Applied Soft Computing Journal 154 (2024) 111395

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Perceive, Reason, and Align: Context-guided cross-modal correlation learning for image–text retrieval

Zheng Liu [a,b,*], Xinlei Pei [a,b], Shanshan Gao [a,b,c], Changhao Li [a,b], Jingyao Wang [a,b], Junhao Xu [a,d]

[a] School of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan, 250014, Shandong, China
[b] Shandong Provincial Key Laboratory of Digital Media Technology, Ji'nan, 250014, Shandong, China
[c] Shandong China-U.S. Digital Media International Cooperation Research Center, Ji'nan, 250014, Shandong, China
[d] School of Computer Science and Technology, Shandong University, Qingdao, 266237, Shandong, China

ARTICLE INFO

Keywords:
Image–text retrieval
Context-perceived embedding
Inter-modal correlation
Intra-modal correlation
Hybrid loss

ABSTRACT

Due to the inconsistency in feature representations between different modalities, namely "Heterogeneous gap", it remains a persistent challenge to correlate images and texts. Existing studies on image–text retrieval (ITR) mainly emphasize on inter-modal correlation learning through aligning instances or their patches from different modalities. However, it is hard to break through performance bottlenecks of ITR without powerfully supporting from intra-modal correlation. Unfortunately, few studies have sufficiently considered two critical tasks in intra-modal correlation learning: (1) intricate contextual information perceiving, and (2) intrinsic semantic relationships reasoning. Therefore, in this paper, we propose the Context-guided Cross-modal Correlation Learning (CCCL) framework for ITR under a novel paradigm: "Perceive, Reason, and Align". Firstly, in the stage of "Perceive", the context-guided mechanism based on the self-attention and gate mechanism is proposed to fully discover contextual information within modalities, eliminating unnecessary interactions between local-level patches. Secondly, in the stage of "Reason", graph convolutional network with the residual structure is used to uncover relationships among patches within each modality to make reasonable inferences. Thirdly, in the stage of "Align", to achieve precise inter-modal alignment, the complementarity between different modalities from both global-level and local-level is effectively mined and fused. Finally, to optimize our proposed CCCL framework, the hybrid loss is constructed by combining the cross-modal coherence term with the cross-modal alignment term. Our approach yields highly competitive results on two publicly available ITR datasets, that is, Flickr30K and MS-COCO.

1. Introduction

In recent years, with the explosive growth of multimedia data on the internet, there is an increasing requirement for efficient and accurate ways to retrieve information from such data [1]. To enable more efficient and effective information retrieval, cross-modal retrieval is proposed to search for relevant data across different modalities [2]. The significance of cross-modal retrieval research lies in its ability to bridge the semantic gap between different modalities [3,4]. As vision and language are two important media for human beings to understand the real world, ITR is an important domain in cross-modal retrieval [5,6]. Researchers have conducted extensive studies to connect visual modality and language modality [7–11]. ITR aims to search and find images that are relevant to a given text query or retrieve texts that are relevant to a given image query, and it has become an important research area in computer vision and natural language processing [12,13].

The difficulty faced by ITR is the "Heterogeneous gap" [1,14], which refers to the inconsistent feature representations between different modalities, such as images and texts. Note that earlier research of ITR mainly focused on mapping image and text features at the global-level to a common embedding space [15–17]. However, this approach primarily focused on global-level instances and overlooked the detailed semantic information present in local-level patches. To address this limitation and capture a more complete understanding of the semantics, recent studies have shifted their focus to local-level patches [18,19]. To bridge the "Heterogeneous gap" between images and texts, it is of paramount importance to thoroughly uncover the correlations among instances and patches, whether they exist within or between modalities. Furthermore, these correlations should be seamlessly integrated within a unified retrieval framework.

* Corresponding author at: School of Computer Science and Technology, Shandong University of Finance and Economics, Ji'nan, 250014, Shandong, China.
E-mail addresses: liuzheng@sdufe.edu.cn (Z. Liu), peixinlei@mail.sdufe.edu.cn (X. Pei), gss_sdufe@sdufe.edu.cn (S. Gao), lichanghao99@mail.sdufe.edu.cn (C. Li), wangjingyao@mail.sdufe.edu.cn (J. Wang), 202018140235@mail.sdufe.edu.cn (J. Xu).

https://doi.org/10.1016/j.asoc.2024.111395
Received 17 August 2023; Received in revised form 30 December 2023; Accepted 9 February 2024
Available online 14 February 2024

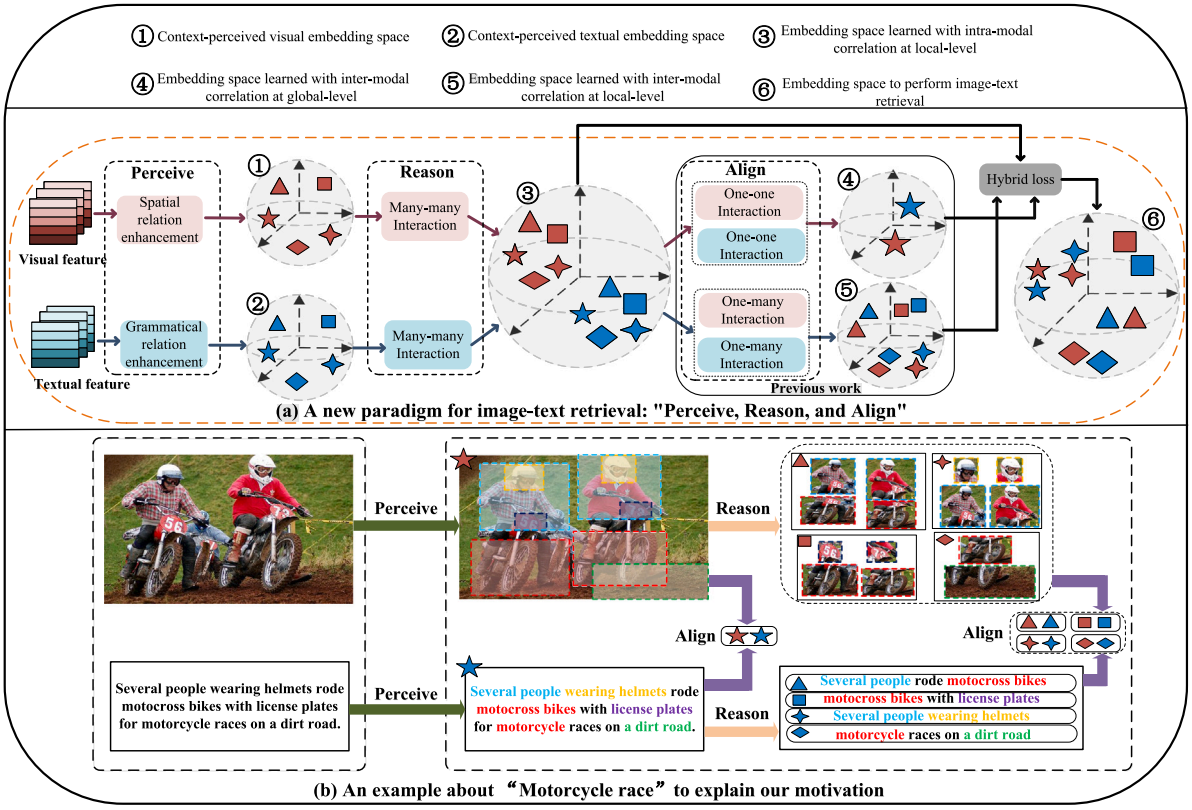


Fig. 1. Illustration of our motivation.

In general, cross-modal correlation learning is the core task of ITR, and it consists of two challenges: (1) **Challenge 1**: Inter-modal correlation learning, and (2) **Challenge 2**: Intra-modal correlation learning. To tackle challenge 1, instances or patches from different modalities should be precisely aligned to capture the pairwise correlations, namely "Align". To address challenge 2, within a single modality, contextual information and semantic relationships existing in patches of image and text should be deeply perceived and accurately reasoned, respectively, namely "Perceive" and "Reason". We propose a novel ITR paradigm - "Perceive, Reason, and Align". More specifically, "Perceive" refers to the perception of intricate contextual information by capturing spatial relations and grammatical relations. "Reason" represents the inference of intrinsic semantic relationships through many-to-many interactions among local-level elements within one modality. "Align" involves the discovery of pairwise correlations at both global and local levels through one-to-one and one-to-many interactions across different modalities. In general, "Perceive" and "Reason" contribute to the accurate estimation of intra-modal correlation, significantly enhancing the "Align" process and resulting in a more precise determination of inter-modal correlation. Our investigation shows that most of existing studies on ITR primarily focus on the "Align" process in challenge 1 [20–22]. However, the "Perceive" and "Reason" processes in challenge 2 are not currently receiving adequate attention, and have not yet been solved well. Therefore, "Perceive" and "Reason" are two critical issues in ITR that urgently require resolution:

- **Problem 1 (Perceive)**: The intricate contextual information from both images and texts has not been fully perceived.
  Contextual information in images refers to the position of image patches, while in texts, it pertains to word sequences and grammatical relations. Without the guidance of contextual information, both inter-modal and intra-modal interactions cannot be comprehensively captured.
- **Problem 2 (Reason)**: The intrinsic semantic relationships within local-level patches have not been effectively reasoned.
  Patches of images are obtained through uniform blocking or salient object detection, while patches of texts can be generated by segmenting a text into several sentences or words. Without reasoning out the intrinsic semantic relationships between local-level patches, sophisticated semantic relationships cannot be completely caught.

In summary, "Align" has been widely used in existing studies, while "Perceive" and "Reason" have not yet been given sufficient attention. Therefore, to simultaneously address challenge 1 and challenge 2, we present a novel paradigm for ITR, that is, "**Perceive**, **Reason**, and **Align**". Our proposed ITR paradigm offers an efficient approach to bridge the "Heterogeneous gap". This effectiveness stems from the following two factors: (1) Perceiving intricate contextual information yields more valuable cues for precise reasoning out intrinsic semantic relationships, and (2) Achieving alignment between images and texts heavily depends on accurately estimating intra-modal correlation, which is drawn from "Perception" and "Reasoning". Thus, we propose a **C**ontext-guided **C**ross-modal **C**orrelation **L**earning (CCCL) framework to enhance the performance of ITR.

To illustrate the motivation of CCCL, we provide an image–text pair describing the semantic concept "motorcycle race" to illustrate our proposed paradigm for ITR in Fig. 1, in which semantically related global-level instances and local-level patches are accurately aligned, greatly benefiting from perceiving contextual information and reasoning semantic relationships. Particularly, Fig. 1 also indicates that CCCL correctly aligns the red and the blue triangle. In addition, greatly benefiting from the "Perceive" and "Reason" operations, four image patches corresponding to the red triangles clearly illustrate the visual contents of two people riding motocross bikes, and the sentence "Several people rode motocross bikes" marked as the blue triangle achieves a complete semantic description.

The main contributions of this paper are as follows:

- **The cross-modal correlation learning framework** is proposed to simultaneously achieve two goals: (1) learning intra-modal correlation based on perceiving intricate contextual information and reasoning intrinsic semantic relationships via many-many interaction within one modality, and (2) learning inter-modal correlation based on aligning instances (patches) via one-one interaction (one-many interaction) across different modalities.
- **The context-guided mechanism** is introduced to adaptively learn context-perceived visual embedding and context-perceived textual embedding, which are utilized to guide cross-modal correlation learning. By employing self-attention and gate mechanism, the context-perceived cell is designed to capture contextual information by eliminating unnecessary interactions between local-level patches.
- **The hybrid loss**, which includes the cross-modal coherence term along with the cross-modal alignment term, is proposed to optimize our CCCL framework. The former term is utilized to mitigate the disparities in image–text similarity learned by CCCL. Meanwhile, the latter term employs the bidirectional triplet loss with a hard sample mining strategy to optimize the fused image–text similarity and ensure precise alignment between modalities.

The rest of this paper is organized as follows. In Section 2, we briefly summarize some prior works on ITR. In Section 3, we present the proposed CCCL framework. In addition, experimental results and analyses are discussed in Section 4, where we perform a quantitative comparison of CCCL framework with previous models. Finally, we conclude this work and point out our future works in Section 5.

2. Related works

In this section, we will summarize the latest advancements in ITR from two perspectives: (1) Correlation learning, and (2) Correlation optimization.

2.1. Correlation learning for ITR

According to the strategies for establishing the correlation between images and texts, existing methods for correlation learning in ITR can be categorized into three types: (1) Inter-modal correlation learning, (2) Intra-modal correlation learning, and (3) Fusion of inter-modal and intra-modal correlation learning.

2.1.1. Inter-modal correlation learning

Inter-modal correlation learning aims to explore the complex interaction relationships between different modalities. In ITR, inter-modal correlation learning plays a crucial role in improving the retrieval performance.

Faghri et al. proposed VSE++ to integrate hard negative mining technology into the ranking loss to improve the retrieval performance and training speed [15]. Karpathy et al. embedded image regions and words into the public space for alignment [22]. Nam et al. extracted image and text features through the attention mechanism and aggregate multiple local similarities to calculate the final result [23]. Lee et al. designed the stacked cross-attention mechanism (SCAN) to design two embedding spaces for obtaining all possible potential alignments, which is a pioneering work in cross-modal retrieval [20]. Chen et al. constructed a visual-based space and a textual-based space, then used a semantic consistency constraint to learn these two spaces simultaneously [24].

To adaptively regulate the degree to which the information from the other modality is fused with the original features, Wang et al. suggested the Cross-modal Adaptive Message Passing (CAMP) model, which combines the Cross-modal Message Aggregation and Cross-modal Gated Fusion modules [13]. Chen et al. proposed a recurrent attention memory iteration-based image–text matching model (IMRAM) with a memory distillation unit to gradually explore the complex interaction relationships of the inter-modality [12]. Furthermore, Zhang et al. proposed a negative-aware attention framework (NAAF), which explicitly utilizes the positive effects of image–text correct matching and the negative effects of mismatching to learn the cross-modal similarity [25].

In summary, inter-modal correlation learning is an important research direction in multi-modal systems, and various techniques have been proposed to explore and leverage the complex interaction relationships between different modalities in ITR tasks. Nonetheless, only focusing on learning inter-modal correlation fails to capture the important semantic elements existing in fine-grained data within each data modality. This limitation results in a decline in retrieval performance.

2.1.2. Intra-modal correlation learning

Intra-modal correlation learning strives to mine the complex interaction relationships within the same modality. In ITR, intra-modal correlation learning can effectively capture fine-grained clues and subsequently enhance the representation of the entire instance.

Wang et al. proposed a position focused attention network (PFAN) to integrate the spatial properties of visual regions and enhance the position perception ability of image features [26]. Gu et al. combined the two generative frameworks into the traditional text–visual feature embedding to calculate the detailed similarity between the two modalities [27]. Wu et al. proposed SAEM with self-attention layers to capture fragment relationships in images or texts [28]. These layers comprise sub-layers of multi-head self-attention and position-wise feed-forward networks, which amalgamate fragment information into embeddings for visual and textual content. Chen et al. [29] proposed a framework VSE∞ with a generalized pooling operator to automatically determine the optimal strategy for various features, achieving comparable performance to more complex cross-modal interaction models. Li et al. [30] developed an effective image–text embedding network (VSRN++), which uses semantic relationship information to enhance image and text features through graph convolution. It performs global semantic reasoning and progressively refines the representation of the entire instance by selecting discriminative information.

To sum up, existing methods involving intra-modal correlation learning can conduct a thorough analysis of complex interaction relationships within the same modality and effectively capture fine-grained clues to enhance the performance of ITR. Particularly, there are two main problems in existing intra-modal correlation learning based ITR: (1) Although some types of contextual information have been exploited, the intra-modal information flow cannot be adaptively weighted and uninformative interactions cannot be effectively suppressed as well, and (2) For the text modality, the word sequences are commonly used as the context, however, this strategy is too simple. Grammatical dependency between words is often overlooked.

2.1.3. Fusion of inter-modal and intra-modal correlation learning

The fusion of inter-modal and intra-modal correlation learning enables the integration of different types of data sources and thus improves the accuracy and robustness of ITR.

Wang et al. proposed the SGM model to represent images and text as scene maps and match object and relationship nodes in different scene graphs [31]. While this approach leverages inter-modal correlations to facilitate matching, it may not capture fine-grained intra-modal relationships between objects and properties. Liu et al. explicitly modeled intra-modal objects, relationships, and properties as structured phrases and learned fine-grained alignment through node-level matching (GSMN) [32]. GSMN jointly infers the correspondence of different modal structured phrases through structure-level matching and fusion of neighborhood information and leverages both intra-modal and inter-modal correlations to improve matching accuracy. Zhang et al. proposed the Context-aware attention network (CAAN),

which can aggregate global inter- and intra-modal interaction to capture latent semantic relations and selectively focus on important local fragments [33]. Furthermore, CAAN leverages both inter-modal and intra-modal correlations to capture more comprehensive cross-modal interaction. Wei et al. [34] proposed a Transformer-based multi-modality cross-attention network (MMCA) for jointly modeling intra- and inter-modality relationships. Diao et al. proposed the Similarity Graph Reasoning and Attention Filtration (SGRAF) network for image–text matching [35]. SGRAF acquires vector-based similarity representations for delineating local and global alignments and introduces the SGR module, employing a graph convolutional neural network, to deduce relation-aware similarities encompassing both alignment types.

Fusion of inter-modal and intra-modal correlation learning can obtain better performance than any one of them. However, previous studies have not effectively utilized contextual information to guide cross-modal correlation learning. Without the guidance of contextual information, existing methods cannot overcome the performance bottlenecks of ITR. In addition, previous studies cannot optimize image–text similarity via simultaneously achieving two important goals: (1) Image–text similarity between matched image–text pairs is larger than that between mismatched ones, and (2) The divergence of different embedding spaces should be constrained to keep semantic consistency, because excessive difference between different embedding spaces may degrade the precision of cross-modal correlation estimation. Therefore, in response to the challenges revealed by previous studies, we introduce a novel context-guided cross-modal correlation learning framework that aims to effectively bridge the heterogeneous gap and significantly promote the performance of ITR.

2.2. Correlation optimization for ITR

The purpose of correlation optimization is to utilize a loss function to optimize the image–text similarity. This ensures that semantically related samples exhibit closer distances, while unrelated samples are effectively separated. In previous works, various correlation optimization methods have been proposed for different tasks. These methods can be broadly categorized into two classes: (1) Non-hybrid loss function and (2) Hybrid loss function.

2.2.1. Non-hybrid loss function

In relevance optimizing, the key challenge lies in designing an effective loss function. For instance, comparative loss [36] pulls positive instances closer while maintaining a fixed separation for all negatives. However, imposing the fixed distance on all negatives may pose strict limitations. This inspired the proposal of Ladder Loss [37], which aims to solve the problem of discontinuity and consequently enhance the stability and convergence of the optimization process. In addition, the goal of triplet loss [38] is learning the embedding space by defining a minimum difference between the distances of anchor-positive and anchor-negative pairs. After that, quadruplets are incorporated into recent approaches such as histogram loss [39] and PDDM [40]. Recently, Wang et al. proposed the multi-similarity loss to gather and assign weights for informative pairs [41]. Frome et al. first attempted to map images and sentences into a common embedding space [42]. They employed an unweighted triplet loss to promote the clustering of semantically related instances. Faghri et al. introduced a hard triplet loss that capitalizes on more difficult negative instances within a mini-batch [15].

2.2.2. Hybrid loss function

In contrast to the non-hybrid loss function, the hybrid loss function not only takes into account the feature representation between images and texts but also incorporates high-level semantic consistency information. Hence, the intricate relationship between images and texts can be more comprehensively captured, facilitating the effective fusion of semantic information present in both images and texts.

In recent years, a multitude of hybrid loss functions have been employed in ITR tasks. For example, Xu et al. introduced the CASC method, which integrates cross-modal attention loss and semantic label prediction loss [43]. This approach combines global semantic coherence, multi-label prediction, and local attention region-word alignment within a unified framework. Zhang et al. proposed a hybrid cross-modal similarity loss, which converts intra-modal semantic correlations into cross-modal similarities for training a unified subspace learning model [44]. Wang et al. combined adversarial loss and embedding loss to achieve optimal feature representation [45]. Zhang et al. proposed a GDL hybrid loss by combining two adversarial losses and semantic constraint loss [46]. To uphold both global and local semantic consistency of samples within the embedding space, Liu et al. introduced a Consistent Multimodal Contrastive (CMC) loss, which incorporates intra-modal and inter-modal ranking losses concurrently during the training process [47].

Generally speaking, many hybrid loss functions involve hyperparameters that require fine-tuning, and their sensitivity to these hyperparameters may impact their overall performance and generalization. Additionally, the scalability of some hybrid loss functions to large datasets and diverse domains may be challenging, and there is a requirement for approaches that demonstrate robustness across different scenarios. In this work, our objective is to introduce a hybrid loss for the ITR task, taking into account the challenges highlighted above.

3. The proposed method

As mentioned above, there are three crucial tasks to be solved in CCCL: (1) Perceiving contextual information, (2) Reasoning semantic relationships within each modality, and (3) Aligning instances or their patches across different modalities.

3.1. Overview of CCCL

This section presents the framework of CCCL, illustrated in Fig. 2, which consists of three distinct components: (1) Context-perceived Embedding Learning, (2) Cross-modal Correlation Learning, and (3) Hybrid Loss. First, context-perceived embedding learning aims to perceive contextual information in images and texts, and obtain context-perceived visual and textual feature representations. In particular, the Stanford CoreNLP [48], a natural language processing toolkit developed by Stanford University, is utilized to parse the semantic dependencies between words, which facilitates the generation of more refined feature representations. Second, since the global-level instance lacks the intra-modal correlation, our proposed CCCL framework contains three branches: (1) Inter-modal correlation learning at global-level, (2) Inter-modal correlation learning at local-level, and (3) Intra-modal correlation learning at local-level. Third, to balance the cross-modal coherence and alignment, a hybrid loss is used to fuse and optimize the image–text similarities learned from the component of cross-modal correlation learning.

CCCL aims to acquire cross-modal correlation at various levels, specifically targeting global-level and local-level information. These terms delineate different scales or scopes of information, and related definitions are outlined below:

Definition 1 (*Global-level Information*)**.** Global-level information represents the overall characteristics or properties of the entire instance (e.g. image, text).

Definition 2 (*Local-level Information*)**.** Local-level information focuses on specific details or fine-grained features within a small part of the instance (e.g. region of an image, word of a text).

Furthermore, cross-modal correlation encompasses both inter-modal correlation and intra-modal correlation, each defined as follows:

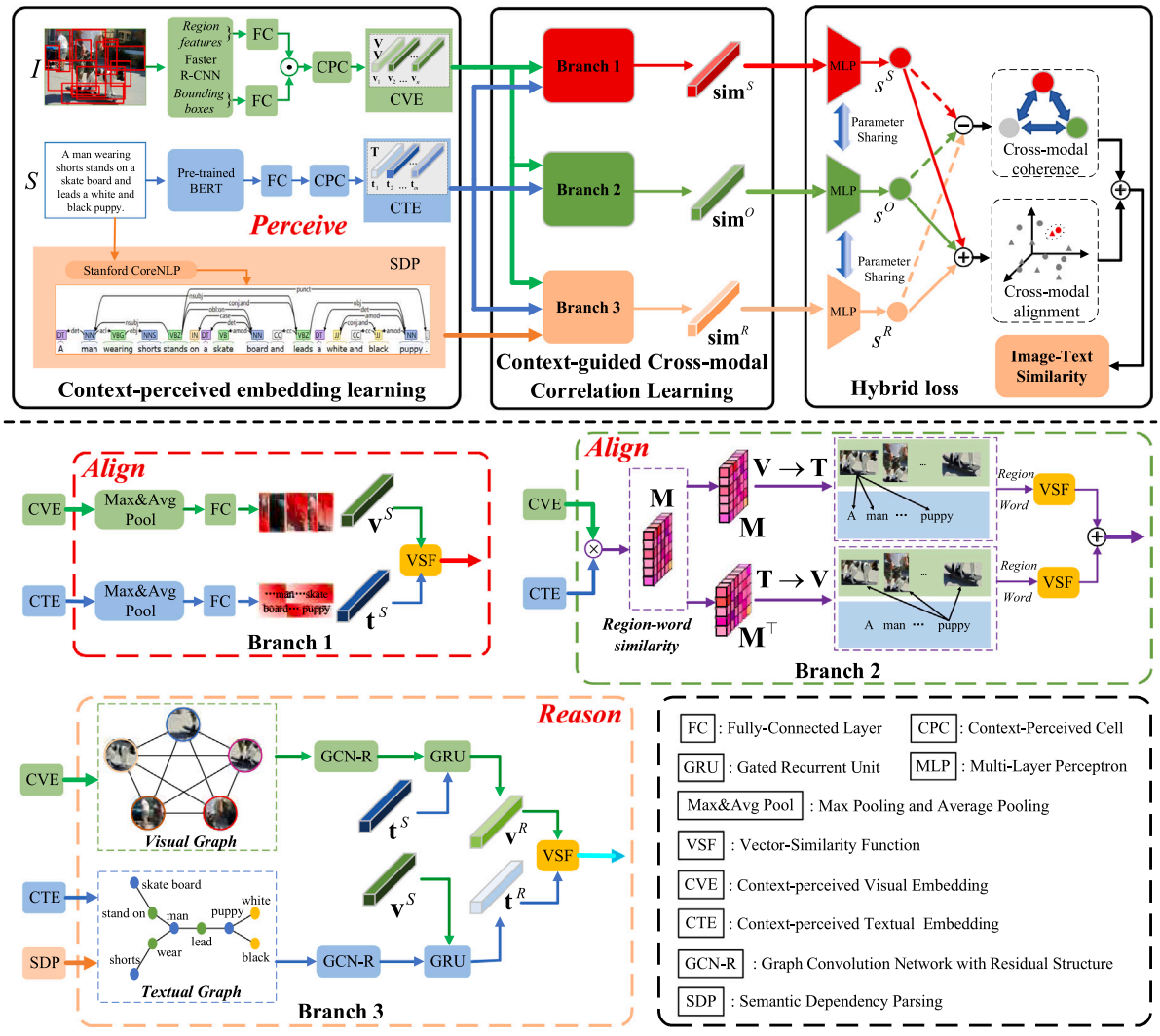


Fig. 2. The CCCL framework comprises three components: (1) Context-perceived embedding learning, (2) Context-guided cross-modal correlation learning, and (3) Hybrid loss. Notably, the cross-modal correlation learning module consists of three branches: (1) Branch 1: Inter-modal correlation learning at global-level, (2) Branch 2: Inter-modal correlation learning at local-level, and (3) Branch 3: Intra-modal correlation learning at local-level.

Definition 3 (*Inter-modal Correlation*)**.** Inter-modal correlation learning delves into intricate interactions across different modalities, including one-to-one interaction and one-to-many interaction. It is an important way to bridge the heterogeneous gap across global-level and local-level.

Definition 4 (*Intra-modal Correlation*)**.** Intra-modal correlation learning is dedicated to unraveling intricate relationships within a single modality. Its primary objective is to capture sophisticated semantic dependencies, thereby achieving a more holistic comprehension of connections between diverse patches in one instance.

The following section provides a comprehensive exposition of each component, accompanied by detailed descriptions.

3.2. Context-perceived embedding learning

Contextual information provides useful clues for cross-modal correlation learning. Existing studies have successfully used contextual information in the task of ITR [20,26,28] and image captioning [49]. We introduce the Context-Perceived Cell (CPC), which utilizes the self-attention mechanism [35] and the gate mechanism to fully exploit intra-modal complementary semantic relationships and effectively capture contextual information within each modality. By suppressing uninformative interactions between fine-grained features, we obtain visual context-perceived representations of regions and textual context-perceived representations of words for both image and text modalities.

3.2.1. Context-perceived cell

Fig. 3 shows how the context-perceived cell combines the self-attention mechanism and the gate mechanism to obtain a context-perceived representation of fine-grained data within one modality in an adaptive way.

Let $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L\} \in \mathbb{R}^{L\times d}$ denotes input feature sequence of context-perceived cell, where L is the sequence length, and d is feature dimension. To derive three distinct input feature sets for the self-attention mechanism, we introduce three fully-connected layers:

$$\begin{aligned}\mathbf{Q} &= \mathbf{Y}\mathbf{W}_{\mathbf{Q}}\\ \mathbf{K} &= \mathbf{Y}\mathbf{W}_{\mathbf{K}}\\ \mathbf{V} &= \mathbf{Y}\mathbf{W}_{\mathbf{V}}\end{aligned} \tag{1}$$

where $\mathbf{W}_{\mathrm{Q}} \in \mathbb{R}^{d\times d_K}$, $\mathbf{W}_{\mathrm{K}} \in \mathbb{R}^{d\times d_K}$, and $\mathbf{W}_{\mathrm{v}} \in \mathbb{R}^{d\times d_V}$ are the weight matrices to be trained. The self-attention mechanism mines the intra-modal contextual information by computing the dot product similarity

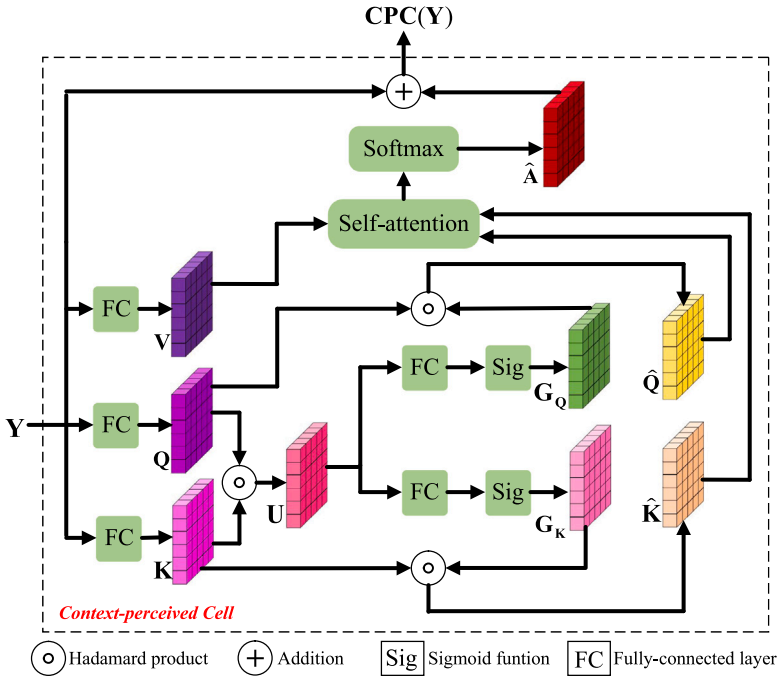


Fig. 3. The internal structure of context-perceived cell.

between $\mathbf{Q}$ and $\mathbf{K}$, which is defined as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_K}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{A}$ contains the intra-modal contextual information output by the original self-attention mechanism, however, $\mathbf{Q}$ and $\mathbf{K}$ may contain noise interference. To effectively capture complementary semantic information from fine-grained patches, we leverage the gate mechanism to adaptively modulate the intra-modal information flow and suppress irrelevant interactions within one modality.

$\mathbf{Q}$ and $\mathbf{K}$ are initially fused as follows.

$$\mathbf{U} = \mathbf{Q} \circ \mathbf{K} \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{L \times d_K}$ is the result of fusion, and $\circ$ represents the Hadamard product. Then, we obtain the gating mask matrix $\mathbf{G_Q}$ and $\mathbf{G_K}$ of $\mathbf{Q}$ and $\mathbf{K}$ by the fully-connected layer and sigmoid function, respectively.

$$\begin{aligned} \mathbf{G}_{\mathrm{Q}} &= sigmoid\left(\mathcal{P}(\mathbf{U}, \mathbf{W}_{\mathrm{Q}}^{G}, \mathbf{b}_{\mathrm{Q}}^{G})\right) \\ \mathbf{G}_{\mathrm{K}} &= sigmoid\left(\mathcal{P}(\mathbf{U}, W_{\mathrm{K}}^{G}, \mathbf{b}_{\mathrm{K}}^{G})\right) \end{aligned} \tag{4}$$

where $sigmoid(\cdot)$ represents the sigmoid function, and it aims to output a value within $[0, 1]$. In addition, the fully-connected layer achieves the linear transformation through the projection function $\mathcal{P}(x, y, z) = xy + z$. Particularly, $\mathbf{W}_{\mathrm{Q}}^{G}, \mathbf{W}_{\mathrm{K}}^{G} \in \mathbb{R}^{d_K * d_K}$ and $\mathbf{b}_{\mathrm{Q}}^{G}, \mathbf{b}_{\mathrm{K}}^{G} \in \mathbb{R}^{1 * d_K}$ are the weight matrix and bias that need to be learned in the fully-connected layer. Finally, the obtained gating mask matrix is used to regulate the information flow of $\mathbf{Q}$ and $\mathbf{K}$ as follows.

$$\begin{aligned} \hat{\mathbf{Q}} &= \mathbf{Q} \circ \mathbf{G}_{\mathrm{Q}} \\ \hat{\mathbf{K}} &= \mathbf{K} \circ \mathbf{G}_{\mathrm{K}} \end{aligned} \tag{5}$$

We use the updated $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}$ to effectively learn the intra-modal correlation, Eq. (2) is updated as follows.

$$\hat{\mathbf{A}} = \text{softmax}\left(\frac{\hat{\mathbf{Q}}\hat{\mathbf{K}}^T}{\sqrt{d_K}}\right)\mathbf{V} \tag{6}$$

where $d_K = d_V = d$, and $\hat{A}$ contains more meaningful contextual information within one modality. Hence, the principle of Context-Perceived Cell can be described as follows.

$$CPC(\mathbf{Y}) = \hat{\mathbf{A}} + \mathbf{Y} \tag{7}$$

3.2.2. Context-perceived visual embedding learning

We utilized Faster R-CNN [50], a pre-trained object detection model with bottom-up attention [51], on the Visual Genome [52] dataset to extract the top n most confident salient regions from each image I. We employed ResNet-101 [53] to extract the features of the aforementioned regions, yielding a set of feature vectors denoted as $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\} \in \mathbb{R}^{n \times d_Q}$, where $\mathbf{q}_i$ refers to feature vector of the ith region, and d_Q represents the dimensionality of the features. Then we map them into a d-dimensional common embedding space via a fully-connected layer:

$$\hat{\mathbf{Q}} = \mathcal{P}(\mathbf{Q}, \mathbf{W}_{\mathbf{Q}}, \mathbf{b}_{\mathbf{Q}}) \tag{8}$$

where $\mathbf{W}_{\mathbf{Q}} \in \mathbb{R}^{n \times d_Q}$ and $\mathbf{b}_{\mathbf{Q}}$ are the weight matrix and the bias to be learned. In particular, $\hat{\mathbf{Q}} = \{\hat{\mathbf{q}}_1, \hat{\mathbf{q}}_2, \ldots, \hat{\mathbf{q}}_n\} \in \mathbb{R}^{n \times d}$ is the updated feature vectors of regions in image I.

In contrast to previous studies that focused on relative position feature vectors of regions, our approach aims to utilize absolute position feature vectors to capture the global spatial complementary relationships between different regions. Inspired by PFAN [26], which incorporates prior object positions to facilitate the learning of a more robust visual–text joint embedding, we endeavor to enhance the visual embedding by acquiring absolute position features for each image region. Suppose that the coordinates of the points on the top-left corner and the bottom-right corner of image I are represented as $\left(x^{tl}, y^{tl}\right)$ and $\left(x^{br}, y^{br}\right)$, respectively. The position feature vectors of all regions in image I are defined as $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\} \in \mathbb{R}^{n \times 6}$, where the six-tuple $\mathbf{s}_i = \left(\frac{x_i^{tl}}{x^{br}-x^{tl}}, \frac{y_i^{tl}}{y^{br}-y^{tl}}, \frac{x_i^{br}}{x^{br}-x^{tl}}, \frac{y_i^{br}}{y^{br}-y^{tl}}, \frac{x_i^{br}-x_i^{tl}}{y_i^{br}-y_i^{tl}}, \frac{(x_i^{br}-x_i^{tl})\times(y_i^{br}-y_i^{tl})}{(x^{br}-x^{tl})\times(y^{br}-y^{tl})}\right)$ is corresponding to the ith region. Specifically, $\left(x_i^{tl}, y_i^{tl}\right)$ is coordinate of the point on the top-left corner of the ith region, and $\left(x_i^{br}, y_i^{br}\right)$ is coordinate of the point on the bottom-right corner of the ith region, respectively.

Then, we define $\hat{s}_i \in \mathbb{R} \times d$ as the absolute position feature of the region obtained through a fully-connected layer and the sigmoid function.

$$\hat{\mathbf{s}}_i = sigmoid\left(\mathcal{P}(\mathbf{W}_S, \tilde{\mathbf{s}_i}, \mathbf{b}_S)\right) \tag{9}$$

where $\mathbf{W}_S \in \mathbb{R}^{d \times 6}$ and $\mathbf{b}_S$ denote the weight matrix and bias to be learned. Furthermore, $\hat{\mathbf{S}} = \{\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \ldots, \hat{\mathbf{s}}_n\} \in \mathbb{R}^{n \times d}$ is the updated position vector features of all regions in image I.

To effectively extract the complementary semantic information among diverse regions, we propose the context-perceived cell to integrate the features of each region with its corresponding position features. To further enhance the contextual understanding of the visual region, we employ the context-perceived cell to capture the contextual information. Furthermore, the spatial-enhanced visual context-perceived representation $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \in \mathbb{R}^{n \times d}$ is defined as follows.

$$\mathbf{V} = CPC\left(\hat{\mathbf{Q}} \circ \hat{\mathbf{S}}\right) \tag{10}$$

3.2.3. Context-perceived textual embedding learning

We utilize the pre-trained BERT model, as introduced in [54], to process textual data and obtain bidirectional feature representations that capture contextual information. Given a text Z consisting of m words, we first apply the WordPiece tokenizer to tokenize the sentence. Then, we utilize BERT to extract the corresponding word features, which are represented as $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m\} \in \mathbb{R}^{m \times d_E}$, where $\mathbf{e}_j$ denotes the feature vector of the jth word. Afterwards, we map $\mathbf{E}$ into a d-dimensional common embedding space via fully-connected layer:

$$\hat{\mathbf{E}} = \mathcal{P}(\mathbf{E}, \mathbf{W}_{\mathbf{E}}, \mathbf{b}_E) \tag{11}$$

where $\mathbf{W}_E \in \mathbb{R}^{d_E \times d}$ and $\mathbf{b}_E$ are the weight matrix and bias to be learned, $\hat{\mathbf{E}} = \{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_m\} \in \mathbb{R}^{m \times d}$ is the updated word features of text Z. The context-perceived cell is employed to augment the fusion of contextual data derived from word sequences within the text, and the sequential-enhanced textual context-perceived embedding $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m\} \in \mathbb{R}^{m \times d}$ is obtained as follows.

$$\mathbf{T} = CPC\left(\hat{\mathbf{E}}\right) \tag{12}$$

Particularly, $\mathbf{t}_j$ in $\mathbf{T}$ is corresponding to the jth word in text Z.

3.3. Context-guided cross-modal correlation learning

Guided by the context-perceived embedding learned from Section 3.2, we design three branches in the CCCL framework to capture the inter-modal and intra-modal correlations at both global and local levels.

3.3.1. Branch 1: Inter-modal correlation learning at global-level with one-one interaction

In branch 1, we aim to learn a feature vector for each instance of different modalities, which reflects inter-modal correlation between an image and a text at global-level, that is, one-one interaction. We conduct Max Pooling and Average Pooling on the visual embedding $\mathbf{V}$ (see Eq. (10)) of image I and the textual embedding $\mathbf{T}$ (see Eq. (12)) of text Z. The resulting Max Pooling features $\tilde{\mathbf{v}}$ and $\tilde{\mathbf{t}}$ are both vectors of dimensionality d, which highlight the importance of discriminative features. On the other hand, the resulting Average Pooling features $\bar{\mathbf{v}}$ and $\bar{\mathbf{t}}$ are also vectors of dimensionality d, which ensure the integrity of intra-modal information. Afterwards, we combine them as follows.

$$\begin{aligned} \hat{\mathbf{v}} &= \tilde{\mathbf{v}} + \bar{\mathbf{v}} \\ \hat{\mathbf{t}} &= \tilde{\mathbf{t}} + \bar{\mathbf{t}} \end{aligned} \tag{13}$$

Next, we employ a fully-connected layer to learn the fused feature vector $\mathbf{v}^{\mathbf{S}} \in \mathbb{R}^d$ for image I, and $\mathbf{t}^{\mathbf{S}} \in \mathbb{R}^d$ for text Z.

$$\begin{aligned} \mathbf{v}^{\mathbf{S}} &= \mathcal{P}(\mathbf{W}_{\mathrm{v}}, \hat{\mathbf{v}}, \mathbf{b}_{\mathrm{v}}) \\ \mathbf{t}^{\mathbf{S}} &= \mathcal{P}(\mathbf{W}_{\mathrm{t}}, \hat{\mathbf{t}}, \mathbf{b}_{\mathrm{t}}) \end{aligned} \tag{14}$$

where $\mathbf{W}_{\mathrm{v}}, \mathbf{W}_{\mathrm{t}} \in \mathbb{R}^{d*d}$ and $\mathbf{b}_{\mathrm{v}}, \mathbf{b}_{\mathrm{t}}$ are the weight matrix and bias of fully-connected layer.

To capture more detailed complementary relationships between different modalities, we have drawn inspiration from SGRAF [35] and incorporated the Vector Similarity function in CCCL. Assume that there are vectors $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{b} \in \mathbb{R}^d$, the vector similarity between them is defined as:

$$\mathrm{VSF}(\mathbf{a}, \mathbf{b}, \mathbf{W}) = \frac{\mathbf{W}|\mathbf{a} - \mathbf{b}|^2}{\left\|\mathbf{W}|\mathbf{a} - \mathbf{b}|^2\right\|_2} \tag{15}$$

where $|g|^2$ and $\|g\|_2$ denote Element-wise square and l_2-norm, $\mathbf{W} \in \mathbb{R}^{P*d}$ is the weight matrix to be learned. Especially, the dimension of the similarity vector learned in the three branches of CCCL is set to P.

Finally, we compute the similarity sim^S between image I and text Z via learning the inter-modal correlation between image I and text Z at global-level.

$$\mathrm{sim}^S = \mathrm{VSF}\left(\mathbf{v}^S, \mathbf{t}^S, \mathbf{W}^S\right) \tag{16}$$

3.3.2. Branch 2: Inter-modal correlation learning at local-level with one-many interaction

As local-level patches can offer important and complementary semantic information, branch 2 is developed to fully capture the inter-modal correlation among image regions and words via the cross-attention mechanism, thereby learning the similarity between image and text at local-level. In particular, this branch investigates the interaction between one patch in one modality and many patches in another modality (one-many interaction).

By utilizing the visual embedding $\mathbf{V}$ and textual embedding $\mathbf{T}$, we obtain the region-word similarity matrix $\mathbf{M}$, where M_{ij} represents the cosine similarity between the ith region and jth word. To fully explore the correlation between regions and words, we apply the cross-attention mechanism in both directions, $\mathbf{V} \rightarrow \mathbf{T}$ and $\mathbf{T} \rightarrow \mathbf{V}$. In the $\mathbf{V} \rightarrow \mathbf{T}$ direction, we learn a fusion vector $\mathbf{u}_i^T = \sum_{j=1}^m \omega_{ij} \mathbf{t}_j$ for each region by combining all the words in a text, where ω_{ij} denotes the cross-attention weight.

$$\omega_{ij} = \frac{\exp\left(\lambda \bar{M}_{ij}\right)}{\sum_{j=1}^m \exp\left(\lambda \bar{M}_{ij}\right)} \tag{17}$$

where $\overline{\mathbf{M}}_{ij} = \frac{\mathrm{relu}(\mathbf{M}_{ij})}{\sqrt{\sum_{i=1}^m \left(\lambda \overline{\mathbf{M}}_{ij}\right)}}$ is the result of normalizing the elements in $\mathbf{M}$ along the column dimension. The similarity between the ith region and its corresponding word fusion vector $\mathbf{u}_i^T$ is computed using the vector similarity function $\mathrm{VSF}(\cdot)$. Subsequently, the local-level similarity between an image I and text Z in the $\mathbf{V} \rightarrow \mathbf{T}$ direction is determined by taking the average of the similarities between all regions and their corresponding word fusion vectors:

$$\mathrm{sim}_{V \rightarrow T}^O = \frac{1}{n} \sum_{i=1}^n \mathrm{VSF}\left(\mathbf{v}_i, \mathbf{u}_i^T, \mathbf{W}_{V \rightarrow T}^O\right) \tag{18}$$

Likewise, the similarity between the image I and text Z in the $\mathbf{T} \rightarrow \mathbf{V}$ direction can be computed by manipulating the word-region matrix $\mathbf{M}^T$:

$$\mathrm{sim}_{T \rightarrow V}^O = \frac{1}{m} \sum_{j=1}^m \mathrm{VSF}\left(\mathbf{t}_j, \mathbf{u}_j^V, \mathbf{W}_{T \rightarrow V}^O\right) \tag{19}$$

To ultimately determine the similarity sim^O between the image I and text Z, we compute the sum of $sim_{V \rightarrow T}^O$ and $sim_{T \rightarrow V}^O$, which involves aligning the image regions and words to learn the intra-correlation at the local-level:

$$\mathrm{sim}^O = \mathrm{sim}_{V \rightarrow T}^O + \mathrm{sim}_{T \rightarrow V}^O \tag{20}$$

3.3.3. Branch 3: Intra-modal correlation learning at local-level with many-many interaction

In contrast to the two aforementioned branches that concentrate on inter-modal correlation, this subsection investigates the acquisition of intra-modal correlation based on exploring the interaction within many patches (many-many interaction). To capture the interdependence among local-level patches, we construct a visual graph for each

image and a textual graph for each text. We subsequently employ graph convolutional networks to transfer and update information between nodes in the graphs. Finally, we use a Gated Recurrent Unit (GRU) [55] model to integrate and reason the relationship between different modalities.

Construction of Visual Graph. We employ the context-perceived visual embedding $\mathbf{V}$ to represent each image as a fully-connected undirected graph $G_1 = (V_1, E_1)$. Here, V_1 denotes the set of nodes in the graph, and E_1 denotes the set of edges. Each node in the graph corresponds to an image region and is represented by the feature vector $\mathbf{v}_i$. All nodes are interconnected by edges, facilitating information transfer between all image regions. To accurately capture the interrelationships between different image regions, we compute the affinity between them as the weight of the edge in the graph:

$$\mathbf{W}_{ij}^{V} = \left(\mathbf{W}_1 \mathbf{v}_i\right)^{\mathrm{T}} \left(\mathbf{W}_2 \mathbf{v}_j\right) \tag{21}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d*d}$ is the projection matrix to be trained. Furthermore, W_{ij}^{V} denotes the weight of the edge connecting the ith and jth nodes in the visual graph, from which we can derive the edge weight matrix $\mathbf{W}_e^{V} \in \mathbb{R}^{n \times n}$.

Construction of Textual Graph. In this work, we utilize Stanford CoreNLP to perform syntactic dependency parsing of the input text, thereby constructing a text graph $G_2 = (V_2, E_2)$. The nodes in the graph are represented using the context-perceived textual embedding $\mathbf{T}$, while the syntax dependency matrix $\mathbf{W}^{D}$ between words is obtained from Stanford CoreNLP. Specifically, W_{ij}^{D} is set to 1, if there is a grammatical dependency between the ith word and jth word, and 0 otherwise. To further capture the intrinsic semantic relationships between words, we compute an affinity matrix $\mathbf{W}^{T}$ between nodes in the text graph. Specifically, $\mathbf{W}_{ij}^{T}$ represents the similarity between ith word and jth word.

$$\mathbf{W}_{ij}^{T} = \left(\mathbf{W}_1' \mathbf{t}_i\right)^{\mathrm{T}} \left(\mathbf{W}_2' \mathbf{t}_j\right) \tag{22}$$

where $\mathbf{W}_1', \mathbf{W}_2'$ is the projection matrix to be learned. To construct the edge weight matrix $\mathbf{W}_e^{T} \in \mathbb{R}^{m \times m}$ of the text graph, we combine the affinity matrix $\mathbf{W}^{T}$ and the grammatical dependency matrix $\mathbf{W}^{D}$:

$$\mathbf{W}_e^{T} = \mathbf{W}^{T} \circ \mathbf{W}^{D} \tag{23}$$

The Graph Convolutional Network (GCN) [56,57] is a powerful tool for modeling and processing graph data. It uses an edge weight matrix to aggregate the information of neighboring nodes when updating node features, allowing them to capture potential relationships between different nodes. To fully utilize the complementarity between fine-grained features within a modality, we employ GCN to infer the correlations between nodes in the graph.

We construct the graph $G = (V, E)$ composed of N nodes, where the node feature matrix $\mathbf{H} \in \mathbb{R}^{N \times D}$ and the edge weight matrix $\mathbf{W}_e \in \mathbb{R}^{N \times N}$ are defined accordingly. To update the node features, we adopt the graph convolution operation, which considers not only the node's own features but also the features of its neighboring nodes.

$$\hat{\mathbf{H}} = \mathrm{f}\left(\mathbf{H}, \mathbf{W}_e\right) = \mathbf{H} + \mathbf{W}_R \left(\mathbf{W}_e \mathbf{H} \mathbf{W}_G\right) \tag{24}$$

where $\mathbf{W}_G \in \mathbb{R}^{D \times D}$ is the parameter matrix to be trained, and $\mathbf{W}_R \in R^{N \times N}$ refers to the residual structure matrix. Note that the updated node feature matrix $\hat{\mathbf{H}} \in \mathbb{R}^{N*D}$ contains the reasoning clues between nodes. Therefore, the reasoning processes for node relationship on visual graph $G_1 = (V_1, E_1)$ and textual graph $G_2 = (V_2, E_2)$ are as follows:

$$\begin{aligned} \hat{\mathbf{V}} &= \mathrm{f}\left(\mathbf{V}, \mathbf{W}_e^{V}\right) \\ \hat{\mathbf{T}} &= \mathrm{f}\left(\mathbf{T}, \mathbf{W}_e^{T}\right) \end{aligned} \tag{25}$$

Utilizing Eq. (25), the updated visual feature $\hat{\mathbf{V}} = \left\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_n\right\} \in \mathbb{R}^{n \times d}$ and the updated textual feature $\hat{\mathbf{T}} = \left\{\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_2, \ldots, \hat{\mathbf{t}}_m\right\} \in \mathbb{R}^{m \times d}$ can be obtained with the intra-modal enhanced neighbor relationships.

To further enhance the discriminative fine-grained interaction information within each modality and eliminate redundant parts, we combine feature vectors (i.e., $\mathbf{v}^{S} \in \mathbb{R}^{d}$ and $\mathbf{t}^{S} \in \mathbb{R}^{d}$) learned from Branch 1 with the neighbor relationship enhancement features of the other modality. That is, the set of features $\hat{\mathbf{V}}^{*} = \left\{\mathbf{t}^{S}, \hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_n\right\} \in \mathbb{R}^{(n+1) \times d}$ and $\hat{\mathbf{T}}^{*} = \left\{\mathbf{v}^{S}, \hat{\mathbf{t}}_1, \hat{\mathbf{t}}_2, \ldots, \hat{\mathbf{t}}_m\right\} \in \mathbb{R}^{(m+1) \times d}$ are generated. Then, we input the integrated features $\hat{\mathbf{V}}^{*}$ and $\hat{\mathbf{T}}^{*}$ into a GRU model to fuse both intra-modal and inter-modal correlations.

$$\begin{aligned} \mathbf{v}^{R} &= \mathrm{GRU}^{V}\left(\hat{\mathbf{V}}^{*}\right)_{n+1} \\ \mathbf{t}^{R} &= \mathrm{GRU}^{T}\left(\hat{\mathbf{T}}^{*}\right)_{m+1} \end{aligned} \tag{26}$$

where $\mathbf{v}^{R} \in \mathbb{R}^{d}$ and $\mathbf{t}^{R} \in \mathbb{R}^{d}$ are the features of the last hidden state of the GRU, and they are used as the global relation vector of the image and the global relation vector of the text, respectively.

Finally, we compute the similarity between $\mathbf{v}^{R}$ and $\mathbf{t}^{R}$ by learning the intra-modal correlation as follows.

$$\mathrm{sim}^{R} = \mathrm{VSF}\left(\mathbf{v}^{R}, \mathbf{t}^{R}, \mathbf{W}^{R}\right) \tag{27}$$

3.4. The hybrid loss

To fully leverage the complementary semantic information across different modalities, this section aims to optimize the cross-modal similarity learned from different branches. Cross-modal similarity essentially relies on shared semantic information between different modalities, which should remain consistent across different embedding spaces. Consequently, inspired by [24], we propose the cross-modal coherence term to constrain the divergence of cross-modal similarity learned from different branches. Additionally, to ensure that the cross-modal similarity between matched image and text is higher than that between mismatched pairs, we employ the cross-modal alignment term to achieve effective alignment between different modalities. By combining these two terms, the hybrid loss is developed to attain the accurate optimization of cross-modal similarity.

In particular, we use three Multilayer Perceptrons (MLP) [58] with shared parameters to map the similarity vector sim^{S}, sim^{O}, and sim^{R} to scalar S^{S}, S^{O}, and S^{R}, respectively. The MLP used in CCCL consists of two fully-connected neural networks. The cross-modal similarity is generated by the softmax activation function that is connected by the final layer. Here, s^{in} and s^{out} denote the input and output of the MLP, respectively. The learning process is as follows:

$$s^{out} = sigmoid\left(\mathbf{W}_2 max\left(\left(\left(\mathbf{W}_1 s^{in} + b_1\right), 0\right)\right) + b_2\right) \tag{28}$$

where $\mathbf{W}_1 \in \mathbb{R}^{P*P}, \mathbf{W}_2 \in \mathbb{R}^{1*P}, b_1, b_2$ are the weight matrix and bias to be trained.

3.4.1. Cross-modal coherence term

As cross-modal similarity computation depends on connecting different modalities based on their shared semantic information, it is essential that the embedding spaces learned by our proposed three branches exhibit minimal divergence. Therefore, we propose a cross-modal coherence term to constrain the differences between different embedding spaces. Specifically, the difference between the image–text similarity x and y is defined as:

$$D(x, y) = \sqrt{(x - y)^2} \tag{29}$$

Hence, we can delineate the distinctions among the three types of image–text similarity (*i.e.*, S^{S}, S^{O}, S^{R}) for a given image I and text Z as follows:

$$D(I, S) = D\left(S^{S}, S^{O}\right) + D\left(S^{S}, S^{R}\right) + D\left(S^{O}, S^{R}\right) \tag{30}$$

For a mini-batch $\left\{\left(I_i, S_i\right)\right\}_{i=1}^{B}$ in the training process, the cross-modal coherence term is defined as:

$$\mathcal{L}_c = \sum_{i=1}^{B} \sum_{j=1}^{B} D\left(I_i, S_j\right) \tag{31}$$

Table 1
Implementation details.

Parameter name	Notation	Value
Visual features dimension	d_F	2048
Number of visual regions	n	36
Text feature dimension	d_E	768
Dimension of the common embedding space	d	1024
Interval coefficient in the triplet loss function	Δ	0.2
Vector dimension P in the vector similarity function (VSF)	P	256
Balance factor in the hybrid loss	λ	0.5 (Flickr30K), 0.75 (MS-COCO)
Batch size	B	128
Learning rate	lr	0.0002
Epoch	$epoch$	30

3.4.2. Cross-modal alignment term

To achieve semantic alignment across multiple modalities, we propose to measure the overall image–text similarity $F(I, Z)$ between image I and text Z via averaging the three aforementioned similarities:

$$F(I, Z) = \frac{1}{3}\left(S^S + S^O + S^R\right) \tag{32}$$

Subsequently, we optimize CCCL using the hinge-based bidirectional triplet loss function [59], while leveraging the hard negative sample mining technique to enhance computational efficiency. The cross-modal alignment term is defined as follows:

$$\begin{aligned}
&\mathcal{L}_t = \sum_{i=1}^{B}\left\{\left[\Delta - F\left(I_i, Z_i\right) + F\left(I_i, \hat{Z}_j\right)\right]_+ + \left[\Delta - F\left(Z_i, I_i\right) + F\left(Z_i, \hat{I}_j\right)\right]_+\right\} \\
&s.t. \hat{Z}_j = \arg\max_{Z_j, j\neq i} F\left(I_i, Z_j\right) \\
&\quad \hat{I}_j = \arg\max_{I_j, j\neq i} F\left(Z_i, I_j\right)
\end{aligned} \tag{33}$$

where the size of the mini-batch (denoted as $\{(I_i, Z_i)\}_{i=1}^{B}$) is B, and $F\left(I_i, Z_j\right)$ represents the similarity between the ith image and the jth text. Specifically, $\hat{Z}_j$ denotes the negative sample that exhibits the highest degree of similarity to the query I_i within the current mini-batch. Likewise, $\hat{I}_j$ represents the negative sample that shows the largest similarity to the query Z_i within the current mini-batch.

Subsequently, we integrate the aforementioned two terms to formulate the hybrid loss as follows:

$$\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_c \tag{34}$$

where λ is the balance factor.

4. Experiment

In this section, we present a series of experiments conducted on benchmark datasets, namely MS-COCO [60] and Flickr30K [61], to evaluate the performance of CCCL and compare it with various recently proposed state-of-the-art baselines. Our experiments include a detailed parameter analysis, as well as ablation experiments to demonstrate the effectiveness of CCCL. Furthermore, we also provide the attention visualization and retrieval examples of CCCL.

4.1. Datasets and evaluation metrics

Flickr30K[1] is a widely used benchmark dataset consisting of 31,783 images sourced from Flickr. Each image in Flickr30K is accompanied by five human-annotated sentences, providing rich linguistic context. Flickr30K is divided into three subsets: 29,783 images for training, 1000 images for validation, and another 1000 images for testing.

[1] http://shannon.cs.illinois.edu/DenotationGraph/

The MS-COCO[2] dataset consists of 123,287 images, each annotated with five sentences. We divide it into three subsets: 113,287 images for training, 5000 for validation, and an additional 5000 for testing. Our evaluation involves a challenging setting called MS-COCO (5K), where our method is directly tested on the full set of 5000 images. To ensure result reliability, we employ a 5-fold validation for the 1000 testing images and average the outcomes to derive comprehensive performance metrics.

We employ standard evaluation metrics such as Recall@K (R@K, where K is 1, 5, 10) and R@sum. R@K represents the proportion of ground truth instances within the top-K retrieved lists.

$$\text{Recall@K} = \frac{1}{N}\sum_{i=1}^{N} r_K \tag{35}$$

The testing set comprises a total of N instances, where r_K is defined as 1 if the ground-truth result is among the top-K returned results, and 0 otherwise. We exhibit the recall rates at the top 1 result (R@1), top 5 results (R@5), and top 10 results (R@10).

$$R@sum = \underbrace{R@1 + R@5 + R@10}_{\text{image-to-text}} + \underbrace{R@1 + R@5 + R@10}_{\text{text-to-image}} \tag{36}$$

Furthermore, we employ R@sum, which represents the cumulative sum of all R@K values in both the image-to-text and text-to-image directions.

4.2. Implementation details

We provide a more detailed implementation of our method in Table 1, including the experimental parameter settings, and the details of network training.

Parameter settings. The dimension of visual features is set to 2048 ($d_F = 2048$), and the number of visual regions is 36 ($n = 36$). To generate the original word embeddings with dimension $d_E = 768$, we utilize the basic version of the pre-trained BERT, which consists of 110M parameters, 12 layers, 12 attention heads, and a total of 768 hidden units. In our CCCL model, we set the dimension of the common embedding space and the GRU hidden layer feature dimension in branch 3 to 1024 ($d = 1024$). Furthermore, we set the interval coefficient in the triplet loss function to 0.2 (i.e. $\Delta = 0.2$). In addition, we discuss how to determine the dimensions of the vector similarity P and the balance factor λ of the hybrid loss in Section 4.4.

Network training details. We train our proposed model using the PyTorch library on a single NVIDIA GeForce RTX 3090 GPU. Moreover, we utilize the Adam optimizer [62] with a mini-batch size of 128 (i.e. $B = 128$), and train the model for 30 epochs. For the first 15 epochs, the learning rate is set to 0.0002 and then decayed by 10% for the remaining epochs, i.e., 0.00002. To determine the optimal model, we use a validation set at the end of each epoch and select the model with the maximum R@sum value.

Table 2
The Recall@K of CCCL and other methods on the Flickr30K test set.

Method	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Inter-modal correlation learning based methods							
VSE++	0.529	0.791	0.872	0.396	0.696	0.795	4.080
SCAN*	0.674	0.903	0.958	0.486	0.777	0.852	4.650
IMRAM	0.741	0.930	0.966	0.539	0.794	0.872	4.842
NAAF	**0.819**	**0.961**	**0.983**	0.610	0.853	0.906	5.132
Intra-modal correlation learning based methods							
PFAN*	0.700	0.918	0.950	0.504	0.787	0.861	4.720
VSE∞	0.765	0.942	0.977	0.564	0.834	0.899	4.981
VSRN++	0.792	0.946	0.975	0.606	0.856	0.914	5.089
Fusion of inter-modal and intra-modal correlation learning based methods							
SGM	0.718	0.917	0.955	0.535	0.796	0.865	4.786
MMCA	0.742	0.928	0.964	0.548	0.814	0.878	4.874
GSMN*	0.764	0.943	0.973	0.574	0.823	0.890	4.968
SGRAF	0.778	0.941	0.974	0.585	0.830	0.888	4.996
CGMN	0.779	0.938	0.968	0.599	0.851	0.906	5.041
w/o B-1	0.770	0.947	0.971	0.642	0.902	0.918	5.150
w/o B-2	0.704	0.881	0.938	0.596	0.863	0.910	4.892
w/o B-3	0.713	0.879	0.949	0.610	0.887	0.917	4.955
CCCL(ours)	0.772	0.950	0.976	**0.644**	**0.902**	**0.926**	**5.170**

Table 3
The Recall@K of CCCL and other methods on the MS-COCO 1K test set.

Method	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Inter-modal correlation learning based methods							
VSE++	0.646	0.891	0.957	0.520	0.831	0.920	4.764
SCAN*	0.727	0.948	0.984	0.588	0.884	0.948	5.082
IMRAM	0.767	0.956	0.985	0.617	0.891	0.950	5.166
NAAF	0.805	**0.965**	0.988	0.641	0.907	0.965	5.272
Intra-modal correlation learning based methods							
PFAN*	0.765	0.963	**0.990**	0.616	0.896	0.952	5.182
VSE∞	0.785	0.960	0.987	0.617	0.903	0.956	5.208
VSRN++	0.779	0.960	0.985	0.641	0.910	0.961	5.236
Fusion of inter-modal and intra-modal correlation learning based methods							
SGM	0.734	0.938	0.978	0.575	0.873	0.943	5.041
MMCA	0.748	0.956	0.977	0.616	0.898	0.952	5.147
GSMN*	0.784	0.964	0.986	0.633	0.901	0.957	5.225
SGRAF	0.796	0.962	0.985	0.632	0.907	0.961	5.243
CGMN	0.768	0.954	0.983	0.638	0.907	0.957	5.207
w/o B-1	0.810	0.959	0.985	0.692	0.922	0.968	5.336
w/o B-2	0.735	0.897	0.924	0.667	0.918	0.966	5.107
w/o B-3	0.752	0.911	0.945	0.678	0.923	0.969	5.178
CCCL(ours)	**0.815**	0.962	**0.990**	**0.701**	**0.924**	**0.973**	**5.365**

4.3. Performance comparison

To assess the effectiveness of CCCL, we conduct a thorough performance analysis using evaluation metrics such as Recall@K (where K=1, 5, and 10) in comparison with several other ITR methods on two commonly used benchmark datasets, namely Flickr30K and MS-COCO. For the related methods discussed in previous works, we utilize the experimental results presented in their original papers and emphasize the best results in bold while underlining suboptimal results. The unreported experimental results are denoted by the symbol '–' while the integrated model is indicated by the symbol '*'.

Table 2 presents the comparative retrieval performance of CCCL with other methods on the Flickr30K test set. The results indicate that CCCL exhibits the most outstanding performance in the T→I direction, with an improvement of 3.4%, 4.6%, and 1.2% over suboptimal results for R@1, R@5, and R@10 indexes, respectively. Moreover, in the I→T direction, the R@5 index value of CCCL is slightly suboptimal, with only 1.1% lower than the optimal NAAF. Nevertheless, the overall performance of CCCL in both retrieval directions is the best, with a 3.8% improvement in R@sum over suboptimal results.

Tables 3 and 4 report the retrieval performance of CCCL and other comparison methods on the MS-COCO (1K) and MS-COCO (5K) test sets. Upon analyzing the tables, we arrive at the following observations:

- Similar to the results on Flickr30K, CCCL attains the best performance in the T→I direction on both test sets. On the MS-COCO (1K) test set, there is a 6.0%, 1.4%, and 0.8% improvement over the suboptimal results for R@1, R@5, and R@10 indexes, respectively. On the MS-COCO (5K) test set, the indexes values of R@1, R@5, and R@10 increase by 4.7%, 5.4%, and 4.4%, respectively.
- On the MS-COCO (1K) test set, CCCL achieves the highest R@1 and R@10 index results in the I→T direction, with improvements of 1.0% and 0.2%, respectively, compared to suboptimal results. However, for R@5, it attains suboptimal results, with only a 0.3% deviation from the NAAF.

[2] https://cocodataset.org/

Table 4
The Recall@K of CCCL and other methods on the MS-COCO 5K test set.

Method	I→T			T→I			R@sum
	R@1	R@5	R@10	R@1	R@5	R@10	
Inter-modal correlation learning based methods							
VSE++	0.413	0.692	0.812	0.303	0.591	0.724	3.534
SCAN*	0.504	0.822	0.900	0.386	0.693	0.804	4.109
IMRAM	0.537	0.832	0.910	0.397	0.691	0.798	4.155
NAAF	**0.589**	**0.852**	**0.920**	0.425	0.709	0.814	4.309
Intra-modal correlation learning based methods							
PFAN*	–	–	–	–	–	–	–
VSE∞	–	–	–	–	–	–	–
VSRN++	0.547	0.829	0.909	0.420	0.722	0.827	4.254
Fusion of inter-modal and intra-modal correlation learning based methods							
SGM	0.500	0.793	0.879	0.353	0.649	0.765	3.939
MMCA	0.540	0.825	0.907	0.387	0.697	0.808	4.164
GSMN*	–	–	–	–	–	–	–
SGRAF	0.578	–	0.916	0.419	–	0.813	–
CGMN	0.534	0.813	0.896	0.412	0.719	0.824	4.198
w/o B-1	0.547	0.833	0.912	0.470	0.774	0.865	4.401
w/o B-2	0.513	0.801	0.882	0.461	0.759	0.864	4.280
w/o B-3	0.518	0.818	0.891	0.466	0.764	0.865	4.322
CCCL(ours)	0.555	0.837	0.918	**0.472**	**0.776**	**0.871**	**4.429**

- CCCL exhibits the highest R@sum on both test sets, with the results significantly improved by 9.3% and 12% compared to suboptimal results.

In conclusion, CCCL demonstrates the competitive retrieval performance on two benchmark datasets, which highlights its effectiveness in mining complementary semantic information within modality and improving the performance of ITR through our proposed three-branch network. By carefully analyzing the aforementioned experimental results, the following observations can be made:

- **Inter-modal correlation learning based methods**
  SCAN significantly outperforms VSE++ by capturing a well-contained fine-grained alignment using a cross-attention mechanism. IMRAM, an improved method proposed based on SCAN, aggregates high-order interaction information iteratively, and all evaluation index values are higher than SCAN. Moreover, NAAF is the first framework to explicitly use the positive effects of region-word correct matching and the negative effects of incorrect matching. NAAF proposes an iterative optimization method with a negative mining strategy, which explicitly drives more negative effects of mismatched segments, resulting in a more comprehensive and explanatory cross-modal similarity. NAAF achieves the best performance in the I→T direction, and the results in the T→I direction are also among the top performers.
- **Intra-modal correlation learning based methods**
  PFAN improves the contextual connection of visual regions by merging relative positional features with region features themselves. This method achieves the highest R@10 results on the MS-COCO 1K dataset, confirming its effectiveness. VSE∞ integrates fine-grained features using a simple pooling strategy and achieves excellent performance. Compared with other methods, the model structure of VSE∞ is simpler and easier to train. VSRN++, on the other hand, outperforms PFAN and VSE∞ significantly on both datasets and even achieves suboptimal performance on multiple metrics. This is mainly due to the following three factors: (1) VSRN++ fully utilizes the correlation between intra-modal regions and words through regional and word relation reasoning and learns more discriminative global features. (2) The cross-modal matching and generating objectives are jointly optimized to standardize the reasoning process of visual semantic information. (3) Contextual word features are extracted using the pre-trained BERT model.
- **Fusion of inter-modal and intra-modal correlation learning based methods**
  When constructing graphs for images and texts, GSMN and CGMN only need to identify whether there are interactions between different nodes, instead of using the scene diagram model which is prone to information loss. This may be one of the reasons why GSMN and CGMN outperform SGM. Additionally, MMCA improves upon methods that only use inter-modal correlations (such as VSE++ and SCAN) and methods that only use intra-modal correlations (such as PFAN). SGRAF combines attention mechanism and graph model, enabling us to learn interaction relations between local and global alignments using graph convolutions. Overall, its performance is better than most of the aforementioned methods across all metrics. These results demonstrate that considering both inter-modal and intra-modal correlations are beneficial for learning richer semantic representations, which in turn improves cross-modal retrieval.
- **Our proposed CCCL**
  CCCL is capable of effectively leveraging the inter-modal alignment and intra-modal reasoning, which is one of the main reasons why it outperforms most of the other methods. Moreover, CCCL has several advantages over the methods that explore both inter-modal alignment and intra-modal reasoning simultaneously: (1) By incorporating the absolute position information of the image region with the region feature itself, the context-perceived cell is utilized to fully extract the spatial context information between the regions, and the position information of the region is inputted to the visual graph as prior knowledge, making the intra-modal reasoning of the image smoother. (2) Through the three-branch network of CCCL, the complementary relationships of image and text modalities at different level of granularities are integrated in all aspects. These three branches well complement each other. In each branch, the cross-modal similarity vector is learned to capture the more detailed complementary relationships between different modalities. (3) In the hybrid loss, we innovatively propose a cross-modal coherence term, which is used to control the differences of image–text similarities of three branches, thus ensuring the effectiveness of retrieval.

4.4. Parameter sensitivity analysis

CCCL includes two groups of hyperparameters: (1) the vector dimension P in the vector similarity function (VSF) and (2) the balance

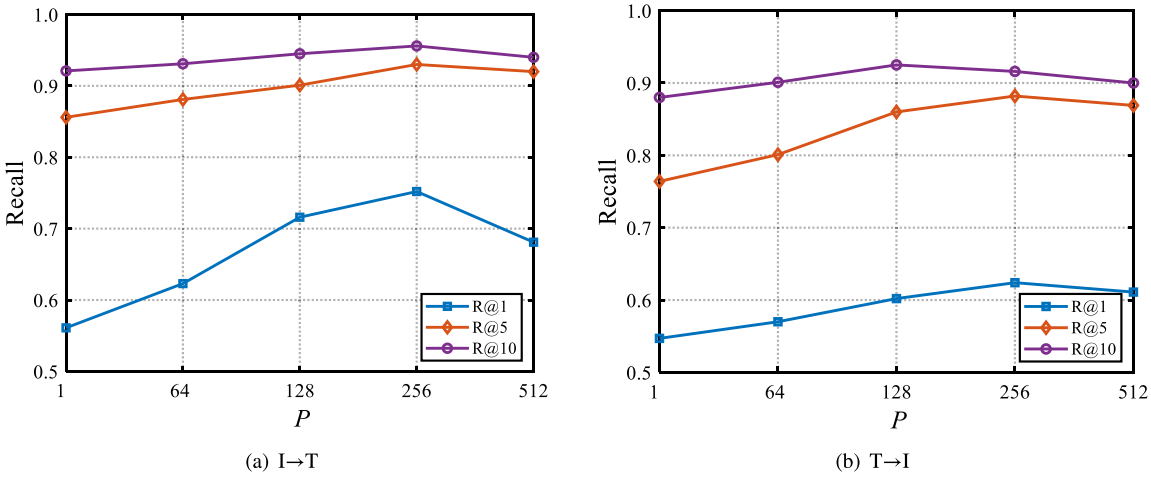


Fig. 4. P effect on the performance of ITR on the Flickr30K test set.

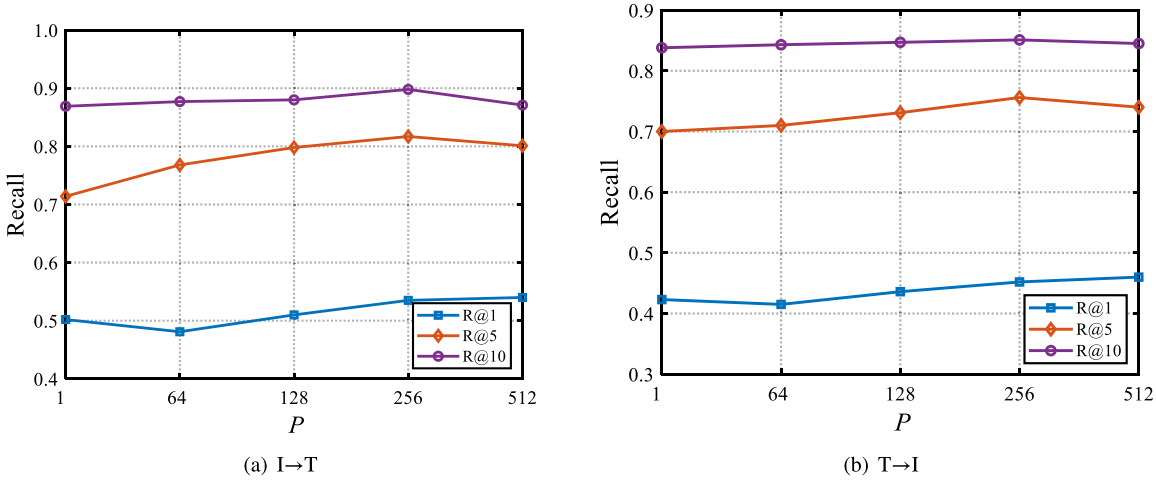


Fig. 5. P effect on the performance of ITR on the MS-COCO 5K test set.

factor λ in the hybrid loss. Initially, we set λ to 0 and determine the optimal value of P on both Flickr30K and MS-COCO datasets. Next, we fix the value of P and further determine the optimal value of λ on the same datasets.

In order to determine the optimal value of P, we consider a range of values for P in $\{1, 64, 128, 256, 512\}$, and plot the retrieval performance variation of Flickr30K and MSCOCO 5K in Figs. 4 and 5. By comprehensively comparing R@1, R@5, and R@10, we observe that when $P = 256$, the values are consistently the highest on both datasets. Hence, we select $P = 256$ as the optimal value.

Through careful observation and analysis of the changing trend of the curve, we have made the following findings regarding the optimal value of P. Firstly, the value of P should not be too small as the similarity vector with a small dimension may fail to capture semantic correlations. Secondly, when P is set to 128 or 256, the retrieval performance does not differ much, but when P is set to 512, the retrieval performance is reduced. Therefore, it is suggested that the value of P should not be too large as similarity vectors with a large dimension are more prone to introducing noise and learning irrelevant cross-modal correlations. Finally, it is observed that R@1 is more affected by P in the two retrieval directions of I→T and T→I, while R@5 and R@10 are relatively less affected.

To determine the optimal value of λ on both datasets, we set the range of λ to $\{0.25, 0.50, 0.75, 1.00\}$. The changes in retrieval performance of CCCL on Flickr30K and MS-COCO 5K are reported in Table 5. By comprehensively comparing R@1 and R@5 in both retrieval directions, we determine the optimal values as 0.5 and 0.75 for Flickr30K and MS-COCO, respectively.

Through the above observations, we have found that: (1) by setting an appropriate value for λ, the cross-modal coherence term can effectively limit the difference of similarity and improve the retrieval performance. (2) In most cases, the value of λ should not be too large or too small, because the cross-modal coherence term may only play an auxiliary role.

4.5. Ablation study

In this section, to validate each component of CCCL, we perform a series of ablation experiments using different model configurations on Flickr30K in Table 6. Specifically, we mainly explore the influence of the CPC component, the multi-branch alignment, and the key modules involved in the objective function.

4.5.1. Ablation study 1: context-perceived cell

To evaluate the effectiveness of the proposed CPC for mining contextual information from fine-grained features, we conduct ablation experiments by removing the context-perceived cell separately from the image and text modalities, resulting in the "w/o V-CPC" and "w/o T-CPC" models, respectively. We also remove CPC to obtain the third ablation model, "w/o CPC", where we directly use the three-branch

Table 5
The Recall@K of CCCL and other methods on the MS-COCO 5K test set.

λ	Flickr30K				MS-COCO 5K			
	I→T		T→I		I→T		T→I	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
0.25	0.766	0.936	0.625	0.887	**0.571**	0.833	0.461	0.767
0.50	**0.772**	0.950	**0.644**	**0.902**	0.563	**0.840**	0.458	0.772
0.75	0.751	**0.961**	0.640	0.893	0.555	0.837	**0.472**	**0.776**
1.00	0.747	0.944	0.638	0.895	0.545	**0.831**	0.466	0.768

Table 6
Ablation model design and experimental results on the Flickr30K dataset.

Method	I→T			T→I		
	R@1	R@5	R@10	R@1	R@5	R@10
Context-perceived cell						
w/o V-CPC	0.741	0.928	0.954	0.628	0.876	0.910
w/o T-CPC	0.753	0.944	0.968	0.635	0.899	0.917
w/o CPC	0.727	0.916	0.948	0.615	0.859	0.901
w/o POS	0.769	0.942	0.972	0.640	0.900	0.923
w/o BERT	0.764	0.939	0.968	0.637	0.891	0.916
Three-branch network						
w/o B-1	0.770	0.947	0.971	0.642	0.902	0.918
w/o B-2	0.704	0.881	0.938	0.596	0.863	0.910
w/o B-3	0.713	0.879	0.949	0.610	0.887	0.917
w/o VSF	0.755	0.938	0.967	0.636	0.899	0.923
Branch 1						
w/o A-Pool	0.759	0.939	0.968	0.633	0.888	0.924
w/o M-Pool	0.768	0.947	0.972	0.639	0.896	0.924
Branch 3						
w/o S-NLP	0.747	0.943	0.967	0.629	0.868	0.907
w/o C-GRU	0.768	0.948	0.973	0.635	0.889	0.921
Hybrid loss						
w/o SC-loss	0.757	0.936	0.973	0.630	0.879	0.920
w/o MLP	0.766	0.941	0.976	0.626	0.887	0.911
CCCL	**0.772**	**0.950**	**0.976**	**0.644**	**0.902**	**0.926**

network instead of using the CPC to process the features mapped by the fully-connected layer. In Table 6, we observe that removing CPC may cause varying degrees of degradation in model performance, which is further reduced when the CPC of both modalities is removed simultaneously. Furthermore, the contextual information complements each other and plays different roles in cross-modal similarity learning, indicating that CPC can effectively suppress intra-modal useless interactions and capture contextual information from fine-grained features. Moreover, "w/o V-CPC" performs worse than "w/o T-CPC", suggesting that contextual image information relies more on the CPC to capture. The contextual visual relationship plays a more crucial role in cross-modal similarity learning, possibly due to the more complex and rich complementary semantics in image regions.

In addition, we further explore the influence of the characteristics of the two modalities themselves of the input CPC on the model performance. For the images, we remove the operation of fusing the absolute position information and only extract the visual features, obtaining the ablation model "w/o POS"; For the text, we remove the pre-trained BERT model and use the Bi-GRU to extract the word features, getting the ablation model "w/o BERT". Both ablation models have less performance compared to CCCL, suggesting that: (1) The location information of the region may assist the CPC to capture the contextual relationship more comprehensively, and have a positive impact on the reasoning relations within each modality, and (2) Pre-trained BERT to extract word context-dependent bidirectional feature representations may be more effective than the end-to-end processing of text using Bi-GRU.

4.5.2. Ablation study 2: three-branch network

To assess the effectiveness of the proposed three-branch network and vector similarity functions, we design four ablation models in Table 6. Specifically, we remove Branch 1 (w/o B-1), Branch 2 (w/o B-2), Branch 3 (w/o B-3), or the vector similarity function (w/o VSF) in all three branches. Our experimental results reveal the following conclusions:

- Removing any branch may negatively impact the model's performance, with Branch 2 and Branch 3 having a greater impact than Branch 1. The reason for this is that Branch 1 captures some global-level information that can also be learned by the other two branches. While, Branch 2 and Branch 3 focus on mining potential alignments and effectively reasoning out objects, attributes, and relationships. Therefore, their roles in cross-modal similarity learning are complementary and difficult to replace.
- Compared to the CCCL model, the "w/o VSF" model's performance is reduced by approximately 2% on average, demonstrating that vector similarity functions can capture cross-modal correlations in more detail and learn more accurate image–text similarity.
- The effective utilization of inter-modal alignment and intra-modal reasoning is a key factor that distinguishes CCCL from most existing methods. This is primarily attributed to its three-branch network architecture, where Branch 1 and Branch 2 integrate complementary relationships at different granularity levels between image and text modalities, capturing more detailed complementary information across modalities, respectively. Branch 3 incorporates spatial context information extracted by context-perceived cells, integrating position information of region as prior knowledge into the visual graph. This enhancement significantly improves the accuracy of intra-modal reasoning. Experimental results demonstrate the synergy of CCCL's three branch networks,

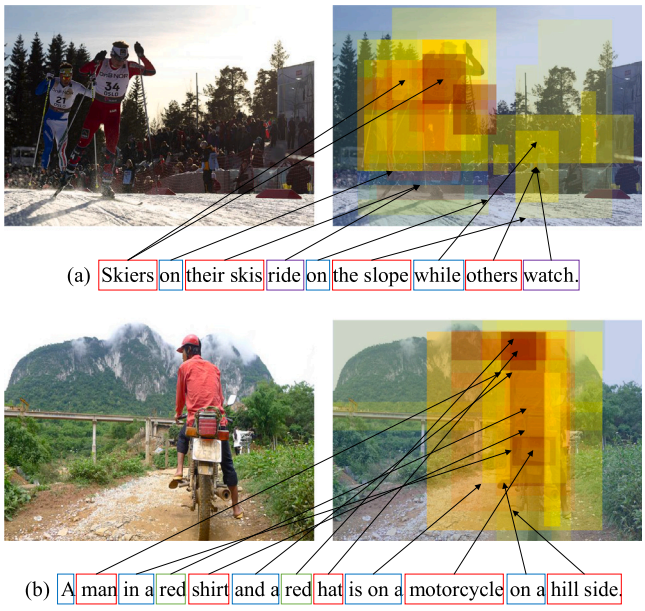


Fig. 6. Examples of context attention visualization for image regions.

indicating that the simultaneous consideration of inter-modal and intra-modal correlations facilitates the learning of richer semantic representations.

To validate the effectiveness of critical steps in Branch 1 and Branch 3, we design an ablation model that comprises the following experiments: (1) "w/o A-Pool" which solely utilizes max pooling instead of the combination of max and average pooling in Branch 1. (2) "w/o M-pool" which adopts the opposite configuration of the former, and only employs average pooling in Branch 1. (3) "w/o S-NLP" which constructs the textual graph in Branch 3 as a fully-connected graph, similar to a visual graph, rather than leveraging StandFord CoreNLP tool for syntax analysis of sentences. (4) "w/o C-GRU" which excludes the utilization of global-level fused feature vectors $\mathbf{v}^{\mathbf{S}}$ and $\mathbf{t}^{\mathbf{S}}$ learned by Branch 1 when GRU is used for global relation reasoning in Branch 3. By comparing the performance of these four ablation models with CCCL, we can draw the following conclusions:

- Combining two pooling strategies in Branch 1 proves to be effective. "w/o A-pool" performs worse than "w/o M-pool", indicating that discriminative features may play a more significant role in describing global semantics.
- The performance of "w/o S-NLP" significantly degrades, indicating that constructing intra-modal relation reasoning by structuring fully-connected graphs for text is unreasonable. This may be because there is a natural syntactic dependency between words, and fully-connected graphs destroy this relationship, leading to false correlation between different objects and making relational reasoning more challenging.
- The performance of "w/o C-GRU" demonstrates that the global-level fused feature vectors $\mathbf{v}^{\mathbf{S}}$ and $\mathbf{t}^{\mathbf{S}}$ learned by Branch 1 can effectively guide the GRU used in Branch 3 to achieve global relation reasoning.

4.5.3. Ablation study 3: hybrid loss

To demonstrate the effectiveness of the proposed cross-modal coherence term, we design the "w/o SC-loss" representation, which only utilizes the cross-modal alignment term function without the cross-modal coherence term. The experimental results in Table 6 reveal that the cross-modal coherence term can effectively eliminate the differences in image–text similarity and further enhance the performance of ITR. Moreover, we employ a fully-connected layer instead of a parameter-shared Multi-Layer Perceptron to learn image–text similarity in the ablation model "w/o MLP". The experimental results demonstrate that MLP ensures the correlation of cross-modal correlations captured by our proposed three-branch network, enabling the learning of more accurate cross-modal similarity.

4.6. Qualitative results and analysis

To demonstrate that CCCL effectively captures contextual information within modalities, we use the image as an example. We obtain the visual feature $\hat{\mathbf{V}} = \{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_n\} \in \mathbb{R}^{n \times d}$ from Branch 3 and the global relation vector $\mathbf{v}^R \in \mathbb{R}^d$ of the image to visually display the contextual attention of the region [30]. Specifically, we first calculate the inner product similarity between each image region feature $\hat{\mathbf{v}}_i$ and $\mathbf{v}^R$. We then rank the inner product similarities in descending order, where the ranking of $\hat{\mathbf{v}}_i$ is denoted as $rank_i$. The score for each region according to the ranking $score_i$ is defined as $score_i = \mu(n - rank_i)^2$, where n is the number of regions (in our case, $n = 36$), and μ is used to emphasize the highly ranked regions. We set $\mu = 100$ in our experiments. Finally, the final attention score at the location of each pixel is obtained by summing the scores of all the regions it belongs to.

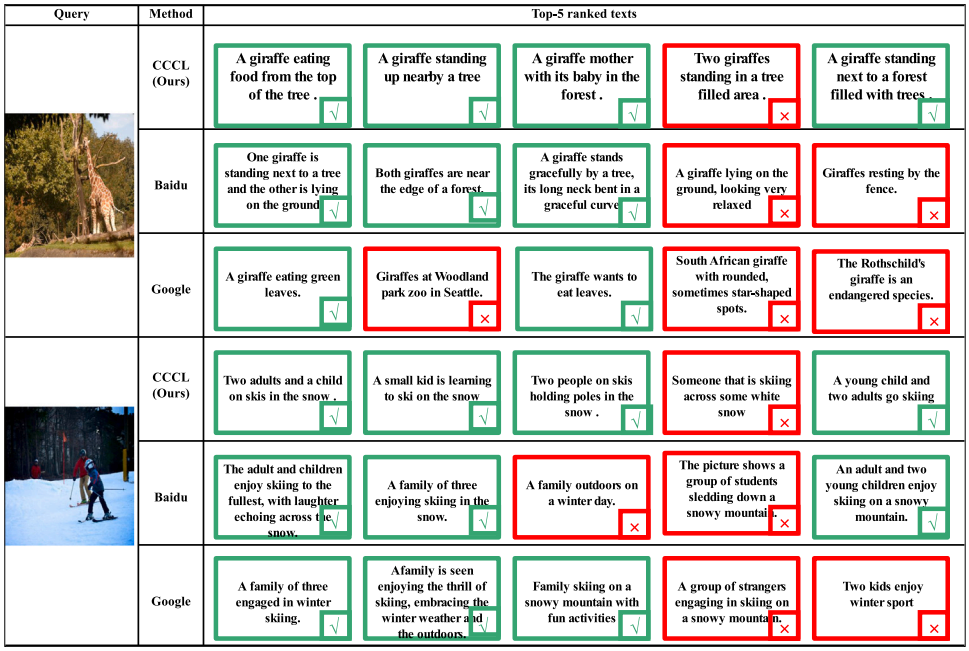
In Fig. 6, we present two original images containing a large number of visual objects and intricate relationships, along with corresponding visualizations of regional attention. Furthermore, we provide the corresponding sentence for each image. Upon analyzing the attention visualization of the context relation of image regions, we can observe that the image representation generated by CCCL accurately captures the key objects and their contextual relationships. For instance, in Fig. 6(a), the contextual relationship among the object "Skiers", the phrases "on their skis", and "ride on the slope" is well represented. Similarly, in Fig. 6(b), the contextual relationships between the object "man", and the phrases "in red shirt and a red hat" and "on a motorcycle" are also well captured.

Furthermore, in Fig. 7, we provide the top five retrieved results for both I→T and T→I corresponding to specific queries. For each direction of the ITR task, we select two queries representing distinct semantic concepts. Specifically, we compare our method with Baidu[3] and Google.[4] Note that the true matches are marked in green rectangles with check marks, while the incorrect retrieved items are indicated by red rectangles and cross marks.

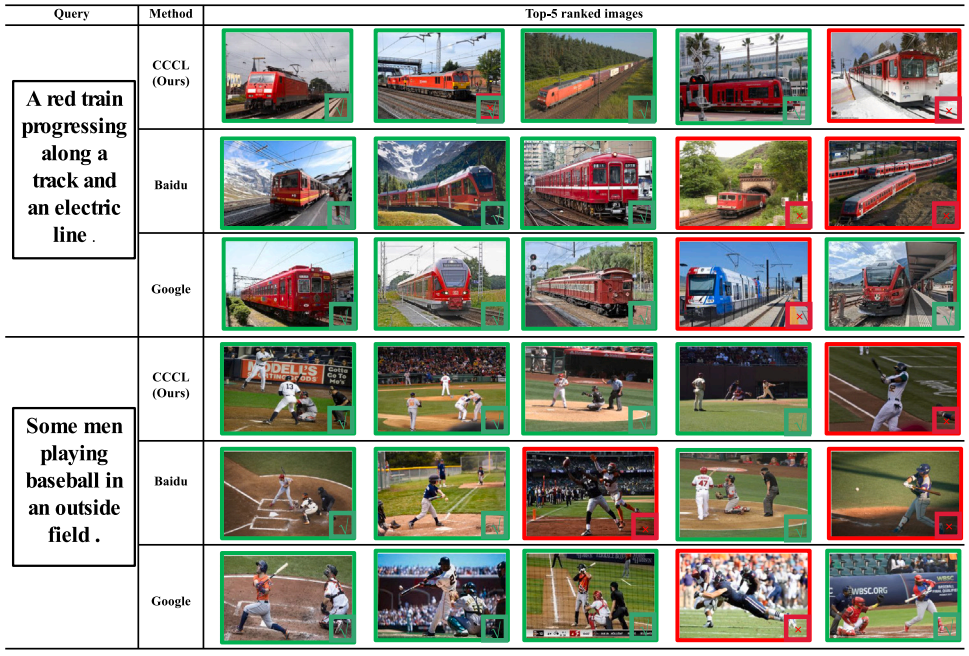
After a comprehensive analysis of the top five retrieved results in both I→T and T→I directions, as shown in Fig. 7, our observations are centered on three key aspects. Firstly, all of the top three items retrieved by CCCL are accurate, while there are some inaccuracies in the top three items retrieved by Baidu and Google. Secondly, CCCL exhibits superior performance among the top five search results compared to Baidu and Google. Thirdly, CCCL, along with Baidu and Google, effectively captures semantic elements. Even the incorrectly retrieved items may contain correct semantic elements, such as "giraffe", "skiing", "red train", and "baseballer".

The reason why CCCL performs better than Baidu and Google lies in the following aspects. Firstly, CCCL completely and correctly perceives complex contextual information by iteratively reasoning out the intrinsic semantic relationships within one modality to learn intra-modal correlations. Secondly, CCCL demonstrates proficiency in learning inter-modal correlations through one-to-one and one-to-many interactions across various modalities. However, due to the complexity of the calculation process of CCCL, it requires more search time compared to Baidu and Google.

[3] https://www.baidu.com
[4] https://www.google.com



(a) I→T



(b) T→I

Fig. 7. Examples of the retrieval results at I→T and T→I on the MS-COCO dataset with CCCL.

5. Conclusion

In this paper, we propose the CCCL framework for ITR to achieve three research objectives: (1) Perceiving intricate contextual information, (2) Reasoning out intrinsic semantic relationships, and (3) Aligning instances or their patches. Particularly, CCCL framework achieves the above three objectives through three steps. In step 1, it incorporates self-attention and gate mechanism to adaptively learn context-perceived patch embeddings for each modality. In step 2, it deeply mines the intra-modal correlation to make connections, draw inferences, and form associations between different patches of instance within the same modality. In step 3, it learns more complete inter-modal alignment from both global and local levels. Experimental results prove the effectiveness of CCCL.

In future research, we plan to explore (1) Cross-scale alignment, aiming to dynamically match specific image regions to words, phrases, or even entire sentences, and (2) Adaptive assignment of importance values for different branches based on visual and textual contents.

CRediT authorship contribution statement

Zheng Liu: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration,

Writing – original draft, Writing – review & editing. **Xinlei Pei:** Software, Writing – original draft, Writing – review & editing. **Shanshan Gao:** Funding acquisition, Methodology, Supervision, Validation. **Changhao Li:** Software, Writing – original draft, Writing – review & editing. **Jingyao Wang:** Software, Writing – original draft, Writing – review & editing. **Junhao Xu:** Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by Humanities and Social Sciences Project of Education Ministry, China (20YJA870013), Natural Science Foundation of Shandong Province, China (ZR2019MF016, ZR2020MF037), Scientific Research Studio in Colleges and Universities of Ji'nan City, China (202228105, 2021GXRC092), Introduction and Education Plan of Young Creative Talents in Colleges and Universities of Shandong Province, Innovation Team of Youth Innovation Science and Technology Plan in Colleges and Universities of Shandong Province, China (2020KJN007).

We greatly appreciate the authors who provide source codes of the compared methods, and also express gratitude for the anonymous reviewers for thoroughly reading the paper and providing thoughtful comments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.

[2] Y. Peng, X. Huang, Y. Zhao, An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2017) 2372–2385, http://dx.doi.org/10.1109/TCSVT.2017.2705068.

[3] X. Liu, Y.-m. Cheung, Z. Hu, Y. He, B. Zhong, Adversarial tri-fusion hashing network for imbalanced cross-modal retrieval, IEEE Trans. Emerg. Top. Comput. Intell. 5 (4) (2021) 607–619, http://dx.doi.org/10.1109/TETCI.2020.3007143.

[4] X. Xu, F. Shen, Y. Yang, H.T. Shen, X. Li, Learning discriminative binary codes for large-scale cross-modal retrieval, IEEE Trans. Image Process. 26 (5) (2017) 2494–2507, http://dx.doi.org/10.1109/TIP.2017.2676345.

[5] P. Kaur, H.S. Pannu, A.K. Malhi, Comparative analysis on cross-modal information retrieval: A review, Comp. Sci. Rev. 39 (2021) 100336, http://dx.doi.org/10.1016/j.cosrev.2020.100336.

[6] X. Xu, H. Lu, J. Song, Y. Yang, H.T. Shen, X. Li, Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval, IEEE Trans. Cybern. 50 (6) (2020) 2400–2413, http://dx.doi.org/10.1109/TCYB.2019.2928180.

[7] A. Karpathy, A. Joulin, L.F. Fei-Fei, Deep fragment embeddings for bidirectional image sentence mapping, in: Advances in Neural Information Processing Systems, vol. 27, 2014.

[8] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Hierarchical multimodal lstm for dense visual-semantic embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1881–1889, http://dx.doi.org/10.1109/ICCV.2017.208.

[9] Y. Peng, J. Qi, X. Huang, Y. Yuan, CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network, IEEE Trans. Multimed. 20 (2) (2017) 405–420, http://dx.doi.org/10.1109/TMM.2017.2742704.

[10] Y. Liu, Y. Guo, E.M. Bakker, M.S. Lew, Learning a recurrent residual fusion network for multimodal matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4107–4116, http://dx.doi.org/10.1109/ICCV.2017.442.

[11] Y. Song, M. Soleymani, Polysemous visual-semantic embedding for cross-modal retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1979–1988, http://dx.doi.org/10.1109/CVPR.2019.00208.

[12] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, J. Han, 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, 2020, pp. 12655–12663, http://dx.doi.org/10.1109/CVPR42600.2020.01267.

[13] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, J. Shao, Camp: Cross-modal adaptive message passing for text-image retrieval, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5763–5772, http://dx.doi.org/10.1109/ICCV.2019.00586.

[14] Q. Cheng, X. Gu, Bridging multimedia heterogeneity gap via graph representation learning for cross-modal retrieval, Neural Netw. 134 (2021) 143–162, http://dx.doi.org/10.1016/j.neunet.2020.11.011.

[15] F. Faghri, D.J. Fleet, J.R. Kiros, S. Fidler, VSE++: Improving visual-semantic embeddings with hard negatives, in: Proceedings of the British Machine Vision Conference, BMVC, 2018, http://dx.doi.org/10.48550/arXiv.1707.05612.

[16] L. Wang, Y. Li, J. Huang, S. Lazebnik, Learning two-branch neural networks for image-text matching tasks, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 394–407, http://dx.doi.org/10.1109/TPAMI.2018.2797921.

[17] N. Sarafianos, X. Xu, I.A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5814–5824, http://dx.doi.org/10.1109/ICCV.2019.00591.

[18] C. Liu, Z. Mao, A.-A. Liu, T. Zhang, B. Wang, Y. Zhang, Focus your attention: A bidirectional focal attention network for image-text matching, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, 2019, pp. 3–11, http://dx.doi.org/10.1145/3343031.3350869.

[19] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, W.-Y. Ma, Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6609–6618, http://dx.doi.org/10.1109/CVPR.2019.00677.

[20] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 201–216, http://dx.doi.org/10.1007/978-3-030-01225-0_13.

[21] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, Visual semantic reasoning for image-text matching, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4654–4662, http://dx.doi.org/10.1109/ICCV.2019.00475.

[22] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 664–676, http://dx.doi.org/10.1109/TPAMI.2016.2598339.

[23] H. Nam, J.-W. Ha, J. Kim, Dual attention networks for multimodal reasoning and matching, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2156–2164, http://dx.doi.org/10.1109/CVPR.2017.232.

[24] H. Chen, G. Ding, Z. Lin, S. Zhao, J. Han, Cross-modal image-text retrieval with semantic consistency, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1749–1757, http://dx.doi.org/10.1145/3343031.3351055.

[25] K. Zhang, Z. Mao, Q. Wang, Y. Zhang, Negative-aware attention framework for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15640–15649, http://dx.doi.org/10.1109/CVPR52688.2022.01521.

[26] Y. Wang, H. Yang, X. Bai, X. Qian, L. Ma, J. Lu, B. Li, X. Fan, PFAN++: Bi-directional image-text retrieval with position focused attention network, IEEE Trans. Multimed. 23 (2021) 3362–3376, http://dx.doi.org/10.1109/TMM.2020.3024822.

[27] J. Gu, J. Cai, S.R. Joty, L. Niu, G. Wang, Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7181–7189, http://dx.doi.org/10.1109/CVPR.2018.00750.

[28] Y. Wu, S. Wang, G. Song, Q. Huang, Learning fragment self-attention embeddings for image-text matching, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2088–2096, http://dx.doi.org/10.1145/3343031.3350940.

[29] J. Chen, H. Hu, H. Wu, Y. Jiang, C. Wang, Learning the best pooling strategy for visual semantic embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15784–15793, http://dx.doi.org/10.1109/CVPR46437.2021.01553.

[30] K. Li, Y. Zhang, K. Li, Y. Li, Y. Fu, Image-text embedding learning via visual and textual semantic reasoning, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2023) 641–656, http://dx.doi.org/10.1109/TPAMI.2022.3148470.

[31] S. Wang, R. Wang, Z. Yao, S. Shan, X. Chen, Cross-modal scene graph matching for relationship-aware image-text retrieval, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1497–1506, http://dx.doi.org/10.1109/WACV45572.2020.9093614.

[32] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, Graph structured network for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10918–10927, http://dx.doi.org/10.1109/CVPR42600.2020.01093.

[33] Q. Zhang, Z. Lei, Z. Zhang, S.Z. Li, Context-aware attention network for image-text retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3533–3542, http://dx.doi.org/10.1109/CVPR42600.2020.00359.

[34] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10938–10947, http://dx.doi.org/10.1109/CVPR42600.2020.01095.

[35] H. Diao, Y. Zhang, L. Ma, H. Lu, Similarity reasoning and filtration for image-text matching, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, (no. 2) 2021, pp. 1218–1226, http://dx.doi.org/10.1609/aaai.v35i2.16209.

[36] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'06, vol. 2, IEEE, 2006, pp. 1735–1742, http://dx.doi.org/10.1109/CVPR.2006.100.

[37] M. Zhou, Z. Niu, L. Wang, Z. Gao, Q. Zhang, G. Hua, Ladder loss for coherent visual-semantic embedding, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, (no. 07) 2020, pp. 13050–13057, http://dx.doi.org/10.1609/aaai.v34i07.7006.

[38] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823, http://dx.doi.org/10.1109/CVPR.2015.7298682.

[39] E. Ustinova, V. Lempitsky, Learning deep embeddings with histogram loss, Adv. Neural Inf. Process. Syst. 29 (2016).

[40] C. Huang, C.C. Loy, X. Tang, Local similarity-aware deep feature embedding, in: Advances in Neural Information Processing Systems, vol. 29, 2016.

[41] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5017–5025, http://dx.doi.org/10.1109/CVPR.2019.00516.

[42] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems, vol. 26, 2013.

[43] X. Xu, T. Wang, Y. Yang, L. Zuo, F. Shen, H.T. Shen, Cross-modal attention with semantic consistence for image–text matching, IEEE Trans. Neural Netw. Learn. Syst. 31 (12) (2020) 5412–5425, http://dx.doi.org/10.1109/TNNLS.2020.2967597.

[44] C. Zhang, J. Song, X. Zhu, L. Zhu, S. Zhang, Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 17 (1s) (2021) 1–22, http://dx.doi.org/10.1145/3412847.

[45] B. Wang, Y. Yang, X. Xu, A. Hanjalic, H.T. Shen, Adversarial cross-modal retrieval, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 154–162, http://dx.doi.org/10.1145/3123266.3123326.

[46] L. Zhang, L. Chen, C. Zhou, X. Li, F. Yang, Z. Yi, Weighted graph-structured semantics constraint network for cross-modal retrieval, IEEE Trans. Multimed. (2023) 1–14, http://dx.doi.org/10.1109/TMM.2023.3282894.

[47] C. Liu, Y. Zhang, H. Wang, W. Chen, F. Wang, Y. Huang, Y.-D. Shen, L. Wang, Efficient token-guided image-text retrieval with consistent multimodal contrastive training, IEEE Trans. Image Process. 32 (2023) 3622–3633, http://dx.doi.org/10.1109/TIP.2023.3286710.

[48] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard, D. McClosky, The stanford corenlp natural language processing toolkit, in: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.

[49] T. Yao, Y. Pan, Y. Li, T. Mei, Hierarchy parsing for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2621–2629, http://dx.doi.org/10.1109/ICCV.2019.00271.

[50] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[51] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086, http://dx.doi.org/10.1109/CVPR.2018.00636.

[52] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D.A. Shamma, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, Int. J. Comput. Vis. 123 (1) (2017) 32–73, http://dx.doi.org/10.1007/s11263-016-0981-7.

[53] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[54] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, http://dx.doi.org/10.48550/arXiv.1810.04805, arXiv preprint arXiv:1810.04805.

[55] R. Dey, F.M. Salem, Gate-variants of Gated Recurrent Unit (GRU) neural networks, in: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems, MWSCAS, IEEE, 2017, pp. 1597–1600, http://dx.doi.org/10.1109/MWSCAS.2017.8053243.

[56] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, http://dx.doi.org/10.48550/arXiv.1609.02907, arXiv preprint arXiv:1609.02907.

[57] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2021) 4–24, http://dx.doi.org/10.1109/TNNLS.2020.2978386.

[58] J. Tang, C. Deng, G.-B. Huang, Extreme learning machine for multilayer perceptron, IEEE Trans. Neural Netw. Learn. Syst. 27 (4) (2016) 809–821, http://dx.doi.org/10.1109/TNNLS.2015.2424995.

[59] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344, http://dx.doi.org/10.1109/CVPR.2016.149.

[60] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[61] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78, http://dx.doi.org/10.1162/tacl_a_00166.

[62] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, http://dx.doi.org/10.48550/arXiv.1412.6980, arXiv preprint arXiv:1412.6980.