Cross-Modal Attention With Semantic Consistence for Image–Text Matching

Xing Xu, Tan Wang, Yang Yang, *Member, IEEE*, Lin Zuo, Fumin Shen, and Heng Tao Shen, *Senior Member, IEEE*

***Abstract*— The task of image–text matching refers to measuring the visual-semantic similarity between an image and a sentence. Recently, the fine-grained matching methods that explore the local alignment between the image regions and the sentence words have shown advance in inferring the image–text correspondence by aggregating pairwise region-word similarity. However, the local alignment is hard to achieve as some important image regions may be inaccurately detected or even missing. Meanwhile, some words with high-level semantics cannot be strictly corresponding to a single-image region. To tackle these problems, we address the importance of exploiting the global semantic consistence between image regions and sentence words as complementary for the local alignment. In this article, we propose a novel hybrid matching approach named Cross-modal Attention with Semantic Consistency (CASC) for image–text matching. The proposed CASC is a joint framework that performs cross-modal attention for local alignment and multilabel prediction for global semantic consistence. It directly extracts semantic labels from available sentence corpus without additional labor cost, which further provides a global similarity constraint for the aggregated region-word similarity obtained by the local alignment. Extensive experiments on Flickr30k and Microsoft COCO (MSCOCO) data sets demonstrate the effectiveness of the proposed CASC on preserving global semantic consistence along with the local alignment and further show its superior image–text matching performance compared with more than 15 state-of-the-art methods.**

***Index Terms*— Cross-modal attention, cross-modal retrieval, hybrid alignment, image–text matching.**

I. Introduction

With the rapidly development of multimedia technology and dramatic volume of different modalities data, exploring the relationship between vision and language has attracted great interest among researchers in their intersections and derived various practical research hotspots, such as image-sentence matching [2]–[4], cross-modal retrieval [5]–[7], cross-modal hashing [8], [9], image captioning [10]–[12], and visual question answering [13], [14]. In this article, we focus on the problem of image-sentence retrieval, i.e., given a sentence query to retrieve the visually related images, namely, image retrieval, and given an image query to obtain matched sentence descriptions, namely, sentence retrieval. The core issue in image-sentence retrieval is how to accurately measure the semantic similarity between visual and textual input and then further discover the full latent semantic correspondence between an image and a sentence. During the last decade, a bundle of prior works has made great progress on image–text matching. One of the common ways is to learn a common embedding space for maximizing the correlation of bimodal input data, where the pairwise image–text similarity can be easily calculated via the projected common representations. This pipeline has been widely studied in the related problems of cross-modal retrieval [7], [15]–[18] and cross-modal hashing [8], [19], [20], where the former focuses on learning real-valued common representations, while the latter seeks for binary codes in the specific common Hamming space [21]–[23] to improve the retrieval efficiency. However, the primary drawback of these methods is that they limitedly consider the entire image and text for globally semantic alignment, which is too coarse to dig into the details of the image regions or text words. Later, the fine-grained matching methods have been proposed to achieve more interpretable image–text matching by capturing the fine-grained interplay between vision and language. In these approaches, image patches are detected and image–text similarity scores are aggregated by all or salient region-word pairs with simple correspondence or sophisticated attention mechanism [1], [24]–[27]. For example, the latest work of stacked cross attention network (SCAN) [1] discovers the latent alignment using both image regions and words in sentences as context via attention across modalities, which produces more accurate image–text similarity for matching.

Although these fine-grained matching methods have shown promising capability on image–text matching via maximizing the overall similarity aggregated by pairwise image regions and text words, they inevitably bring severe problems. On the one hand, the local alignment highly relies on the accuracy of the detected image regions, and different regions would interfere with each other. Taking the sentence retrieval task on

Manuscript received December 20, 2018; revised June 10, 2019 and October 15, 2019; accepted January 11, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61976049 and Grant 61632007 and in part by the Sichuan Science and Technology Program, China, under Grant 2019ZDZX0008, Grant 2019YFG0003, and Grant 2018GZDZX0032. *(Xing Xu and Tan Wang contributed equally to this work.) (Corresponding author: Heng Tao Shen.)*

Xing Xu, Tan Wang, Yang Yang, Fumin Shen, and Heng Tao Shen are with the Center for Future Multimedia, University of Electronic Science and Technology of China, Chengdu 610051, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: xing.xu@uestc.edu.cn; wangt97@hotmail.com; dlyyang@gmail.com; fumin.shen@gmail.com; shenhengtao@hotmail.com).

Lin Zuo is with the School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: linzuo@uestc.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2967597

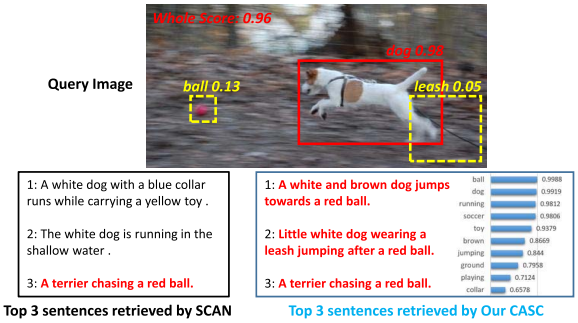


Fig. 1. Typical example of the sentence retrieval results obtained by the latest fine-grained matching method SCAN [1] and our proposed CASC. The top-three retrieved sentences of the two methods are, respectively, presented in the two rectangles. In particular, the most related semantic labels with their similarity scores to the image can also be predicted by our CASC.

an image containing an object of dog for example (in Fig. 1), when using the latest fine-grained matching approach SCAN on this image, it can exactly detect the object dog (in the red solid rectangle) and get a high similarity score (the red number 0.98) between the patch dog and the word "dog"; however, it fails to predict some informative words, such as "ball," "leash," and "jumps" (in yellow dot rectangles) and provides a low grade (the yellow numbers 0.05 and 0.13) as they cannot be easily detected and represented from pixel-level image compared to the dog. Thus, the importance of these words would be weakened, even despised in the whole similarity measure leading to a wrongly high aggregated similarity score (the red number 0.96), which means the interference between pairwise image regions and words is ignored by SCAN when performing local alignment. On the other hand, the missing word may also correlate to multiple regions even the whole image, and it may present diverse appearances in different images. Therefore, only using fine-grained matching is quite insufficient and defective, and it is rational to address a global semantic consistence between image regions and words as complementary for the local alignment. Considering that the widely used image–text data sets, such as Flickr30k [28] and Microsoft COCO (MSCOCO) [29], lack of manually annotated label information, an effective scheme is to extract the meaningful words from text as the semantic labels to further guide the preservation of global semantic consistence.

Motivated by jointly combining local alignment and global semantic consistence, we propose a novel Cross-modal Attention with Semantic Consistence approach dubbed Cross-modal Attention with Semantic Consistency (CASC) for image–text matching. As illustrated in Fig. 2, the proposed CASC is a hybrid matching approach that consists of two branches: cross-modal attention for local alignment and multilabel prediction for global semantic consistence. Specifically, after mapping different modality features into a common semantic subspace, the local alignment is accomplished via cross-modal attention on both the image regions and text words by calculating the local similarity mutually from both the image-to-text (I2T) and text-to-image (T2I) directions, rather than separately utilized in previous SCAN method. The two directions ensure more sufficient local alignment. In addition, the semantic consistence is preserved by multilabel classification rather than directly using global features to bridge the whole image and the extracted labels from the five associated sentences. In this way, it provides a global similarity constraint upon local alignment and measures the similarity between unimodal sentences explicitly, which is beneficial to fully incorporate complementary correspondence between all the regions and words. Furthermore, what can be easily neglected is that in our CASC method, an image and its five associated sentences share a common label set. In this way, the similarity between unimodal text that has been usually ignored in previous approaches can be explicitly explored in our CASC method. Finally, these two branches are integrated into a joint learning framework, leading to more accurate image–text similarity measurements on both local and global aspects.

Our primary contributions in this article are as follows.

1) We propose the novel CASC that combines the global semantic consistence with the multilabel prediction with the local coattention region-word alignment in a joint framework.
2) We develop an efficient semantic label extraction scheme that directly constructs label dictionary from available text modality data without additional labor cost.
3) Extensive experiments on two benchmark data sets show advantages of the proposed CASC on integrating the local alignment scores and preserving the global semantic consistence. It also achieves a more accurate image–text similarity measure and gains significant matching performance compared with state-of-the-art methods.

The rest of this article is organized as follows. In Section II, we first briefly introduce the related studies on image–text matching. Then, in Section III, we present the details of the proposed CASC. Section IV discusses the experiments and the ablation study. Finally, we make a conclusion in Section V.

II. Related Work

A. Global Matching Methods

The global matching methods aim to learn mapping matrices to project heterogeneous data into common embedding space and generate global common representations for each instance (image or text), by which the pairwise image–text similarity or distance can be calculated directly using cosine similarity or the Euclidean distance. Particularly, canonical correlation analysis (CCA) [15] has been a well-tested and representative embedding technique for decades. It learns a linear transformation to project two modalities into a common space where their correlations are maximized.

Driven by the successful developments of deep neural network (DNN) on extracting powerful visual and textual features, researchers adopt DNN to learn common space for image–text retrieval [18], [30]–[35]. The DNN-based approaches typically construct a multipathway network, where each pathway is for the data of one media type and the multiple pathways are linked at the joint layer to model cross-media correlation. For example, Ngiam *et al.* [30] proposed bimodal autoencoders (bimodal AEs) to extend the restricted Boltzmann machine (RBM), which models the correlation by mutual reconstruction between different media types. Andrew *et al.* proposed deep CCA (DCCA) [32] to extend the traditional CCA with a deep network structure,

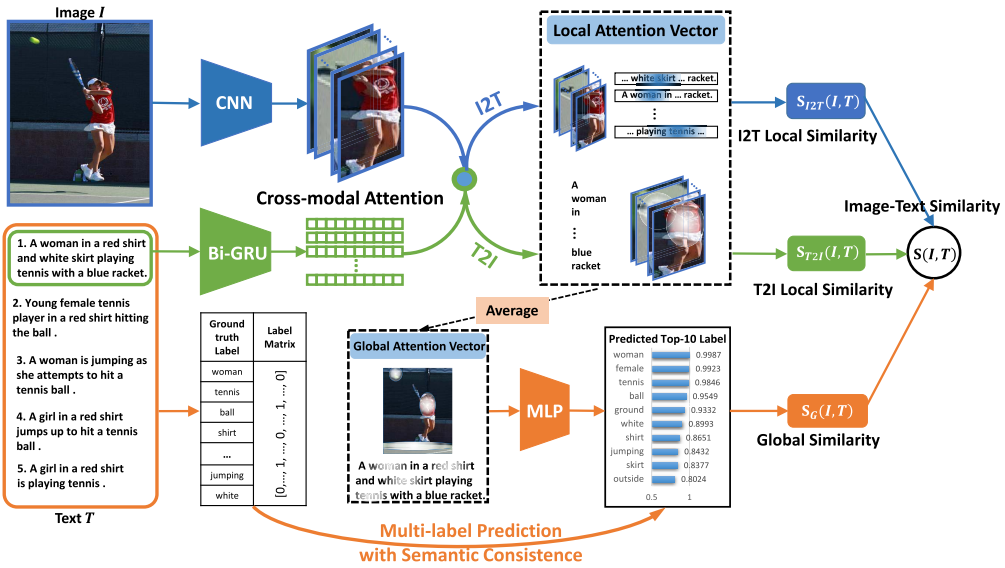


Fig. 2. Overview architecture of our proposed CASC. It mainly consists of two parts: cross-modal attention for local alignment and multilabel prediction for global semantic consistence. For the former part, CASC attends differentially to important image regions and words with each other as the context for inferring the local image–text similarity as I2T and T2I independently. For the latter, a semantic label prediction is adopted to alleviate the interference between patches and words and leads to more accurate global representation. Finally, two branches are combined via a joint framework for learning and optimization with a tradeoff parameter.

which maximizes correlation on the top of two subnetworks. Feng *et al.* [31] jointly modeled cross-media correlation and reconstruction information to perform correspondence autoencoder (Corr-AE). Peng *et al.* [35] recently proposed cross-modal correlation learning (CCL) that utilizes fine-grained information and adopts a multitask learning strategy that obtains better performance.

Recently, a generative adversarial network (GAN) [36], which contains a generator module and a discriminator module with adversarial learning strategy, has shown its promising ability to model the data distribution in the DNN models. Several cross-modal GAN (CM-GAN) approaches have been developed, including adversarial cross-modal retrieval (ACMR) [34], generative cross-modal feature learning network (GXN) [37], CM-GAN [38], cycleMatch [39], and recipe retrieval GAN (R2GAN) [40]. These approaches mainly utilize adversarial learning schemes to improve the effect of common embedding learning by introducing a discriminator to predict the modality type of the embedding feature.

The above-mentioned shallow and DNN-based methods typically learn real-valued common representations in the common embedding space, which is less efficient for computing the similarity of pairwise image–text instances on large-scale data. As mentioned in Section I, inspired by unimodal hashing approaches [22], [23], [41], cross-modal hashing methods, which also include shallow methods [8], [22] and DNN-based methods [5], [19], [41], have been proposed to learn binary common representations in a common space for more efficient Hamming distance computation and less storage requirement. The primary difference in these hashing methods is that they employ additional binary constraints during learning common embedding space, and more complex algorithms, such as iterative quantization and discrete optimization, are required.

Generally, the above-mentioned global matching methods, either learning real-valued or binary common representations, limitedly consider the whole image and text for the globally semantic alignment that is considerably coarse to result in inferior matching accuracy.

B. Fine-Grained Matching Methods

In a sense, sentence descriptions are weak annotations, where words in a sentence correspond to some particular, but unknown regions in the image. Therefore, inferring the latent correspondence between image regions and words is a key to more interpretable image–text matching by capturing the fine-grained interplay between vision and language. The fine-grained methods [1], [24], [37], [42], [43] usually detect image regions at object region level and simply aggregate similarity of all possible pairs of image regions and words in the sentence to infer the global image–text similarity. Karpathy and Fei-Fei [24] proposed a method of detecting and encoding image regions at the object level and then inferring the image–text similarity by aggregating the similarities from all possible region-word pairs. Niu *et al.* [42] presented a model that maps noun phrases within sentences and objects in images into a shared embedding space on top of full sentences and whole images embeddings. Recently, attention-based models have been proposed for the image–text matching problem. Huang *et al.* [44] developed a context-modulated attention scheme to selectively attend to a pair of instances appearing in both the image and sentence. Similarly, Nam *et al.* [45] proposed a dual attentional network (DAN) to capture the fine-grained interplay between vision and language through multiple steps. Lee *et al.* [1] presented a model that attend differently to important image regions and words with each other as the context for similarity measuring by using an attention mechanism.

C. Hybrid Matching Methods

Although the above-mentioned fine-grained matching methods attempt to explore local alignment between image patches

and words, they ignore important relation information in the context of these fine-grained patches, which can provide rich complementary hints for cross-media correlation learning. Recently, some works [44], [46]–[49] put forward to integrate local alignment and global cues to further explore the correlation of fine-grained patches. Huang *et al.* addressed the importance of learning semantic concepts and orders (SCO) in improving the image and text representation [46]. They developed a series of model components in terms of multiregional multilabel CNN, gated fusion unit, and joint matching and generation learning to add the semantic-order concepts into training. Qi *et al.* [48] constructed a cross-media relation attention network (CRAN) to explore multilevel alignment, which contains three subnetworks for global, local, and relation alignments, respectively. Wang *et al.* [49] formulated a global representation learning task with both intramodal and intermodal relative similarities and further developed a coattention learning procedure to fully exploit different levels of visual–linguistic relations. Huang *et al.* [50] recently proposed a bidirectional spatial-semantic attention network (BSSAN) that leverages both the relation of word to regions and visual object to words in a holistic deep framework for more effective matching. Ma *et al.* [51] also considered the bidirectional cues of the cross-modality correlations and intramodality similarities, and they proposed a local and global multimodal deep matching (MDM) model to match the local and global representations of image and sentence.

The proposed CASC also belongs to hybrid matching; however, it differs from the above approaches in several aspects, especially the following.

1) Most hybrid methods [46], [48], [49], [51] applied image patches and sentence words vectors as local features and their average pooling vectors or the outputs of the last fully connected layer as global representations, which means the two aspects are analogous and insufficient.
2) The proposed CASC incorporates the local attention mechanism and global alignment in a joint framework, which is more effective than SCO [46] that performs alignment separately without attention. Moreover, our semantic concepts focus on label prediction that can be simpler and more efficient than a semantic-order constraint in SCO.
3) The CASC is advantaged to leverage the freely extracted semantic labels from the text modality data to improve the global semantic consistence for visual objects, while both SCO and CRAN rely on the relative positions of object regions to ensure the order or correlation of objects.

III. Our CASC Approach

In this section, we depict our CASC approach from four aspects: 1) cross-modal representation learning that maps different modality features into a common semantic subspace; 2) cross-modal attention that performs local attention alignment on informative image regions and text words mutually; 3) semantic consistence that extracts informatively semantic concepts and equips a global similarity constraint upon local alignment; and 4) joint optimization on the final objective function and the testing phase for matching query data.

A. Cross-Modal Representation Learning

We first introduce the formal definition of cross-modal image–text data set as $\mathcal{O} = \{(I_n, T_n)\}_{n=1}^{N}$ that has N image–text pairs, where images $\mathcal{I} = \{I_n\}_{n=1}^{N}$ and texts $\mathcal{T} = \{T_n\}_{n=1}^{N}$ have totally N instances in each modality. Given a query of any modality, the goal of image–text matching is to measure the cross-modal similarity of the pth and qth instance of image I_p and text T_q and match relevant instances of another modality. To simplify the notations, we denote I and T as a single instance of the image modality and the text modality, respectively. Unlike previous global matching approaches that represent an image or text instance using a global feature vector in a common subspace, the proposed CASC considers more comprehensive representation for both image and text modalities to achieve more accurate matching.

1) Image Representation: Given an image I, we aim to represent it with a set of multiple features $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, $\mathbf{v}_i \in \mathbb{R}^h$, $i \in [1, k]$, where each image feature $\mathbf{v}_i$ encodes a region in an image contains k regions. Multiple region-level features are more accurate to describe an image than a single global feature. In particular, we follow [23] and [52] to select salient regions to extract features, and we utilize the pretrained Faster R-CNN [53] model in conjunction with ResNet-101 on Visual Genomes [54] data set. In order to learn feature representations with rich semantic meanings, instead of predicting the object classes, the model predicts attribute classes and instance classes, where the instance classes contain objects and other salient regions that are difficult to localize.

For each selected region i in I, we denote $\mathbf{f}_i \in \mathbb{R}^{d_v}$ as the convolutional feature extracted by ResNet-101 after average pooling from this region, where d_v is the dimension of the region feature vector. To obtain the cross-modal representation for image modality, we further add a fully connect layer to transform $\mathbf{f}_i$ to a h-dimensional embedding vector, as

$$\mathbf{v}_i = \mathbf{W}_v \mathbf{f}_i + b_v, \quad i \in [1, k] \tag{1}$$

where $\mathbf{W}_v \in \mathbb{R}^{h \times d_v}$ is the transformation matrix and b_v represents the bias coefficient. Therefore, the representation of an image I_p is a set of semantic vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, $\mathbf{v}_i \in \mathbb{R}^h$, where each $\mathbf{v}_i$ encodes a selected salient region and k is the number of region.

2) Text Representation: Similar to the region-level representation for image modality, here, we also extract word-level features for text representation. To fully incorporate the semantic context in a text sentence, we employ an RNN to embed the words along with their context. Specifically, for a text sentence T, we represent it with a one-hot vector of its index in the word vocabulary and then embed the word into a d_t-dimensional embedding vector through an embedding matrix.

Thus, T is initially represented by multiple embedding vectors $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_l\}$, $\mathbf{e}_j \in \mathbb{R}^{d_t}$, $j \in [1, l]$, where l denotes the number of words in a sentence. We aim to further map the

initial representation E to the same h-dimensional semantic subspace, resulting in cross-modal representations for both image and text modality. Therefore, we additionally use a bidirectional GRU [33] to map the embedding vectors $\mathbf{E}$ to the semantic vector space. The bidirectional GRU contains a forward GRU and a backward GRU that read words from $\mathbf{e}_1$ to $\mathbf{e}_l$ and reversely. The final word feature $\mathbf{t}_j$ for $\mathbf{e}_j$ in the semantic vector space can be formulated as

$$\mathbf{t}_j = \overleftrightarrow{GRU}(\mathbf{e_j}, \mathbf{W}_t), \quad j \in [1, l] \tag{2}$$

where $\mathbf{W}_t$ denotes the parameters in bi-GRU unit. Therefore, the complete representation of a text T_q consists of a set of semantic vectors $\{\mathbf{t}_1, \ldots, \mathbf{t}_l\}$, $\mathbf{t}_j \in \mathbb{R}^h$, where each $\mathbf{t}_j$ encodes a word and l is the number of words.

B. Cross-Modal Attention

The existing SCAN method is the first to match the image and text with a local staked cross attention mechanism, which has been proven effective since it accurately measures the interaction between fine-grained image patch and sentence word. In our CASC, we follow this effective attention module to construct our base model of local alignment. Similarly, the proposed CASC attends differentially to image patches and words using both as context to each other to measure the similarity of an image–text pair. Specifically, given an image I with k region features $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and a text T with l word features $\{\mathbf{t}_1, \ldots, \mathbf{t}_l\}$, we first compute the cosine similarity for each pairwise region and word as

$$s_{i,j} = \frac{\mathbf{v}_i^T \mathbf{t}_j}{\|\mathbf{v}_i\| \|\mathbf{t}_j\|}, i \in [1, k], \quad j \in [1, l] \tag{3}$$

where $s_{i,j}$ represents the similarity between the ith region and the jth word, and $\|\cdot\|$ means the L_2-norm on the number of image regions.

Then, as illustrated in Fig. 2, the cross-modal attention in CASC can be divided into two parts: I2T attention and T2I attention. Unlike the SCAN that takes the attention mechanism from these two directions, I2T and T2I, individually, here we average them for performance balancing in our CASC approach.

1) I2T Attention: For the I2T attention, it attends to words in the sentence according to each image region and then determines the importance of the image regions to the sentence by comparing each image region to the corresponding attended word vector. With the similarity between each pairwise image patch and sentence word $s_{i,j}$, we compute the weighted combination of word representations as the word attention vector $\mathbf{a}_i^T$ to attend on words for each image region

$$\alpha_{i,j} = \text{softmax}(\lambda_1 s_{i,j}); \qquad \mathbf{a}_i^T = \sum_{j=1}^{l} \alpha_{i,j} \mathbf{t}_j \tag{4}$$

where λ_1 controls the smoothness of the softmax function. Given the word attention features, we can just determine the importance of each image region to the whole sentence, as if a region is not described in the sentence, its feature would not be similar to the word attention vectors. Specifically, the similarity $S_{I2T}(\mathbf{v}_i, \mathbf{a}_i^T)$ between the ith region and the sentence is computed according to the word attention vector $\mathbf{a}_i^T$ and each image region feature $\mathbf{v}_i$, as

$$S_{I2T}(\mathbf{v}_i, \mathbf{a}_i^T) = \frac{\mathbf{v}_i^\top \mathbf{a}_i^T}{\|\mathbf{v}_i\| \|\mathbf{a}_i^T\|}. \tag{5}$$

Finally, the similarity between image I and T via I2T attention can be accumulated averagely by attending each image region to the whole sentence, as $S_{I2T}(I, T) = (1/k) \sum_{i=1}^{k} S_{I2T}(\mathbf{v}_i, \mathbf{a}_i^T)$.

2) T2I Attention: For the T2I attention direction, we first attend to image regions for each word and determine the importance of each word to the image attention vector. Similarly, to attend on image regions with respect to each word, we compute the weighted combination of image region features as the image attention vector $\mathbf{a}_j^V$

$$\alpha'_{i,j} = \text{softmax}(\lambda_2 s_{i,j}); \qquad \mathbf{a}_j^V = \sum_{i=1}^{k} \alpha'_{i,j} \mathbf{v}_i \tag{6}$$

where λ_2 balances the smoothness of the softmax function same as λ_1, and $s'_{i,j}$ is similar as the definition of $s_{i,j}$ in (3), whereas $s'_{i,j}$ is normalized according to the number of words. Likewise, we can measure the similarity between the jth word and the image attention vector as

$$S_{T2I}(\mathbf{t}_j, \mathbf{a}_j^V) = \frac{\mathbf{t}_j^\top \mathbf{a}_j^V}{\|\mathbf{t}_j\| \|\mathbf{a}_j^V\|}. \tag{7}$$

Finally, the similarity between image I and T via T2I attention can be accumulated by attending each word to the image, as $S_{T2I}(I, T) = (1/l) \sum_{j=1}^{l} S_{T2I}(\mathbf{t}_j, \mathbf{a}_j^V)$.

3) Local Alignment With Cross-Modal Attention: We aim to incorporate the full local alignment using both image regions and words in a sentence, where the alignment can be assessed by the image–text similarity measure that we have derived earlier. In previous studies, the hinge-based triplet loss [45], [55] that addresses a common ranking objective for local alignment is achieved by setting a considerable margin between the positive and negative image–text pairs. In practice, the negative samples are chosen from a minibatch rather than the entire pairs for computational efficiency, which deteriorates the ranking objective.

In the proposed CASC, we focus on the hardest negatives in a minibatch following the latest work [1], [55]. Specifically, we use the triplet ranking loss with the hardest negative samples to maximize the similarity of positive pairs and minimize the similarity of negative pairs. Given a positive pair (I, T), the hardest negatives are selected by $I_- = \text{argmax}_{M \neq I} S(M, T)$ and $T_- = \text{argmax}_{N \neq T} S(I, N)$, where M and N are samples in image and text modality, respectively. With the predefined margin σ, then the triplet loss for image–text pair (I, T) in the proposed CASC can be formulated as

$$\begin{aligned} L_{\text{loc}}^I(I, T) &= [\sigma - S_{I,T} + S_{I,T_-}]_+ \\ L_{\text{loc}}^T(I, T) &= [\sigma - S_{I,T} + S_{I_-,T}]_+ \end{aligned} \tag{8}$$

where $S_{I,T}$ and S_{I,T_-} denote the similarity between the matched image–text pair and the similarity between the hardest negative samples, respectively. The operator $[z]_+ = \max(0, z)$

compares the tolerance value with zero. Finally, we define our triplet loss for local alignment on N image–text pairs as

$$L_{\text{loc}}(I,T)=\sum_{n=1}^{N}L_{\text{loc}}^{I}(I_n,T_n)+L_{\text{loc}}^{T}(I_n,T_n). \tag{9}$$

C. Global Semantic Consistence

1) Semantic Labels Extraction: Since the existing data sets, e.g., Flickr30k and MSCOCO, do not explicitly provide supervised label information for the pairwise image and text, one effective solution is building a label vocabulary based on the attached sentences of images, where the meaningful words are extracted as the semantic labels. This strategy also bypasses the labor cost of manually defining appropriate semantic labels and assigning them to each of the image–text pairs.

Specifically, considering that a large number of unique words in the sentences refer to various concepts, such as objects, properties, quantities, and actions with different frequencies, we can select the top-K frequently occurred words in all the sentences to build the label dictionary to ensure its diversity. Here, K is an integer that relies on the frequency distribution of a specific data set. Then, each image–text pair can be assigned with one or more semantic labels to represent the high-level semantics. In Section IV-A2, we depict the detailed label extraction procedure in our experiments.

2) Semantic Consistence Preservation: Unlike previous global matching methods that directly extract global features for image and text instances, here, we utilize the learned attention vectors of image regions and sentence words to generate the global feature for two modalities. In particular, given an image–text pair (I,T) and the learned attention vectors $\{\mathbf{a}_j\}^V$, $j\in[1,l]$ [in (6)] of image regions and $\{\mathbf{a}_i\}^T$, $i\in[1,k]$ of words [in (4)], the global attention vector for I_p and T_q is, respectively, the averaged vector for image regions and text words, as

$$\mathbf{A}^I=\frac{1}{l}\sum_{j=1}^{l}\mathbf{a}_j^V;\quad \mathbf{A}^T=\frac{1}{k}\sum_{i=1}^{k}\mathbf{a}_i^T. \tag{10}$$

It is expected that both the two global attention vectors $\mathbf{A}^I$ and $\mathbf{A}^T$ are semantically consistent with the semantic labels of (I_p,T_q).

Intuitively, associating the two vectors in (10) with semantic labels can be cast to a multilabel classification problem. Thus, in our CASC, we learn a nonlinear mapping from the global attention vectors to the semantic labels used two fully connected layers with tanh activation. For an image–text pair (I,T), its ground-truth label vector is $\mathbf{Y}=\{y_1,\ldots,y_c,\ldots,y_C\}$, where $y_c=1$ denotes the presence of the cth label for the image–text pair and C denotes the label amount. The predicted targets using the global attention vectors of image and text are $\hat{\mathbf{Y}}^I$ and $\hat{\mathbf{Y}}^T$, respectively. As the label distribution is unbalanced in the built label dictionary, we further employ the positive weighting scheme for each label in the multilabel prediction procedure. In particular, p_c is a positive weight value, which is computed as the ratio of negative and positive samples of label c. The positive weight scheme is effective to boost the recall of labels with low frequency in label dictionary. Then, predicting targets $\hat{\mathbf{Y}}^I$ and $\hat{\mathbf{Y}}^T$ can be derived as a set of binary classification task for each label on both image and text samples, as

$$\begin{aligned}L_{\text{glob}}(I,T)=&-\sum_{c=1}^{C}\left[p_c y_c\log\hat{y}_c^I+(1-y_c)\log\left(1-\hat{y}_c^I\right)\right]\\&-\sum_{c=1}^{C}\left[p_c y_c\log\hat{y}_c^T+(1-y_c)\log\left(1-\hat{y}_c^T\right)\right].\end{aligned} \tag{11}$$

D. Optimization and Testing

1) Optimization: The final objective in our CASC method is a combination of the cross-modal attention loss and the semantic label prediction loss. Specifically, for all the N image–text pairs, the proposed CASC optimizes the final loss function with the ADAM [56] optimizer as

$$\mathcal{L}=\sum_{n=1}^{N}L_{\text{loc}}(I_n,T_n)+\alpha\sum_{n=1}^{N}L_{\text{glob}}(I_n,T_n) \tag{12}$$

where α is a constant value that balances the impact of two loss terms. In practice, the model parameters involved in (12) can be iteratively optimized by the widely used stochastic gradient descent (SGD) algorithm similar to [23] and [45], which also integrate several different loss terms.

2) Testing: In most existing work, the image–text similarity on the testing phase is measured by the local alignment of image regions and text words, which contains the I2T attention similarity $S_{I2T}(\cdot)$ and the T2I attention similarity $S_{T2I}(\cdot)$. However, this similarity measure is inaccurate since the global semantic consistence is not preserved. In our CASC, we further consider the global semantic similarity to account for the predicted labels and the ground-truth ones. For example, given a query image I' and a text T in the database, where T is associated with ground-truth label vector $\mathbf{Y}_T$. We first predict the label vector $\hat{\mathbf{Y}}'$ for I', then we can compute the cosine similarity between $\hat{\mathbf{Y}}'$ and $\mathbf{Y}_T$ as $S_G(I',T)=\mathbf{Y}_T^{\top}\hat{\mathbf{Y}}'/(\|\mathbf{Y}_T\|\|\hat{\mathbf{Y}}'\|)$. The final image–text similarity for the query image I' and the text T is a weighted combination of the three similarity terms, as

$$S(I',T)=\beta(S_{I2T}(I',T)+S_{T2I}(I',T))+(1-\beta)S_G(I',T), \tag{13}$$

where $\beta\in[0,1]$ is a tradeoff coefficient that controls the importance of each term in computing the image–text similarity score, and the detailed assessment of β on the matching performance is provided in Section IV.

IV. Experiment

A. Experimental Setup

1) Data Sets: Here, we briefly introduce the two data sets in our experiment and their widely used experimental protocols as follows.

1) Flickr30k [57] contains 31 783 images collected from the Flickr website. Each image is manually annotated by five sentences. Following [58], [59], we use the

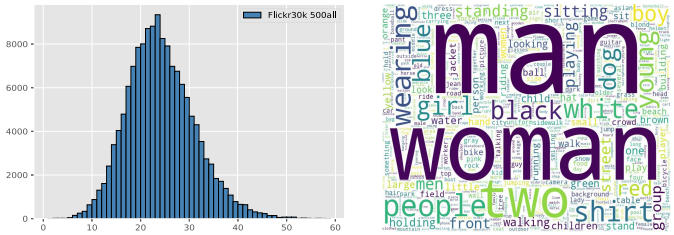


Fig. 3. Illustration of (a) label distribution and (b) word cloud of the selected top-500 semantic concepts on Flickr30k data set.

public training, validation and test split that contain 28 000, 1000, and 1000 images, respectively.

2) MSCOCO [29] consists of 123 287 images, and each one is associated with five sentences.

We use the public training and validation splits [1], with 113 287 and 5000 images, respectively. We report results by either averaging over fivefolds of 1000 test images or directly evaluating on the full 5000 test images.

2) Implementation Details: Our proposed CASC is implemented by PyTorch [60], and all the experiments are conducted on a workstation with an NVIDIA 1080 Ti GPU. The implementation code has been publicly available at https://github.com/Wangt-CN/Code_CASC. In representation learning, for each image, we extract 36 patches and each dimension is 2048 and then separately feed them into a linear fully connected layer to a 1024-D common space. For each sentence, we set the max length for all sentences as the maximum length in the minibatch and use zero-padding when a sentence is not enough. The word embedding size is set to 300, and then, a bidirectional GRU is used to encode the sentence word and the number of hidden units is 1024. Furthermore, the L_2-norm was added after image and sentence encoding to normalize the features. For multilabel prediction, we deploy two-layer feed-forward neural networks activated by tanh function to nonlinearly projecting, that is $(1024 \rightarrow 500 \rightarrow 512)$.

For the global semantic label extraction, as each image is associated with five text sentences on both Flickr30k and MSCOCO data sets, we select the words of nouns, adjectives, verbs, and numbers in the sentences that are all meaningful semantic labels to build the label dictionary. To ensure that each image has at least one label based on the label dictionary, we aggregate the corresponding five text sentences as a united text for each image and choose appropriate K to control the size of the label dictionary. Fig. 3 demonstrates the general distribution of all the labels in the Flickr30k data set and the word cloud of the selected top $K = 500$ words of all types. In addition, in this section, we also consider different compositions of label vocabulary that only preserving nouns, verb, or adjective to investigate the impact of different types of labels on our proposed CASC method.

During training, we set $\lambda_1 = 9$, $\lambda_2 = 4$, the L_{loc} margin $\sigma = 0.2$, and α is empirically set to be 10 to make L_{loc} and L_{glob} numerically closer. We trained our model by two independent Adam optimizer for L_{loc} and L_{glob} with a minibatch size of 128 in all our experiments. The initial learning rate for L_{loc} is set to 0.0002 for Flickr30k and 0.0005 for MSCOCO, while for L_{glob}, the initial learning rate is 0.01 empirically, decayed by 10 every 10 epoch. Besides, to fairly consider all training data in multilabel classification, we set the positive weight p_c by adding weights to positive examples, which is set to be the ratio between negative and positive examples. Considering the setting of our global semantic label vocabulary, in Section IV-D1, we explicitly assess the effect of using different label types on our CASC method. Here, we empirically choose 500 nouns (denoted as "500 nn") as the semantic labels for our CASC to compare with other methods.

3) Compared Methods and Evaluation Metric: We conduct two kinds of image–text matching tasks: 1) sentence retrieval, i.e., retrieving ground-truth sentences given a query image (I2T) and 2) image retrieval, i.e., retrieving ground-truth images given a query text (T2I). The commonly used evaluation metric for these tasks is the "recall at K" (R@K) defined as the recall rate at the top K results to the query, and usually, $K = \{1, 5, 10\}$. We also used "mR" by averaging all the recall scores to evaluate the overall performance for both image and sentence retrievals. We also used "mR" score proposed in [46] for additional evaluation, which averages all the recall scores of R@K to assess the overall performance for both sentence retrieval and image retrieval tasks.

We compare our proposed model with several recent state-of-the-art models to verify its effectiveness. As mentioned in Section II, these compared methods can be divided into three groups.

1) *Global Matching Methods:* DCCA [32], DSPE [16], DPC [62], 2WayNet [61], VSE++ [55], TBNN [27], and cycleMatch [39].
2) *Fine-Grained Matching Methods:* DVSA [24], CCL [35], sm-LSTM [44], DAN [45], and SCAN [1].
3) *Hybrid Matching Methods:* SCO [46], CRAN [48], JGCAR [49], BSSAN [50], and MDM [51].

Note that for fair and objective comparison purpose, feature representation of both image and text modalities in all compared methods is exactly following [23] and [46], except for DCCA, CCL, and CRAN methods that use VGG features, we directly report their results in the original articles. In addition, we take a baseline of our proposed CASC, termed CASC-base, into account. The CASC-base only performs local alignment without global semantic consistence, which can also be considered as a fine-grained approach based on SCAN. It directly averages the two directions, i.e., SCAN (I2T) and SCAN (T2I), to balance the performance of image retrieval and sentence retrieval. To systematically validate the effectiveness, we also conduct an experiment with two variants of our CASC: CASC (I2T) and CASC (T2I), which use the I2T branch or T2I branch, respectively.

B. Evaluation of Ablation Models

To systematically evaluate the contributions of different model components (i.e., local alignment, global constraint, and their combination) in our method, we use the two data sets MSCOCO (1000 testing) and Flickr30k as test beds and design various ablation models, as shown in Tables I and II, which are defined as follows: 1) "Local" denotes only using local alignment; 2) "Local Attention" means applying local alignment

TABLE I
COMPARISON OF DIFFERENT MODEL COMPONENTS OF OUR PROPOSED CASC ON MSCOCO (1000 TESTING) DATA SET

Index	Ablation Model				Sentence Retrieval			Image Retrieval			mR	Evaluation Time (s)
	Local	Local Attention	Global	Global Label	R@1	R@5	R@10	R@1	R@5	R@10		
①	√				68.7	93.1	97.5	52.2	85.2	94.0	81.8	158.2
②		√			69.6	94.0	98.3	53.3	86.4	94.1	82.6	162.9
③			√		66.3	91.4	96.3	50.4	84.2	91.9	80.1	143.5
④				√	23.1	48.4	77.3	18.5	41.1	62.3	45.1	45.9
⑤		√	√		71.6	95.2	98.4	57.9	89.0	95.	84.6	188.1
⑥	√		√		71.7	95.5	98.7	58.1	89.2	95.4	84.8	183.8
⑦	√			√	71.9	95.7	98.7	58.3	89.7	95.9	85.0	160.2
⑧		√		√	**72.3**	**96.0**	**99.0**	**58.9**	**89.8**	**96.0**	**85.3**	163.4

TABLE II
COMPARISON OF DIFFERENT MODEL COMPONENTS OF OUR PROPOSED CASC ON FLICKR30K DATA SET

Index	Flickr30k						
	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10	
①	62.5	87.4	92.9	43.1	73.2	82.5	73.6
②	63.7	88.6	93.7	44.0	73.9	82.6	74.4
③	61.3	86.2	91.7	42.2	72.5	82.1	72.7
④	19.7	43.1	75.2	13.8	37.6	58.9	41.4
⑤	68.0	89.4	94.6	58.8	76.2	85.9	78.8
⑥	67.7	89.2	94.7	59.1	76.5	86.1	78.9
⑦	68.2	89.9	95.3	59.9	77.8	86.2	79.6
⑧	**68.5**	**90.6**	**95.9**	**60.2**	**78.3**	**86.3**	**80.0**

with attention mechanism (CASC-base); 3) "Global" refers to adopting typical global constraint with an average local feature that can be easily obtained; 4) "Global Label" represents just using global constraint with the semantic label; 5) "Local Attention + Global" is to combine the two best individual modules (i.e., local attention mechanism 2 and typical global constraint 3); 6) "Local + Global" is to combine 1 and typical global constraint 3; 7) "Local + Global Label" refers to combining 1 and 4; 8) "Local Attention + Global Label" denotes integrating 2 and 4, which is our proposed CASC.

The retrieval results of each ablation modules on two data sets are shown in Tables I and II. We can observe that, first, from the result of 3, we can see only using the global feature for alignment has achieved comparable results with sufficient training data on MSCOCO data set. A silent performance drop exists since the local feature contains much more information and details. However, the performance of 4 deteriorates severely. The potential reason is that the independent global label can be too weak for supervised alignment and matching, which is more suitable for coarse tasks, such as classification. We can also observe that hybrid methods can often get higher performance than the individual module, which implicitly incorporates the advantages of both local and global methods to predict more accurate similarity. However, combining the two best individual modules (i.e., 5) cannot bring the best hybrid performance. The potential reason is that the image–text matching process from both global and local may interact with each other, leading to the performance drop. Moreover, the attention mechanism has a beneficial impact on the fine-grained local alignment for it can attend to different regions with different queries. Despite the matching performance, the time consuming and model efficiency can also be an important aspect. From Table I, we can see that our CASC achieves the best performance with a moderate time consuming, which benefits from the efficient design of our global label constraint, especially compared with 5. Please note that methods such as 1, 3, and 4 consume lower time; however, their results still have a great margin compared with our method since they are too simple.

C. Comparison With the State of the Arts

Table III shows the overall sentence retrieval and image retrieval results of our model and the counterparts on the Flickr30k and MSCOCO (1000 testing) data sets. Besides, Table IV reports the two kinds of retrieval results on the MSCOCO (5000 testing) data set. The best scores in the two tables are highlighted in bold font. Note that in Table III, three multirows denote the global matching methods, the fine-grained matching methods, and the hybrid matching methods, respectively. For our proposed CASC, we have mentioned using a noun (nn) for the type of label because noun can get the best performance. The detailed analysis can be referred to Section IV-D1. From Table III, we can obtain following observations.

1) Our proposed CASC is superior to both global matching approaches and fine-grained matching approaches that without attention mechanism. The reason is that local alignment with attention mechanism aims to focus on certain aspects of data sequentially and aggregate essential information over time to infer results, which benefits the model to capture the full latent alignment between vision and semantic. Similar results have been shown in Section IV-B.
2) The proposed CASC also outperforms DAN and SCAN that use attention mechanisms for discovering alignments between all possible image region-word pairs. Different from these methods, our CASC adds the global-level semantic alignment via multilabel prediction on the basis of attention mechanism. It simultaneously uses the coattention of image patches and words as reference and the ground-truth semantic label in the matched sentence as supervision for inferring the image–text similarity, especially improving the matching performances on complex queries. Moreover, compared

TABLE III

OVERALL COMPARISON RESULTS OF TWO RETRIEVAL TASKS ON FLICKR30K AND MSCOCO (1000 TESTING) DATA SETS. THE COMPARED METHODS IN THE TOP THREE PANELS ARE GLOBAL, FINE-GRAINED, AND HYBRID MATCHING METHODS, RESPECTIVELY

Methods	Flickr30k							MSCOCO						
	Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
DCCA [32] (2013)	27.9	56.9	68.2	26.8	52.9	66.9	49.9	22.5	34.6	45.5	19.2	30.4	41.3	32.3
DSPE [16] (2016)	40.3	68.9	79.9	29.7	60.1	72.1	58.5	50.1	79.7	89.2	39.6	75.2	86.9	70.1
VSE++ [55] (2017)	41.3	69.0	77.9	31.4	59.7	71.2	58.4	57.2	85.1	93.3	45.9	78.9	89.1	74.9
2WayNet [61] (2017)	49.8	67.5	-	36.0	55.6	-	-	55.8	75.2	-	39.7	63.3	-	-
DPC [62] (2017)	55.6	81.9	89.5	39.1	69.2	80.9	69.4	65.6	89.8	95.5	47.1	79.9	90.0	78.0
TBNN [27] (2019)	43.2	71.6	79.8	31.7	61.3	72.4	60.0	54.9	84.0	92.2	43.3	76.4	87.5	73.1
CycleMatch [39] (2019)	58.6	83.6	91.6	43.6	75.3	84.2	72.8	61.1	86.8	94.2	47.9	80.9	90.9	76.9
CCL [35] (2018)	37.7	69.4	81.1	37.3	68.4	80.0	62.3	36.3	47.2	62.4	31.1	40.6	59.7	46.2
DVSA [24] (2015)	22.2	48.2	61.4	15.2	37.7	50.5	39.2	38.4	69.9	80.5	27.4	60.2	74.8	58.5
sm-LSTM [44] (2017)	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
DAN [45] (2017)	55.0	81.8	89.0	39.4	69.2	79.1	68.9	65.1	89.2	95.4	56.7	80.2	90.0	79.4
SCAN (T2I)[1] (2018)	61.8	87.5	93.7	45.8	74.4	83.0	74.4	70.9	94.5	97.8	56.4	87.0	93.9	83.4
SCAN (I2T)[1] (2018)	67.9	89.0	94.4	43.9	74.2	82.8	75.4	69.2	93.2	97.5	54.4	86.0	93.6	82.3
CRAN [48] (2018)	38.1	70.8	82.8	38.1	71.1	82.6	63.9	-	-	-	-	-	-	-
JGCAR [49] (2018)	44.9	75.3	82.7	35.2	62.0	72.4	62.1	52.7	82.6	90.5	40.2	74.8	85.7	71.1
SCO [46] (2017)	55.5	82.0	89.3	41.1	70.5	80.1	69.8	69.9	92.9	97.5	56.7	87.5	94.8	83.2
BSSAN [50] (2019)	44.6	74.9	84.3	33.2	62.6	72.9	62.1	56.0	82.6	91.3	41.8	76.7	88.5	72.8
MDM [51] (2019)	44.9	75.4	84.4	34.4	67.0	77.7	64.0	54.7	84.1	91.9	44.6	79.6	90.5	74.2
CASC-base	63.7	88.6	93.7	44.0	73.9	82.6	74.4	69.6	94.0	98.3	53.3	86.4	94.1	82.6
CASC (T2I)	62.2	87.8	93.9	46.2	74.1	82.7	74.5	71.1	94.2	97.7	56.8	87.2	94.3	83.6
CASC (I2T)	68.4	88.9	94.5	44.0	74.1	82.6	75.4	69.5	93.3	97.7	54.5	86.1	93.8	82.5
CASC (ResNet-152)	68.3	**90.7**	95.8	**50.3**	78.2	**86.5**	**78.3**	**72.5**	95.6	98.7	58.8	**89.9**	95.7	85.2
CASC (ResNet-101)	**68.5**	90.6	**95.9**	50.2	**78.3**	86.3	**78.3**	72.3	**96.0**	**99.0**	**58.9**	89.8	**96.0**	**85.3**

TABLE IV

COMPARISON RESULTS OF IMAGE RETRIEVAL AND SENTENCE RETRIEVAL ON MSCOCO (5000 TESTING) DATA SET

Method	MSCOCO					
	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DCCA [32] (2013)	6.9	21.1	31.8	6.6	20.9	32.2
CCL [35] (2018)	18.6	47.4	62.5	19.6	46.9	62.3
CRAN [48] (2018)	23.0	52.0	66.0	21.1	48.9	64.5
VSE++ [55] (2017)	41.3	69.2	81.2	30.3	59.1	72.4
DPC [62] (2017)	41.2	70.5	81.1	25.3	53.4	66.4
GXN [37] (2018)	42.0	-	84.7	31.7	-	74.6
SCO [46] (2017)	42.8	72.3	83.0	33.1	62.9	75.5
SCAN [1] (2018)	46.4	77.4	87.2	34.4	63.7	75.7
CASC (Ours)	**47.2**	**78.3**	**87.4**	**34.7**	**64.8**	**76.8**

with SCAN or our variate CASC that using individual I2T or T2I branch, our CASC also gains an improvement. The reason is that in our semantic constraint, two branches bring both image and sentence attention representation into the shared label prediction, which helps a lot in image–text matching.

3) Comparing to the hybrid methods, i.e., CRAN, JGCAR, SCO, and MDM, our CASC also gains a notable improvement, which indicates that our semantic label is more efficient than the typically used global constraint that easily obtained by averaging local features.
4) For our baseline model CASC-base that is similar to the current state-of-the-art method SCAN but averaging the two directions of I2T and T2I for cross-modal retrieval, we can see a slight performance drop comparing to SCAN. The probable reason is that the I2T and T2I attention directions may interact with each other with a simple averaging. However, when combining CASC-base with the global semantic constraint, a remarkable improvement is achieved by CASC. It again demonstrates the effectiveness of our CASC on exploiting the semantic labels to preserve the global semantic consistence and the local alignment procedure in CASC-base in a much more effective way.
5) We also tried to change the backbone of ResNet-101 to the more commonly used ResNet-152. As shown in the last two rows of Table III, the performance improvement is slight, and the potential reason is that the ResNet-101 have provided a quite enough prior detection and the usage of ResNet-152 may cause more consumption.

Moreover, we also present several typical visual results of retrieved sentences and images by the three models: the state-of-the-art model SCAN, CASC-base, and CASC in Fig. 4. For our proposed CASC, we also visualize the top-10 labels and their predicted scores from the image to validate its effectiveness. For example, as shown in Fig. 4, the CASC-base fails to capture the full latent visual-semantic alignment. For example in text query, with local attention alignment, the model can retrieval the black dog and water or dog and grass, but it cannot retrieval the correct image from the whole sentence probably because there exists interference between patches and words. SCAN generally obtains better matches than CASC-base, as it obtains more effective local alignment of image regions and textual words in either I2T or T2I directions. Nevertheless, CSAC achieves the best matches comparing with CASC-base and SCAN, as it can incorporate most of the elements in the

Fig. 4. Comparison of retrieval examples obtained by three models: SCAN, CASC-base, and CASC. The ground-truth sentence and images are marked as red, while some results sharing similar meanings as ground-truth are marked with an underline.

TABLE V
COMPARISON OF DIFFERENT TYPES OF SEMANTIC LABELS USED IN THE PROPOSED CASC ON FLICKR30K DATA SET

Types	Flickr30k						MSCOCO					
	Sentence Retrieval			Image Retrieval			Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
nn	68.5	**90.6**	**95.9**	**50.2**	**78.3**	**86.3**	**72.3**	96.0	**99.1**	**58.9**	**89.8**	**96.1**
adj	55.3	81.9	90.7	40.8	69.9	79.5	67.2	91.0	94.4	52.9	85.1	92.7
verb	65.3	89.6	95.4	48.0	77.4	85.0	70.4	93.3	96.6	55.3	87.7	94.2
nn+adj	**69.7**	90.4	**95.9**	49.4	78.0	85.9	72.1	**96.2**	98.8	58.7	89.2	95.8
nn+verb	69.1	90.3	95.7	49.7	78.0	86.2	72.2	95.9	98.9	58.8	89.0	95.9
adj+verb	64.6	88.7	95.1	46.3	76.1	84.2	70.1	92.9	96.1	54.8	87.2	93.4
all	68.1	90.3	95.6	50.0	78.2	85.8	71.9	95.7	98.4	58.5	89.1	95.6

images and sentences as a global constraint to alleviate the problem of interference.

D. Further Analysis

1) Analysis on Label Type: To qualitatively validate the effectiveness of our proposed CASC, especially its semantic label constraint, in this experiment we first investigate the different combinations of label types used for building the label dictionary in terms of our global constraint.

We first set the dictionary size as 500 and choose meaningful labels from noun (as "nn"), adjective (as "adj"), verb, and their combination (as "nn + adj ", "all," and so on). As shown in Table V, we can observe that the label type of 500 nn, 500 nn + adj, 500 nn + verb, and 500 all have similar retrieval performance, and the possible reason is that they all contain the primary semantic information (noun words) of the sentences to ensure the accuracy of multilabel prediction. Note that the case of 500 nn also contains some words, such as "white" and "playing," that can be considered the action and color information that does benefits the sentence semantic preservation. On the contrary, the cases of 500 adj and 500 verb fail to improve performance and even lead to a huge drop. It is probable that these types of words generate inferior semantic representation. In addition, selecting only adjective or verb words may not ensure that every image has at least one label.

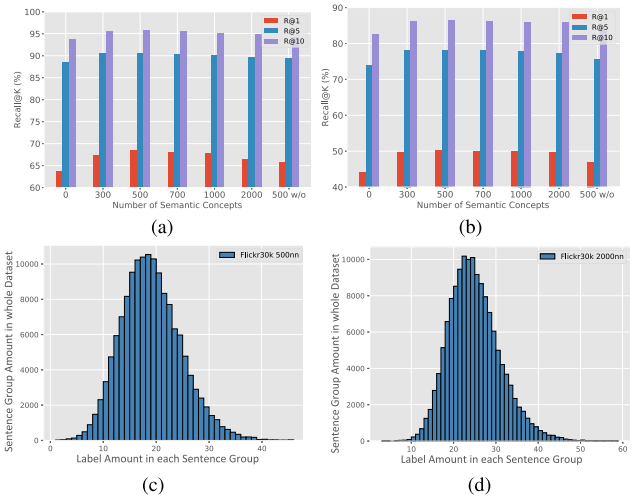


Fig. 5. (a) and (b) Comparison in terms of Recall@K using different label amounts. (c) and (d) Different label distributions on Flickr30k data set. (a) Recall@K of sentence retrieval. (b) Recall@K of image retrieval. (c) Semantic label distribution of 500 nn. (d) Semantic label distribution for 2000 nn.

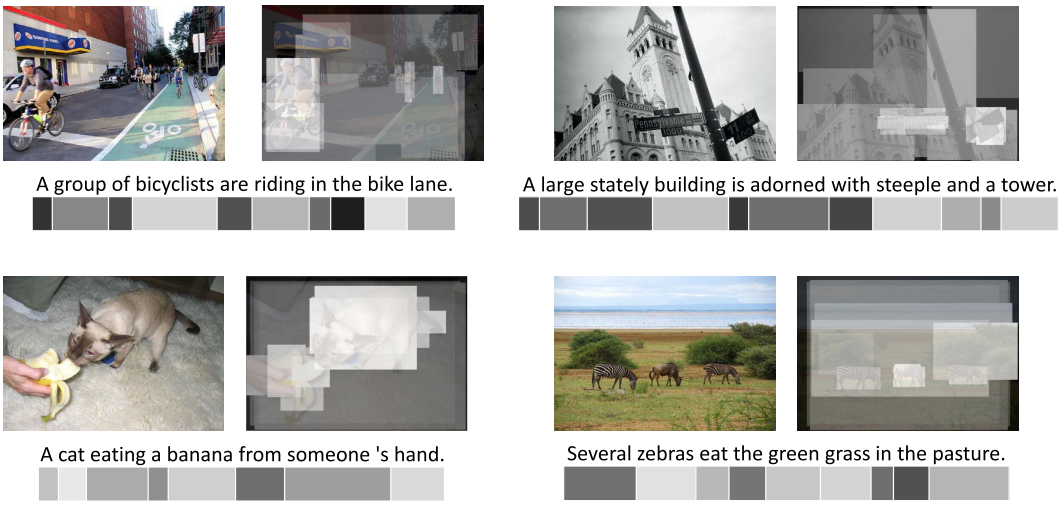


Fig. 6. Typical examples of attended image patches and sentence words of the images and their retrieved sentences by our CASC. The brightness denotes the attention weight, indicating the importance of the regions obtained by the cross-modal attention mechanism for local alignment. The lighter areas represent the attended instances, best viewed in colors.

2) Analysis on Label Amount: In this experiment, we investigate the effect of the label amount on our CASC. Specifically, we take Flickr30k as a test bed to evaluate the retrieval performance of our CASC by varying the number of semantic labels, where the result is shown in Fig. 5. Our proposed CASC ("500 labels") gains 5.2% relative improvement on sentence retrieval (R@1) and 14.1% improvement on image retrieval (R@1) comparing with no global label constraint ("0 label"). It shows the great effectiveness of the label prediction comparing to only focusing on the main part of the image or sentence for image and text matching. Besides, without the positive weighting scheme ("500 w/o"), the proposed CASC fails to achieve the best performance since the semantic label distribution is not balanced for effective multilabel classification.

Moreover, we can also observe that the retrieval performance of our CASC is robust with a various number of labels (e.g., from 300 to 2000). The potential reason is that the distributions of label frequency are quite similar with a different number of labels, as a typical example of 500 nn and 2000 nn shown in Fig. 5, which may lead to similar performance. Note that the x-axis denotes how many labels does a sentence group (five sentences attached to the same image) contains, while the y-axis represents how many sentence groups in the whole data set have the same number of labels. In practice, around 300 to 500 semantic labels are enough for our CASC to achieve considerable retrieval performance, which is efficient with much less labor cost.

3) Visualization of the Cross-Modal Attention: Cross-modal attention plays an important role in the proposed CASC. First, it is applied for local alignment by taking both image patches and sentence words as a context via the two directions of I2T and T2I.

Second, as shown in the overall architecture Fig. 2, the semantic multilabel prediction in CASC is also based on the global attention representation by averaging local attention from the two directions. To explicitly explore the effectiveness of the attention mechanism used in CASC, we directly visualize the global attention representations of typical images and their corresponding descriptions in Fig. 6. Note that in Fig. 6, the region-level features of images and sentences are represented with generated bounding boxes using Faster R-CNN and word-level bi-GRU, respectively. Then, for each bounding box (word) feature, the local attention mechanism will assign an attention weight with the corresponding caption (image).

Finally, we aggregate them together for attention visualization. From Fig. 6, we can see that our local alignment with the fine-grained attention can accurately attend the most important instance and infer the interplay between patches and words, which is beneficial for the further global semantic label prediction.

TABLE VI

TRAINING AND EVALUATION TIME OF DIFFERENT COMPETITORS WITH OUR PROPOSED CASC ON MSCOCO (1000 TESTING) DATA SET

Method	DAN	SCO	SCAN	CASC-base	CASC
mR (%)	79.4	83.2	83.4	82.6	85.3
Training Time (h)	55.4	61.2	56.2	58.8	59.1
Evaluation Time (s)	152.1	169.7	159.7	162.9	163.4

4) Analysis on Model Parameters: Since our CASC contains a set of parameters, such as λ_1, λ_2, α, β, and σ, for further investigating the effect of these hype-parameters, we conduct different experiments for parameter tuning during training. First, for the weighted similarity β in (13), we sample the value of the Recall@1 in both image and sentence retrievals when changing β from 0 to 1 by 0.1 every step, and the results are shown in Fig. 7. We can observe that 0.8 is the best weight for both on Flickr30k and MSCOCO that well balances the local and global similarities. In addition, we also apply the weighted similarity to the training for validation and model selection; however, it cannot gain additional improvement. The potential reason is that the global semantic similarity in our CASC has certain consistence with the local similarity, and it has less effect when using the two types of similarity metrics in the validation and evaluation processes. Second, λ_1 and λ_2 are also important for the

TABLE VII

MATCHING RESULTS OF IMAGE RETRIEVAL AND SENTENCE RETRIEVAL ON FLICKR30K AND MSCOCO (1000 TESTING) DATA SETS BY COMBINING OUR CASC WITH FINE-GRAINED METHODS SM-LSTM [44], DAN [45], AND SCAN [1]

Methods	Flickr30k							MSCOCO						
	Sentence Retrieval			Image Retrieval			mR	Sentence Retrieval			Image Retrieval			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
sm-LSTM	42.5	71.9	81.5	30.2	60.4	72.3	59.8	53.2	83.1	91.5	40.7	75.8	87.4	72.0
DAN	55.0	81.8	89.0	39.4	69.2	79.1	68.9	65.1	89.2	95.4	56.7	80.2	90.0	79.4
SCAN (T2I)	61.8	87.5	93.7	45.8	74.4	83.0	74.4	70.9	94.5	97.8	56.4	87.0	93.9	83.4
SCAN (I2T)	67.9	89.0	94.4	43.9	74.2	82.8	75.4	69.2	93.2	97.5	54.4	86.0	93.6	82.3
CASC-sm	42.8	72.5	82.1	30.6	61.2	73.4	60.4	54.1	83.8	92.2	41.4	76.6	88.4	72.9
CASC-DAN	57.9	83.3	90.2	41.1	71.8	80.3	70.8	65.7	90.3	95.8	57.8	80.9	90.4	80.2
CASC-SCAN (T2I)	62.2	87.8	93.9	46.2	75.1	82.7	74.7	71.1	95.2	98.7	56.8	87.2	94.3	83.8
CASC-SCAN (I2T)	68.4	89.6	94.9	44.0	74.6	83.4	75.8	69.5	93.7	98.2	54.6	86.7	94.1	82.8
CASC	**68.5**	**90.6**	**95.9**	**50.2**	**78.3**	**86.3**	**78.3**	**72.3**	**96.0**	**99.0**	**58.9**	**89.8**	**96.0**	**85.3**

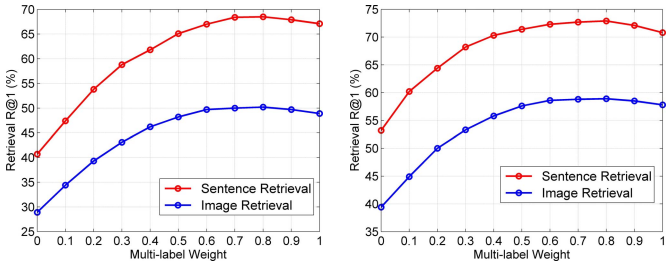


Fig. 7. Retrieval score R@1 with different chosen similarity weights β on (a) Flickr30k and (b) MSCOCO data sets.

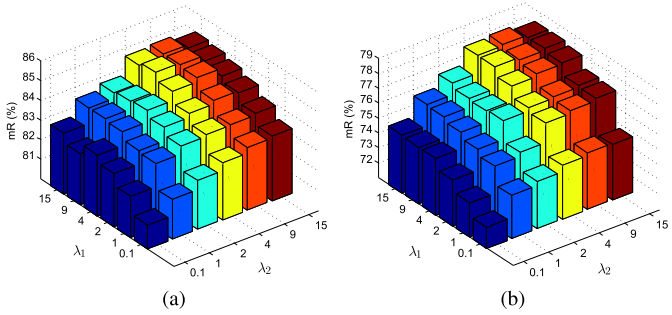


Fig. 8. Evaluation of the softmax smoothness parameters on MSCOCO and Flickr30k data sets. (a) Different values of λ_1 and λ_2 on MSCOCO. (b) Different values of λ_1 and λ_2 on Flickr30k.

smoothness in (4) and (7). As they influence the image and word attention vector simultaneously, we plot the bar figures on both data sets in Fig. 8. We can see that λ_1 and λ_2 in the range of [4, 9] achieve satisfying results on both MSCOCO and Flickr30k. Note that the value "1" represents the standard softmax function without smoothness. Moreover, the margin value σ is empirically selected following the previous work [1], [45], and we also try other values that also prove that 0.2 is the best. Besides, the value of α in (12) is also empirically selected to balance the two parts of loss (local loss and global loss) for stable gradient descent during the training stage.

5) Analysis on Time Consumption: In this section, we show the model training and evaluating time consumption on the MSCOCO data set to compare the complexity of our CASC and recent state-of-the-art approaches. As shown in Table VI, our CASC achieves better performance and a similar time

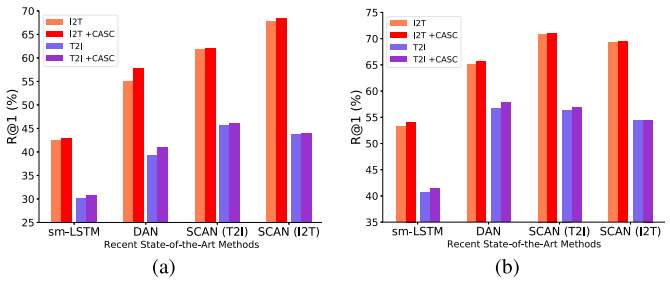


Fig. 9. Evaluation of applying our global semantic label prediction scheme to the recent methods sm-LSTM [44], DAN [45], and SCAN [1] on MSCOCO and Flickr30k data sets. (a) Our global semantic constraint applied to state-of-the-art methods on Flickr30k. (b) Our global semantic constraint applied to state-of-the-art methods on MSCOCO.

cost with the currently best model SCAN. Moreover, comparing with other costly methods, such as SCO, our CASC outperforms it on both matching score and model complexity. Furthermore, the time consumption of CASC slightly increases compared with CASC-base, which also demonstrates the efficiency of our global label prediction module in CASC.

6) Analysis on Global Label Constraint: To further investigate the effectiveness of our global label prediction scheme, we try to integrate our semantic label constraint to three recent fine-grained matching methods, i.e., sm-LSTM, DAN, and SCAN. The integrated baselines are denoted as CASC-sm, CASC-DAN, CASC-SCAN (T2I), and CASC-SCAN (I2T), respectively. As shown in Fig. 9 and Table VII, we can observe that combining with semantic label prediction, all these fine-grained methods achieve a considerable improvement on matching performance. Therefore, it further validates the effectiveness of the global label constraint in our CASC approach, which can help to obtain more accurate similarity measurement by preserving the global semantic consistence.

V. CONCLUSION

In this article, we proposed a novel semantic-enhance image–text matching model termed CASC, which creatively combines the global semantic consistence using multilabel prediction with the local image region-word alignment in a joint framework. The proposed CASC simultaneously uses the cross-modal attention of image patches and words as a reference and the semantic labels in matched sentences as supervision for inferring more accurate image-sentence similarity.

Furthermore, the semantic labels can be directly extracted from the text modality data, efficiently bypassing the labor-costly annotating process. Extensive experimental results on two standard data sets demonstrated the superiority of the proposed CASC compared with a bundle of the state-of-the-art methods on image–text matching task. For the future direction, we will explore the capability of our CASC on preserving and transferring the semantic consistence across different data sets.

REFERENCES

[1] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 106–125.

[2] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.

[3] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, *arXiv:1412.6632*. [Online]. Available: https://arxiv.org/abs/1412.6632

[4] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, "Matching images and text with multi-modal tensor fusion and re-ranking," in *Proc. 27th ACM Int. Conf. Multimedia (MM)*, 2019, pp. 12–20.

[5] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5292–5303, Nov. 2018.

[6] L. Jin, K. Li, Z. Li, F. Xiao, G.-J. Qi, and J. Tang, "Deep semantic-preserving ordinal hashing for cross-modal similarity search," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1429–1440, May 2019.

[7] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2019.2928180.

[8] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1059–1071, May 2018.

[9] M. Hu, Y. Yang, F. Shen, N. Xie, and H. T. Shen, "Hashing with angular reconstructive embeddings," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 545–555, Feb. 2018.

[10] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.

[11] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[12] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, Jul. 2019.

[13] S. Antol *et al.*, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2425–2433.

[14] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," 2017, *arXiv:1704.03162*. [Online]. Available: https://arxiv.org/abs/1704.03162

[15] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Multimedia Conf.*, 2010, pp. 251–260.

[16] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5005–5013.

[17] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4127–4136.

[18] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.

[19] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[20] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[21] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, Jul. 2018.

[22] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3034–3044, Dec. 2018.

[23] E. Yang, T. Liu, C. Deng, and D. Tao, "Adversarial examples for Hamming space search," *IEEE Trans. Cybern.*, early access, doi: 10.1109/TCYB.2018.2882908.

[24] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, Apr. 2017.

[25] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1908–1917.

[26] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2630–2636.

[27] L. Wang, Y. Li, J. Huang, and S. Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 394–407, Feb. 2019.

[28] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2641–2649.

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.

[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[31] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 7–16.

[32] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[33] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "RNN Fisher vectors for action recognition and image annotation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 833–850.

[34] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 154–162.

[35] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[36] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7181–7189.

[38] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 1–24, Feb. 2019.

[39] Y. Liu, Y. Guo, L. Liu, E. M. Bakker, and M. S. Lew, "CycleMatch: A cycle-consistent embedding network for image-text matching," *Pattern Recognit.*, vol. 93, pp. 365–379, Sep. 2019.

[40] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11477–11486.

[41] X. Zhou *et al.*, "Graph convolutional network hashing," *IEEE Trans. Cybern.*, to be published.

[42] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Hierarchical multimodal LSTM for dense visual-semantic embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1899–1907.

[43] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using Fisher vectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4437–4446.

[44] Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 2, no. 6, pp. 2310–2318.

[45] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 299–307.

[46] Y. Huang, Q. Wu, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6163–6171.

[47] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 707–723.
[48] J. Qi, Y. Peng, and Y. Yuan, "Cross-media multi-level alignment with relation attention network," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 892–898.
[49] S. Wang, Y. Chen, J. Zhuo, Q. Huang, and Q. Tian, "Joint global and co-attentive representation learning for image-sentence retrieval," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 1398–1406.
[50] F. Huang, X. Zhang, Z. Zhao, and Z. Li, "Bi-directional spatial-semantic attention networks for image-text matching," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2008–2020, Apr. 2019.
[51] L. Ma, W. Jiang, Z. Jie, and X. Wang, "Bidirectional image-sentence retrieval by local and global deep matching," *Neurocomputing*, vol. 345, pp. 36–44, Jun. 2019.
[52] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," 2017, *arXiv:1707.07998*. [Online]. Available: https://arxiv.org/abs/1707.07998
[53] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[54] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.
[55] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*. [Online]. Available: https://arxiv.org/abs/1707.05612
[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[57] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Dec. 2014.
[58] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: https://arxiv.org/abs/1411.2539
[59] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2623–2631.
[60] *Pytorch Open Source Toolkit*. Accessed: Jul. 10, 2018. [Online]. Available: https://github.com/pytorch/pytorch
[61] A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4601–4611.
[62] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, "Dual-path convolutional image-text embedding with instance loss," 2017, *arXiv:1711.05535*. [Online]. Available: https://arxiv.org/abs/1711.05535

Xing Xu received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015.

He is currently with the Center for Future Media and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests mainly focus on multimedia information retrieval, pattern recognition, and computer vision.

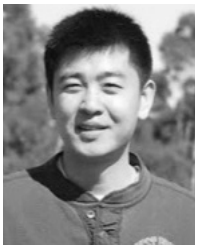
Tan Wang is currently with the Center for Future Media, the School of Computer Science and Engineering, and the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia content analysis, computer vision, and social media analysis.

Yang Yang (Member, IEEE) received the bachelor's degree from Jilin University, Changchun, China, in 2006, the master's degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from The University of Queensland, Brisbane, QLD, Australia, in 2012, all in computer science.

He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Lin Zuo received the bachelor's, master's, and Ph.D. degrees from the University of Electronic Science and Technology of China, Chengdu, China, in 1999, 2006, and 2011, respectively.

She is currently an Associate Professor with the School of Information and Software Engineering, University of Electronic Science and Technology of China. Her major research interests are computational intelligence, information acquisition, and processing optimization technology.



Fumin Shen received the bachelor's degree from Shandong University, Jinan, China, in 2007, and the Ph.D. degree from the Nanjing University of Science and Technology, Nanjing, China, in 2014.

He is currently with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His major research interests include computer vision and machine learning.

Heng Tao Shen (Senior Member, IEEE) received the B.Sc. (Hons.) and Ph.D. degrees in computer science from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively.

He is currently a Professor, the Dean of the School of Computer Science and Engineering, the Executive Dean of the AI Research Institute, and the Director of Center for Future Media, University of Electronic Science and Technology of China, Chengdu, China. He has published more than 250 peer-reviewed articles.

Dr. Shen received seven best paper awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and the Best Paper Award-Honourable Mention from the ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval (ACM SIGIR) 2017. He has served as the General Co-Chair for the ACM Multimedia 2021 and the TPC Co-Chair for the ACM Multimedia 2015. He is also an Associate Editor of the *ACM Transactions on Data Science*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.