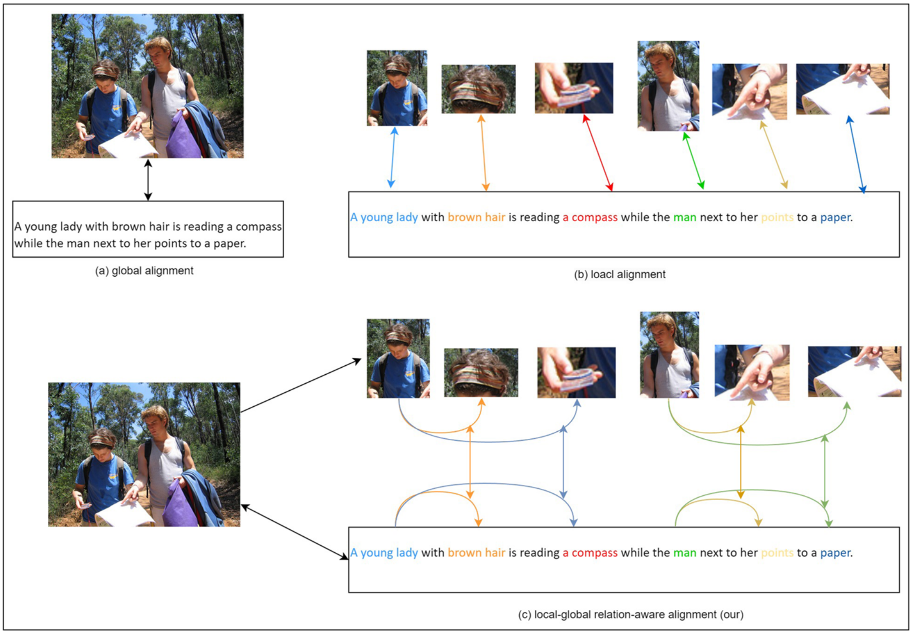
Extended author information available on the last page of the article

followed by measuring cross-modal semantic similarity as relevance based on resulting alignments.

As can bee seen from Fig. 1, alignment methods can be primarily categorized into three types: global alignment, local alignment, and global-local alignment.

Global alignment methods initially gained prominence in the field of image-text matching, aiming to capture global semantic relationships and consistency to establish comprehensive image-text correspondences and improve matching accuracy. These methods often employ deep neural networks [1–4], including traditional networks [5, 6] and recurrent neural networks [7, 8], to handle global information. However, they may overlook the local information of images and text, thereby limiting their ability for fine-grained matching. To enhance the accuracy of fine-grained matching, local alignment methods have received widespread attention. These methods [9–12] focus on the local regions of images and text to better handle cases of partial matches, ensuring effective matching even when images and text have different lengths. Recent research [13–16] has concentrated on methods for identifying region-word correspondences to enhance the details of matching. However, local alignment methods may disregard overall semantic consistency, especially when there is strong global correlation between images and text. To overcome the limitations of global and local alignment methods, global-local fusion methods [17–19] have been proposed, combining global and local matching to consider both global and local perspectives during the matching process, enhancing the flexibility and adaptability of matching.

Both global alignment methods and local alignment methods typically rely on predefined feature representations to measure the similarity between images and text. However, these feature representations may not fully capture the semantic information of images and text, leading to potential decreases in matching performance in certain scenarios. To address this issue, attention-based matching methods have been proposed [20–22]. Zhang


A young lady with brown hair is reading a compass while the man next to her points to a paper.
A young lady with brown hair is reading a compass while the man next to her points to a paper.
A young lady with brown hair is reading a compass while the man next to her points to a paper.

(a) global alignment

(b) loacl alignment

(c) local-global relation-aware alignment (our)

Fig. 1 Illustration of different feature alignment architectures

et al. [20] introduced a unified context-aware attention network that selectively focuses on critical local segments by aggregating global context. Wang et al. [21] proposed a consensus-aware visual-semantic embedding model that incorporates shared commonsense knowledge between modalities into image-text matching. Zhang et al. [22] introduced an innovative negative-aware attention framework, which explicitly considers both the positive impact of matching segments and the negative impact of mismatched segments to jointly infer the similarity between images and text. These attention-based matching methods help capture fine-grained correspondences between images and text, enhance the alignment process, and improve the performance of image-text matching tasks by selectively attending to relevant regions or words.

To address the challenges in fine-grained matching tasks, including issues related to global semantic consistency and the concentration of cross-modal attention errors, this paper introduces a Dual perception Attention and local-global Similarity Fusion framework. What sets this framework apart is its improved integration of global and local information, enabling the system to better understand the subtle yet crucial correlations between images and text. The primary innovations of this approach lie in the introduction of a Dual-Perception Attention mechanism, which enhances the precision of both global and local attention, and a novel Local-Global Similarity Fusion method, ensuring the accuracy of fine-grained matching. By applying this framework, we achieve promising results on two datasets, Flickr30K and MS-COCO, outperforming other methods and enhancing the performance and accuracy of fine-grained matching tasks. This work not only offers new ideas and methods for the fine-grained matching domain but also underscores the critical role of global and local information in cross-modal matching, serving as an inspiration for improving tasks involving the matching of images and text. The major contributions of this work are summarized as follows.

a) It combine two types of similarity matching, global and local, to establish a more accurate correspondence between images and text. By integrating both global and local information during the matching process, we can better capture the semantic relationships between the two modalities.
b) Dual perception attention mechanisms are employed to learn the relationship between images and text, determine the degree of matching and mismatch, and leverage the influence of positive and negative attention to infer image-text similarity.
c) Extensive experiments on two benchmarks, Flickr30K and MS-COCO, show that DASF outperforms compared methods. The Analyses also well demonstrate the superiority and reasonableness of the proposed method.

2 Related work

In recent years, the field of image-text matching has seen significant advancements, with research primarily falling into two categories: Firstly, global-level matching, which focuses on learning global alignment by representing the entire image or text as a holistic feature to measure their similarity; second, local-level matching, which emphasizes fine-grained alignment between local segments, inferring overall image-text similarity by analyzing the correlations between all word-region pairs. There is also a third approach that combines both global and local matching.

2.1 Global alignment methods

Global alignment methods aim to capture the semantic relationships and consistency between images and text by considering the overall information. For instance, [2, 23] proposed the global piping method, which establishes global correspondences to achieve image-text matching. Subsequent studies [24, 25] have focused on enhancing the two-stream network architecture to achieve improved alignment of global features. Recent global alignment-driven approaches, as seen in a pretrain-then-finetune paradigm [26], have demonstrated the ability to yield satisfactory outcomes. This success can be attributed to the increased scale of pre-training data. However, the performance of the aforementioned methods that solely rely on global alignment, such as the global piping method, is often limited due to the tendency of smoothing fine-grained image details in the text descriptions.

2.2 Local alignment methods

Local alignment methods establish correspondences between regions or patches in the image and words in the sentence, enabling finer-grained matching. They serve as a complementary approach to global alignment, addressing its limitations and enhancing the overall performance of image-text matching.

In image-text matching tasks, some popular approaches [9, 27] involve learning semantic alignments between image regions and text words. However, due to the semantic complexity, these approaches may not well catch the optimal fine-grained correspondences. For one thing, attending to local components selectively is a solution for searching for an optimal local alignment. Chen et al. [28] learned to associate local components with an iterative local alignment scheme. Zhang et al. [20] noticed that an object or a word might have different semantics under the different global contexts and proposed to adaptively select informative local components based on the global context for the local alignment. After that, some approaches with the same goal as the above have been successively proposed with either designing an alignment guided masking strategy [29]. Diao et al. [18] developed an attention filtration technique. For another thing, achieving the local correspondence in a comprehensive manner is also a pathway to approximate an optimal local alignment. Ji et al. [30] proposed a step-wise hierarchical alignment network that achieves the local-to-local, global-to-local and global-to-global alignments.

Other than these, there is another type of local alignment, the relation-aware local alignment that can promote fine-grained alignment. Wei et al. [31] explored the intra-modal relation for facilitating inter-modal alignment. In addition, some approaches [14, 18, 32] model the image or text data as a graph structure with the edge conveying the relation information, and infer relation-aware similarities with both the local and global alignments by the graph convolutional network. The SGR module proposed by Diao et al. [18] supports the flow of information between local and global comparisons, providing a more comprehensive understanding of interactions and enhancing similarity predictions.

2.3 Attention mechanism

Attention mechanisms play a crucial role in various tasks, including natural language processing, computer vision, and machine learning. Adopting the vanilla attention mechanism [10–12, 22, 33–35] is a trivial way to explore the semantic region/patch-word correspondences. DAN [33], for instance, introduced dual attention networks, allowing focused attention on specific regions within images and words in text across multiple stages. Lee et al. [10] employed stacked cross attention, enabling either image-to-text or text-to-image attention at any given time. Wang et al. [34] proposed cross-modal adaptive message passing, directing attention to fragments. In terms of visual relationships among regions, a recent approach [22] propose a novel Negative-Aware Attention Framework , which explicitly exploits both the positive effect of matched fragments and the negative effect of mismatched fragments to jointly infer image-text similarity.

Additionally, several recent methods have extended the widely used BERT [36] architecture to jointly learn visual and textual representations. These methods [37, 38] either employ a single-stream model to fuse textual and visual data as input or opt for a two-stream model to independently process each modality before merging them. Leveraging the self-attention module inherent to BERT, they have achieved state-of-the-art performance.

3 Methodology

In this section, we elaborate on the Dual perception Attention and local-global Similarity Fusion framework for Image-Text Matching into cross-modal relevance measurement. As illustrated in Fig. 2, The proposed DASF is composed of three modules. Firstly, the way to learn visual and textual representations and extend the semantic of detected image regions is introduced in Section 3.1. Secondly, we perform local-global similarity matching in Section 3.2. Thirdly, we perform dual perception attention to measure image-text similarity, using both negative and positive effects in Section 3.3. Finally, our the objective function for training is mentioned in Section 3.4.

Notations Formally, for image - text of(U, V), the text feature of text representation for word $U = \left\{u_i \mid i \in [1, m], u_i \in \mathbb{R}^d\right\}$, the image is expressed as regional visual characteristics $V = \left\{v_j \mid j \in [1, n], v_j \in \mathbb{R}^d\right\}$, which m and n respectively represents the number of words and area; d is the dimension of the feature representation.

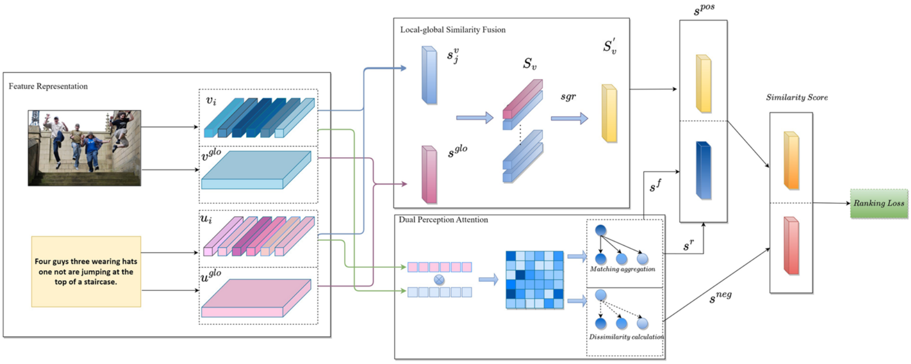


Fig. 2 Flowchart of the proposed approach. Mainly including feature representation, global-local similarity fusion, and dual perception attention effects

3.1 Feature representation

Text Representation In order to capture the nuanced interplay between vision and language, we extract text semantic information at the word level. To achieve this, we map the one-hot encoding $\{w_1, w_2, \ldots, w_m\}$ of words in the text T to distributed representations using a learnable word embedding layer, denoted as $t_i = W_e w_i$. To enrich the text representation with contextual semantics, we employ a bidirectional GRU that encodes both forward and backward information. This encoding process can be summarized as follows:

$$\overrightarrow{f_i} = \overrightarrow{GRU}\left(t_i, \overrightarrow{f_{i-1}}\right), i \in [1, m] \tag{1}$$

$$\overleftarrow{f_i} = \overleftarrow{GRU}\left(t_i, \overleftarrow{f_{i+1}}\right), i \in [1, m] \tag{2}$$

where $\overrightarrow{f_i}$ and $\overleftarrow{f_i}$ represent the hidden states from the forward and backward GRU, respectively. Moreover, the context enhanced word representation u_j is defined as the mean of bi-directional hidden states:

$$u_i = \frac{\overrightarrow{f_i} + \overleftarrow{f_i}}{2}, i \in [1, m] \tag{3}$$

The average feature of the whole text T can be expressed as: $u_{av} = \frac{1}{m}\sum_{i=1}^{m} u_i$. Moreover, under the action of the attention mechanism, u_{av} is used as the basis for the query, and u_{glo} can be encoded as:

$$u^{glo} = \frac{\sum_{i=1}^{m} w_i u_i}{\| \sum_{i=1}^{m} w_i u_i \|_2} \tag{4}$$

where the attention weight w_i is the normalized similarity between u_i and the query u_{av}.

Image representation We employ the Faster R-CNN framework [39], a deep model commonly used for object detection, along with ResNet-101, a deep convolutional neural network widely utilized for semantic segmentation and image classification, as the underlying architecture for implementing bottom-up attention. This model enables us to detect salient regions within an image I and encode their corresponding visual representations a_j. We then transform each a_j into a d-dimensional vector v_j using linear projection:

$$v_j = W_v a_j + b_j \tag{5}$$

The image I can be represented as a set of visual vectors $\left\{v_j | j = 1, 2, \ldots, n\right\}$, where n is the total number of regions in image I. Each visual vector v_j belongs to $\mathbb{R}^d$ dimensional space.

The average feature of the whole image I can be expressed as: $v_{av} = \frac{1}{n}\sum_{j=1}^{n} v_j$. Moreover, under the action of the attention mechanism, v_{av} is used as the basis for the query. Similarly, the global representation v_{glo} of the full image I is represented in the same way as u_{glo}:

$$v^{glo} = \frac{\sum_{j=1}^{n} w_j v_j}{\| \sum_{j=1}^{n} w_j v_j \|_2} \tag{6}$$

where the attention weight w_j is the normalized similarity between v_j and the query v_{av}.

3.2 Cross-modal similarity fusion

To describe the detailed correspondence between vision and language and achieve visual-semantic alignment across different modalities, we utilize a normalized distance-based representation to capture the semantic similarity between heterogeneous modalities.

Specifically, the local semantic similarity s_j^v between image region v_j and its semantically matched relevant words in the text as:

$$s_j^v = \frac{W_s^v \left| v_j - a_j^u \right|^2}{\| W_s^v \left| v_j - a_j^u \right|^2 \|_2} \tag{7}$$

$$a_j^u = \sum_{i=1}^{m} a_{ij} u_i \tag{8}$$

$$a_{ij} = \frac{e^{(\lambda \hat{e}_{ij})}}{\sum_{j=1}^{n} e^{(\lambda \hat{e}_{ij})}} \tag{9}$$

where $W_s^v \in \mathbb{R}^{k \times d}$is a learnable parameter matrix, the text context a_j^u is attended by region v_j. $\hat{c}_{ij} = \left[c_{ij}\right]_+ / \sqrt{\sum_{i=1}^{m} \left[c_{ij}\right]_+^2}$, where c_{ij} represent the cosine similarity between word u_i and region v_j . The semantic s_j^v is queried by image region v_j .

Moreover, the semantic similarity s_{glo} between whole image and full text could be measured by:

$$s_{glo} = \frac{W_s^g \left| v^{glo} - u^{glo} \right|^2}{\| W_s^g \left| v^{glo} - u^{glo} \right|^2 \|_2} \tag{10}$$

where $W_s^g \in \mathbb{R}^{k \times d}$ is a learnable parameter matrix.

During the matching process, we seek to eliminate local semantic similarities contributed by region-word pairs that are locally matched but not truly referenced in the global textual context, which can be referred to as unreliable region-word pairs. We adopt the following approach in the overall measurement of cross-modal relevance: we multiply the semantic similarity of each region query, denoted as s_n^v, by its corresponding coefficient, c_n. Consequently, we combine the global semantic similarity and the scaled local similarities. This combination effectively amalgamates global and local information, facilitating the extraction of meaningful matching characteristics.

$$S_v = concat\left[s^{glo}, c_1 s_1^v, \cdots, c_n s_n^v\right] \tag{11}$$

$$S_v^{'} = sgr(S_v) \tag{12}$$

where,sgr represents similarity graph reasoning framework.

3.3 Dual perception attention

To ensure the significance of mismatched segments is not overlooked, our approach involves a simultaneous focus on both mismatched and matched fragments within image-text pairs. This is achieved by utilizing distinct attention masks within the negative and positive attention mechanisms, allowing for a precise assessment of their impacts. To initiate this process, we

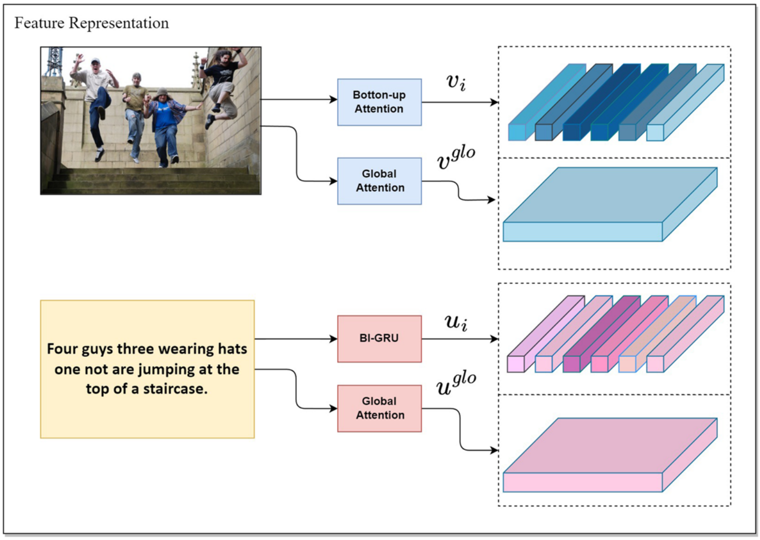


Fig. 3 Flowchart of the feature representation

begin by computing semantic relevance scores between all words and regions (Figs. 3,4 and 5).

$$s_{ij} = \frac{u_j u_i^T}{\| u_j \| \| u_i \|}, i \in [1, m], j \in [1, n] \tag{13}$$

With the aim of effectively harnessing non-matching segments to meaningfully diminish the overall similarity of mismatched image-text pairs, we identify segments within the textual modality that lack corresponding matched image regions as non-matching segments. These segments are assigned a certain level of significance in the process. We use the maximum cross-modal similarity between a segment and all segments from the other modality to reflect

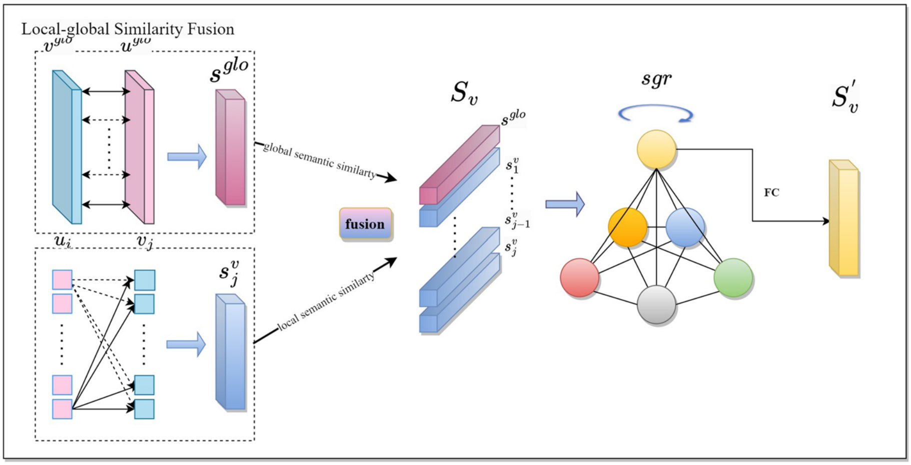


Fig. 4 Flowchart of the cross-modal similarity fusion

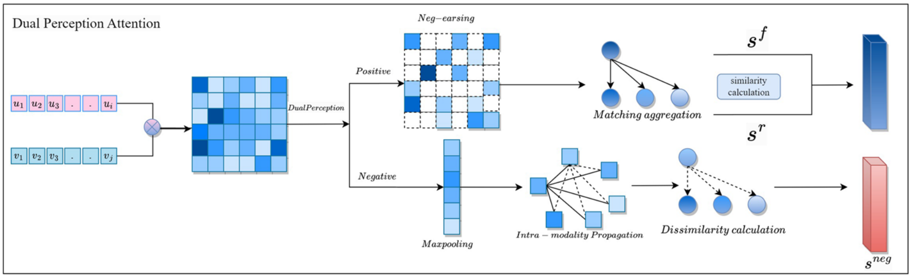

Dual Perception Attention
u_1 u_2 u_3 . . u_t
v_1 v_2 v_3 . . v_j
DualPerception
Positive
Negative
Neg−earsing
Maxpooling
Matching aggregation
similarity calculation
s^f
s^r
Intra − modality Propagation
Dissimilarity calculation
s^{neg}

Fig. 5 Flowchart of the dual perception attention

their level of match. Thus, we employ the maximum pooled similarity between each word segment u_i, where $i \in [1, m]$, and all image regions v_j, where $j \in [1, n]$.

$$s_j = \max_i \left(\left\{s_{ij} - t_k\right\}_{i=1}^{n}\right) \tag{14}$$

Therefore, the negative impact, or dissimilarity, of the i-th word in an image-text pair can be measured as follows:

$$s_i^{neg} = s_i \odot \mathrm{Mask}_{\mathrm{neg}}\left(s_i\right) \tag{15}$$

where, the function $Mask_{neg}(\cdot)$ represents a mask that equals 1 when the input is negative, and 0 otherwise. The symbol $\odot$ denotes the dot-product operation.

Given that semantically similar word segments are expected to share similar matching relationships, we account for the intra-textual semantic relationships among word segments, enhancing the accuracy of negative effect measurement. Consequently, we carry out intra-modality propagation to determine the matching degree of each word:

$$\hat{s}_j = \sum_{l=1}^{m} w_{jl}^{intra} s_l, \; s.t. \; w_{jl}^{intra} = \mathrm{softmax}_\lambda\left(\left\{\frac{u_j u_l^T}{\| u_j \| \| u_l \|}\right\}_{l=1}^{m}\right) \tag{16}$$

where, w_{jl}^{intra} signifies the semantic relationship between the i-th and l-th word segments, with λ being a scaling factor. During the inference, we replace $\hat{s}_i$ with the enhanced s_i .

We measure the similarity of image-text pairs from two perspectives. Firstly, we focus on the cross-modal shared semantics, which involves aggregating matched image regions for each query word to quantify the degree of similarity of the matched fragments.

$$w_{ji}^{\mathrm{inter}} = \mathrm{softmax}_\lambda\left(\left\{\mathrm{Mask}_{pos}\left(s_{ji} - t_k\right)\right\}_{i=1}^{n}\right) \tag{17}$$

where, w_{ji}^{inter} represents the semantic relationship between the word u_j and the image region v_i. $Mask_{pos}(\cdot)$ functions as a mask that equals the input when it's positive, and $-\infty$ otherwise. This is used to erase attention weights for unrelated image regions, specifically when the difference between s_{ij} and t_k is less than zero, making the weight effectively zero.

For the i-th word, the shared semantics corresponding to the image can be combined as: $\hat{v}_j$. Using this weighted image feature, the similarity of u_j is quantified by:

$$s_j^f = u_j \hat{v}_j^T / \left(\| u_j \| \| \hat{v}_j \|\right) \tag{18}$$

Additionally, we employ the high correlation score s_{ij} to indicate the similarity between words and regions. Simultaneously, we calculate the weighted similarity s_j^r of word u_i as using the corresponding correlation scores.

$$s_j^r = \sum_{i=1}^{n} w_{ij}^{\text{relev}} \, s_{ij} \tag{19}$$

Therefore, the positive impact of the matched fragments in the image-text pair (U, V) can be assessed as follows:

$$s^{pos} = s_j^f + s_j^r + S_v^{'} \tag{20}$$

Finally, the similarity of the image-text pair (U, V) can be comprehensively determined by the combined positive and negative effects:

$$(U, V) = \frac{1}{m} \sum_{i=1}^{m} \left(s_j^{neg} + s_j^{pos} \right) \tag{21}$$

3.4 Objective function

Following the previous method, after obtaining the final representations V and U of the image modality and the text modality respectively, this paper uses the triplet loss as the objective function to supervise the matching and learning process in the latent space. When using text as a query, we sample both matched and mismatched images in each mini-batch. The loss function tries to find the hardest negative items in the mini-batch, which form triples with the positive items and the ground truth query, and the similarity in the formed positive pairs should be a bound β higher than the similarity in the negative pairs. Similarly, when using image as a query, the selected negative samples should be texts that do not match the given query, and the formed positive and negative pairs should also satisfy the above bound β. The loss function L is defined as follows:

$$L = \sum_{(U,V)} \left[\beta - S(U, V) + S\left(U, V'\right)\right]_+ + \left[\beta - S(U, V) + S\left(U', V\right)\right]_+ \tag{22}$$

where $sim(\cdot)$ is the similarity function, and here we use cosine similarity. $I^{'}$ and $T^{'}$ are hard negative samples, $[\cdot]_+$ is equivalent to $max[\cdot, 0]$.

3.5 Ensemble and re-ranking scheme

Similar to previous approaches during the testing phase, the model encodes images and texts into visual and text feature vectors. Cosine similarity is employed to generate a similarity matrix for all test images and texts. In the ensemble approach, the similarity scores from two trained models are averaged and incorporated in the final ranking process. The obtained similarities between queries and search terms are then ranked, simplifying the retrieval of results. For the purpose of calculating the single-peak text similarity, the text-to-image reordering necessitates an additional text encoding path. However, in our experiments, we solely employ the image-to-text reordering, while maintaining the original results for text-to-image reordering.

4 Experiments

4.1 Data sets

This paper conducts performance evaluation on two publicly available datasets: Flickr30K [40] and Microsoft COCO [41]. The **Flickr30K** dataset comprises 31,783 images, with each image having 5 corresponding captions. Among these, 29,000 images are allocated to the training set, while the remaining 1,000 are divided equally between the test and validation sets. The **MSCOCO** dataset consists of 123,287 images, also paired with 5 captions each. Of these, 113,287 images are designated for training, 5,000 for validation, and an additional 5,000 for testing. Given the considerable size of the MSCOCO dataset, this paper obtains the final result by either averaging the performance over 1,000 test images five times or directly testing the entire set of 5,000 images.

4.2 Evaluation indicators

The evaluation records are captured through the computation of recall at K (R@K), where the overall recall proportion (Rsum) can be calculated using the subsequent formula:

$$\begin{aligned} Rsum = R@1 + R@5 + R@10(image\, retrieval) \\ +R@1 + R@5 + R@10(text\, retrieval) \end{aligned} \tag{23}$$

4.3 Evaluation of baseline methods

This subsection compares the DASF model with representative techniques in the same field, and the comparison does not include large pre-trained models due to the limitation of the experimental environment. We compare with three types of models: global matching methods, local matching methods and multi-level matching methods. The global models include VSE++ [2] and MTFN [42], the local area matching methods include SCAN [10], PFAN [11], GSMN [14], VSRN++ [43] ,HREM [44] and CGMN [45], and the multi-level matching methods include MDM [46], CASC [47], SGRAF [18] and NAAF [22]. We obtain the results of the comparative methods by running the source code provided in the original paper or by citing the experimental results reported in the original paper. "-" indicates that the corresponding result is not shown in the cited work. The details of the above methods are as follows:

Improving Visual-Semantic Embeddings (VSE++) [2] enhances the standard multimodal embedding loss function by incorporating hard negative samples, ranking loss, and fine-tuning with data augmentation.

Multi-modal Tensor Fusion and Re-ranking (MTFN) [42] proposed a new multi-modal tensor fusion network, which uses the tensor fusion of rank to learn image-text similarity function.

Stacked Cross Attention (SCAN) [10] is used to infer the fine-grained semantic relationship between prominent objects in images and words in sentences, find the interaction between vision and language, and infer the similarity of text and text.

Position Focused Attention Network (PFAN) [11] uses the combination of location text cues and relational attention to live valuable location features and imposes the relationship between regional images and texts.
Multi Modal Deep Matching (MDM) [46] uses the local and global representations of images and texts to combine cross-modal correlation and intra-modal similarity to investigate the matching relationship between images and texts at a deeper level.
Cross-Modal Attention With Semantic Consistence (CASC) [47] directly extracts semantic labels from the sentence corpus, makes high-level semantic words correspond to a single image region, and uses local alignment and multi-label prediction to achieve global semantic consistency.
Graph Structured Matching Network (GSMN) [14] explicitly models objects, relations and attributes as a structured phrase, learns the correspondence between objects, relations between attributes, and the fine-grained correspondence between structured phrases, and realizes node-level matching and structure-level matching.
Similarity Graph Reasoning and Attention Filtration (SGRAF) [18] adopts the similarity graph inference (SGR) module to learn the local and global relational perception similarity. Then, the similarity attention filtering model is designed to selectively focus on meaningful and representative alignments and eliminate noise interference.
Cross-modal Graph Matching Network (CGMN) [45] Spatial and non-fully connected graphs are explicitly constructed for each image, with object region features serving as graph nodes.
Visual and Textual Semantic Reasoning (VSRN++) [43] uses regional or word relation reasoning to integrate semantic relation information into visual and text features and conduct global semantic reasoning to select discriminative information and gradually increase the representation of the whole scene.
Novel Hierarchical Relation Modeling (HREM) [44] explicitly captures both fragment and instance-level relations to learn discriminative and robust cross-modal embeddings.
Negative-Aware Attention Framework (NAAF) [22] uses both matching and mismatching fragments to jointly infer image-text similarity. Using an iterative optimization method, it effectively enhances the discriminative negative effects by thoroughly exploring mismatched fragments.
Cross-modal confidence-aware network (CMCAN) [48] combines the inferred confidence levels with local semantic similarity to refine the measurement of relevance between images and text. This integration allows for more accurate relevance assessment.

4.4 Experimental details

This section provides the detailed model setup and training parameter settings of DASF in the experiments. The experiments are performed on NVIDIA A100 GPU with batch size set to 400 for MSCOCO and 330 for Flickr30K,with 16 and 20 training epoches on different datasets. The Adam optimizer is used to train the model, the learning rate is set to 0.0005 at the beginning, decaying by 0.1 every 10 epochs. The feature dimension d is set to 1,024. The scaling parameter λ is set to 20,and the margin hyperparameter γ is selected as 0.2.

4.5 Results on flickr30K

The outcomes reported in Table 1 for the Flickr30K dataset highlight the superiority of DASF over other methods. The bold entries indicate the best results in the current indicator, and

Table 1 Results on Flick30K

Methods	Image-To-Text			Text-To-Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
VSE++ [2]	52.9	80.5	87.2	39.6	70.1	79.5	409.8
MTFN [42]	65.3	88.3	93.3	52.0	80.1	86.1	465.1
SCAN [10]	67.9	89.0	94.4	43.9	74.2	82.8	452.2
PFAN [11]	67.6	90.0	93.8	45.7	74.7	83.6	455.4
MDM [46]	44.9	75.4	84.4	34.4	67.0	77.7	384.0
CASC [47]	68.5	90.6	95.9	50.2	78.3	86.3	469.8
GSMN [14]	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SGRAF [18]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
CGMN [45]	77.9	93.8	96.8	59.9	85.1	90.6	504.1
VSRN++ [43]	79.2	94.6	97.5	60.6	**85.6**	**91.4**	508.9
CMCAN [48]	79.5	95.6	97.6	60.9	84.3	89.9	507.8
HREM [44]	81.4	96.5	**98.5**	60.9	85.6	91.3	514.3
NAAF [22]	81.9	96.1	98.3	61.0	85.3	90.6	513.2
DASF(ours)	**83.6**	**96.6**	98.3	**61.2**	84.7	**91.4**	**515.8**

the same meaning applies to the following table. Following the utilization of the re-ranking scheme, the model proposed in this paper achieves the most favorable results across various indicators. Based on the experimental findings, DASF secures high scores of 83.6, 96.6, and 98.3 for R@1, R@5, and R@10 in text retrieval, respectively. For image retrieval, the corresponding scores are 61.2, 84.7, and 91.4. The model only falls short in R@5 for image retrieval.

Table 2 Results on MSCOCO 1K

Methods	Image-To-Text			Text-To-Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
VSE++ [2]	64.6	90.0	95.7	52.0	84.3	92.0	478.6
MTFN [42]	74.3	94.9	97.9	60.1	89.1	95.0	511.3
SCAN [10]	70.9	94.5	97.8	56.4	87.0	93.9	500.5
PFAN [11]	75.8	95.9	99.0	61.0	89.1	95.1	515.9
MDM [46]	54.7	84.1	91.9	44.6	79.6	90.5	445.4
CASC [47]	72.3	96.0	99.0	58.9	89.8	96.0	512.0
GSMN [14]	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SGRAF [18]	79.6	96.2	98.5	63.2	90.7	96.1	524.5
CGMN [45]	76.8	95.4	98.3	63.8	90.7	97.5	520.7
VSRN++ [43]	77.9	96.0	98.5	64.1	91.0	96.1	523.6
CMCAN [48]	81.2	96.8	98.7	**65.4**	91.0	96.2	529.3
HREM [44]	81.2	96.5	**98.9**	63.7	90.7	96.0	527.1
NAAF [22]	80.5	96.5	98.5	64.1	90.7	96.5	527.2
DASF(ours)	**82.8**	**97.1**	98.8	65.0	**91.05**	**96.9**	**531.6**

4.6 Results on MSCOCO

According to the results on MSCOCO shown in Tables 2 and 3, We can see that, DASF performs the best,in most cases. Text retrieval on the 1K test set achieves 82.8 in R@1, which exceeds VSRN++ 4.9, CMCAN 1.6, and NAAF 2.3 respectively. It reaches 97.1 and 98.8 in R@5 and R@10, respectively, which is also superior to the above models. For image retrieval, the three metrics are 65.0, 91.05 and 96.9 respectively. On the 5K test set, we obtain similar conclusions. On the 5K test set, except for R@1 in both text retrieval and image retrieval, where it falls slightly behind CMCAN, it achieves the best results in all other metrics.

4.7 Time complexity

In score-based methods, cross-modal interaction evaluates the relationship between queries and samples by calculating similarity scores. The time complexity of this approach is $O(N^2)$. Specifically, assuming there is one query and a set of N samples, the time complexity of score-based query retrieval is $O(N)$. This implies that as the sample set increases, retrieval time grows quadratically. In contrast, the time complexity of embedding-based query retrieval is $O(1)$, meaning that even with a large sample set, retrieval time remains at a constant level.

Therefore, to enhance performance, score-based methods sacrifice retrieval speed when assessing the relationship between queries and samples. This trade-off means that when dealing with large-scale datasets, more time may be required for retrieval, but more accurate similarity assessments can be obtained. This is a common trade-off in cross-modal interactions. Our approach aims primarily at achieving high accuracy, thus sacrificing some retrieval time. As shown in Fig. 6, we present a comparison of recent retrieval methods in terms of retrieval accuracy and retrieval speed. We perform all methods on the whole Flickr30K test set.

Table 3 Results on MSCOCO 5K

Methods	Image-To-Text			Text-To-Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
VSE++ [2]	41.3	71.1	81.2	30.3	59.4	72.4	355.7
MTFN [42]	48.3	77.6	87.3	35.9	66.1	76.1	391.3
SCAN [10]	46.4	77.4	87.2	34.4	63.7	75.7	384.0
PFAN [11]	-	-	-	-	-	-	-
MDM [46]	-	-	-	-	-	-	-
CASC [47]	47.2	78.3	87.4	34.7	64.8	76.8	389.2
GSMN [14]	-	-	-	-	-	-	-
SGRAF [18]	57.8	-	91.6	41.9	-	81.3	-
CGMN [45]	53.4	81.3	89.6	41.2	71.9	82.4	419.8
VSRN++ [43]	54.7	82.9	90.9	42.0	72.2	82.7	425.4
CMCAN [48]	**61.5**	-	92.9	**44.0**	-	82.6	-
HREM [44]	-	-	-	-	-	-	-
NAAF [22]	58.9	85.2	92.0	42.5	70.9	81.4	430.9
DASF(ours)	58.2	**85.9**	**92.9**	41.8	**72.6**	**82.9**	**434.3**

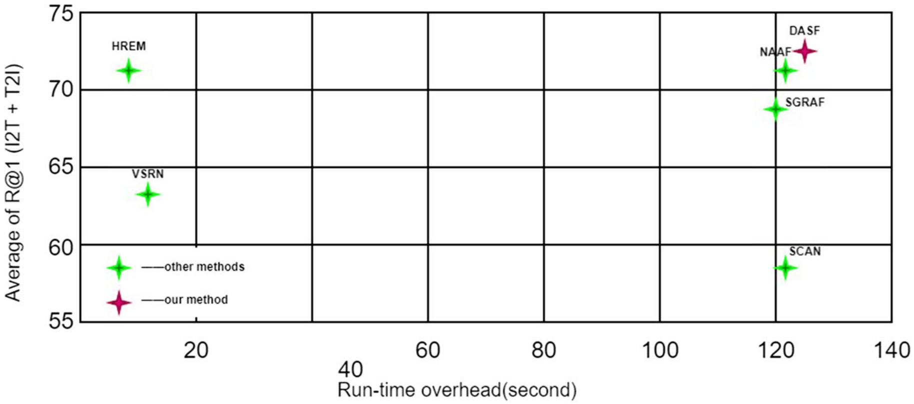


Fig. 6 The comparison between accuracy and speed for cross-modal retrieval

4.8 Ablation experiments

In order to demonstrate the effectiveness of the DASF model in cross-modal matching, we have provided a summary of the impact of key components within the model on the overall framework in Table 4. We conducted experiments in two directions: image retrieval and text retrieval, on the Flickr30K dataset. In this context, DASF-Full represents the average performance of both models, while the remaining entries correspond to individual model tests.

- ***DASF-no-sgr***: removes similarity graph reasoning in the whole framework.
- ***DASF-no-s$_{glo}$***: removes global matching structure in the whole framework.
- ***DASF-no-s$_j^v$***: removes local matching structure in the whole framework.
- ***DASF-only-S$_v$***: Keep only the global-local similarity fusion and train without the dual perceptual attention part
- ***DASF-full***: Keep all components identical to the original model, and it represents the average performance of the two models.

The comparative results for variants of DASF are shown in Table 4. From the results, we find that our method outperforms the other four variants on the Flickr30K dataset. Specifically, the largest drop in performance occurs when dual perceptual attention is removed, and only local-global similarity is used. Using only local or global matching also results in a noticeable decrease in performance. The impact of similarity graph reasoning is relatively minimal compared to other model variants but still demonstrates a discernible decrease in performance. These results validate the effectiveness of our architecture.

Table 4 Results of ablation experiments

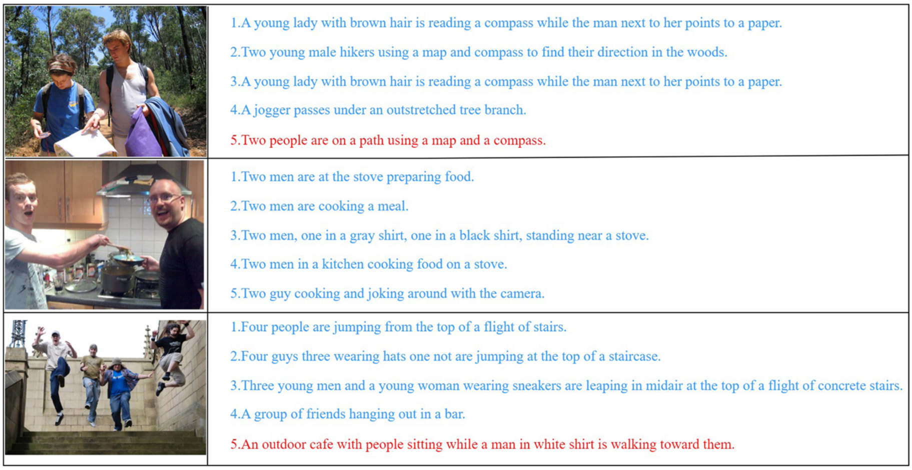
Methods	Image-To-Text			Text-To-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
$DASF w/osgr$	79.8	94.1	97.8	58.4	82.2	89.8
$DASF w/os_{glo}$	78.8	93.9	96.9	58.1	81.6	88.8
$DASF w/os_j^v$	72.2	90.5	93.5	53.8	78.2	85.6
$DASF\text{-}only\text{-}S_v$	70.8	87.6	90.7	52.3	76.8	83.6
DASF-full	**83.6**	**96.6**	**98.3**	**61.2**	**84.7**	**91.4**

4.8.1 Analysis of ablation results

1. When the sgr module is removed from the DASF, the model's performance exhibits a noticeable decline due to the absence of continuous updates from the nodes, which results in the loss of inference related to local-global relationship-aware similarity. This observation reaffirms the effectiveness of similarity graph reasoning.
2. The removal of the global matching structure results in insufficient consideration of global information between the entire image and text, leading to a decrease in semantic consistency and the inability of the model to achieve high performance. This observation confirms the significance of global matching for this model.
3. Neglecting local matching significantly impacts the model's performance due to inadequate inference of word-region correlations. Experimental results indicate that the performance decline is more pronounced when the local matching structure is removed compared to the removal of the global matching structure. This suggests that in this model, local matching takes precedence over global matching.
4. When the model removes the dual-perception network, neglecting negative perception attention leads to the model's inattention to non-matching segments' influence. On the other hand, ignoring positive perception attention results in a reduced ability of the model to focus on important regions. Experimental results confirm the effectiveness of the dual-perception network in this model.

4.9 Case study

Figures 7 and 8 illustrates the results of DASF on Flickr30K dataset for image retrieval and text retrieval, respectively. In the three image retrieval text tasks, the first four matching results of each task are correct.It is evident that the DASF model, when dealing with complex scenes, significantly enhances its performance by effectively distinguishing and utilizing matching and non-matching segments in complex scene matching, thanks to the support of



1.A young lady with brown hair is reading a compass while the man next to her points to a paper.
2.Two young male hikers using a map and compass to find their direction in the woods.
3.A young lady with brown hair is reading a compass while the man next to her points to a paper.
4.A jogger passes under an outstretched tree branch.
5.Two people are on a path using a map and a compass.

1.Two men are at the stove preparing food.
2.Two men are cooking a meal.
3.Two men, one in a gray shirt, one in a black shirt, standing near a stove.
4.Two men in a kitchen cooking food on a stove.
5.Two guy cooking and joking around with the camera.

1.Four people are jumping from the top of a flight of stairs.
2.Four guys three wearing hats one not are jumping at the top of a staircase.
3.Three young men and a young woman wearing sneakers are leaping in midair at the top of a flight of concrete stairs.
4.A group of friends hanging out in a bar.
5.An outdoor cafe with people sitting while a man in white shirt is walking toward them.

Fig. 7 The image retrieval task demonstration. The top five retrieval results are sorted in descending order. The figure above shows image retrieval text tasks, and the correct results are marked in blue font, and the incorrect results are marked in red font

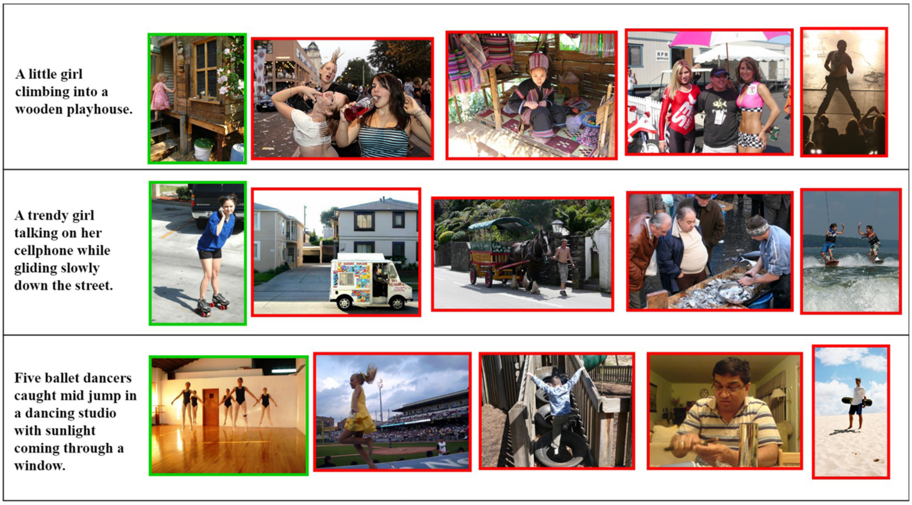

A little girl climbing into a wooden playhouse.
A trendy girl talking on her cellphone while gliding slowly down the street.
Five ballet dancers caught mid jump in a dancing studio with sunlight coming through a window.

Fig. 8 The text retrieval task demonstration. The top five retrieval results are sorted in descending order. The above figure shows the text retrieval image tasks, and the image with green border are the correct retrieval results

dual perception attention during the matching process. From the results of three text retrieval image tasks, it can be seen that out of the first five matching results, all obtained the correct results in the first position. It can be observed that in cases where region features are partially similar, the model's simultaneous consideration of both global and local matching, along with the integration of global and local information during the matching process, ensures the completeness of information. Consequently, the model can obtain more accurate answers, even in similar scenes, with higher priority given to the correct answers compared to other incorrect ones.

5 Conclusion

In this paper, we propose a Dual perception Attention and local-global Similarity Fusion framework. We combine two types of similarity matching, global and local, to establish a more accurate correspondence between images and text by considering both global semantics and local details during the matching process. Simultaneously, we integrate dual perception attention mechanisms to learn the relationship between images and text, determining the degree of matching and better considering contextual and semantic information. Experiments show that our model outperforms previous methods on the image-text matching task on two widely used datasets MSCOCO and Flickr30K. Ablation experiments also demonstrate the effectiveness of each individual modules of the model, as well as the effectiveness of the model as a whole. In the future, we look forward to applying this framework learning to more cross-modal tasks, such as image captioning and visual question answering.

Acknowledgements The work is partially supported by the National Social Science Fund of China (19BYY076).

Data Availibility Statement The datasets used and analyzed during the current study are all public datasets, and the Flicker30k can be found in the http://shannon.cs.illinois.edu/DenotationGraph/data/index.html. The MSCOCO can be found in the https://cocodataset.org/.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Wang B, Yang Y , Xu X , Hanjalic A , Shen HT (2017) Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on multimedia, pp 154–162
2. Faghri F , Fleet DJ , Kiros JR , Fidler S (2017) Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612
3. Dutton B (2020) Adversarial canonical correlation analysis. arXiv preprint arXiv:2005.10349
4. Kiros R , Salakhutdinov R , Zemel RS (2014) Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539
5. Simonyan K , Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
6. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
7. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
8. Cho K , Van Merriënboer B , Gulcehre C , Bahdanau D , Bougares F , Schwenk H , Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
9. Ji Z , Wang H , Han J , Pang Y (2019) Saliency-guided attention network for image-sentence matching. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5754–5763
10. Lee K-H , Chen X , Hua G , Hu H , He X (2018) Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 201–216
11. Wang Y , Yang H , Qian X , Ma L , Lu J , Li B , Fan X (2019) Position focused attention network for image-text matching. arXiv preprint arXiv:1907.09748
12. Liu C , Mao Z , Liu A-A , Zhang T , Wang B , Zhang Y (2019) Focus your attention: A bidirectional focal attention network for image-text matching. In: Proceedings of the 27th ACM international conference on multimedia, pp 3–11
13. Ge X , Chen F , Xu S , Tao F , Jose JM (2023) Cross-modal semantic enhanced interaction for image-sentence retrieval. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1022–1031
14. Liu C , Mao Z , Zhang T , Xie H , Wang B , Zhang Y (2020) Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10921–10930
15. Li Z , Guo C , Feng Z , Hwang J-N , Xue X (2022) Multi-view visual semantic embedding. In: IJCAI, vol 2, p 7
16. Pan Z , Wu F , Zhang B (2023) Fine-grained image-text matching by cross-modal hard aligning network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19275–19284
17. Li Z, Ling F, Zhang C, Ma H (2020) Combining global and local similarity for cross-media retrieval. IEEE Access 8:21847–21856
18. Diao H, Zhang Y, Ma L, Lu H (2021) Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence vol 35, pp 1218–1226
19. Wen K, Gu X, Cheng Q (2020) Learning dual semantic relations with graph attention for image-text matching. IEEE Trans Circuits Syst Video Technol 31(7):2866–2879
20. Zhang Q , Lei Z , Zhang Z , Li SZ (2020) t-aware attention network for image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3536–3545

21. Wang H , Zhang Y , Ji Z , Pang Y , Ma L (2020) Consensus-aware visual-semantic embedding for image-text matching. In: Computer Vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16, pp 18–34. Springer
22. Zhang K , Mao Z , Wang Q , Zhang Y (2022) Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15661–15670
23. Wang L , Li Y , Lazebnik S (2016) Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
24. Sarafianos N , Xu X , Kakadiaris IA (2019) Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5814–5824
25. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen Y-D (2020) Dual-path convolutional image-text embeddings with instance loss. ACM Trans Multimed Comput Commun Appl (TOMM) 16(2):1–23
26. Jia C , Yang Y , Xia Y , Chen Y-T , Parekh Z , Pham H , Le Q , Sung Y-H , Li Z , Duerig T (2021) Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning, pp 4904–4916. PMLR
27. Huang Y , Wu Q , Song C , Wang L (2018) Learning semantic concepts and order for image and sentence matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6163–6171
28. Chen H , Ding G , Liu X , Lin Z , Liu J , Han J (2020) Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12655– 12663
29. Zhuge M , Gao D , Fan D-P , Jin L , Chen B , Zhou H , Qiu M , Shao L (2021) Kaleido-bert: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12647– 12657
30. Ji Z , Chen K , Wang H (2020) Step-wise hierarchical alignment network for image-text matching. arXiv preprint arXiv:2106.06509
31. Wei X , Zhang T , Li Y , Zhang Y , Wu F (2020) Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10941– 10950
32. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10313– 10322
33. Nam H , Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 299– 307
34. Wang L, Li Y, Huang J, Lazebnik S (2018) Learning two-branch neural networks for image-text matching tasks. IEEE Trans Pattern Anal Mach Intell 41(2):394–407
35. Zhu Z, Zhang D, Li L, Li K, Qi J, Wang W, Zhang G, Liu P (2023) Knowledge-guided multi-granularity gcn for absa. Inform Process Manag 60(2):103223
36. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
37. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Adv Neural Inform Process Syst 32
38. Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F , Gan Z , Cheng Y , Liu J (2020) Uniter: Universal image-text representation learning. In: European conference on computer vision, pp 104–120. Springer
39. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inform Process Syst 28
40. Plummer BA , Wang L , Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S (2015) Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641– 2649
41. Lin T-Y, Maire M, Belongie S , Hays J , Perona P , Ramanan D , Dollár P , Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, pp 740– 755. Springer
42. Wang T , Xu X , Yang Y , Hanjalic A , Shen HT , Song J (2019) Matching images and text with multi-modal tensor fusion and re-ranking. In: Proceedings of the 27th ACM international conference on multimedia, pp 12– 20
43. Li K , Zhang Y , Li K , Li Y , Fu Y (2022) Image-text embedding learning via visual and textual semantic reasoning. IEEE Trans Pattern Anal Mach Intell
44. Fu Z , Mao Z , Song Y , Zhang Y (2023) Learning semantic relationship among instances for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 15159– 15168
45. Cheng Y, Zhu X, Qian J, Wen F, Liu P (2022) Cross-modal graph matching network for image-text retrieval. ACM Trans Multimed Comput Commun Appl (TOMM) 18(4):1–23

46. Ma L, Jiang W, Jie Z, Wang X (2019) Bidirectional image-sentence retrieval by local and global deep matching. Neurocomputing 345:36–44
47. Xu X, Wang T, Yang Y, Zuo L, Shen F, Shen HT (2020) Cross-modal attention with semantic consistence for image-text matching. IEEE Trans Neural Netw Learn Syst 31(12):5412–5425
48. Zhang H, Mao Z, Zhang K, Zhang Y (2022) Show your faith: Cross-modal confidence-aware network for image-text matching. In: Proceedings of the AAAI Conference on Artificial Intelligence vol 36, pp 3262–3270

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

Authors and Affiliations

Xiangyu Shui[1] · Zhenfang Zhu[1] · Yun Liu[1] · Hongli Pei[1] · Kefeng Li[1] · Huaxiang Zhang[2]

Xiangyu Shui
xiangyuuuus@gmail.com

Yun Liu
YunnLiu1120@hotmail.com

Hongli Pei
peihongli@sdjtu.edu.cn

Kefeng Li
205073@sdjtu.edu.cn

Huaxiang Zhang
Huaxzhang@163.com

[1] Shandong Jiaotong University, School of Information Science and Electrical Engineering, Jinan 250357, Shandong Province, China

[2] Shandong Normal University, School of Information Science and Engineering, Jinan 250358, Shandong Province, China