# Context-Aware Attention Network for Image-Text Retrieval

Qi Zhang[1,2] Zhen Lei[1,2*] Zhaoxiang Zhang [1,2] Stan Z. Li[3]
[1] NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Center for AI Research and Innovation, Westlake University, Hangzhou, China.
{qi.zhang,zhaoxiang.zhang}@ia.ac.cn, {zlei,szli}@nlpr.ia.ac.cn

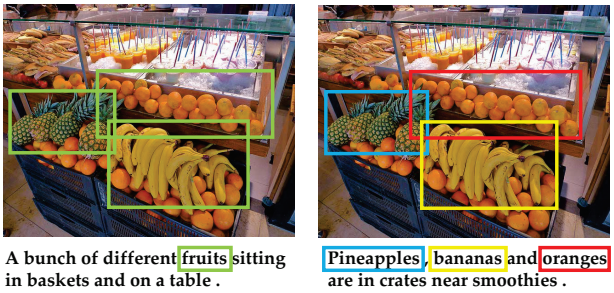
## Abstract

*As a typical cross-modal problem, image-text bi-directional retrieval relies heavily on the joint embedding learning and similarity measure for each image-text pair. It remains challenging because prior works seldom explore semantic correspondences between modalities and semantic correlations in a single modality at the same time. In this work, we propose a unified Context-Aware Attention Network (CAAN), which selectively focuses on critical local fragments (regions and words) by aggregating the global context. Specifically, it simultaneously utilizes global inter-modal alignments and intra-modal correlations to discover latent semantic relations. Considering the interactions between images and sentences in the retrieval process, intra-modal correlations are derived from the second-order attention of region-word alignments instead of intuitively comparing the distance between original features. Our method achieves fairly competitive results on two generic image-text retrieval datasets Flickr30K and MS-COCO.*

## 1. Introduction

Associating vision with language and exploring the relations between them have attracted great interest in the past decades. Many tasks have efficiently combined these two modalities and made significant progress, such as visual question answering (VQA) [1, 2, 33, 25], image captioning [1, 9], and person search with natural language [22, 23]. Image-text bidirectional retrieval [40, 44] is one of the most popular branches in the field of cross-modal research. It aims to retrieve images given descriptions or find sentences from image queries. Due to the large discrepancy between these two modalities, the main challenge is how to learn joint embeddings and accurately measure the image-text similarity.

While describing a target image, people tend to make frequent references to salient objects and depict their attributes and actions. Based on the observation, some approaches [15, 16, 33] map regions in images and words in sentences into a latent space and explore alignments between them. Although validating the efficacy of exploring region-word correspondences, they ignore the different importance of each local fragment. Recently, attention-based methods [19, 20, 26, 41] have taken steps toward attending differently to the specific regions and words and shown very promising results in the image-text retrieval task. SCAN [19] is a typical one to decide the importance of fragments based on fragments from another modality, aiming to discover full region-word alignments. Nevertheless, it ignores semantic correlations (common or exclusive attributes, categories, scenes *etc*.) between fragments in a single modality. Furthermore, some works [20, 26] have been proposed to either learn visual relation features with pre-trained neural scene graph generators or eliminate irrelevant fragments based on intra-modal relations, which alleviate the problems mentioned above to some extent.




(a) (b)

Figure 1. Illustration of the adaptive retrieval process with different contexts. An image is annotated with two different sentences. The regions highlighted with green in (a) correspond to "fruits" in the left sentence. However, they correspond to "pineapples", "bananas" and "oranges" in the right sentence, highlighted with blue, yellow and red in (b), respectively.

*Corresponding author.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00359

However, most previous attention-based methods [19, 20, 26, 41] ignore the fact that a word or region might have different semantics in different global contexts. Specifically, the global context refers to both interaction and alignments between two modalities (inter-modal context) and semantic summaries and correlations in a single modality (intra-modal context). As shown in Figure 1, people sometimes automatically summarize high-level semantic concepts (such as fruits) based on the relationships between objects in Figure 1(a), and sometimes describe each object separately (such as pineapple, banana, orange) in Figure 1(b). Therefore, it is beneficial to take into account intra-modal and inter-modal contexts simultaneously and perform image-text bidirectional retrieval with adaptation to various contexts.

To address the problems above, we first propose a unified Context-Aware Attention Network (CAAN) to selectively attend to local fragments based on the global context. It formulates the image-text retrieval as an attention process, which integrates both the inter-modal attention to discover all possible alignments between word-region pairs and intra-modal attention to learn semantic correlations of fragments in a single modality. By exploiting the context-aware attention, our model can simultaneously perform image-assisted textual attention and text-assisted visual attention. As a result, the attention scores assigned for fragments aggregate the context information.

Instead of intuitively using feature-based similarities, we further propose Semantics-based Attention (SA) to explore latent intra-modal correlations. Our semantics-based attention is formulated as the second-order attention of region-word alignments, which explicitly considers interactions between modalities and effectively utilizes region-word relations to infer the semantic correlations in a single modality. It is aware of the current input pair, and the comprehensive context from the image-text pair can directly influence the computation of each other's responses in the retrieval process. Therefore, it achieves the actual adaptive matching according to the given context.

In summary, the main contributions of our work are listed as follow:

- We propose a unified Context-Aware Attention Network to adaptively select informative fragments based on the given context from a global perspective, including semantic correlations in a single modality and possible alignments between region and words.
- We propose the Semantics-based Attention to capture latent intra-modal correlations. It is the interpretable second-order attention of region-word alignments.
- We evaluate our proposed model on two benchmark datasets Flickr30K [46] and MS-COCO [24] and it achieves fairly competitive results.

## 2. Related Work

Most existing methods for image-text retrieval either embed whole images and full sentences into a shared space or consider latent correspondences between local fragments. Some recent approaches further adopt the attention mechanism to focus on the most important local fragments.

### 2.1. Image-Text Retrieval

**Global embeddings based methods.** A common solution is to learn joint embeddings for images and sentences. DeViSE [10] made the first attempt to unify image features and skip-gram word features by a linear mapping. Wang *et al*. [39] combined the bi-directional ranking constraints with neighborhood structure preservation constraints in a single modality. Li *et al*. [22] used identity-level annotations and a two-stage framework to learn better feature representations. More recent works focus on the design of objective functions. Zheng *et al*. [47] learned the dual-path convolutional image-text embeddings with the proposed instance loss.

Although these methods have achieved a certain degree of success, image-text retrieval remains challenging due to a lack of detailed understanding of the fine-grained interplay between images and sentences.

**Local fragments based methods.** Different from the methods above, many efforts have been devoted to addressing the problem of image-text retrieval on top of local fragments. DVSA [16] first adopted R-CNN to detect salient objects and inferred latent alignments between words in sentences and regions in images. Ma *et al*. [30] proposed to learn relations between images and fragments composed from words at different levels. sm-LSTM [13] attempted to jointly predict instance-aware saliency maps for both images and sentences and use their similarities within several timesteps. HM-LSTM [33] exploited hierarchical relations between sentences and phrases, and between whole images and image regions, to jointly establish their representations. Huang *et al*. [14] proposed a semantic-enhanced image and sentence matching model, which learns semantic concepts and organizes them in a correct semantic order.

In this paper, we adopt the same local fragments based strategy to consider the contents of images and text at a finer level instead of using a rough overview.

### 2.2. Attention Mechanism

Attention mechanism recently has gained popularity and been applied to various applications, including image classification [31, 38], image captioning [29, 43] and question answering [36, 42, 45]. Benefiting from its great power, many attention-based methods have been proposed in the image-text retrieval task. DAN [32] introduced Dual Attention Networks to attend to specific regions in images and words in text through multiple steps. SCAN [19] used

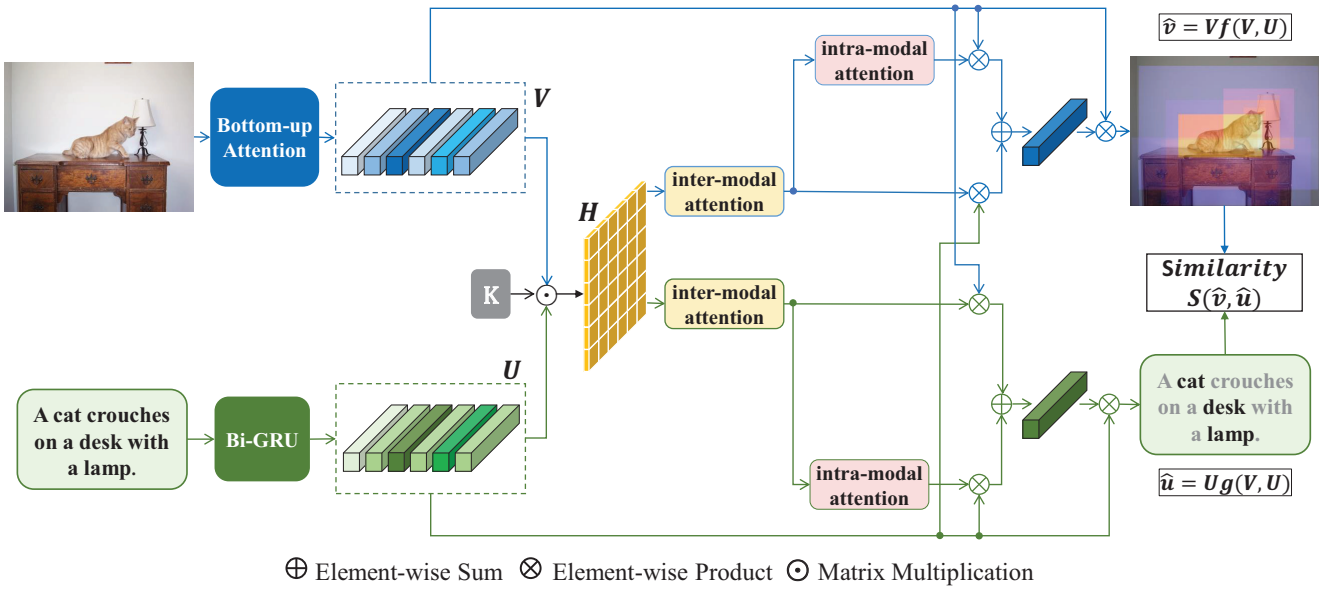



Figure 2. The pipeline of our proposed context-aware attention network (CAAN). It consists of three modules, (a) extracting and encoding regions in images and words in sentences, (b) context-aware attention with adaptation to the dynamic global context and (c) joint optimization of the final representations with the bi-directional ranking loss.

Stacked Cross Attention to perform either image-to-text attention or text-to-image attention at a time. CAMP [41] proposed Cross-Modal Adaptive Message Passing to attend to fragments. Considering visual relations between regions, recent approach [20] adopted cross-modal attention and learned visual relation features with pretrained neural scene graph generators.

In addition to methods above, there some recent methods extend the popular BERT [5] architecture to jointly learn visual and textual representations. These methods [21, 4, 28] either use a single-stream model to fuse textual and visual data as input, or take a two-stream model to process each modality separately and then fuse them. Benefiting from the self-attention module of BERT, they have achieved the state-of-the art performance.

## 3. Method

In this section, we will present an overview of our proposed Context-Aware Attention Network (CAAN). As shown in Figure 2, given an image-text pair, we first embed regions in images and words in sentences into a shared space. Concretely, the bottom-up attention [1] is utilized to generate image regions and their representations. Meanwhile, we encode words in sentences along with the sentence context. In the association module, we perform our context-aware attention network on the extracted features of local fragments, which captures semantic alignments between region-word pairs and semantic correlations between fragments in a single modality. Finally, the model is trained with image-text matching loss.

Next, we will introduce details of our proposed method from the following aspects: 1) visual representations, 2) textual representations, 3) context-aware attention network for global context aggregation, 4) objective function to optimize image-text retrieval.

### 3.1. Visual Representations

Given an image, we observe that people tend to make frequent references to salient objects and describe their actions and attributes, *etc*. Instead of extracting the global CNN feature from a pixel-level image, we focus on local regions and take advantage of bottom-up attention [1]. Following [1, 19, 20], we detect objects and other salient regions in each image utilizing a Faster R-CNN [34] model in conjunction with ResNet-101 [12] in two stages, which is pretrained on Visual Genome [18]. In the first stage, the model uses greedy non-maximum suppression with an IoU threshold to select the top-ranked box proposals. In the second stage, the extracted features of those bounding boxes are obtained after the mean-pooled convolutional layer. The features are used to predict both instance and attribute classes, in addition to refining bounding boxes. For each region $i$, $x_i$ denotes the original mean-pooling convolutional feature with 2048 dimensions. The final feature $v_i$ is transformed by a linear mapping of $x_i$ into a D-dimensional vector as follows:

$$v_i = W_x x_i + b_i. \tag{1}$$

Therefore, the target image $v$ can be presented as a set of features of selected ROIs with the highest class detection confidence scores.

### 3.2. Textual Representations

In order to discover region-word correspondences, words in sentences are mapped into the same D-dimensional space as image regions. Instead of processing each word individually, we consider to encode the word and its context at a time. Given one-hot encodings $W = \{w_1, ..., w_m\}$ of $m$ input words in a sentence, we first embed them into 300-dimensional vectors by the word embedding layer as $x_i = W_e w_i$, where $W_e$ is a parametric matrix learned end-to-end. We then feed vectors into a bi-directional GRU [3, 35], which is written as:

$$\overrightarrow{h_i} = \overrightarrow{GRU}(x_i, \overrightarrow{h_{i-1}}), i \in [1, m], \tag{2}$$

$$\overleftarrow{h_i} = \overleftarrow{GRU}(x_i, \overleftarrow{h_{i+1}}), i \in [1, m], \tag{3}$$

where $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote hidden states from the forward and backward directions, respectively. The final word embedding $u_i$ is the mean of bi-directional hidden states, which collects the context centered in the word $w_i$:

$$u_i = \frac{\overrightarrow{h_i} + \overleftarrow{h_i}}{2}, i \in [1, m]. \tag{4}$$

### 3.3. Context-aware Attention

#### 3.3.1 Formulation

The attention mechanism aims to focus on the most pertinent information of the corresponding task rather than using all available information equally. We first provide a general formulation of attention mechanism designed for the cross-modal retrieval problem. For image $v$ and text $u$, their feature maps are formulated as $V = [v_1, ..., v_n]$ and $U = [u_1, ..., u_m]$, respectively. We define the attention process for image-text retrieval as:

$$\hat{v} = Vf(V, U) = \sum_{i=1}^{n} [f(V, U)]_i v_i, \tag{5}$$

$$\hat{u} = Ug(V, U) = \sum_{j=1}^{m} [g(V, U)]_j u_j, \tag{6}$$

where $f(\cdot)$ and $g(\cdot)$ are attention functions to calculate scores for each local fragment $v_i$ and $u_j$, respectively. The final image and text features $\hat{v}$ and $\hat{u}$ are computed as the weighted sum of local fragments. Following [**?**, 29], we calculate similarities between region-word pairs for the target image and text. The similarity matrix $H$ is written as:

$$H = \tanh(V^T K U), \tag{7}$$

where $K \in \mathbb{R}^{d \times d}$ is a weight matrix. Attentive Pooling Networks [6] performs column-wise and row-wise max-pooling based on the assumption that the importance of each fragment is represented as its maximal similarity over fragments of another modality. It is an alternative version of the proposed attention process when $f(V, U)$ becomes the softmax computation after applying row-wise max-pooling operation on $H$. Furthermore, we not only calculate the similarity matrix but use it as a feature to predict the attention map. To be more specific, the importance score of a fragment is decided by all the relevant fragments, taking into account intra-modal correlations in a single modality and inter-modal alignments between all region-word pairs. Based on the consideration, the normalized attention function $f(V, U)$ for regions can be formulated as follows:

$$\widetilde{f}(V, U) = \tanh(H^v V^T Q_1 + H^{uv} U^T Q_2), \tag{8}$$

$$f(V, U) = \text{softmax}(W^v \widetilde{f}(V, U)), \tag{9}$$

where $W^v \in \mathbb{R}^z$ is a projection vector. $Q_1, Q_2 \in \mathbb{R}^{d \times z}$ are parametric matrices to do dimension-wise fusion. $H^v \in \mathbb{R}^{n \times n}$ is the attention matrix capturing intra-modal correlations for regions. $H^{uv} \in \mathbb{R}^{n \times m}$ is the attention matrix for word-to-region re-weighting. Likewise, the normalized attention function $g(V, U)$ for words is written as follows:

$$\widetilde{g}(V, U) = \tanh(H^u U^T Q_3 + H^{vu} V^T Q_4), \tag{10}$$

$$g(V, U) = \text{softmax}(W^u \widetilde{g}(V, U)), \tag{11}$$

where $Q_3, Q_4 \in \mathbb{R}^{d \times z}$ and $W^u \in \mathbb{R}^z$ are learned weights.

The designed attention functions $f(V, U)$ and $g(V, U)$ selectively attend to those informative fragments according to the global context, applying both inter-modal attention and intra-modal attention.

#### 3.3.2 Inter-modal Attention: $H^{uv}, H^{vu}$

The matrix $H$ calculates similarities of local region-word pairs. Following [15, 19, 20], we threshold the similarities to zero and normalize them to obtain alignment scores. The word-to-region attention $H^{uv}$ is computed as:

$$H_{ij}^{uv} = \frac{[H_{ij}]_+}{\sqrt{\sum_{k=1}^{n} [H_{kj}]_+^2}}, \tag{12}$$

where $[x]_+ \equiv max(0, x)$. Each element $H_{i,j}^{uv}$ in the word-to-region attention matrix $H^{uv}$ represents the relative pairwise correspondences of two local fragments region $v_i$ and word $u_j$. Similarity, the region-to-word attention $H^{vu}$ is computed as:

$$H_{ij}^{vu} = \frac{[H_{ij}]_+}{\sqrt{\sum_{k=1}^{m} [H_{ik}]_+^2}}, \tag{13}$$

Both $H^{uv}$ and $H^{vu}$ infer fine-grained interplay between images and sentences by aligning regions and words.

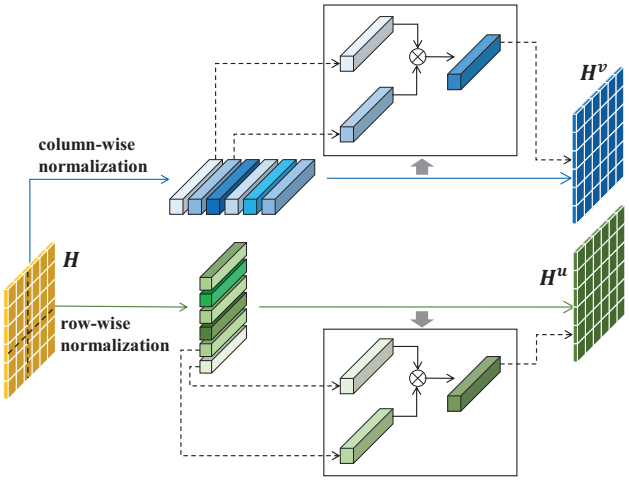



Figure 3. Detailed illustration of the semantics-based intra-modal attention process. The intra-modal affinity matrices $H^v$ and $H^u$ are designed to capture latent region-to-region and word-to-word relations, respectively. They are calculated by fully utilizing the inter-modal alignments.

#### 3.3.3 Intra-modal Attention: $H^v, H^u$

Next, we will discuss two versions of $H^v$ and $H^u$, which model intra-modal correlations from two different perspectives.

**Feature-based attention (FA).** A natural choice of measuring intra-modal correlations is to calculate feature similarities. That is, the intra-modal attention matrices $H^v$ and $H^u$ can be defined as:

$$H^v = V^T M_1 V, \tag{14}$$

$$H^u = U^T M_2 U, \tag{15}$$

where $M_1, M_2 \in \mathbb{R}^{d\times d}$ are learned weight parameters. When they are equal to identity matrices, elements in $H^v$ and $H^u$ denote dot-product similarities between local fragments in a single modality. The matrix product of a learned matrix and its transpose is another alternative version, which projects $U$ into a new space. It not only allows the calculated intra-modal attention matrices to represent the cosine similarities between normalized features, but also preserves the model capacity.

However, it ignores that the semantic summary (intra-modal context) in one modality varies for different queries. Therefore, the semantic correlation mining between fragments in a single modality should be conducted in an interactive way.

**Semantics-based attention (SA).** Considering the interactions and message passing across two modalities in the retrieval process, we propose the semantics-based attention to explore intra-modal correlations based on region-word relations. In our work, we use the interpretable second-order attention of inter-modal alignments. The detailed procedure of SA is illustrated in Figure 3. The intra-modal attention matrices $H^v$ and $H^u$ are defined as:

$$H^v = \begin{bmatrix} \mathrm{norm}(H^{uv}_{1\cdot}) \\ \mathrm{norm}(H^{uv}_{2\cdot}) \\ \vdots \\ \mathrm{norm}(H^{uv}_{n\cdot}) \end{bmatrix} \begin{bmatrix} \mathrm{norm}(H^{uv}_{1\cdot}) \\ \mathrm{norm}(H^{uv}_{2\cdot}) \\ \vdots \\ \mathrm{norm}(H^{uv}_{n\cdot}) \end{bmatrix}^T, \tag{16}$$

$$H^u = \begin{bmatrix} \mathrm{norm}(H^{vu}_{1\cdot}) \\ \mathrm{norm}(H^{vu}_{2\cdot}) \\ \vdots \\ \mathrm{norm}(H^{vu}_{m\cdot}) \end{bmatrix} \begin{bmatrix} \mathrm{norm}(H^{vu}_{1\cdot}) \\ \mathrm{norm}(H^{vu}_{2\cdot}) \\ \vdots \\ \mathrm{norm}(H^{vu}_{m\cdot}) \end{bmatrix}^T, \tag{17}$$

where $\mathrm{norm}(\cdot)$ means the $l_2$-normalized operation on the input vector. As the $i$-th row of the inter-modal attention matrix $H^{uv}$, $H^{uv}_{i\cdot}$ is considered to be the word-to-region affinity distribution or response vector for all words with respect to the given $v_i$. It measures the distance between $v_i$ and the entire word features set $\{u_1, ..., u_m\}$. Therefore, each element $H^v_{ij}$ is the cosine similarity of two region-word response vectors $H^{uv}_{i\cdot}$ and $H^{uv}_{j\cdot}$. The intra-modal attention matrix $H^v$ calculates pairwise relations of any two affinity distributions.

The intra-modal summaries and correlations are related to the global context in the retrieval process, and they implicitly contain both statistics and semantic information, $i.e.$ co-existence, dependencies and affiliation. When two regions $v_i$ and $v_j$ have similar responses to the same sentence, they are viewed as a high-correlated pair. Accordingly, SA focuses more on region $v_i$ in the process of assigning attention scores with respect to region $v_j$. It comprehensively takes into account the similarity of two responses, which models the relationship between the movement of similarities of fragments between two modalities.

To summarize, the adaptive intra-modal attention process is driven by the global semantic information. It requires discrimination on semantics based on the given context rather than original context-free features.

### 3.4. Objective Function

The hinge-based bi-directional ranking loss [8, 16, 19] is the most popular objective function for image-text retrieval, which can be formulated as follows:

$$\begin{aligned} L(\hat{v}, \hat{u}) = \sum_{\hat{v}^-, \hat{u}^-} \{ & max[0, m - S(\hat{v}, \hat{u}) + S(\hat{v}, \hat{u}^-)] \\ & + max[0, m - S(\hat{v}, \hat{u}) + S(\hat{v}^-, \hat{u})]\}, \end{aligned} \tag{18}$$

where m is a margin constraint, $(\hat{v}, \hat{u}^-)$ and $(\hat{v}^-, \hat{u})$ are negative pairs. $S(\cdot)$ is a matching function, which is defined as the inner product in our experiments. The objective function attempts to pull positive image-text pairs close

| Methods | MS-COCO 5-fold 1K Test Images | | | | | | Flickr30K 1K Test Images | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 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 |
| (R-CNN, AlexNet) | | | | | | | | | | | | |
| DVSA [16] | 38.4 | 69.9 | 80.5 | 27.4 | 60.2 | 74.8 | 22.2 | 48.2 | 61.4 | 15.2 | 37.7 | 50.5 |
| (VGG) | | | | | | | | | | | | |
| VQA-A [25] | 50.5 | 80.1 | 89.7 | 37.0 | 70.9 | 82.9 | 33.9 | 62.5 | 74.5 | 24.9 | 52.6 | 64.8 |
| sm-LSTM [13] | 53.2 | 83.1 | 91.5 | 40.7 | 75.8 | 87.4 | 42.5 | 71.9 | 81.5 | 30.2 | 60.4 | 72.3 |
| 2WayNet [7] | 55.8 | 75.2 | - | 39.7 | 63.3 | - | 49.8 | 67.5 | - | 36.0 | 55.6 | - |
| (ResNet) | | | | | | | | | | | | |
| RRF-Net [27] | 56.4 | 85.3 | 91.5 | 43.9 | 78.1 | 88.6 | 47.6 | 77.4 | 87.1 | 35.4 | 68.3 | 79.9 |
| VSE++ [8] | 64.6 | 90.0 | 95.7 | 52.0 | 84.3 | 92.0 | 52.9 | 80.5 | 87.2 | 39.6 | 70.1 | 79.5 |
| DAN [32] | - | - | - | - | - | - | 55.0 | 81.8 | 89.0 | 39.4 | 69.2 | 79.1 |
| DPC [47] | 65.6 | 89.8 | 95.5 | 47.1 | 79.9 | 90.0 | 55.6 | 81.9 | 89.5 | 39.1 | 69.2 | 80.9 |
| GXN [11] | 68.5 | - | 97.9 | 56.6 | - | 94.5 | 56.8 | - | 89.6 | 41.5 | - | 80. |
| SCO [14] | 69.9 | 92.9 | 97.5 | 56.7 | 87.5 | 94.8 | 55.5 | 82.0 | 89.3 | 41.1 | 70.5 | 81.1 |
| (Faster-RCNN, ResNet) | | | | | | | | | | | | |
| SCAN-single [19] | 70.9 | 94.5 | 97.8 | 56.4 | 87.0 | 94.8 | 67.9 | 89.0 | 94.4 | 43.9 | 74.2 | 82.8 |
| R-SCAN [20] | 70.3 | 94.5 | 98.1 | 57.6 | 87.3 | 93.7 | 66.3 | 90.6 | 96.0 | 51.4 | 77.8 | 84.9 |
| CAMP [41] | 72.3 | 94.8 | 98.3 | 58.5 | 87.9 | 95.0 | 68.1 | 89.7 | 95.2 | 51.5 | 77.1 | 85.3 |
| BFAN-single [26] | 73.7 | 94.9 | - | 58.3 | 87.5 | - | 64.5 | 89.7 | - | 48.8 | 77.3 | - |
| CAAN (ours) | **75.5** | **95.4** | **98.5** | **61.3** | **89.7** | **95.2** | **70.1** | **91.6** | **97.2** | **52.8** | **79.0** | **87.9** |

Table 1. Results of the cross-modal retrieval on MS-COCO 5-fold 1K test set and Flickr30K 1K test setre. The best performance is denoted with bold text. '-': the result is not provided.

and push negative ones away. Despite widely used in the cross-modal task, it suffers from high redundancy and slow convergence caused by the random triplet sampling process. Rather than summing over all the negative pairs in a mini-batch, bi-directional ranking loss with the hardest negatives is often adopted for computational efficiency. It focuses on the hardest samples which are the negative ones closest to positive pairs. Given a positive pair $(\hat{v}, \hat{u})$, the hardest negatives are formulated as $v_h = \arg\max_{p \neq \hat{v}} S(p, \hat{u})$ and $u_h = \arg\max_{k \neq \hat{u}} S(\hat{v}, k)$. Therefore, the bi-directional ranking loss with the hardest negatives is written as:

$$\begin{aligned} L_{hard}(\hat{v}, \hat{u}) = max[0, m - S(\hat{v}, \hat{u}) + S(\hat{v}, \hat{u}_h^-)] \\ + max[0, m - S(\hat{v}, \hat{u}) + S(\hat{v}_h, \hat{u})]. \end{aligned} \tag{19}$$

# 4. Experiments

## 4.1. Datasets and Metrics

**Datasets.** We evaluate our model on the Flickr30K [46] and MS-COCO [24] datasets. Flickr30K contains 31,000 images and each image is associated with five sentences. We adopt the same protocol in [8, 16] to split the dataset into 1,000 test images, 1,000 validation images, and 29,000 training images. MS-COCO contains 123,287 images and each is annotated with five descriptions. In [16], MS-COCO is split into 82,783 training images, 5000 validation images and 5,000 test images. We follow [8, 19] to use other 30,504 images as part of the training set, which were originally in the validation set but have been left out in the split.The experiments are conducted on both 5K and 1K test images, where the result of 1K test images is reported by averaging over 5-fold on the full 5K test images.

**Evaluation Metrics.** We use R@K and mR to evaluate our models. R@K is the percentage of correct matchings in the top-K lists. R@1, R@5 and R@10 are adopted in the experiments. mR is the mean value of R@K (K=1,5,10).

## 4.2. Implementation Details

The Adam optimizer [17] is employed for optimization. In the MS-COCO, we set the initial learning rate to 0.0005 for the first 10 epochs and then decay it by 10 times in the following 10 epochs. In the Flickr30K, the learning rate is 0.0002 in the first 15 epochs, and reduced to 0.00002 in the next 15 epochs. The best model is chosen based on the sum of recalls on the validation set.

## 4.3. Quantitative Results

### 4.3.1 Comparisons with non-BERT Methods

We compare our model with several recent state-of-the-art non-BERT methods on the MS-COCO and Flickr30K datasets. As shown in Table 1, CAAN outperforms other methods by a large margin. For fair comparisons, we only report single model results of SCAN [19] and BFAN [26]

| Methods | MS-COCO 5K Test Images | | | | | |
|---|---|---|---|---|---|---|
| | Sentence Retrieval | | | Image Retrieval | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| (R-CNN, AlexNet) | | | | | | |
| DVSA [16] | 16.5 | 39.2 | 52.0 | 10.7 | 29.6 | 42.2 |
| (VGG) | | | | | | |
| VQA-A [25] | 23.5 | 50.7 | 63.6 | 16.7 | 40.5 | 53.8 |
| (ResNet) | | | | | | |
| VSE++ [8] | 41.3 | 69.2 | 81.2 | 30.3 | 59.1 | 72.4 |
| GXN [11] | 42.0 | - | 84.7 | 31.7 | - | 74.6 |
| SCO [14] | 42.8 | 72.3 | 83.0 | 33.1 | 62.9 | 75.5 |
| (Faster-RCNN, ResNet) | | | | | | |
| PVSE [37] | 45.2 | 74.3 | 84.5 | 32.4 | 63.0 | 75.0 |
| SCAN-single [19] | 46.4 | 77.4 | 87.2 | 34.4 | 63.7 | 75.7 |
| R-SCAN [20] | 45.4 | 77.9 | 87.9 | 36.2 | 65.6 | 76.7 |
| CAMP [41] | 50.1 | 82.1 | 89.7 | 39.0 | 68.9 | 80.2 |
| CAAN (ours) | **52.5** | **83.3** | **90.9** | **41.2** | **70.3** | **82.9** |

Table 2. Comparisons of the cross-modal retrieval results on the MS-COCO full 5K test set.

on the two datasets rather than using the ensemble version. On the 1K test set, CAAN gives R@10=98.5 and 95.2 with image and text as queries, respectively. It achieves the performance with R@1=61.3 for image retrieval, which is a 3% relative improvement compared to the current state-of-the-art non-BERT methods, $i.e.$, BFAN-single [26]. On the Flickr30K dataset, CAAN achieves better R@1 at 70.1 and 52.8 with sentence and image retrieval, respectively. The results on the MS-COCO 5K test set are summarized in Table 2. CAAN significantly outperforms the current non-BERT methods on all metrics, which verifies the effectiveness of our proposed method. As illustrated in the section 3.3, our introduced attention process explores both region-word alignments and semantic correlations in a single modality. The performance gain compared with other non-BERT methods demonstrates the superior to consider the specific context in the adaptive retrieval process.

| | Sentence Retrieval | | | Image Retrieval | | |
|---|---|---|---|---|---|---|
| | R@1 | R@10 | mR | R@1 | R@10 | mR |
| ViLBERT†[28] | - | - | - | 45.5 | 85.0 | 69.1 |
| UNITER†[4] | - | - | 83.3 | - | - | 73.9 |
| ViLBERT‡[28] | - | - | - | 58.2 | 91.5 | 78.2 |
| Unicoder-VL†[21] | 73.0 | 94.1 | 85.4 | 57.8 | 88.9 | 76.3 |
| CAAN (ours) | 70.1 | 97.2 | 86.3 | 52.8 | 87.9 | 73.2 |
| UNITER‡[4] | - | - | 92.2 | - | - | 83.1 |
| Unicoder-VL‡[21] | **86.2** | **99.0** | **93.8** | **71.5** | **94.9** | **85.8** |

Table 3. Comparisons with BERT-based methods on the Flickr30k dataset. CAAN (ours) is the baseline model, which uses Faster R-CNN pre-trained on Visual Genome, without pre-training the language model. † indicates methods using both pre-trained visual features and language model (BERT) initialization with text-only data. ‡ indicates methods pre-trained with extra out-of-domain (Vision-Language) data.

#### 4.3.2 Comparisons with BERT-based Methods

We additionally make comparisons with other BERT-based methods, which achieve the state-of-art performance on the Flickr30K and MS-COCO datasets. As shown in Table 3, our method has fairly comparable results compared with the BERT-based methods, even without introducing and fine-tuning on a pre-trained language model.

Besides, our method is much faster and smaller, compared to BERT-based ones. Taking ViLBERT as an example, computing similarity between a text-image pair takes around 0.5 s, while ours is around 45 $\mu$s, using 1 GTX1080Ti. ViLBERT has parameters of 275 M, while ours is only 11 M. Considering the speed and model size requirements of the real-world scenes, our method is more convenient and practical for deployment and application.

| | Image Query | | Sentence Query | |
|---|---|---|---|---|
| | R@1 | R@10 | R@1 | R@10 |
| baseline | 58.1 | 90.0 | 42.0 | 79.7 |
| baseline+IA | 60.6 | 92.4 | 45.2 | 81.5 |
| baseline+FA | 62.3 | 93.2 | 46.6 | 83.0 |
| baseline+SA | 64.5 | 93.8 | 48.8 | 83.4 |
| baseline+IA+FA | 62.6 | 93.0 | 45.0 | 82.9 |
| CAAN | **70.1** | **97.2** | **52.8** | **87.9** |

Table 4. Results of ablation studies on the Flickr30K test set.

### 4.4. Ablation Studies on Attention Mechanism

In this section, we perform ablation studies to quantify the effect of our proposed attention mechanism, including intra-modal and inter-modal attention. We first provide the baseline model with bottom-up attention [1], denoted as "baseline" in Table 4. It takes the average of all local features as final representations. We can see that it achieves a fairly competitive result compared to the methods extracting global features shown in Table 1. It shows the reasonability to focus on local fragments rather than using a rough overview of a whole image or a full sentence.

**Baseline with Inter-modal Attention.** We implement inter-modal attention in the baseline model, denoted as "baseline+IA" in Table 4. It achieves R@1=60.6 and 45.2 with image and text as queries, respectively. Compared with "baseline", CAAN demonstrates its effectiveness of considering full alignments between region-word pairs.

**Baseline with Intra-modal Attention.** Table 4 illustrates the impact of preforming intra-modal attention. Both "baseline+FA" and "baseline+SA" use only relations of fragments in a single modality. The difference between them is the way to measure fragment affinities. Although "baseline+FA" introduces additional parameters $M_1$ and $M_2$ to fit data, "baseline+SA" still achieves better results, which shows the superior of inferring semantic correlations by


(a) Inter-modal Alignments (b) Intra-modal Correlations

Figure 4. Visualization of the attention weights of each image region with respect to sentence query on MS-COCO and Flickr30K datasets. The left sub-figure (a) shows the qualitative examples of text-to-image retrieval with different sentences. The right sub-figure (b) compares "baseline+IA+FA" and our CAAN, which shows that the similar semantics shared by different objects affect the attention process. It is beneficial to consider both inter-modal alignments and intra-modal correlations in an interactive way. (Best viewed in color)

adaptively measuring the distance of response vectors instead of original features.

**Baseline with both Inter-modal and Intra-modal Attention.** We further integrate inter-modal and intra-modal attention into the baseline modal. Results are denoted as "baseline+IA+FA" and "CAAN" shown in Table 4. "baseline+IA+FA" even has a slightly worse result compared to "baseline+FA". It shows that without careful designs, combing inter-modal alignments and intra-modal correlations might hurt the performance. While "CAAN" outperforms "baseline+IA+FA" and "baseline+SA", indicating that it is a better solution to consider the global context and conduct semantic correlation mining in an interactive way.

## 5. Visualization

To better understand the effectiveness of our proposed model, we visualize the attention assignment of the text-to-image retrieval process in Figure 4. For the qualitative examples in Figure 4(a), we can observe that attention weights are assigned to different regions for different image-text pairs. As shown in the first row of Figure 4(a), the region "bottle" receives more attention in the left sub-figure while the region "bags" is the focus in the right sub-figure. It indicates that our model infers inter-modal alignments based on the global context. For the qualitative examples in Figure 4(b), we provide comparisons with "baseline+FA+IA". As shown in the second row of Figure 4(b), the region "boy" is assigned more attention weight with the proposed CAAN compared with the model "baseline+IA+FA". It is notable that different objects with similar semantics affect the matching process.

## 6. Conclusion

In this paper, we propose a unified Context-Aware Attention Network (CAAN) to formulate the image-text retrieval as an attention process to selectively focus on the most informative local fragments. By incorporating intra-modal and inter-modal attention, our model aggregates the context information of alignments between word-region pairs (inter-modal context) and semantic correlations between fragments in a single modality (intra-modal context). Furthermore, we perform the semantic-based attention to model intra-modal correlations, which is the interpretable second-order attention of region-word alignments. The model demonstrates its effectiveness by achieving fairly competitive results on the Flickr30K and MS-COCO datasets.

## 7. Acknowledgements

This work has been partially supported by the Chinese National Natural Science Foundation Projects #61872367, #61876178, #61806196, #61806203, #61976229.

## 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*, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[6] Cícero Nogueira dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*, 2016.

[7] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *CVPR*, 2017.

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2017.

[9] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

[10] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

[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*, 2018.

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

[13] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal LSTM. In *CVPR*, 2017.

[14] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, 2018.

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

[16] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018.

[20] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*, 2019.

[21] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019.

[22] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.

[23] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017.

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

[25] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *ECCV*, 2016.

[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 *ACMMM*, 2019.

[27] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, 2017.

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[29] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.

[30] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *ICCV*, 2015.

[31] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NeurIPS*, 2014.

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 2017.

[33] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal LSTM for dense visual-semantic embedding. In *ICCV*, 2017.

[34] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[35] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 1997.

[36] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[37] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, 2019.

[38] Marijn F. Stollenga, Jonathan Masci, Faustino J. Gomez, and Jürgen Schmidhuber. Deep networks with internal selective attention through feedback connections. In *NeurIPS*, 2014.

[39] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.

[40] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *arXiv preprint arXiv:1704.03470*, 2017.

[41] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. CAMP: cross-modal adaptive message passing for text-image retrieval. *arXiv preprint arXiv:1909.05506*, 2019.

[42] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[44] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *CVPR*, 2015.

[45] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014.

[47] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.0553*, 2017.