Learning Semantic Relationship among Instances for Image-Text Matching

Zheren Fu[1], Zhendong Mao[1,2,*], Yan Song[1], Yongdong Zhang[1,2]
[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
fzr@mail.ustc.edu.cn, {zdmao, songyan, zhyd73}@ustc.edu.cn

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.01455

Abstract

Image-text matching, a bridge connecting image and language, is an important task, which generally learns a holistic cross-modal embedding to achieve a high-quality semantic alignment between the two modalities. However, previous studies only focus on capturing fragment-level relation within a sample from a particular modality, e.g., salient regions in an image or text words in a sentence, where they usually pay less attention to capturing instance-level interactions among samples and modalities, e.g., multiple images and texts. In this paper, we argue that sample relations could help learn subtle differences for hard negative instances, and thus transfer shared knowledge for infrequent samples should be promising in obtaining better holistic embeddings. Therefore, we propose a novel hierarchical relation modeling framework (HREM), which explicitly capture both fragment- and instance-level relations to learn discriminative and robust cross-modal embeddings. Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art ones by 4%-10% in terms of rSum. Our code is available at https://github.com/CrossmodalGroup/HREM.

1. Introduction

Image-text matching bridges the semantic gap between visual and textual modalities and is a fundamental task for various multi-modal learning applications, such as cross-modal retrieval [22] and text-to-image synthesis [17]. The critical challenge is accurately and efficiently learning cross-modal embeddings and their similarities for images and texts, to achieve a high-quality semantic alignment. In general, existing image-text matching methods can be classified into two paradigms. The first *embedding-based* matching [4, 10, 20, 35] separately encodes the whole images and texts into a holistic embedding space, then globally measures the semantic similarity of the two modalities. The second *score-based* matching [3, 7, 19, 27] applies the cross-modal interaction between visual and textual local features, then learns a cumulative similarity score.

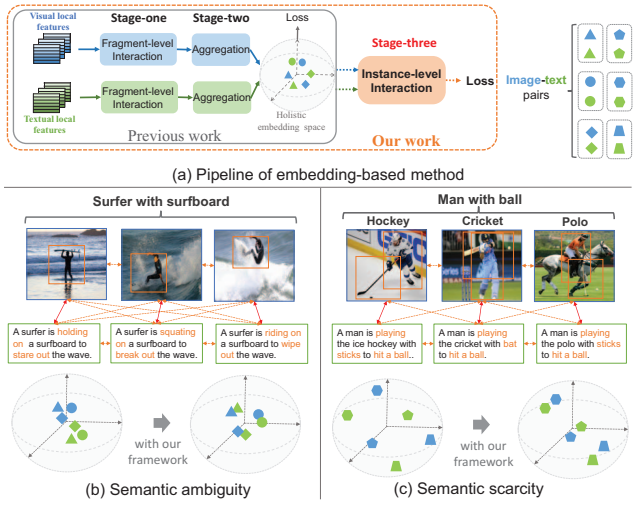


Figure 1. Illustration of our motivation. Sample relation modeling improves the holistic representation of cross-modal learning. Colors and shapes indicate different modalities and image-text pairs, respectively. Orange elements mark effective interactions: (a) The pipeline of the previous and our work, we add the cross-modal relation interaction between samples. (b) For the identical theme of "surfer with surfboard", specific behaviors exist subtle differences, like "hold/squat/ride on the surfboard" and "stare/break/wipe out the wave". Our method distinguishes these hard negative samples from semantic ambiguities. (c) For similar themes under "man play a ball", corresponding behaviors usually are semantic similar, like "play the hockey/cricket/polo" all need to "hit the ball" with "sticks/bats". Our method improves learning embeddings on these infrequent samples with semantic scarcities for themselves.

Recently, embedding-based methods have served as the mainstream solution owing to both accuracy and efficiency in image-text matching, which contains *two steps* as shown in Fig. 1 (a): (1) Capturing the intra-modal relation between visual fragments (*e.g.,* regional features) or textual

*Corresponding author.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.01455

fragments (*e.g.,* word features) independently, then enhancing the semantic representation of local features. (2) Aggregating relation-enhanced local features of two modalities into the holistic embedding space. For the first step, most use the graph neural network [5, 20, 21] or attention module [35, 45, 46] to capture semantic relations and enhance the local features of two modalities, respectively. Some work further exploits the spatial relation [45] for visual regions or grammatical relation [28] for textual words. For the second step, they design pooling functions [4] or sequence encoders [21] to aggregate local features and get holistic embeddings. Existing embedding-based methods follow the principle of separately encoding images and texts by two branches as Fig. 1 (a). In each branch, these methods only focus on the fragment-level relation modeling and local features interaction within one sample, *e.g.*, the region features inside one image (or the word features inside one text). In this way, the instance-level relation modeling and global embeddings interaction among different samples and modalities, *e.g.*, holistic embeddings of multiple images and texts, are entirely overlooked.

Consequently, existing embedding-based methods directly use global embeddings to compute loss function, *e.g.*, hard negative triplet loss [10] on random mini-batch, which is insufficient to exploit the manifold structure of holistic embedding space [39]. First, they fail to learn subtle semantic discrepancies among different samples (*e.g.*, similar behaviors with an identical theme as shown in Fig. 1 (b)), then can not distinguish hard negative samples with semantic ambiguities because of the heterogeneity of visual and textual semantics. Second, they are unable to transfer shared knowledge from diverse samples (*e.g.*, different samples that contain similar behaviors with similar themes as shown in Fig. 1 (c)), then can not effectively learn on these infrequent samples with semantic scarcities. Therefore, it is expected that a framework should precisely capture the sample relationship to learn better cross-modal embeddings, while does not break the principle of embedding-based methods, *i.e.*, independently encodes embeddings without modality interaction at the inference stage.

In doing so, we propose a **H**ierarchical **RE**lation **M**odeling framework (HREM) that, for the first time to our knowledge, explicitly captures both fragment-level and instance-level relations to learn holistic embeddings jointly. Therefore, HREM learns not only contextual semantics among intra-modal fragments to enhance local features, but also the associated semantics among inter-modal instances to distinguish hard negative samples and improve learning on infrequent samples. As illustrated in Fig. 1 (a) and Fig. 2, we propose a novel step (*i.e.*, the "stage-three") to exactly capture the semantic relationship of cross-modal samples. First, we propose a novel cross-embedding association graph, which explicitly identifies the connection relation and learns the relevance relation between batch samples with fragment-level semantic matching. Next, we propose two relation interaction mechanisms, which explore inter-modal and intra-modal relations synchronously or asynchronously with our improved attention modules to obtain enhanced embeddings. Consequently, HREM only needs to capture the instance-level relation for training, then encode multi-modal embeddings independently at the inference stage, to achieve high accuracy and efficiency for image-text matching.

To summarize, the major contributions are as follows: (1) We propose a hierarchical relation modeling framework (HREM) for image-text matching. To the best of our knowledge, this is the first work that explicitly captures both fragment-level relations within modality and instance-level relations across modalities. (2) We propose a novel cross-embedding association graph by identifying the connection relation and learning the relevance relation. (3) We propose two relation interaction mechanisms to learn the relation-enhanced embeddings. (4) HREM outperforms all state-of-the-art methods for image-text retrieval on two widely used benchmarks, Flickr30K and MS-COCO, by 4%-10% rSum.

2. Related Work

2.1. Image-Text Matching

According to how the cross-modal interaction is implemented, image-text matching methods are divided into two categories, *embedding-based* and *score-based* matching. **Embedding-based.** It independently encodes images and sentences into a holistic embedding space by two branches, where the semantic similarity is calculated by cosine similarity [10]. Existing work usually utilizes the GCN [20, 44] or self-attention layer [45, 46] to capture semantic relations between fragments inside one sample and enhance contextual semantics of local features, then propose a particular aggregator to learn global embeddings. For example, VSRN [20] proposes a semantic reasoning network to learn local visual features with key scene concepts. CVSE [43] proposes a consensus-aware module to integrate commonsense knowledge into local features for two modalities. GPO [4] presents a generalized pooling function to project local features into the global embedding. MV [24] proposes a multi-view encoder to learn multiple embeddings for one image and models intra-class variations.

Score-based. It conducts fine-grained cross-modal interaction and semantic alignment between local fragments, then calculates a cumulative similarity score. [3, 7, 19, 51]. For example, SCAN [19] proposes an attention mechanism for cross-modal interaction between visual and textual fragments. IMRAM [3] proposes an iterative network for multiple steps of cross-modal interaction. NAAF [51] measures the similarity and dissimilarity degrees via two matching

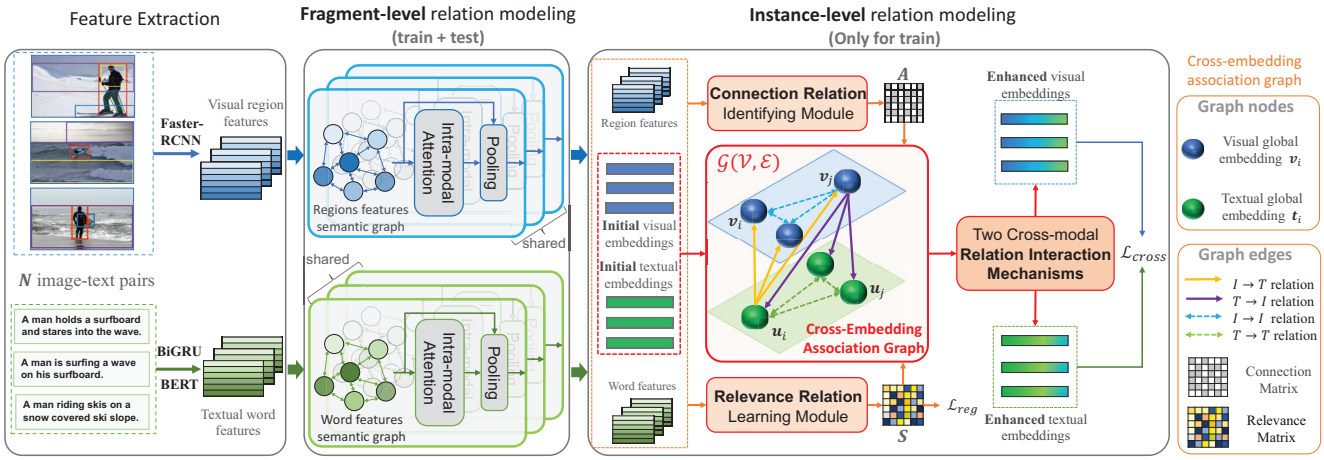


Figure 2. Overview of our hierarchical relation modeling framework (HREM). Given N image-text pairs ($N = 3$ in this Figure), we first capture the fragment-level relations and learn the relation-enhanced local features for each image or text independently, then aggregate local features by pooling operation to get global embeddings $\{\boldsymbol{v}_i, \boldsymbol{u}_i\}_{i=1}^{N}$. Next, we propose a novel cross-embedding association graph to capture instance-level relations by identifying the connection relation and learning the relevance relation between samples. Finally, we propose two cross-modal relation interaction mechanisms to get relation-enhanced embeddings and compute the final loss function.

mechanisms to infer the overall similarity jointly.

Nonetheless, existing embedding-based methods only capture the relation between local fragments within each sample, In addition, we explore the relation across different samples and modalities to learn better holistic embeddings.

2.2. Sample Relationship Learning

Among data scarcity and diversity, sample relations have been explicitly explored for representation enhancement [23, 49]. A simple way is to generate input as new samples from existing samples [13], such as mix-up [49], and cut-mix [48]. Some work designs the delicate objective function for output, such as various metric learning losses [12, 14]. Another way is using the knowledge distillation between samples, such as transferring invariant knowledge for zero-shot learning [32] and domain generalization [33]. Recently, some methods focus on the view of batch interaction [8, 9, 16]. IBC [39] constructs a fully connected graph for mini-batch samples and classifies each sample employing a message passing network. HIST [25] constructs a hypergraph to formulate higher-order relations between samples. However, these methods study the primary vision task, *e.g.*, classification. We capture the sample relationship for cross-modal learning and bridge semantic discrepancies between visual and textual modalities.

3. The Proposed Method

The overview of HREM is depicted in Fig. 2. We first introduce feature extraction in Sec. 3.1 and fragment-level relation modeling in Sec. 3.2. Then we introduce instance-level relation modeling in Sec. 3.3. Finally, we describe the optimization in Sec. 3.4 and discussion in Sec. 3.5.

3.1. Feature Extraction

Visual Representation. Given an image I, we use the bottom-up-attention network [1] to extract the salient regions by the Faster-RCNN [37] and get the region features by the pre-trained ResNet-101 [15]. Then we add a fully-connect (FC) layer to map each region to a d-dimensional local feature. We denote as $R = \{\boldsymbol{r}_1, \cdots, \boldsymbol{r}_{n_r}\} \in \mathbb{R}^{n_r \times d}$, which is the visual fragments and local features for the image I, n_r is the number of region features.

Textual Representation. Given a sentence T, we use the sequence models, bi-directional gated recurrent unit (Bi-GRU) [38], or pre-trained BERT [6] to extract the set of word features. We also add an FC layer to keep the same dimension with images. We denote as $C = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{n_c}\} \in \mathbb{R}^{n_c \times d}$, which is the textual fragments and local features for the text, n_c is the number of word features.

3.2. Fragment-level Relation Modeling

To capture contextual information between fragments and enhance local features of two modalities introduced by Sec. 3.1, we propose fragment-level relation modeling for visual regions and textual words, respectively.

Visual Regions. We construct a semantic relation graph between visual regions within one image, and propose the relation interaction module to learn the contextual semantics for region features. The graph nodes are region features R, and edges are their semantic relation. Specifically, we use the graph attention network [42], *e.g.*, self-attention layer [41], to capture the semantic relation and learn relation-enhanced local features. It first maps original

features to queries and key-value pairs, then use a weighted sum of the values as outputs, where the weighting depends on scaled dot-product between queries and keys:

$$Att\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_k}}\right)\boldsymbol{V}, \quad (1)$$

where Att is called the scaled dot-product attention [29], d_k is the dimension of input features. In this case, we build up a fully-connected graph, and the semantic relation is implicitly expressed in the attention weight. After initial region features $R=\{\boldsymbol{r}_1,\cdots,\boldsymbol{r}_{n_r}\}\in\mathbb{R}^{n_r\times d}$ pass through the self-attention layer with the feed-forward network to learn contextual information, we obtain the relation-enhanced region features $R^V=\{\boldsymbol{r}_1^V,\cdots,\boldsymbol{r}_{n_r}^V\}\in\mathbb{R}^{n_r\times d}$.

Finally, we aggregate initial and enhanced region features to get the global visual embedding $\boldsymbol{v}\in\mathbb{R}^d$, by using the maximum pooling and average pooling, with the parameter β to control the ratio of two representations.

$$\boldsymbol{v}=\beta\cdot MaxPool(R)+(1-\beta)\cdot AvgPool(R^V), \quad (2)$$

Textual Words. Similarly, we construct a semantic relation graph between textual words within one sentence, and use the relation interaction module to enhance contextual information for the word features. The nodes are word features $C=\{\boldsymbol{c}_1,\cdots,\boldsymbol{c}_{n_c}\}\in\mathbb{R}^{n_c\times d}$ and the edges are their semantic relation. We also use the self-attention layer to implement, hence we get relation-enhanced word features $C^U=\{\boldsymbol{c}_1^U,\cdots,\boldsymbol{c}_{n_c}^U\}\in\mathbb{R}^{n_c\times d}$, then aggregate them to get the global textual embedding $\boldsymbol{u}\in\mathbb{R}^d$, like Eq. (2).

$$\boldsymbol{u}=\beta\cdot MaxPool(C)+(1-\beta)\cdot AvgPool(C^U), \quad (3)$$

3.3. Instance-level Relation Modeling

We propose the instance-level relation modeling for multiple images and texts to learn better cross-modal embeddings obtained by Sec. 3.2. Given N image-text pairs and their embeddings $\{\boldsymbol{v}_i,\boldsymbol{u}_i\}_{i=1}^N$, we propose a novel cross-embedding association graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, where the nodes are the embeddings $\mathcal{V}=\{\boldsymbol{v}_1,...,\boldsymbol{v}_N,\boldsymbol{u}_1...,\boldsymbol{u}_N\}\in\mathbb{R}^{2N\times d}$, the edges $\mathcal{E}$ are pairwise semantic relations.

3.3.1 Cross-Embedding Association Graph

The critical challenge is how to construct the pairwise relation accurately. Without loss of generality, we divide the relation into two parts: *connection* and *relevance*.

We use the matrix $\boldsymbol{A}\in\mathbb{R}^{2N\times 2N}$ to represent the connection relation, *i.e.*, whether exists an associated edge between nodes. We use the matrix $\boldsymbol{S}\in\mathbb{R}^{2N\times 2N}$ to represent the relevance relation, *i.e.*, the degree of semantic association between nodes. Further, we divide these matrices into two patterns and four blocks: intra-modal relation (**I**mage-to-**I**mage $I\to I$, **T**ext-to-**T**ext $T\to T$) and inter-modal relation (**I**mage-to-**T**ext $I\to T$, **T**ext-to-**I**mage $T\to I$), the shape of each block is equal to $\mathbb{R}^{N\times N}$.

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{I\to I} & \boldsymbol{A}_{I\to T}\\ \boldsymbol{A}_{T\to I} & \boldsymbol{A}_{T\to T}\end{bmatrix},\ \boldsymbol{S}=\begin{bmatrix}\boldsymbol{S}_{I\to I} & \boldsymbol{S}_{I\to T}\\ \boldsymbol{S}_{T\to I} & \boldsymbol{S}_{T\to T}\end{bmatrix}, \quad (4)$$

Connection Relation Identifying. The simple idea is to build a fully connected graph like Sec. 3.2 and previous work [5,45] that elements of $\boldsymbol{A}$ are all 1, However, the fully-connection easily likely leads to over-smoothing and noisy association [39] for nodes. We need to identify the effective connection and filter out unassociated semantics.

Therefore, we construct the neighbor space of embedding to identify the connection relation. If two embedding nodes are close, their semantic information usually overlaps, they probably have a semantic connection [25].

$$\begin{aligned}(\boldsymbol{A}_{I\to I})_{ij}&=1 \text{ if } \boldsymbol{v}_j\in\mathcal{N}_{intra}\left(\boldsymbol{v}_i\right) \text{ else } 0,\\ (\boldsymbol{A}_{T\to T})_{ij}&=1 \text{ if } \boldsymbol{u}_j\in\mathcal{N}_{intra}\left(\boldsymbol{u}_i\right) \text{ else } 0,\end{aligned} \quad (5)$$

where $\mathcal{N}_{intra}$ is the neighbor space of embeddings for intra-modal connection. We select the first τN (τ is range from 0 to 1) nearest single-modal samples ranked by the embedding similarity to construct $\mathcal{N}_{intra}$ in Eq. (5). Correspondingly, we define $\mathcal{N}_{inter}$ for inter-modal connection:

$$\begin{aligned}(\boldsymbol{A}_{I\to T})_{ij}&=1 \text{ if } \boldsymbol{u}_j\in\mathcal{N}_{inter}\left(\boldsymbol{v}_i\right) \text{ else } 0,\\ (\boldsymbol{A}_{T\to I})_{ij}&=1 \text{ if } \boldsymbol{v}_j\in\mathcal{N}_{inter}\left(\boldsymbol{u}_i\right) \text{ else } 0,\end{aligned} \quad (6)$$

Since the heterogeneity between visual and textual semantics, directly using global embeddings is insufficient to identify the inter-modal connection relation. Following score-based methods [19, 40], we use the fine-grained matching of fragments and local features to measure the inter-modal neighbor space. As shown in Fig. 3, given an image-text pair and its region-word similarity matrix. We first pick up the most matching textual word (or visual region) for each region (or each word), then average these matched scores to express the overall matching values p.

$$\begin{aligned}p_{I\to T}&=\frac{1}{n_r}\sum_{m=1}^{n_r}\max_{n\in[1,n_c]}\left(\boldsymbol{r}_m^{\mathrm{T}}\boldsymbol{c}_n\right),\\ p_{T\to I}&=\frac{1}{n_c}\sum_{m=1}^{n_c}\max_{n\in[1,n_r]}\left(\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{r}_n\right),\end{aligned} \quad (7)$$

where we select the first τN nearest cross-modal samples ranked by the fragment-level matching values Eq. (7) to construct the inter-modal neighbor space $\mathcal{N}_{inter}$ in Eq (6).

Relevance Relation Learning. The relevance relation is the degree of semantic association between two connected nodes. Existing work [16, 39] uses the global embedding similarity to approximate, which is insufficient to bridge the

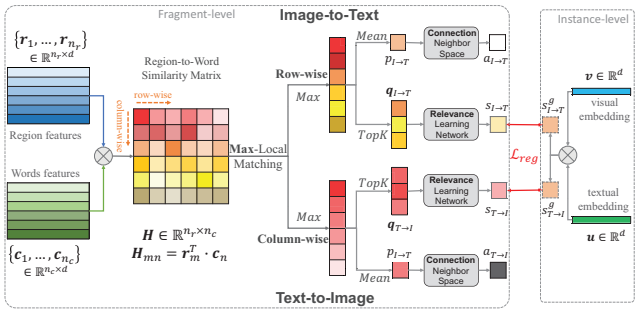


Figure 3. Graph construction of inter-modal relation with fragment-level matching. For every image-text pair, we use corresponding modules to get connection relation $a_{I\to T}$ ($a_{T\to I}$) and relevance relation $s_{I\to T}$ ($s_{T\to I}$), which are elements of connection matrix $\boldsymbol{A}_{I\to T}$ ($\boldsymbol{A}_{T\to I}$) and relevance matrix $\boldsymbol{S}_{I\to T}$ ($\boldsymbol{S}_{T\to I}$).

semantic discrepancy between visual and textual space [19]. In contrast, we explicitly learn the cross-modal relevance relation. As shown in Fig. 3, we employ fine-grained matching like Eq. (7). After picking up the max matching of row-wise and column-wise, we select the top-K scores and concatenate them as the matching vectors $\boldsymbol{q}$.

$$\begin{aligned}\boldsymbol{q}_{I\to T} &= TopK(\{\max_{n\in[1,n_c]}\left(\boldsymbol{r}_m^{\mathrm{T}}\boldsymbol{c}_n\right)\}_{m=1}^{n_r}),\\ \boldsymbol{q}_{T\to I} &= TopK(\{\max_{n\in[1,n_r]}\left(\boldsymbol{c}_m^{\mathrm{T}}\boldsymbol{r}_n\right)\}_{m=1}^{n_c}),\end{aligned} \tag{8}$$

where $\boldsymbol{q}\in\mathbb{R}^K$, then we use an MLP and add the overall matching values to learn the inter-modal relevance (scalar).

$$\begin{aligned}s_{I\to T} &= MLP(\boldsymbol{q}_{I\to T}) + p_{I\to T},\\ s_{T\to I} &= MLP(\boldsymbol{q}_{T\to I}) + p_{T\to I},\end{aligned} \tag{9}$$

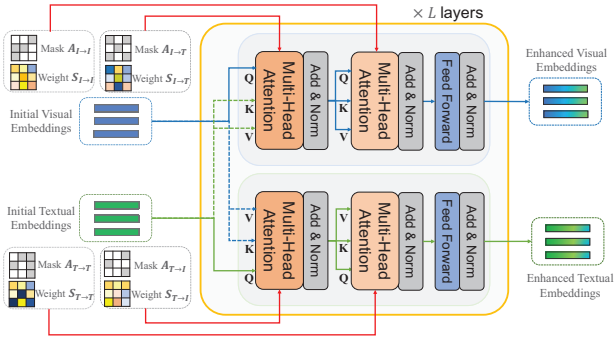
For different image-text pairs, we compute the corresponding relevance values, which are the elements of the inter-modal relevance matrix $\boldsymbol{S}_{I\to T}$ and $\boldsymbol{S}_{T\to I}$. As for the intra-modal relevance relation, we can use the global embeddings to compute the relevance matrix.

$$(\boldsymbol{S}_{T\to T})_{ij} = e^{-\frac{\|\boldsymbol{v}_i-\boldsymbol{v}_j\|_2^2}{\sigma}},\ (\boldsymbol{S}_{I\to I})_{ij} = e^{-\frac{\|\boldsymbol{u}_i-\boldsymbol{u}_j\|_2^2}{\sigma}}, \tag{10}$$

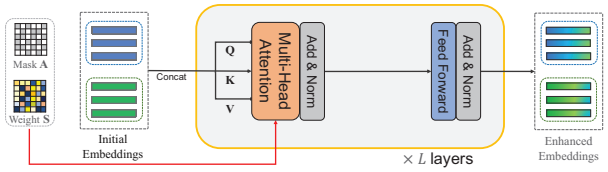
where σ is a positive scalar to control the relevance values (σ=1 for simplicity). Besides, we add a relevance regularization loss to ensure the learned inter-modal relevance matrix does not crash during training [18].

$$\mathcal{L}_{reg} = \mathcal{L}_{kl}\left(\boldsymbol{S}_{I\to T}, \boldsymbol{S}_{I\to T}^{g}\right) + \mathcal{L}_{kl}\left(\boldsymbol{S}_{T\to I}, \boldsymbol{S}_{T\to I}^{g}\right), \tag{11}$$

where $\mathcal{L}_{kl}$ is the KL-divergence loss with softmax function, $(\boldsymbol{S}_{I\to T})_{ij}^{g} = e^{-\frac{\|\boldsymbol{v}_i-\boldsymbol{u}_j\|_2^2}{\sigma}}$, $(\boldsymbol{S}_{T\to I})_{ij}^{g} = e^{-\frac{\|\boldsymbol{u}_i-\boldsymbol{v}_j\|_2^2}{\sigma}}$ are computed by the global embeddings like Eq. (10). We hope the learned inter-modal relevance relation is close to the semantic relation of global embeddings.



(a) Standalone mechanism. Explore inter-modal and intra-modal relationship between visual and textual embeddings asynchronously.



(b) Fusion mechanism. Explore inter/intra-modal relation synchronously.

Figure 4. Two relation interaction mechanisms based on how to explore the inter-modal and intra-modal relations. The connection matrix $\boldsymbol{A}$ and the relevance matrix $\boldsymbol{S}$ are applied to the attention modules wholly or separately, as Eq. (12).

3.3.2 Relation Interaction Mechanisms

After constructing our cross-embedding association graph in Sec. 3.3.1, we design two relation interaction mechanisms to capture the semantic relations between images and texts, where the embeddings are updated by the information interaction process, as shown in Fig. 4.

Fusion Mechanism. We concatenate the visual and textual embeddings as input, as shown in Fig. 4b. The inter-modal and intra-modal relation interactions are conducted synchronously. The embeddings first pass through the multi-head self-attention module for attention diversity. Besides, we adopt the feed-forward network module for relation reasoning, which a multi-layer perceptron implements [41]. It is similar to the fragment-level interaction modules in Sec. 3.2. We also add the residual connection [15] and layer normalization [2] after them.

The connection matrix $\boldsymbol{A}$ is an attention mask matrix for the attention module, where the zero positions are not allowed to attend while the non-zero positions will be unchanged [41]. The relevance matrix $\boldsymbol{S}$ is an extra attention weight matrix as the explicit relation modeling, we use λ to balance $\boldsymbol{S}$ with the original attention weight matrix. Therefore, we revise the basic attention formula Eq. (1) to:

$$Att(\boldsymbol{QKV};\boldsymbol{A},\boldsymbol{S}) = \underset{s.t.\ mask(\boldsymbol{A})}{softmax}\left(\frac{\boldsymbol{QK}^T}{\sqrt{d_k}} + \lambda\boldsymbol{S}\right)\boldsymbol{V}, \tag{12}$$

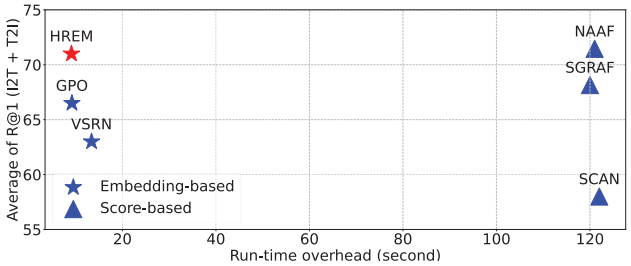


Figure 5. The comparison between accuracy and speed for cross-modal retrieval. We perform all methods (*Region+ BiGRU* based) on the whole Flickr30K test-set with one RTX3090 GPU.

Standalone Mechanism. As shown in Fig. 4a, the visual and textual embeddings are sent into two branches and get relation interaction. The embeddings first get inter-modal relation interaction by the multi-head cross-attention module, where $\boldsymbol{Q}$ and $\boldsymbol{K}, \boldsymbol{V}$ come from two modalities. Then they get intra-modal relation interaction by the multi-head self-attention module, where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ come from the same modality. Finally, the enhanced embeddings are output by following the feed-forward network module.

The connection matrix $\boldsymbol{A}$ and relevance matrix $\boldsymbol{S}$ first are divided into pre-defined four blocks as Eq. (4), then each block is applied to corresponding modules as Eq. (12). Specifically, the inter-modal relation parts perform on the first cross-attention module, and the intra-modal relation parts act on the second self-attention module.

After the L layers of the relation interaction mechanism, we get the final relation-enhanced embeddings of two modalities, $\{\overline{\boldsymbol{v}}_1, ..., \overline{\boldsymbol{v}}_N\}$ and $\{\overline{\boldsymbol{u}}_1 ..., \overline{\boldsymbol{u}}_N\}$.

3.4. Optimization

Neighbor Batch Sampling. To ensure effective relation interaction in Sec. 3.3, we propose a neighbor sampling to replace random sampling for batches at the later training. We use the k-means clustering [30] on visual embeddings, then randomly choose P clusters and select K images from each cluster, batch size $N = P \times K$. Finally, we select one positive text for each image to get N image-text pairs.

Objective Function. We use the triplet loss [10], the similarity score is the cosine similarity between visual embedding $\boldsymbol{v}$ and textual embedding $\boldsymbol{u}$, $s(\boldsymbol{v}, \boldsymbol{u}) = \frac{\boldsymbol{v}^\top \boldsymbol{u}}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{u}\|}$.

$$\mathcal{L} = [\alpha - s(\boldsymbol{v}, \boldsymbol{u}) + s(\boldsymbol{v}, \boldsymbol{u}^-)]_+ + [\alpha - s(\boldsymbol{v}, \boldsymbol{u}) + s(\boldsymbol{v}^-, \boldsymbol{u})]_+, \tag{13}$$

where α represents a margin parameter, $[x]_+ = \max(x, 0)$. $(\boldsymbol{v}, \boldsymbol{u})$ is a positive image-text pair, and $(\boldsymbol{v}, \boldsymbol{u}^-), (\boldsymbol{v}^-, \boldsymbol{u})$ are negative image-text pairs in the batch. We use the distance-weighted sampling [31] for hard negative mining.

We not only use the relation-enhanced embeddings to compute matching loss as Eq. (13), but also add the initial embeddings for matching loss to keep the embedding consistency, since we need to encode embeddings directly without sample interaction at the inference stage.

Table 1. Comparisons of image-text retrieval on MS-COCO 5K test-set. *Region* represents using region features [1] for images. *BiGRU* [38] and *BERT* [6] represent using their word features for texts. *E* and *S* indicate *embedding-based* and *score-based* methods, respectively. $*$ shows the ensemble results of two models.

Type	Method	MS-COCO 5K						rSum
		IMG → TEXT			TEXT → IMG			
		R@1	R@5	R@10	R@1	R@5	R@10	
Region + BiGRU								
S	IMRAM$^*_{2020}$ [3]	53.7	83.2	91.0	39.7	69.1	79.8	416.5
S	UARDA$^*_{2022}$ [50]	56.2	83.8	91.3	40.6	69.5	80.9	422.3
S	NAAF$^*_{2022}$ [51]	58.9	85.2	92.0	**42.5**	70.9	81.4	430.9
E	GPO$_{2021}$ [4]	56.6	83.6	91.4	39.3	69.9	81.1	421.9
E	CGMN$_{2022}$ [5]	53.4	81.3	89.6	41.2	71.9	82.4	419.8
E	MV$_{2022}$ [24]	56.7	84.1	91.4	40.3	70.6	81.6	424.6
E	**HREM (Standalone)**	58.4	85.5	92.4	39.8	70.5	81.0	427.6
E	**HREM (Fusion)**	58.9	85.3	92.1	40.0	70.6	81.2	428.1
E	**HREM (Full)***	**60.6**	**86.4**	**92.5**	41.3	**71.9**	**82.4**	**435.1**
Region + BERT								
S	SSAMT$_{2021}$ [11]	57.7	84.2	90.8	40.8	70.5	80.5	424.5
S	DIME$^*_{2021}$ [36]	59.3	85.4	91.9	43.1	73.0	83.1	435.8
S	DCPA$_{2022}$ [40]	53.5	82.4	90.2	40.4	71.0	82.0	419.5
E	DSRAN$^*_{2021}$ [45]	55.3	83.5	90.9	41.7	72.7	82.8	426.9
E	GPO$_{2021}$ [4]	58.3	85.3	92.3	42.4	72.7	83.2	434.3
E	VSRN++$_{2022}$ [21]	54.7	82.9	90.9	42.0	72.2	82.7	425.4
E	**HREM (Standalone)**	61.8	87.0	93.2	44.0	73.7	83.4	443.1
E	**HREM (Fusion)**	62.3	87.6	93.4	43.9	73.6	83.3	444.1
E	**HREM (Full)***	**64.0**	**88.5**	**93.7**	**45.4**	**75.1**	**84.3**	**450.9**

$$\mathcal{L}_{cross} = \mathcal{L}(\overline{\boldsymbol{v}}, \overline{\boldsymbol{u}}) + [\mathcal{L}(\boldsymbol{v}, \boldsymbol{u}) + \mathcal{L}(\overline{\boldsymbol{v}}, \boldsymbol{u}) + \mathcal{L}(\boldsymbol{v}, \overline{\boldsymbol{u}})], \tag{14}$$

Finally, we combine the cross-embedding matching loss Eq. (14) with the relevance regularization loss Eq. (11).

3.5. Discussion

Inference Stage. Since we may not have the batch data in the actual application, our framework can encode the cross-modal embeddings without sample interaction at the inference stage. The instance-level relation modeling is only for training. Intuitively, when we train the embedding encoding network and the sample interaction network together with the end-to-end manner and the consistent loss in Sec. 3.4, the encoding network will also be improved with the helpful supervision of embedding interaction.

Time Complexity. Two matching methods have different time complexity in cross-modal retrieval. Given N image-text pairs, separate encoding makes the time complexity of embedding-based methods to be $O(2N)$, while cross-modal interaction makes score-based to be $O(N^2)$. Given one query and the set to be retrieved N samples, the time complexity of query retrieval is $O(1)$ for embedding-based, but is $O(N)$ for score-based. Hence *score-based* methods usually sacrifice the retrieval speed for the performance boost. However, our method can achieve both highly accurate and efficient retrieval, as shown in Fig. 5.

Table 2. Comparisons of image-text retrieval performances on Flickr30K and MS-COCO 1K test-set. *Region* represents using Faster-RCNN [37] to extract region features [1] for images. *BiGRU* [38] and *BERT* [6] represent using them to extract word features for texts. We list the existing state-of-the-art *embedding-based* image-text matching methods. $*$ indicates the ensemble results of two models.

Method	Flickr30K 1K							MS-COCO 1K						
	IMG → TEXT			TEXT → IMG			rSum	IMG → TEXT			TEXT → IMG			rSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Region + BiGRU														
VSRN$^*_{2019}$ [20]	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
CVSE$_{2020}$ [43]	73.5	92.1	95.8	52.9	80.4	87.8	482.4	74.8	95.1	98.3	59.9	89.4	95.2	512.7
GPO$_{2021}$ [4]	76.5	94.2	97.7	56.4	83.4	89.9	498.1	78.5	96.0	98.7	61.7	90.3	95.6	520.8
MV$_{2022}$ [24]	79.0	94.9	97.7	59.1	84.6	90.6	505.8	78.7	95.7	98.7	62.7	90.4	95.7	521.9
HREM (Fusion)	79.5	94.3	97.4	59.3	85.1	91.2	506.8	80.0	96.0	98.7	62.7	90.1	95.4	522.8
HREM (Full)*	**81.4**	**96.5**	**98.5**	**60.9**	**85.6**	**91.3**	**514.3**	**81.2**	**96.5**	**98.9**	**63.7**	**90.7**	**96.0**	**527.1**
Region + BERT														
CAMERA$^*_{2020}$ [35]	78.0	95.1	97.9	60.3	85.9	91.7	508.9	77.5	96.3	98.8	63.4	90.9	95.8	522.7
DSRAN$^*_{2021}$ [45]	77.8	95.1	97.6	59.2	86.0	91.9	507.6	78.3	95.7	98.4	64.5	90.8	95.8	523.5
GPO$_{2021}$ [4]	81.7	95.4	97.6	61.4	85.9	91.5	513.5	79.7	96.4	98.9	64.8	91.4	96.3	527.5
VSRN++$_{2022}$ [21]	79.2	94.6	97.5	60.6	85.6	91.4	508.9	77.9	96.0	98.5	64.1	91.0	96.1	523.6
HREM (Fusion)	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7
HREM (Full)*	**84.0**	**96.1**	**98.6**	**64.4**	**88.0**	**93.1**	**524.2**	**82.9**	**96.9**	**99.0**	**67.1**	**92.0**	**96.6**	**534.6**

4. Experiments

4.1. Experimental Setup

Datasets & Metrics. We choose the typical Flickr30K [47] and MS-COCO [26] datasets, where each image is associated with five texts. Flickr30K contains 29,000, 1,000, and 1,014 training, testing, and validation images, respectively. MS-COCO contains 82,738, 5,000, and 5,000 training, testing, and validation images, respectively. The results of MS-COCO are tested on averaging over 5-folds of 1K test images and on the entire 5K test images. Following [4], We evaluate performances by the metric, *R@K* and *rSum*.

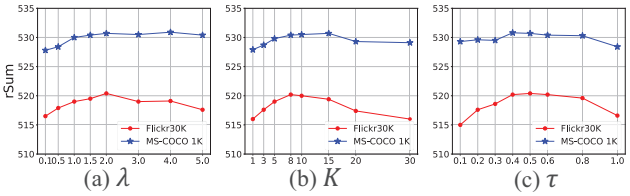
Implementation Details. The embedding dimension $d = 1024$ as previous work [4]. We use the pre-extracted region features [1] for images, and we use the BiGRU [38] with GloVe [34] or BERT-base [6] to extract textual features. The batch size $N = 128$ for Flickr30K and $N = 256$ for MS-COCO. The layer of interaction mechanism $L = 1$, the hyper-parameters as $\beta = 0.8$, $\tau = 0.5$, $\lambda = 1.5$, $K = 10$, and the margin of triplet loss $\alpha = 0.2$.

4.2. Comparison with State-of-the-art Methods

We follow the standard evaluation protocols [51] on two datasets. We first perform two proposed interaction mechanisms individually, then report the ensemble results.

On Flickr30K. Quantitative results on Flickr30K 1K test-set are shown in Tab. 2, where our proposed method outperforms all state-of-the-art embedding-based image-text matching methods [20, 21, 24, 35, 45] with impressive margins for the R@K and rSum. Furthermore, our method has the coincident superiority on different textual encoders [6, 38] and still gets the best on the ensemble results.

On MS-COCO. Performances on MS-COCO 1K and 5K test-set are shown in Tab. 1 and Tab. 2. Our method performs best compared with existing state-of-the-art methods [4, 35, 45] on the 1K test set. Furthermore, on the more extensive database of 5K test set, our method outperforms previous work [21, 50, 51] with larger performance gaps, both embedding-based and score-based methods, which shows the superiority of our method more convincingly.



(a) λ (b) K (c) τ

Figure 6. Effect of hyper-parameters. (a) The weight of relevance matrix λ for Eq. (12), (b) The number of local matching K for Eq. (8), (c) The threshold of neighbor space τ for Eq. (5, 6).

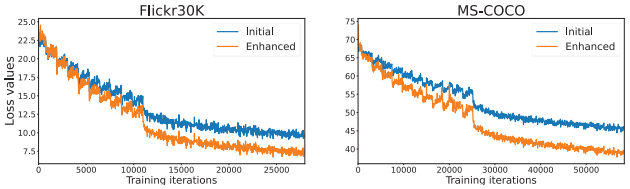


Figure 7. The curve of triplet loss with of hard negative mining. The loss values are computed by enhanced and initial embeddings, which are the first and second item of Eq. (14), respectively.

4.3. Ablation Study & Robustness Analysis

By default, we perform the experiments on our framework of *fusion mechanism* with *Region+BERT* settings.

Table 3. Comparison of different module ablation for our framework. '✓' means retaining it (or otherwise removing it).

(a) The ablation study of hierarchical relation modeling on Flickr30K

Fragment-level		Instance-level		IMG → TEXT		TEXT → IMG	
Visual	Textual	Intra-modal	Inter-modal	R@1	R@5	R@1	T R@5
		✓	✓	81.5	94.9	62.3	86.2
	✓	✓	✓	81.8	95.1	62.4	86.2
✓		✓	✓	83.1	95.9	63.2	87.0
✓	✓			80.1	94.8	60.9	85.3
✓	✓	✓		80.5	94.7	61.2	85.2
✓	✓		✓	82.2	95.5	62.6	86.4
✓	✓	✓	✓	83.3	96.0	63.5	87.1

(b) The ablation study of instance-level relation modeling on Flickr30K

Methods	IMG → TEXT		TEXT → IMG	
	R@1	R@5	R@1	T R@5
w/o connection matrix $\mathbf{A}$	81.4	95.1	61.5	86.3
w/o relevance matrix $\mathbf{S}$	81.7	95.3	61.9	86.5
w/o consistency $\mathcal{L}_{cross}$	81.6	95.6	61.8	86.7
w/o regularization $\mathcal{L}_{reg}$	82.8	95.8	62.9	86.9
w/o neighbor batch sampling	82.8	95.9	63.1	87.0
HREM	83.3	96.0	63.5	87.1

Hyper-parameters. Fig. 6 shows the effect of three critical parameters, which demonstrates the significance of capturing sample relations and the effectiveness of our proposed modules. λ is the attention coefficient for the relevance matrix, thus it should be large enough to provide extra fine-grained supervision, K is related to the number of words in variable sentences, thus it cannot be too large to adapt short texts. τ represents the ratio of connected samples in the batch, it will not be meaningful if close to 0 or 1. Finally, the performances are relatively stable when all parameter values change in a proper range, our method is insensitive to hyper-parameter selection.

Hierarchical Relation. To better verify the effectiveness of hierarchical relation modeling, we provide the ablation study in Tab. 3a. It shows that both instance-level and fragment-level relation modeling help improve the learning of cross-modal embeddings. First, capturing the instance-level relation, especially the inter-modal relation, is more critical than the intra-modal and fragment-level relation. Besides, capturing fragment-level relations for the textual modality seems redundant. We believe that word features have already learned contextual semantics by the textual encoders, *e.g.*, BiGRU [38] and BERT [6]. However, exploring the semantic relations for region features are meaningful as previous work [20, 21, 46]. Finally, modeling all hierarchical relations is also essential for the optimal result.

Instance-level Relation. Tab. 3b shows the ablation of our instance-level relation modeling. First, the connection and relevance matrices are significant in explicitly capturing the cross-modal relation. And the cross-embedding matching loss $\mathcal{L}_{cross}$ is indispensable to keep the embedding consistency at the inference stage. Besides, the regularization loss $\mathcal{L}_{reg}$ can ensure the training stability to improve performances. Finally, the neighbor batch sampling adapts to mine more potential connected samples from mini-batch.

Table 4. R@1 comparison on hard samples with semantic ambiguity and infrequent samples with semantic scarcity in different proportions, selected by the *percentile rank* of average embedding similarity between themselves and other samples in the dataset. '✔' means training with sample relation modeling.

Samples Type	Percentile Rank	Sample Interaction	Flickr30K		MS-COCO 5K	
			I → T	T → I	I → T	T → I
Hard	5%		74.6	32.0	36.4	8.5
		✔	**76.0**	**34.8**	**40.4**	**9.2**
	10%		72.0	38.8	39.2	12.3
		✔	**75.0**	**40.4**	**44.6**	**13.6**
	20%		74.5	45.1	43.6	18.5
		✔	**78.0**	**47.1**	**49.3**	**19.5**
Infrequent	5%		79.2	38.4	50.8	16.7
		✔	**84.0**	**40.8**	**59.2**	**19.4**
	10%		78.0	41.2	53.0	19.8
		✔	**81.2**	**43.6**	**61.0**	**22.6**
	20%		75.0	46.6	54.8	23.9
		✔	**82.0**	**48.4**	**61.0**	**26.2**

Special Samples. Tab. 4 shows the performance comparison of hard samples with semantic ambiguity and infrequent samples with semantic scarcity in the test set (hard and infrequent samples will get high average similarity with other negative samples in datasets). We find our method can improve these special samples significantly. Fig. 7 shows the curve of triplet loss with hard negative mining, the values of enhanced embeddings are lower than the initial, proving the capacity for recognizing hard samples.

Retrieval Speed. We show the trade-off between performance and computation for cross-modal retrieval in Fig. 5. Although our method belongs to the embedding-based image-text matching methods [4, 20], achieves both high accuracy and efficiency, and is more than 10 times faster than the latest score-based methods [7, 19, 51] on Flickr30K.

5. Conclusion

This paper proposes a novel hierarchical relation modeling framework (HREM) for image-text matching. HREM not only captures fragment-level relations within the single modality and sample, but also effectively exploits instance-level relations across different modalities and samples to learn better holistic embeddings. Based on our design, HREM encodes embedding without interaction on sample or modality at the inference stage, thus achieving high efficiency on cross-modal retrieval. Extensive experiments on two benchmarks show the superiority of our method.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62222212, 62121002.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3, 6, 7

[2] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 5

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12652–12660, 2020. 1, 2, 6

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Chang Lian Wang. Learning the best pooling strategy for visual semantic embedding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15784–15793, 2021. 1, 2, 6, 7, 8

[5] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. Cross-modal graph matching network for image-text retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 18(4), mar 2022. 2, 4, 6

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. 3, 6, 7, 8

[7] Haiwen Diao, Ying Zhang, Lingyun Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *ArXiv*, abs/2101.01368, 2021. 1, 2, 8

[8] Ismail Elezi, Jenny Seidenschwarz, Laurin Wagner, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss++: A deeper look into group loss for deep metric learning. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 3

[9] Ismail Elezi, Sebastiano Vascon, Alessandro Torcinovich, Marcello Pelillo, and Laura Leal-Taixé. The group loss for deep metric learning. *ArXiv*, abs/1912.00385, 2020. 3

[10] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 1, 2, 6

[11] Zhihao Fan, Zhongyu Wei, Zejun Li, Siyuan Wang, Haijun Shan, Xuanjing Huang, and Jianqing Fan. Constructing phrase-level semantic labels to form multi-grained supervision for image-text retrieval. *ArXiv*, abs/2109.05523, 2021. 6

[12] Zheren Fu, Yan Li, Zhendong Mao, Quan Wang, and Yongdong Zhang. Deep metric learning with self-supervised ranking. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021. 3

[13] Zheren Fu, Zhendong Mao, Bo Hu, An-An Liu, and Yongdong Zhang. Intra-class adaptive augmentation with neighbor correction for deep metric learning. *IEEE Transactions on Multimedia*, pages 1–14, 2022. 3

[14] Zheren Fu, Zhendong Mao, Chenggang Clarence Yan, An an Liu, Hongtao Xie, and Yongdong Zhang. Self-supervised synthesis ranking for deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:4736–4750, 2022. 3

[15] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3, 5

[16] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7256–7266, 2022. 3, 4

[17] Mengqi Huang, Zhendong Mao, Penghui Wang, Quang Wang, and Yongdong Zhang. Dse-gan: Dynamic semantic evolution generative adversarial network for text-to-image generation. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 1

[18] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Self-taught metric learning without labels. *CVPR*, 2022. 5

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 1, 2, 4, 5, 8

[20] Kunpeng Li, Yulun Zhang, K. Li, Yuanyuan Li, and Yun Raymond Fu. Visual semantic reasoning for image-text matching. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4653–4661, 2019. 1, 2, 7, 8

[21] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Raymond Fu. Image-text embedding learning via visual and textual semantic reasoning. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 2, 6, 7, 8

[22] Pandeng Li, Hongtao Xie, Jiannan Ge, Lei Zhang, Shaobo Min, and Yongdong Zhang. Dual-stream knowledge-preserving hashing for unsupervised video retrieval. In *European Conference on Computer Vision*, pages 181–197. Springer, 2022. 1

[23] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5208–5217, 2021. 3

[24] Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. In *IJCAI*, 2022. 2, 6, 7

[25] Jongin Lim, Sangdoo Yun, Seulki Park, and Jin Young Choi. Hypergraph-induced semantic tuplet loss for deep metric learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 212–222, 2022. 3, 4

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 7

[27] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020. 1

[28] Xin Liu, Yi He, Yiu ming Cheung, Xing Xu, and N. Wang. Learning relationship-enhanced semantic graph for fine-grained image-text matching. *IEEE transactions on cybernetics*, PP, 2022. 2

[29] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *ArXiv*, abs/1508.04025, 2015. 4

[30] J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297, 1967. 6

[31] R. Manmatha, Chaoxia Wu, Alex Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867, 2017. 6

[32] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 953–962, 2021. 3

[33] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *ArXiv*, abs/1904.12347, 2019. 3

[34] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 7

[35] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. Context-aware multi-view summarization network for image-text matching. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 1, 2, 7

[36] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 6

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. 3, 7

[38] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45:2673–2681, 1997. 3, 6, 7, 8

[39] Jenny Seidenschwarz, Ismail Elezi, and Laura Leal-Taix'e. Learning intra-batch connections for deep metric learning. In *ICML*, 2021. 2, 3, 4

[40] Zhan Shi, Tianzhu Zhang, Xiaoyan Wei, Feng Wu, and Yongdong Zhang. Decoupled cross-modal phrase-attention network for image-sentence matching. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, PP, 2022. 4, 6

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 5

[42] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2018. 3

[43] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lingyun Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, 2020. 2, 7

[44] Sijin Wang, Ruiping Wang, Ziwei Yao, S. Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1497–1506, 2020. 2

[45] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:2866–2879, 2021. 2, 4, 6, 7

[46] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. 2, 8

[47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 7

[48] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Young Joon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 3

[49] Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018. 3

[50] Kun Zhang, Zhendong Mao, Anan Liu, and Yongdong Zhang. Unified adaptive relevance distinguishable attention network for image-text matching. *IEEE Transactions on Multimedia*, pages 1–1, 2022. 6, 7

[51] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2022. 2, 6, 7, 8