Negative-Aware Attention Framework for Image-Text Matching

Kun Zhang[1], Zhendong Mao[1]*, Quan Wang[2], Yongdong Zhang[1]
[1]University of Science and Technology of China, Hefei, China; [2]MOE Key Laboratory of Trustworthy Distributed Computing and Service, Beijing University of Posts and Telecommunications, Beijing, China
kkzhang@mail.ustc.edu.cn, {zdmao, zhyd73}@ustc.edu.cn, wangquan@bupt.edu.cn

Abstract

Image-text matching, as a fundamental task, bridges the gap between vision and language. The key of this task is to accurately measure similarity between these two modalities. Prior work measuring this similarity mainly based on matched fragments (i.e., word/region with high relevance), while underestimating or even ignoring the effect of mismatched fragments (i.e., word/region with low relevance), e.g., via a typical LeaklyReLU or ReLU operation that forces negative scores close or exact to zero in attention. This work argues that mismatched textual fragments, which contain rich mismatching clues, are also crucial for image-text matching. We thereby propose a novel Negative-Aware Attention Framework (NAAF), which explicitly exploits both the positive effect of matched fragments and the negative effect of mismatched fragments to jointly infer image-text similarity. NAAF (1) delicately designs an iterative optimization method to maximally mine the mismatched fragments, facilitating more discriminative and robust negative effects, and (2) devises the two-branch matching mechanism to precisely calculate similarity/dissimilarity degrees for matched/mismatched fragments with different masks. Extensive experiments on two benchmark datasets, i.e., Flickr30K and MSCOCO, demonstrate the superior effectiveness of our NAAF, achieving state-of-the-art performance. Code will be released at: https://github.com/CrossmodalGroup/NAAF.

1. Introduction

Image-text matching, which devotes to bridging the semantic gap between these two heterogeneous modalities, is a fundamental task in computer vision (CV) and natural language processing (NLP). This matching task aims to search images for a given textual description or find texts w.r.t. an image query. The critical challenge of image-text matching lies in accurately learning semantic correspondence between images and texts to measure their similarity.

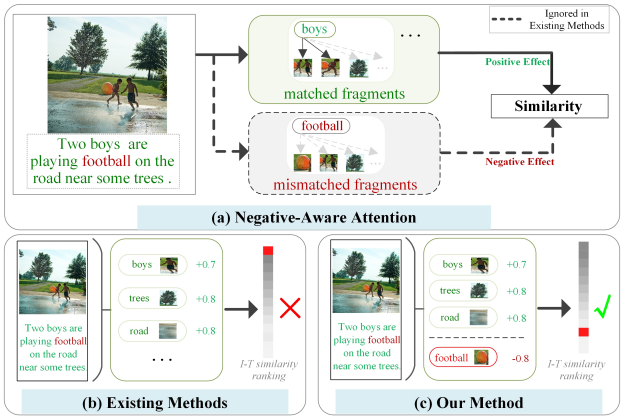

Figure 1. Motivation of the negative-aware attention. (a) Existing methods mainly find matched fragments, *e.g.*, "boys", "trees", to compute image-text (I-T) similarity, while the effect of mismatched fragments, *e.g.*, "football", is weakened or ignored, by the typical LeaklyReLU or ReLU. (b) shows the false-positive problem of existing methods, where the I-T pair can still obtain a high similarity, contributed by most matched fragments, and may rank quite the top as correct. (c) In our method, both mismatched and matched fragments are mined to produce negative and positive effects, respectively, thereby downgrading the false-positive pairs.

Generally, there are two paradigms among existing image-text matching approaches [2, 7, 27]. The first one tends to perform global-level matching, *i.e.*, finding the semantic correspondence between the full text and the whole image [3, 9, 24, 43]. They typically project the holistic images and texts into a common latent space and then match the two modalities. The second paradigm focuses on examining local-level matching, *i.e.*, matching between salient regions in images and words in texts [19, 23]. Local-level matching takes into account fine-grained semantic correspondence between images and texts.

Recently, attention-based local-level matching has been proposed and quickly becomes the mainstream in image-text matching. SCAN [23], as well as its various variants [2, 5, 7, 13, 16, 26, 27, 39], is a representative method of this

*Zhendong Mao is the corresponding author.

kind. The key idea is to discover all word-region alignments by attending to relevant fragments w.r.t. each query fragment from another modality. In summary, matched fragments (*i.e.*, word-region pairs with high relevance scores) will contribute a lot to the final image-text similarity, while the effect of mismatched fragments (*i.e.*, word-region pairs with low relevance scores) will be weakened or even erased, *e.g.*, via a typical LeakyReLU or ReLU that forces negative scores close or exact to zero during the attention process [2, 5, 7, 13, 16, 23, 26, 27, 39]. Although achieving promising performance, these methods completely ignore the crucial role of mismatched textual fragments in proving image-text mismatching, since they describe contents not in the image. (In fact, images usually contain more background object regions, thus we principally focus on mismatched textual fragments, *i.e.*, words.)

Consequently, existing methods, which mainly find the matched fragments while underestimating or neglecting the effect of mismatched ones, will be inevitably prone to produce false-positive matching. Namely, image-text pairs containing many matched fragments but a few mismatched textual fragments (directly indicating image-text mismatching), can still obtain the high similarity and may rank quite the top as correct, which is certainly not a satisfying result (Fig. 1(b)). Therefore, we argue that a reasonable matching framework should simultaneously consider two aspects, *i.e.*, the overall matching score of an image-text pair is determined not only by the positive effect of matched fragments, but also by the negative effect of mismatched ones (*e.g.*, words not mentioned in the image will probably downgrade the overall matching score). For example, as shown in Fig. 1(c), by further emphasizing and mining the negative effect of mismatched fragments w.r.t. "$football$", it will be easy to eliminate this false-positive pair.

To this end, we propose a novel negative-aware attention framework which, for the first time to our knowledge, explicitly considers both positively matched and negatively mismatched fragments to jointly measure image-text similarity (Fig. 1(a)). Different from conventional matching mechanisms that focus on matched fragments unilaterally, our attention framework can effectively mine mismatched textual fragments and use them to accurately reflect how dissimilar the two modalities are. In this sense, we call it negative-aware attention framework (NAAF). As illustrated in Fig. 2, NAAF consists of two modules. (1) We devise a two-branch matching to solve the lack utilization of mismatched fragments, which contains the negative and positive attention with different masks, one to precisely calculate the dissimilarity degrees of mismatched fragments, and the other the similarity degrees of matched ones. (2) We propose a new iterative optimization method to explicitly model and mine the mismatched fragments. Concretely, based on the similarity distributions of mismatched and matched fragments, we first adaptively learn the optimal boundary between them by minimizing the penalty probability of their error overlaps, which can theoretically guarantee the mining accuracy. Then, the learned boundary is integrated into the attention matching process to optimize more discriminative similarity distributions. Such iterative optimizing forcedly separates these two types of distributions as far as possible, enabling maximally mining the mismatched textual fragments. In this way, NAAF not only focuses on matched fragments but also discriminates subtle mismatched ones across modalities towards more accurate image-text matching.

The major contributions of this work are summarized as follows. 1) We propose a novel two-branch matching module, which jointly utilizes both mismatched and matched textual fragments to make accurate image-text matching. To the best of our knowledge, this is the first framework that explicitly exploits both negative effects of mismatched clues and positive effects of matched clues in image-text matching. 2) We propose a novel iterative optimization method with negative mining strategies, which can explicitly drive more negative effects of mismatched fragments, and theoretically guarantee the mining accuracy, yielding more comprehensive and interpretable image-text similarity measurement. 3) Extensive experiments on two benchmarks, *i.e.*, Flickr30K and MS-COCO, show that NAAF outperforms compared methods. Analyses also well demonstrate the superiority and reasonableness of our method.

2. Related Work

Recently, image-text matching has been dramatically developed, where there are general two research lines: global-level matching, which tends to learn the global alignment, *i.e.*, representing the image or text as a holistic feature to measure similarity; and local-level matching, which focuses on the fine-grained alignment between local fragments, *i.e.*, inferring the overall image-text similarity by the relevance of all word-region pairs. NAAF belongs to the latter one.

Global-level matching methods. A general solution in this field is to learn semantic alignment in image-text pairs by mapping them into a shared space and optimizing via a ranking loss. A line of researches focus on different optimization functions [8, 9, 21, 36, 41]. The famous hinge-based triplet loss forces aligned image-text pairs have a higher similarity than misaligned ones [18, 21]. Faghri *et al.* [9] improve the performance of triplet loss by attending to the hardest misaligned pairs. Wang *et al.* [36] consider the external constraint loss that preserves the neighborhood structure in a single modality. Recently, some novel optimization designs are proposed, such as the ladder loss [45], the polynomial loss [40], and the adaptive offline quintuplet loss [4]. In addition to hash search for efficiency [6, 28, 46], many approaches focus on designing specific learning networks for

improving search accuracy [11, 20, 24, 29, 32]. For example, recent work focuses on the aggregation strategy for the holistic presentation of image or text, where a promising way is the generalized pooling operator [3].

Local-level matching methods. Learning semantic alignment between image regions and text words is popular in image-text matching [15, 17, 19]. Karpathy *et al*. [19] first attempt to optimize the most similar region-word pairs for selecting matched semantics. Lots of works [2,5,7,13,16,23,26,27,37–39,42,44] are devoted to discovering full region-word alignments. One of the most typical approaches is attention-based SCAN [23] that attends to the specific regions/words fragments to filter out unmatched information for enhancing semantic alignment learning.

A line of methods [14, 27, 37, 39] focus on mining more information in images and texts to further enhance the association of matched fragments, such as regional position embedding [39] and object relationship information [37]. Shi *et al*. [33] use the knowledge graph of semantic concepts to improve image representation. Another branch works focus on designing more sophisticated models [2, 5, 7, 26, 27, 44], such as the focal attention to eliminate irrelevant fragments [26], the dual-path recurrent neural network that considers the relation between image and objects like the relation between text and words [5], and the iterative recurrent attention for serial multiple-step matching [2]. Recently, Liu *et al*. [27] and Diao *et al*. [7] further employ the graph neural network to enhance the meaningful alignments. Nonetheless, they mainly concentrate on maximizing the effect of matched (*i.e*. aligned) fragments, while underestimating or neglecting the clue role of mismatched ones. In contrast, we explicitly mine the mismatched textual fragments to further exploit both types of clues for joint similarity inference.

3. Method

The overall framework of our proposed NAAF is depicted in Fig. 2. We first extract features of image regions and text words, and then perform negative-aware attention to measure image-text similarity, using both negative and positive effects. In this section, we first introduce the proposed negative-aware attention and elaborate on its modules in Sec. 3.1. Then, we describe the objective function and feature extraction in Sec. 3.2 and Sec. 3.3, respectively.

Notations. Formally, for an image-text pair (U, V), the text is represented as words' textual features $U = \{u_i | i \in [1, m], u_i \in \mathbb{R}^d\}$, and the image is represented as regions' visual features $V = \{v_j | j \in [1, n], v_j \in \mathbb{R}^d\}$, where m and n denote the number of words and regions, respectively; d is the dimension of feature representation.

3.1. Negative-aware Attention

Given an image-text pair, it may contain rich matched and mismatched fragments. Our goal is to take full advantage of the two types of clues to achieve more accurate matching performance. There are mainly two modules in our NAAF framework, which are 1) Discriminative Mismatch Mining (Sec. 3.1.1), aiming to explicitly model and maximally mine mismatched fragments, by minimizing the penalty probability of error overlaps between the matched and mismatched similarity distributions in the training process; and 2) Neg-Pos Branch Matching (Sec. 3.1.2), aiming to precisely calculate the effects of both negative mismatches and positive matches for jointly inferring similarity via the designed two-branch matching, *i.e*., negative and positive attention branches. Next, we will introduce these two modules in detail.

3.1.1 Discriminative Mismatch Mining

Different from existing methods that do not explore the precise similarity boundary of mismatched and matched fragments, which implicitly use empirical fixed zero to distinguish them, *i.e*., via the typical ReLU or LeakyReLU operation [23, 44], we expect to explicitly and adaptively model the similarity distributions of mismatched and matched fragments, aiming to maximally separate them to achieve effective mismatched fragments mining.

To this end, in the training process, regarding the mismatched and matched word-region fragment pairs, we first sample their similarity degrees as:

$$S_k^- = [s_1^-, s_2^-, s_3^-, \ldots, s_i^-, \ldots], \quad (1)$$

$$S_k^+ = [s_1^+, s_2^+, s_3^+, \ldots, s_i^+, \ldots], \quad (2)$$

where S_k^- and S_k^+ are defined as as the sets of mismatched word-region similarity s_i^- and matched word-region similarity s_i^+, respectively. Note that both sets S_k^- and S_k^+ are dynamically updated with index k in training. As a matter of fact, it is challenging for sampling s_i^- and s_i^+, as no matching annotations exist for fragment-level word-region pairs. We solve this problem by the delicately designed sampling strategy, which is described in Sec. 3.1.3.

According to the sampled two sets S_k^- and S_k^+, the mismatched and matched probability distributions about similarity s of word-region fragment pairs can be modeled as: $f_k^-(s) = \frac{1}{\sigma_k^- \sqrt{2\pi}} e^{[-\frac{(s-\mu_k^-)^2}{2(\sigma_k^-)^2}]}$, $f_k^+(s) = \frac{1}{\sigma_k^+ \sqrt{2\pi}} e^{[-\frac{(s-\mu_k^+)^2}{2(\sigma_k^+)^2}]}$, where (μ_k^-, σ_k^-) and (μ_k^+, σ_k^+) are the mean and standard deviation of the two distributions respectively. Assume that there is a boundary t to distinguish whether the similarity of a word-region is mismatched or matched. As illustrated in Fig. 2, the distinction errors are twofold, *i.e*., truly mismatched fragments are distinguished as matched ones (depicted as E_1 in Fig. 2) and vice versa (depicted as E_2 in Fig. 2). We target to learn an optimal boundary that can maximally distinguish the mismatched fragments, while also decrease the error probability, *i.e*., E_1 and E_2,

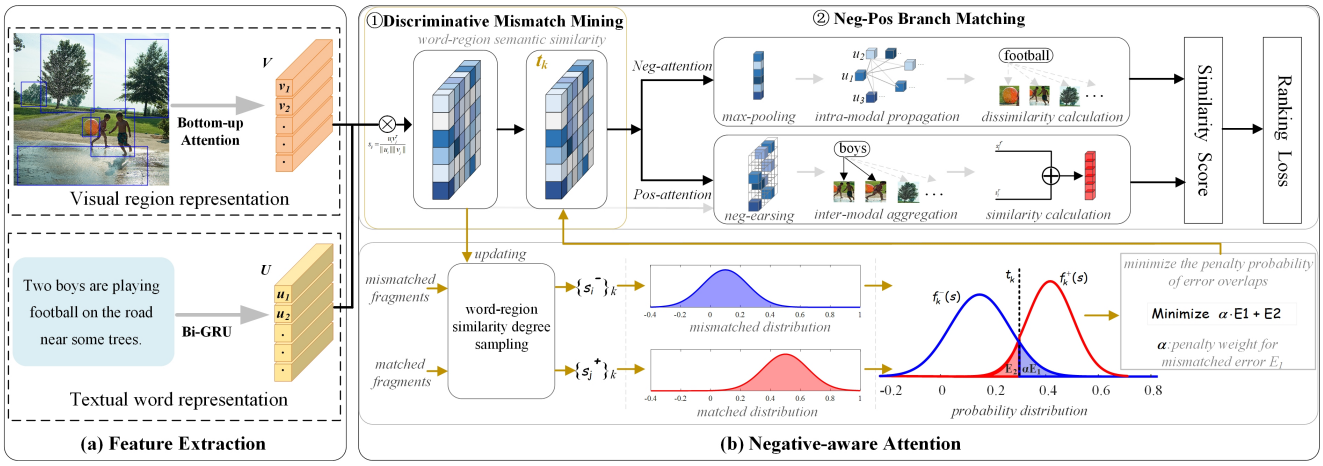

Figure 2. An overview of the proposed framework NAAF, containing two major modules for explicitly exploiting both negatively mismatched and positively matched textual fragments to jointly infer image-text similarity. ① discriminative mismatch mining that focuses on maximally separating the similarity distribution of mismatched fragments from the matched one, while also learning the adaptive boundary between these two distributions, enabling mismatched clues to yield more robust negative effects. ② Neg-Pos branch matching introduces different masks to precisely calculate the positive and negative effects of the two types of fragments to measure the overall similarity.

as much as possible to ensure the accuracy of identification. Hence, the optimal boundary learning can be written as a minimum weighted error probability problem:

$$\begin{aligned} \min_{t} \quad & \alpha \int_{t}^{+\infty} f_k^-(s)ds + \int_{-\infty}^{t} f_k^+(s)ds, \\ \text{s.t.} \quad & t \geq 0, \end{aligned} \tag{3}$$

where t is the decision variable; α is the penalty parameter for distinguishing errors of mismatched fragments; $t \geq 0$ is a sufficient condition for matched fragments.

To find the minimum point of Eq.(3), we search the zero point of its first derivative, truncate at $t \geq 0$ (using $[\cdot]_+ \equiv max(\cdot, 0)$), and obtain the optimal solution as:

$$t_k = [((\beta_2^k{}^2 - 4\beta_1^k\beta_3^k)^{\frac{1}{2}} - \beta_2^k)/(2\beta_1^k)]_+ \tag{4}$$

where $\beta_1^k = (\sigma_k^+)^2 - (\sigma_k^-)^2$, $\beta_2^k = 2(\mu_k^+\sigma_k^{-2} - \mu_k^-\sigma_k^{+2})$, and $\beta_3^k = (\sigma_k^+\mu_k^-)^2 - (\sigma_k^-\mu_k^+)^2 + 2(\sigma_k^+\sigma_k^-)^2 \ln \frac{\sigma_k^-}{\alpha\sigma_k^+}$.

There are two points worth highlighting. (1) During the training process, this explicit boundary t_k, firstly learned from the similarity distributions of mismatched and matched fragments, will then be integrated into the attention matching process to adjust more discriminative similarity distributions, which creates an iterative optimization. In this way, the distribution of mismatched fragments will be maximally separated from the distribution of matched fragments, where the mismatched fragments can yield more robust negative effects. Thus mismatched fragments can also be exploited as meaningful clues to accurately measure image-text similarity. (2) At the end of training, we expect the adaptive learning boundary t_k can simultaneously guarantee the maximum mining of mismatched fragments and avoid misjudgment of matched fragments causing performance degradation. Towards to make the learning boundary converge to this state that enjoys better mining accuracy, we give the theoretical condition that adjusting the initial penalty parameter α satisfies:

$$\alpha^* = \sigma_k^-[\sigma_k^+ exp^{\frac{\beta_4^k}{2(\sigma_k^{+2} - \sigma_k^{-2})}}]^{-1} \tag{5}$$

where $\beta_4^k = [\sigma_k^+(\mu_k^+ - \mu_k^-)/\sigma_k^- - 3(\sigma_k^{+2} - \sigma_k^-)]^2 - (\mu_k^+ - \mu_k^-)^2$, and the specific derivation is in the supplementary material.

3.1.2 Neg-Pos Branch Matching

In contrast to most existing works that merely focus on strengthening the attention of matched fragments to associate cross-modal shared semantics while simply weakening and ignoring mismatched fragments. Our two-branch framework can simultaneously focus on mismatched and matched fragments in the image-text pair, by resorting to different attention masks to precisely measure their effects in the negative and positive attention, respectively. Concretely, we first compute the semantic relevance scores between all words and regions as:

$$s_{ij} = \frac{u_i v_j^{\mathrm{T}}}{\|u_i\|\|v_j\|}, i \in [1, m], j \in [1, n] \tag{6}$$

Negative attention. In this branch, we aim to accurately and effectively utilize the mismatched fragments, making them valuable to downgrade the overall similarity of mismatching image-text pairs. Fragments in the textual modality that have no matched image regions are considered as mismatched. Moreover, compared with adaptively learned

relevance boundary t_k, the maximum cross-modal similarity between a fragment and another modality's all fragments reflects the degree of whether it is mismatched or matched. Therefore, we employ the max-pooling similarity between each word fragment $u_i, i \in [1, m]$ and all image regions $\{v_j\}_{j=1}^n$:

$$s_i = \max_j(\{s_{ij} - t_k\}_{j=1}^n), \tag{7}$$

Thus, the negative effect of i-th word in the image-text pair (U, V), *i.e.*, how dissimilar it is, can be measured as:

$$s_i^{neg} = s_i \odot \text{Mask}_{neg}(s_i) \tag{8}$$

where $\text{Mask}_{neg}(\cdot)$ is the mask that when the input is negative, it equals 1, otherwise it is 0; $\odot$ denotes dot-product.

Moreover, to make a more accurate negative effect measurement, we also consider the intra-semantic relationship of word fragments within the text, since the fragments with similar semantics should have the same matching relationship. Hence, the intra-modal propagation of each word's matching degree is conducted as:

$$\hat{s}_i = \sum_{l=1}^{m} w_{il}^{intra} s_l, \text{ s.t. } w_{il}^{intra} = \text{softmax}_\lambda(\{\frac{u_i u_l^{\mathrm{T}}}{\|u_i\|\|u_l\|}\}_{l=1}^m), \tag{9}$$

where w_{il}^{intra} denotes the semantic relationship between i-th and l-th word fragments, λ is a scaling factor; the enhanced $\hat{s}_i$ is used to replace the s_i in Eq.(8) in the inference.

Positive attention. This branch aims to measure how similar the image-text pair is, where there are two aspects to be considered. We firstly focus on attending to the cross-modal shared semantics, *i.e.*, aggregating matched image regions with respect to each query word, to measure the similarity degree of matched fragments. Concretely, the inter-modal attention weights are calculated by

$$w_{ij}^{inter} = \text{softmax}_\lambda(\{\text{Mask}_{pos}(s_{ij} - t_k)\}_{j=1}^n), \tag{10}$$

where w_{ij}^{inter} is the semantic relationship between word u_i and image region v_j. $\text{Mask}_{pos}(\cdot)$ denotes the mask that when the input is positive, it equals the input, otherwise it is $-\infty$, in which the attention weight of irrelevant image regions, *i.e.*, $s_{ij} - t_k < 0$, will be erased to zero.

For i-th word, the corresponding shared semantics in the image can be aggregated as: $\hat{v}_i = \sum_{j=1}^n w_{ij}^{inter} v_j$. Based on this weighted image feature, the similarity of u_i is measured as $s_i^f = u_i \hat{v}_i^{\mathrm{T}} / (\|u_i\| \|\hat{v}_i\|)$.

In addition, the high relevance scores s_{ij} between words and regions also reflect the degree of similarity, thus we also compute the weighted similarity based on the corresponding relevance scores with respect to the word u_i as: $s_i^r = \sum_{j=1}^n w_{ij}^{relev} s_{ij}$, where the relevance weights are calculated by $w_{ij}^{relev} = \text{softmax}_\lambda(\{\bar{s}_{ij}\}_{j=1}^n)$, in which we have $\bar{s}_{ij} = [s_{ij}]_+ / \sqrt{\sum_{i=1}^m [s_{ij}]_+^2}$. Therefore, the positive effect of the matched fragments in the image-text pair (U, V) can be measured as:

$$s_i^{pos} = s_i^f + s_i^r \tag{11}$$

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

$$S(U, V) = \frac{1}{m} \sum_{i=1}^{m} (s_i^{neg} + s_i^{pos}) \tag{12}$$

3.1.3 Sampling and Updating Strategy

In this section, we describe how to sample and update similarities of mismatched and matched word-region fragment pairs in Eq.(1) and Eq.(2), respectively. Although there is no matching ground-truth about word-region pairs, we solve this by allocating the pseudo word-region similarity via image-text instance-level matching annotations.

Concretely, the devised sampling is built on the simple fact: 1) the truly aligned text of an image should be totally matched to the image, that is, for the textual words, there is at least one matched region in the correct image. Thus, we sample the maximum similarity between word $u_i, i \in [1, m]$ and image regions $\{v_j^+\}_{j=1}^n$ from the correct image as:

$$s_i^+ = \max_j(\{v_j^+ u_i^{\mathrm{T}} / (\|v_j^+\| \|u_i\|)\}_{j=1}^n), \tag{13}$$

and 2) the misaligned text is mismatched to the incorrect image. In fact, with respect to the misaligned word, all regions in the incorrect image are mismatched with it. Yet, we argue that the maximum value of mismatched word-region similarities offers the greatest distinction ability, as it reveals their upper bound. Therefore, for word $u_i, i \in [1, m]$ with image regions $\{v_j^-\}_{j=1}^n$ from the incorrect image, we also sample the largest one as:

$$s_i^- = \max_j(\{v_j^- u_i^{\mathrm{T}} / (\|v_j^-\| \|u_i\|)\}_{j=1}^n), \tag{14}$$

where the update is for each text in a mini-batch. Moreover, to sample the accurate pseudo word-region similarity labels, we devise to decide whether to update s_i^+ and s_i^- at each sampling time, which is based on the correctness of the calculated similarity ranking. Note that sampling and updating operations are performed only in training.

3.2. Objective Function

Following existing approaches [15, 18, 19], the objective function adopted in this paper for end-to-end training is the bi-directional triplet ranking loss, which constrains the similarity of aligned image-text pairs to be higher than that of misaligned ones by a fixed margin. Moreover, we concentrate on optimizing hardest misaligned samples that yield highest loss. Given the ground-truth image-text pair (U, V) and its all umatched pairs $(U, V^{'})$ and

$(U^{'}, V)$, the hardest misaligned samples are selected by $V^{'} = \arg\max_{p \neq V} S(U, p)$ and $U^{'} = \arg\max_{q \neq U} S(q, V)$. Thus, the objective function is written as:

$$L = \sum_{(U,V)} [\gamma - S(U,V) + S(U,V^{'})]_{+} + [\gamma - S(U,V) + S(U^{'},V)]_{+}, \quad (15)$$

where γ is a margin hyperparameter, $[x]_{+} \equiv max(x, 0)$.

3.3. Feature Extraction

Visual Representation. Given an image V, it is represented as a set of salient regions features $[v_1, v_2, \ldots, v_n]$ by using the advantage of bottom-up attention [1]. The salient objects and other regions are detected utilizing a Faster-RCNN pretrained on Visual Genome [22], in which we select the top-K (K=36) proposals. Then, the detected regions are extracted by mean-pooled convolutional features by pre-trained ResNet-101 [12]. A fully connected layer is employed to map each region to a 1024-dimensional feature.

Textual Representation. Given a text U which comprises of m words, we encode each word into a 1024-dimensional vector as $[u_1, u_2, \ldots, u_m]$. Each word is first represented as a one-hot encoding, and embed into a pre-trained GloVe vector [30]. Then the vectors are fed into a bi-directional gated recurrent unit (BiGRU) to integrate the forward and backward contextual information. The final word representation u_i is the average of bi-directional hidden states.

4. Experiments

4.1. Dataset and Implementation Details

Datasets. To validate effectiveness, we conduct extensive experiments on two benchmark datasets. Flickr30K [31] totally has 31,000 images and 155,000 sentences. Following the same protocol in [19], Flickr30K is split into 1,000 test images, 1,000 validation images, and 29,000 training images. MS-COCO [25] contains 123,287 images and 616,435 sentences, and we split it into 5,000 test images, 5000 validation images, and 113,287 training images [18]. The results of MS-COCO is tested on averaging over 5-folds of 1K test images and on the full 5K test images.

Evaluation Metrics. We adopt the commonly used Recall at K (R@K, K=1, 5, 10) and rSum. R@K means the percentage of ground truth in the retrieved top-K lists. rSum is the sum of all R@K in both image-to-text and text-to-image, reflecting the overall matching performance.

Implementation Details. All experiments are conducted on an NVIDIA GeForce RTX 3090Ti GPU. The Adam optimizer is employed for model optimization, with 0.0005 as the initial learning rate and decaying by 10% every 10 epochs. The mini-batch size is set to 128 and 256 for Flickr30K and MSCOCO respectively, with 20 training epoches on both datasets. The feature dimension d is set to 1,024. The scaling parameter λ is set to 20. The initial penalty parameter α is set to 2.0, which will be investigated at Sec. 4.3. The adjusting epoch in training is selected as 15. The margin hyperparameter γ is selected as 0.2.

Methods	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN('18) [23]	67.4	90.3	95.8	48.6	77.7	85.2	465.0
BFAN('19) [26]	68.1	91.4	-	50.8	78.4	-	288.7
CVSE('20) [35]	70.5	88.0	92.7	54.7	82.2	88.6	476.7
DPRNN('20) [5]	70.2	91.6	95.8	55.5	81.3	88.2	482.6
SGM('20) [37]	71.8	91.7	95.5	53.5	79.6	86.5	478.6
IMRAM('20) [2]	74.1	93.0	96.6	53.9	79.4	87.2	484.2
GSMN('20) [27]	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SMFEA('21) [10]	73.7	92.5	96.1	54.7	82.1	88.4	487.5
SHAN('21) [16]	74.6	93.5	96.9	55.3	81.3	88.4	490.0
VSE∞ ('21) [3]	76.5	94.2	97.7	56.4	83.4	89.9	498.1
SGRAF('21) [7]	77.8	94.1	97.4	58.5	83.0	88.8	499.6
NAAF(ours)	**81.9**	**96.1**	**98.3**	**61.0**	**85.3**	**90.6**	**513.2**

Table 1. Quantitative evaluation results on Flickr30K test set. The bests are in bold.

4.2. Comparison Results

We compare our proposed NAAF with the recent state-of-the-art models on the two benchmarks. We belong to the local level paradigm. Same with the compared models, we report the ensemble results by averaging two models' similarity, *i.e.*, whether using the intra-modal relationship. Tab.1 shows the quantitative results of our NAAF approach on Flickr30K. Our NAAF outperforms state-of-the-arts significantly on all evaluation metrics. Specifically, we obtain relative 13.6% improvement on rSum, where the R@1 gains 4.1% and 2.5% improvement at two retrieval directions, respectively. Moreover, compared with the typical SCAN which our model builds on its basis, NAAF obtains 14.5% and 12.4% on R@1 at two directions, respectively, and largely improves rSum by relative 48.2%.

The experimental results on the larger and complicated MS-COCO are shown in Tab.2. We can see that our NAAF outperforms state-of-the-arts in terms of most evaluation metrics. Compared with SHAN and SMFEA, NAAF gains relative improvements of 7.4% and 9.6% on rSum, respectively, and we can achieve competitive results with the state-of-the-arts, getting near 3% improvements on rSum. For the full 5K test datasets, we have nearly 1% improvement in terms of R@1, which is the main concern in practical applications. Compared with the baseline model SCAN, our approach surpasses all its evaluation performance, with relative 8.5% and 3.9% improvements on R@1 in two directions, respectively. The improvements show that the effectiveness of our proposed framework for maximally mining the negative effect of mismatched fragments, and further verifies that jointly using the negative and positive effects of mismatched and matched fragments can obtain more ac-

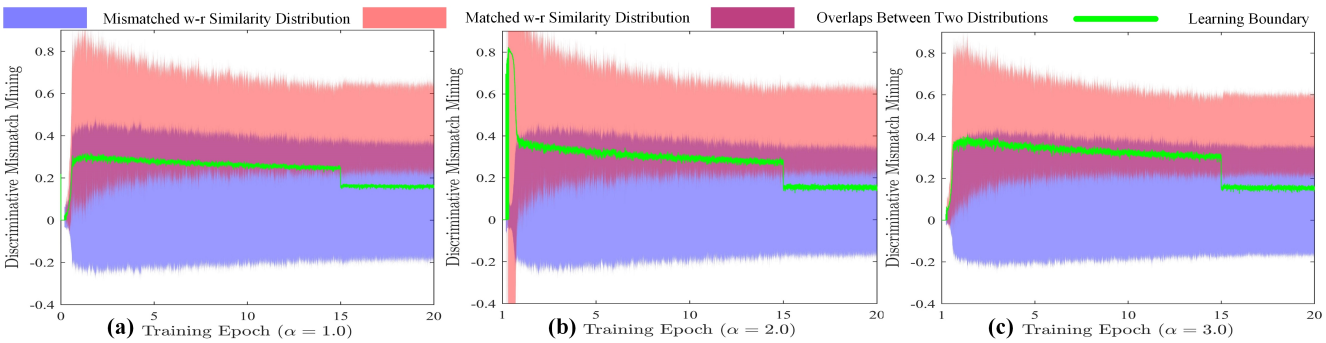

Figure 3. Visualization of the discriminative mismatch mining process with different penalty weights α, where the mismatched word-region (w-r) similarity distribution is explicitly separated, making mismatched fragments yield more robust negative effects. The larger α, the stronger the discriminative ability of mismatch mining, but also the greater discrimination errors for matched fragments. α is adjusted to α^* via Eq.5 in the later stage of training, which guarantees the learning boundary converge to the state with high mining accuracy.

Methods	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
1K Test Set							
SCAN('18) [23]	72.7	94.8	98.4	58.8	88.4	94.8	507.9
BFAN('19) [26]	74.9	95.2	-	59.4	88.4	-	317.9
PVSE('19) [34]	69.2	91.6	96.6	55.2	86.5	93.7	492.8
CVSE('20) [35]	69.2	93.3	97.5	55.7	86.9	93.8	496.4
DPRNN('20) [5]	75.3	95.8	98.6	62.5	89.7	95.1	517.0
SGM('20) [37]	73.4	93.8	97.8	57.5	87.3	94.3	504.1
IMRAM('20) [2]	76.7	95.6	98.5	61.7	89.1	95.0	516.6
GSMN('20) [27]	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SMFEA('21) [10]	75.1	95.4	98.3	62.5	90.1	96.2	517.6
SHAN('21) [16]	76.8	96.3	98.7	62.6	89.6	95.8	519.8
VSE∞('21) [3]	78.5	96.0	98.7	61.7	90.3	95.6	520.8
SGRAF('21) [7]	79.6	96.2	98.5	63.2	90.7	96.1	524.3
NAAF(ours)	**80.5**	**96.5**	**98.8**	**64.1**	**90.7**	**96.5**	**527.2**
5K Test Set							
SCAN ('18) [23]	50.4	82.2	90.0	38.6	69.3	80.4	410.9
PVSE ('19) [34]	45.2	74.3	84.5	32.4	63.0	75.0	374.3
SGM ('20) [37]	50.0	79.3	87.9	35.3	64.9	76.5	393.9
IMRAM ('20) [2]	53.7	83.2	91.0	39.7	69.1	79.8	415.5
SMFEA ('21) [10]	54.2	-	89.9	41.9	-	**83.7**	269.7
SGRAF('21) [7]	57.8	-	91.6	41.9	-	81.3	272.6
NAAF(ours)	**58.9**	**85.2**	**92.0**	**42.5**	**70.9**	81.4	**430.9**

Table 2. Quantitative evaluation results on MS-COCO 1K and 5K test set. The bests are in bold. Because some works did not report 5K results, we compare the reported ones.

curate image-text similarity.

4.3. Ablation Study

We conduct extensive ablation studies on Flickr30K to verify the effectiveness of each component of our NAAF.

The impact of different penalty parameters. In our model, the most sensitive and important parameter is the penalty weight α, which determines the ability to mine mismatched fragments in training. Note that α should be a trade-off between the mining strength of mismatched clues and the misclassification of matched clues. 1) We investigate the matching performance with setting α as 1.0, 2.0 and 3.0, all of which are defaultly equipped with α^*. As shown in Tab.3, it can be seen that NAAF achieves better performance when $\alpha = 2.0$. The higher one ($\alpha = 3.0$) has a better performance than the lower one ($\alpha = 1.0$), where we visualize the mismatch mining of different α in training in Fig.3. 2) If we omit the adjusting of α, it suffers performance degradation, since many matched clues are misjudged, which verifies the necessity of α^*.

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Penalty Parameter for Mismatch Errors						
initial $\alpha = 1.0$	78.0	95.7	98.0	58.9	83.9	89.8
initial $\alpha = 3.0$	78.9	95.3	97.8	59.3	84.1	89.9
initial $\alpha = 2.0^{\ddagger}$	79.6	96.3	98.3	59.3	83.9	90.2
$\ddagger$ w/o adjusting to α^*	78.7	95.4	97.6	59.3	83.4	89.7
Sampling and Updating						
batchsize 32	76.0	95.0	97.5	57.8	83.6	89.7
batchsize 64	78.3	96.0	98.3	58.9	84.1	89.8
batchsize 128‡	79.6	96.3	98.3	59.3	83.9	90.2
$\ddagger$ w/o up. neg. samp.	76.3	94.2	97.2	59.5	83.6	90.1

Table 3. Ablation studies about the penalty parameter (Sec.3.1.1) and the sampling and updating strategy (Sec.3.1.3), which are obtained on Flickr30K. Underlines mean the best model, marked ‡.

The impact of sampling and updating strategy. The sampling of matched and mismatched word-region similarity is crucial to NAAF. In Tab.3, 1) For the batch size, which relates to the number of updates, we find that a larger batch size brings better performance, as the amount of updated data determines the accuracy of the distribution modeling. 2) For the sampling of mismatched words without using maximum (upper) similarity, the performance is severely degraded, verifying the maximum sampling of mismatched word-region has more effective constraint effect.

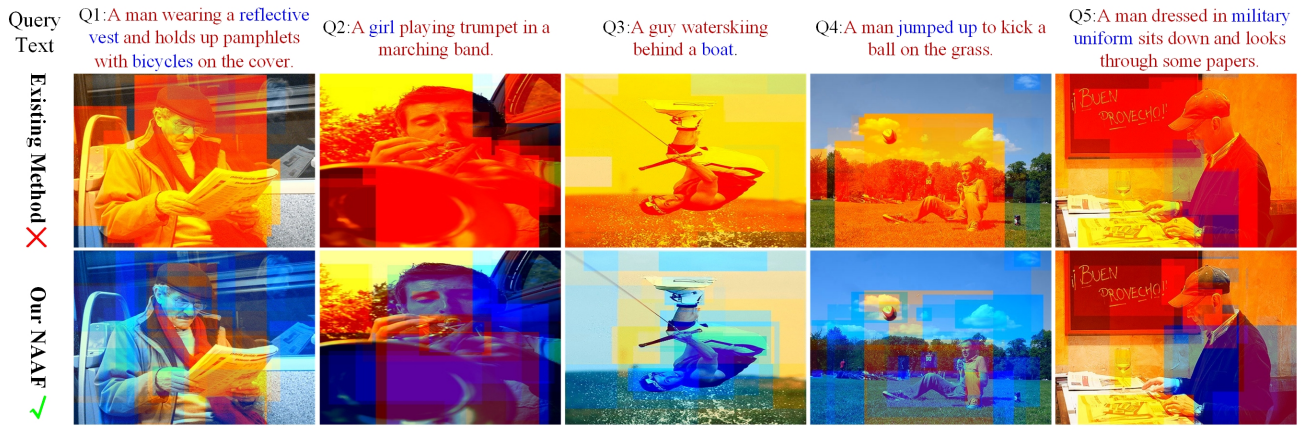

Figure 4. Visual comparison of negative effects of mismatched words (blue) and positive effects of matched words (red) in our NAAF and existing methods. It shows that our NAAF can explicitly mine mismatch regions in the image w.r.t. the mismatched words for highlighting their negative effects, while existing methods mainly concentrate on matched ones, and typically ignore the effects of mismatched words.

Methods	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
NAAF w/o Neg.	66.2	91.0	96.2	53.4	80.0	87.8
NAAF w/o Masks	74.2	93.0	96.4	57.2	82.9	88.4
NAAF w i2t	78.7	94.6	97.7	58.3	84.0	90.1
NAAF w/o GloVe	75.9	93.6	97.7	55.5	81.0	87.9
NAAF w/o Intra	79.3	96.1	98.0	59.2	83.9	90.0
NAAF Full	**81.9**	**96.1**	**98.3**	**61.0**	**85.3**	**90.6**

Table 4. Ablation studies about the model design (Sec.3.1.2), which are obtained on Flickr30K.

The effectiveness of model designing. The 'NAAF-full' denotes the two models' averaging, while others are all single models. As shown in Tab.4, 1) For omitting the negative branch in NAAF, it is obvious that the performance has a major degradation, which verifies the effectiveness of exploiting the negative effects of mismatched clues. 2) When removing the designed masks, the model cannot achieve high performance since it cannot precisely calculate the similarity/dissimilarity in image-text pairs. 3) Adding the i2t direction to NAAF will obtain suboptimal performance, since as we analyzed in the Sec.1, the mismatched regions in the image are meaningless so that the introduction may cause interference. 4) Following SHAN, we use the GloVe embeddings that relatively improve the performance. Note that, for our single model without GloVe, it can also exceed the single SGRAF with nearly 3% rSum improvements. 5) Compared with the single model (with underlines in Tab.3), the performance is slightly decreased without using intra-modal, verifying that the intra-relationship can make more accurate negative effect measurement.

4.4. Visualization and Case Study

To better understand the effectiveness of NAAF, we visualize the word-region similarity comparison of the false-positive, *i.e.*, incorrect candidate but ranking top-1 in existing method [7, 23, 26, 27], where the text is query in Fig. 4. We can see that mismatched words' negative effects, *i.e.*, regions marked in blue, in existing method are negligible. In contrast, our NAAF can correctly capture and highlight the negative effects of mismatched words, hence these false-positive can be eliminated. As for the specific cases, NAAF can accurately locate the mismatched regions w.r.t. the text. In terms of "reflective vest" in Q1 and "military uniform" in Q5, NAAF can accurately focus on the mismatched regions of clothes in the image, showing robust negative effects. Moreover, for a more complex example, the action "jumped up" in Q3 can also be found mismatch in the image, demonstrating the efficiency of mismatched mining.

5. Conclusion

In this paper, we propose a novel negative-aware attention framework for image-text matching. Different from conventional attention, our method can simultaneously focus on both mismatched and matched fragments to explicitly exploit their negative and positive effects, where an efficient iterative optimization is constructed to maximally mine the negative mismatch fragments, yielding discriminative and robust negative effects. Moreover, the two-branch matching mechanism enables respectively measuring the accurate similarity/dissimilarity degrees to jointly infer the overall image-text similarity, solving the neglection of mismatched clues in existing methods. Comprehensive experiments demonstrate the superiority of our NAAF framework.

6. Acknowledgements

This work has been supported by the National Key Research and Development Program of China under Grant 2020YFB1406603.

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. In *CVPR*, pages 6077–6086, 2018. 6

[2] 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. In *CVPR*, pages 12655–12663, 2020. 1, 2, 3, 6, 7

[3] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *CVPR*, pages 15789–15798, 2021. 1, 3, 6, 7

[4] Tianlang Chen, Jiajun Deng, and Jiebo Luo. Adaptive offline quintuplet loss for image-text matching. In *ECCV*, pages 549–565, 2020. 2

[5] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *AAAI*, volume 34, pages 10583–10590, 2020. 1, 2, 3, 6, 7

[6] Hui Cui, Lei Zhu, Jingjing Li, Yang Yang, and Liqiang Nie. Scalable deep hashing for large-scale social image retrieval. In *IEEE Transactions on image processing*, volume 29, pages 1271–1284, 2019. 2

[7] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, 2021. 1, 2, 3, 6, 7, 8

[8] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, pages 4601–4611, 2017. 2

[9] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *arXiv preprint arXiv:1707.05612*, 2017. 1, 2

[10] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *ACM MM*, pages 5185–5193, 2021. 6, 7

[11] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[13] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *IJCAI*, pages 789–795, 2019. 1, 2, 3

[14] Yan Huang and Liang Wang. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In *ICCV*, pages 5774–5783, 2019. 3

[15] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, pages 6163–6171, 2018. 3, 5

[16] Zhong Ji, Kexin Chen, and Haoran Wang. Step-wise hierarchical alignment network for image-text matching. In *IJCAI*, 2021. 1, 2, 3, 6, 7

[17] Zhong Ji, Haoran Wang, Jungong Han, and Yanwei Pang. Saliency-guided attention network for image-sentence matching. In *ICCV*, pages 5754–5763, 2019. 3

[18] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2, 5, 6

[19] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NeurIPS*, 2014. 1, 3, 5, 6

[20] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *ICML*, 2014. 3

[21] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, volume 123, pages 32–73, 2017. 6

[23] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 1, 2, 3, 6, 7, 8

[24] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019. 1, 3

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6

[26] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *ACM MM*, pages 3–11, 2019. 1, 2, 3, 6, 7, 8

[27] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020. 1, 2, 3, 6, 7, 8

[28] Xu Lu, Lei Zhu, Zhiyong Cheng, Liqiang Nie, and Huaxiang Zhang. Online multi-modal hashing with dynamic query-adaption. In *ACM SIGIR*, pages 715–724, 2019. 2

[29] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, pages 2623–2631, 2015. 3

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014. 6

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 6

[32] Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, and Liqiang Nie. Dynamic modality interaction modeling for image-text retrieval. In *ACM SIGIR*, pages 1104–1113, 2021. 3

[33] Botian Shi, Lei Ji, Pan Lu, Zhendong Niu, and Nan Duan. Knowledge aware semantic concept expansion for image-text matching. In *IJCAI*, volume 1, page 2, 2019. 3

[34] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019. 7

[35] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *ECCV*, pages 18–34. Springer, 2020. 6, 7

[36] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016. 2

[37] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020. 3, 6, 7

[38] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *ACM MM*, pages 12–20, 2019. 3

[39] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *IJCAI*, pages 3792–3798, 2019. 1, 2, 3

[40] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *CVPR*, pages 13005–13014, 2020. 2

[41] Yiling Wu, Shuhui Wang, and Qingming Huang. Online asymmetric similarity learning for cross-modal retrieval. In *CVPR*, pages 4269–4278, 2017. 2

[42] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *ACM MM*, pages 2088–2096, 2019. 3

[43] Shiyang Yan, Li Yu, and Yuan Xie. Discrete-continuous action space policy gradient-based attention for image-text matching. In *CVPR*, pages 8096–8105, 2021. 1

[44] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3536–3545, 2020. 3

[45] Mo Zhou, Zhenxing Niu, Le Wang, Zhanning Gao, Qilin Zhang, and Gang Hua. Ladder loss for coherent visual-semantic embedding. In *AAAI*, volume 34, pages 13050–13057, 2020. 2

[46] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, and Huaxiang Zhang. Deep collaborative multi-view hashing for large-scale image search. In *IEEE Transactions on Image Processing*, volume 29, pages 4643–4655, 2020. 2