Show Your Faith: Cross-Modal Confidence-Aware Network for Image-Text Matching

Huatian Zhang[1], **Zhendong Mao**[1*], **Kun Zhang**[1], **Yongdong Zhang**[1, 2]

[1]University of Science and Technology of China, Hefei, China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
{huatianzhang, kkzhang}@mail.ustc.edu.cn, {zdmao, zhyd73}@ustc.edu.cn

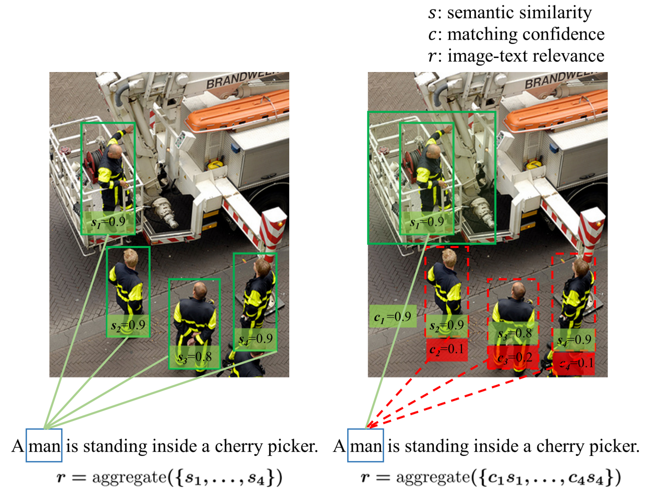
Abstract

Image-text matching bridges vision and language, which is a crucial task in the field of multi-modal intelligence. The key challenge lies in how to measure image-text relevance accurately as matching evidence. Most existing works aggregate the local semantic similarities of matched region-word pairs as the overall relevance, and they typically assume that the matched pairs are equally reliable. However, although a region-word pair is locally matched across modalities, it may be *inconsistent/unreliable* from the global perspective of image-text, resulting in inaccurate relevance measurement. In this paper, we propose a novel Cross-Modal Confidence-Aware Network to infer the matching confidence that indicates the reliability of matched region-word pairs, which is combined with the local semantic similarities to refine the relevance measurement. Specifically, we first calculate the matching confidence via the relevance between the semantic of image regions and the complete described semantic in the image, with the text as a bridge. Further, to richly express the region semantics, we extend the region to its visual context in the image. Then, local semantic similarities are weighted with the inferred confidence to filter out unreliable matched pairs in aggregating. Comprehensive experiments show that our method achieves state-of-the-art performance on benchmarks Flickr30K and MSCOCO.

Introduction

Image-text matching, which refers to image searching given descriptions or text retrieval given image queries, is beneficial to many multi-modal tasks (Anderson et al. 2018)(Xu et al. 2018)(Xu et al. 2019)(Yu et al. 2020) such as image captioning, text-to-image synthesis, visual question answering. The matching aims to bridge vision and language so as to reduce the visual-semantic discrepancy between these two heterogeneous modalities. Despite the remarkable progress in recent years, image-text matching remains the challenge that how to measure image-text relevance accurately as matching evidence.

To explore efficacious approaches to capture cross-modal semantic interplays for image-text relevance measuring, plenty of researches have been done. The common paradigm is to first align vision and language semantically, and then measure cross-modal semantic similarity as relevance based on resulting alignments. There are two main strategies: global aligning based and local aligning based. Global aligning based methods (Wang, Li, and Lazebnik 2016)(Liu et al. 2017) (Gu et al. 2018) (Huang et al. 2018)(Shi et al. 2019)(Li et al. 2019) infer cross-modal semantic similarity directly from the global alignment between the whole image and full text in a common embedding space. Local aligning based methods aggregate the overall relevance from local semantic alignments between detected salient image regions and text words. Recent works mainly probe into local aligning to discover fine-grained visual-semantic similarity at region-word level. (Lee et al. 2018) proposes a stacked


s: semantic similarity
c: matching confidence
r: image-text relevance
BRANDWE
BRANDWE
s_1=0.9
s_2=0.9
s_3=0.8
s_4=0.9
s_1=0.9
c_1=0.9
s_2=0.9
c_2=0.1
s_3=0.8
c_3=0.2
s_4=0.9
c_4=0.1
A man is standing inside a cherry picker.
$r = \text{aggregate}(\{s_1, \dots, s_4\})$
A man is standing inside a cherry picker.
$r = \text{aggregate}(\{c_1 s_1, \dots, c_4 s_4\})$

(a) Existing methods

(b) Our method

Figure 1: Illustration of the necessity of matching confidence. (a) Existing methods typically measure the overall image-text relevance with aggregating local semantic similarities, assuming all matched region-word pairs are reliable. The word “man” in text will align to all man regions, even s_2, s_3, s_4 are not really referred to. (b) Our method further infers the matching confidence to distinguish the reliability of each matched region-word pair from the global perspective, and filters out the unreliable matched pairs (*e.g.*, regions with red box) to achieve more accurate semantic alignment.

[*]Corresponding author.

cross attention network to capture all latent local alignments by attending to image regions and words with each other, which achieves promising performance and inspires a series of works (Wang et al. 2019b) (Hu et al. 2019) (Zhang et al. 2020b) (Chen et al. 2020) (Wu et al. 2019)(Wehrmann, Kolling, and Barros 2020) (Chen and Luo 2020a). They handle sophisticated semantic interactions reasonably between modalities, for obtaining discriminative visual-semantic representations to facilitate cross-modal aligning. (Liu et al. 2020) (Diao et al. 2021) focus on exploring local alignments aggregating mechanisms such as graph convolution and attentional reasoning to enhance meaningful alignments in overall relevance measurement. In general, most local aligning based methods match image regions and text words with associating visual-semantic locally, and aggregate semantic similarities between matched region-word pairs mechanically to measure the overall image-text relevance.

However, local semantic similarities, *i.e.*, relevance of matched region-word pairs, are aggregated by default confidence 1 equally in most existing works, which is ill-considered since the matching confidence, *i.e.*, reliability of matched region-word pairs, depending on the global image-text semantic context, is different from each other. That is, a local region-word is matched across modalities, yet it may be inconsistent/unreliable with the global perspective of image-text. Thus, in order to reveal the real contribution level of local semantic similarities to the overall cross-modal relevance, it is necessary to explicitly indicate the confidence of region-word pairs in matching. Without considering the confidence, the inconsistent region-word pairs will be aggregated indiscriminately and thus interfere with the overall relevance measurement. More seriously, redundant inconsistent region-word pairs may even overwhelm the matched ones, causing the effects of other relatively few matched pairs that are critical for matching are diluted. As shown in Figure 1, the word “man” locally aligns to all man regions in Figure 1(a), even the man who is not standing inside the cherry picker, which results in inaccurate semantic alignment and interferes with the relevance measurement. With taking the matching confidence into account in Figure 1(b), the interferences from the man regions irrelevant to text semantic can be filtered out.

To address the above issues, we propose a novel Cross-Modal Confidence-Aware Network (CMCAN) for image-text matching, which takes the confidence of matched region-word pairs into account and combines it with the local semantic similarities to measure cross-modal relevance accurately. CMCAN infers the matching confidence from the relevance between the semantic of image regions and the complete described semantic in the image, with the text as a bridge. Specifically, the confidence is measured by the inner product between the semantic similarity of the region-text and the semantic similarity of the whole image-text, which are connected by the full text. Moreover, to express the semantic of the image region richly, we extend the region to its visual context in the image. In detail, our method contains three modules: 1) Feature Representing: we first extract the representations of detected image regions and text words for global and local aligning. In order to fully exploit region semantic in the image, we extend each region with its surrounding scene together as its visual context based on the natural neighboring relationship; 2) Matching Confidence Inferring: the matching confidence is inferred from how much semantic similarity between visual context of regions and the full text can be contained in the overall semantic similarity of image-text, since it indicates the relative extent to which regions are described in text from the perspective of the whole image; 3) Cross-Modal Relevance Measuring: we weight each region-queried local semantic similarity with the corresponding inferred confidence, and implement self-attentional reasoning on global similarity with both weighted region-queried local similarities and word-queried local similarities separately to measure the overall image-text relevance.

Our contributions are summarized as follows:

- We propose a novel Cross-Modal Confidence-Aware Network, which is the first time to, for the best of our knowledge, infer the confidence of matched region-word pairs from a global perspective in image-text matching, to filter out the inconsistent local matched region-word pairs to enhance more accurate relevance measurement.
- We propose a delicately designed matching confidence inferring method, which uses the full text as a bridge to measure the faith of whether the regions are really described in the text, relative to the global semantic similarity of the whole image-text.
- The experimental results demonstrate that our method achieves state-of-the-art performance on public benchmarks Flickr30K and MSCOCO.

Related Work

Extensive efforts have been made to align visual-semantic between heterogeneous modalities and measure cross-modal relevance for image-text matching which is more complicated than unimodal retrieval (Cui et al. 2019)(Zhu et al. 2020). (Wang, Li, and Lazebnik 2016)(Liu et al. 2017)(Gu et al. 2018)(Shi et al. 2019)(Li et al. 2019) conform to the global aligning paradigm and mainly focus on exploring the ways of feature fusion or exploiting latent scene semantic to learn more discriminative representations.

To capture fine-grained cross-modal interplays, (Nam, Ha, and Kim 2017)(Huang, Wang, and Wang 2017) attempt to learn region-word level correspondences locally but can only attend to limited alignments because of the high coupling in alignment aggregating. Significantly, (Lee et al. 2018) proposes a stacked cross attention to mine region-word local alignments by attending to image regions and words with each other as context and aggregates the local alignments by average or LogSumExp to measure overall cross-modal relevance. Under the local aligning framework inspired by (Lee et al. 2018), (Wang et al. 2019b) (Hu et al. 2019) (Zhang et al. 2020b) (Chen et al. 2020) (Wu et al. 2019)(Wehrmann, Kolling, and Barros 2020) (Chen and Luo 2020a) aim to design reasonable cross-modal aligning mechanisms to meet visual semantic interactions in order to facilitate the relevance measuring. (Wang et al. 2020) models scene graph (Xu et al. 2020) to describe the natural scene in images. (Liu et al. 2020) introduces relative spatial position of image regions and syntactic dependency tree of text to model the semantic associations between regions and words respectively, and then aggregates local alignment between regions and words, based on graph convolutional network(Kipf and Welling 2017). (Diao et al. 2021) enhances global alignment and local alignments mutually with the help of attentional reasoning. (Chen et al. 2021) discovers that simple pooling can outperform well-designed complex methods in feature aggregating, and automatically learns the best pooling strategy. (Yan, Yu, and Xie 2021) explicitly transforms features from heterogeneous modalities into a common embedding space with attention mechanism, which optimizes attention weights towards evaluation metrics, based on policy gradient.

In summary, the most existing works aggregate fine-grained local semantic similarities or global and local semantic similarities mechanically either by read-out functions or with weights inferred from inter-alignment reasoning to measure cross-modal relevance, without taking the inherent reliability of matched region-word pairs from the global image-text perspective into account. That is, in most existing works, interference from matching relationships that are locally matched but inconsistent with the global perspective are aggregated into the overall image-text relevance without screening.

Methodology

In this section, we elaborate on the matching confidence inferring and how to introduce the inferred confidence into cross-modal relevance measurement. As illustrated in Figure 2, our CMCAN is composed of three modules. *Firstly*, the way to learn visual and textual representations and extend the semantic of detected image regions is introduced in section 3.1. *Secondly*, how to infer the matching confidence of matched region-word pairs from the global image-text perspective is proposed in section 3.2. *Finally*, our vision-language self-attentional reasoning method for measuring cross-modal relevance is presented in section 3.3, and the objective function for training is mentioned in section 3.4.

Feature Representing

Image Representation To extract image regions with expressive visual semantics, bottom-up attention has been widely employed in multi-modal tasks (Zhang et al. 2020a), which imitates human to focus on salient objects or other regions spontaneously. Following (Anderson et al. 2018) (Lee et al. 2018), we utilize Faster R-CNN (Ren et al. 2015) with ResNet-101(He et al. 2016) as backbone to implement the bottom-up attention, which is pretrained on the Visual Genomes dataset (Krishna et al. 2017). Specifically, the Faster R-CNN is utilized to detect salient regions in an image I and encode visual representation $\boldsymbol{x}_i$ for detected image region r_i. Then we transform $\boldsymbol{x}_i$ to a D-dimensional $\boldsymbol{v}_i$ via linear projection:

$$\boldsymbol{v}_i = W_v \boldsymbol{x}_i + b_i \tag{1}$$

The image I can be denoted as $\{\boldsymbol{v}_i | i = 1, 2, \cdots, N,\ \boldsymbol{v}_i \in \mathbb{R}^D\}$, where N is the number of regions in image I.

Further, the global representation $\boldsymbol{v}^{glo}$ of the whole image I is encoded by attention mechanism with the average feature $\boldsymbol{v}_{\text{ave}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i$ as query. Concretely, $\boldsymbol{v}^{glo}$ is aggregated from the detected regions as follows:

$$\boldsymbol{v}^{glo} = \frac{\sum_{i=1}^{N} w_i \boldsymbol{v}_i}{\left\| \sum_{i=1}^{N} w_i \boldsymbol{v}_i \right\|_2} \tag{2}$$

where the attention weight w_i is the normalized similarity between $\boldsymbol{v}_i$ and the query $\boldsymbol{v}_{\text{ave}}$.

Text Representation We extract text semantic information at word level in order to capture the fine-grained interplay between vision and language. We first map one-hot encodings $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M\}$ of words in text T to distributed representations by learnable word embedding layer as $\boldsymbol{t}_j = W_e \boldsymbol{w}_j$. To enhance the text representation with context semantics, we utilize a bi-directional GRU (Bahdanau, Cho, and Bengio 2015) to encode both forward and backward information as follows:

$$\overrightarrow{\boldsymbol{h}_j} = \overrightarrow{\text{GRU}}\left(\boldsymbol{t}_j, \overrightarrow{\boldsymbol{h}_{j-1}}\right), j \in [1, M] \tag{3}$$

$$\overleftarrow{\boldsymbol{h}_j} = \overleftarrow{\text{GRU}}\left(\boldsymbol{t}_j, \overleftarrow{\boldsymbol{h}_{j+1}}\right), j \in [1, M] \tag{4}$$

where $\overrightarrow{\boldsymbol{h}_j}$ and $\overleftarrow{\boldsymbol{h}_j}$ denote hidden states from the forward and backward GRU, respectively. The context enhanced word representation $\boldsymbol{u}_j$ is defined as the mean of bi-directional hidden states:

$$\boldsymbol{u}_j = \frac{\overrightarrow{\boldsymbol{h}_j} + \overleftarrow{\boldsymbol{h}_j}}{2}, j \in [1, M] \tag{5}$$

The text T can be denoted as $\{\boldsymbol{u}_j | j = 1, 2, \cdots, M,\ \boldsymbol{u}_j \in \mathbb{R}^D\}$. Similarly, the global representation $\boldsymbol{u}^{glo}$ of the full text T is represented in the same way as $\boldsymbol{v}^{glo}$ in Eq.2.

Semantic Extending In order to represent the image regions more discriminative from each other, we take a further step on extracting the visual context of each region for semantic extending. Moreover, considering that the surrounding scene of a region usually contains its related semantics, we design to extend a region with its neighboring regions as the visual context. To be specific, for a region $\boldsymbol{v}_i$, we divide its surrounding scene into four equal scopes with $\boldsymbol{v}_i$ as the center, and extract the K nearest detected regions from each scope (*i.e.*, top, bottom, left or right). Then we gather the indexes of all extracted image regions as well as the center, that is idx_i, as:

$$\text{idx}_i = \{\bigcup_{\text{scope}} \text{idx}_{\text{scope}}, i\},\ \text{scope} \in \{\text{top}, \text{bottom}, \text{left}, \text{right}\} \tag{6}$$

where $\text{idx}_{\text{scope}}$ denotes indexes of the K extracted nearest regions in one scope. The surrounding scene of region $\boldsymbol{v}_i$ is disassembled into its neighboring regions indexed by idx_i. Furthermore, we formulate the scene $\boldsymbol{v}_i^{neig}$ as:

$$\boldsymbol{v}_i^{neig} = \frac{\sum_{i \in \text{idx}_i} w_i \boldsymbol{v}_i}{\left\| \sum_{i=1}^{N} w_i \boldsymbol{v}_i \right\|_2} \tag{7}$$

where w_i in Eq.7 shares the same attention weight parameter of region $\boldsymbol{v}_i$ in Eq.2.

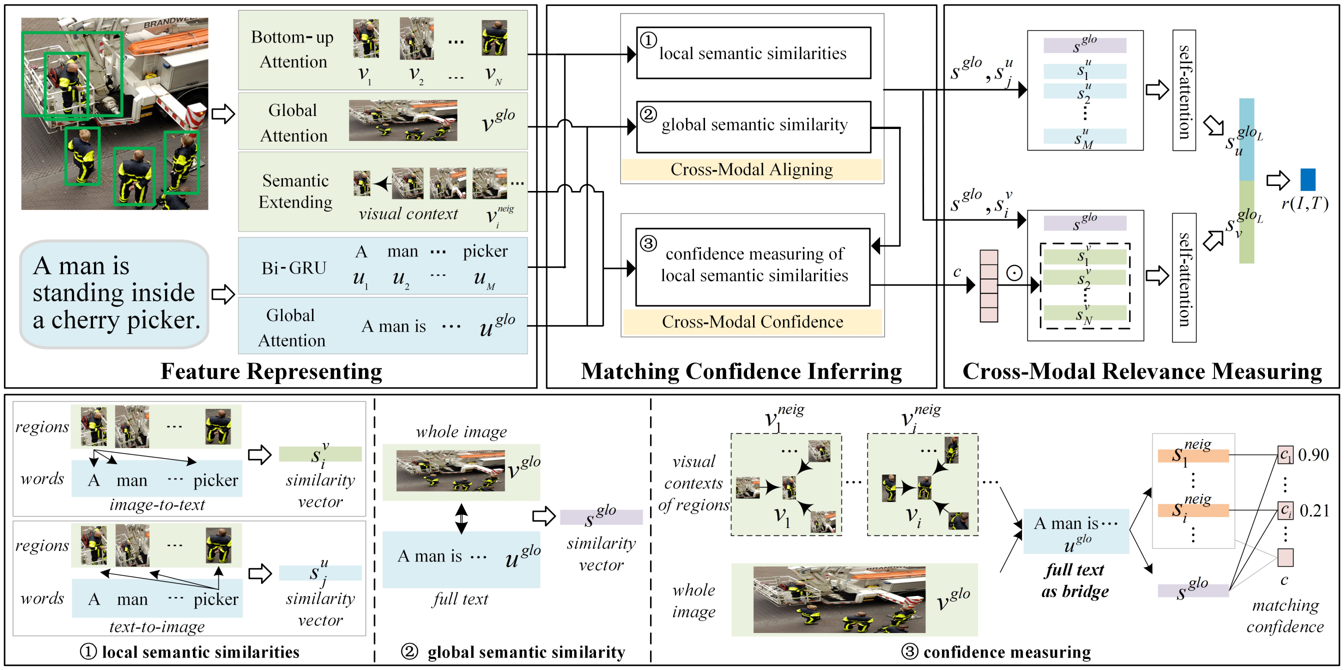


Figure 2: Illustration of our proposed CMCAN. The entire method consists of three modules: feature representing, matching confidence inferring, and cross-modal relevance measuring. The confidence is inferred from the relevance between the visual context of regions and the complete described semantic in the image, with the text as a bridge.

Matching Confidence Inferring

Cross-Modal Aligning To characterize the detailed correspondence between vision and language and align visual semantics across modalities, inspried by (Diao et al. 2021), we embody the semantic similarity between heterogeneous modalities with normalized distance-based representation.

Specfically, the local semantic similarity $\boldsymbol{s}_i^v$ between image region $\boldsymbol{v}_i$ and its semantically matched relevant words in the text is represented as:

$$\boldsymbol{s}_i^v = \frac{W_s^v|\boldsymbol{v}_i - \boldsymbol{a}_i^u|^2}{\left\| W_s^v|\boldsymbol{v}_i - \boldsymbol{a}_i^u|^2 \right\|_2} \tag{8}$$

where $W_s^v \in \mathbb{R}^{P \times D}$ is a learnable parameter matrix. The text context $\boldsymbol{a}_i^u$ is attended by region $\boldsymbol{v}_i$ with $\boldsymbol{a}_i^u = \sum_{j=1}^{M} \alpha_{ij} \boldsymbol{u}_j$ as (Lee et al. 2018), where $\alpha_{ij} = \frac{e^{(\lambda \hat{c}_{ij})}}{\sum_{i=1}^{N} e^{(\lambda \hat{c}_{ij})}}$, $\hat{c}_{ij} = [c_{ij}]_+ / \sqrt{\sum_{j=1}^{M} [c_{ij}]_+^2}$, and c_{ij} is the cosine similarity between region $\boldsymbol{v}_i$ and word $\boldsymbol{u}_j$. That is, the semantic similarity $\boldsymbol{s}_i^v$ is queried by image region $\boldsymbol{v}_i$. Similarly, the semantic similarity $\boldsymbol{s}_j^u$ between word $\boldsymbol{u}_j$ and its matched visual context $\boldsymbol{a}_j^v$ in the image is captured by $\boldsymbol{s}_j^u = \frac{W_s^u|\boldsymbol{u}_j - \boldsymbol{a}_j^v|^2}{\left\| W_s^u|\boldsymbol{u}_j - \boldsymbol{a}_j^v|^2 \right\|_2}$.

We further measure the global semantic similarity $\boldsymbol{s}^{glo}$ between the whole image $\boldsymbol{v}^{glo}$ and full text $\boldsymbol{u}^{glo}$:

$$\boldsymbol{s}^{glo} = \frac{W_s^g|\boldsymbol{v}^{glo} - \boldsymbol{u}^{glo}|^2}{\left\| W_s^g|\boldsymbol{v}^{glo} - \boldsymbol{u}^{glo}|^2 \right\|_2} \tag{9}$$

where $W_s^g \in \mathbb{R}^{P \times D}$ is a learnable parameter matrix.

Cross-Modal Confidence When salient image regions are viewed separately, their visual semantics are fragmented, which leads to a locally aligned region-word that may be inconsistent with the global image-text semantics. The confidence is to show the consistency degree of each region with the global perspective of image-text, which can filter out the inconsistent matched region-word pairs. Specifically, we first extend each region $\boldsymbol{v}_i$ as its visual context $\boldsymbol{v}_i^{neig}$, in order to make the representation of each region more discriminative. The extended visual context can be exploited to verify how much semantic of the image region are described in the global text semantic, which is measured by the alignment between visual context $\boldsymbol{v}_i^{neig}$ and the full text $\boldsymbol{u}^{glo}$ as:

$$\boldsymbol{s}_i^{neig} = \frac{W_s^n|\boldsymbol{v}_i^{neig} - \boldsymbol{u}^{glo}|^2}{\left\| W_s^n|\boldsymbol{v}_i^{neig} - \boldsymbol{u}^{glo}|^2 \right\|_2} \tag{10}$$

where $W_s^n \in \mathbb{R}^{P \times D}$ is a learnable parameter matrix.

Referring to the given text, we have obtained how much semantic of the whole image are described in the global text semantic, namely $\boldsymbol{s}^{glo}$ in Eq.9. Then, bridged by the full text, we measure the matching confidence c_i with the normalized relevance between the global semantic similarity $\boldsymbol{s}^{glo}$ and the corresponding $\boldsymbol{s}_i^{neig}$ as:

$$\epsilon_i = \boldsymbol{w}_n \left(\boldsymbol{s}^{glo} \odot \boldsymbol{s}_i^{neig} \right), \; i = 1, 2, \cdots, N \tag{11}$$

$$\boldsymbol{c} = \sigma \left(\text{LayerNorm} \left([\epsilon_1, \epsilon_2, \cdots, \epsilon_N] \right) \right) \tag{12}$$

where $\boldsymbol{c} = [c_1, c_2, \cdots, c_N]$, $\boldsymbol{w}_n \in \mathbb{R}^{1 \times P}$ is a learnable parameter vector, $\odot$ indicates the element-wise product, σ in-

	Text Retrieval			Image Retrieval			
Method	R@1	R@5	R@10	R@1	R@5	R@10	R@sum
CAMP (Wang et al. 2019b)	68.1	89.7	95.2	51.5	77.1	85.3	466.9
SCAN (Lee et al. 2018)	67.4	90.3	95.8	48.6	77.7	85.2	465.0
SGM (Wang et al. 2020)	71.8	91.7	95.5	53.5	79.6	86.5	478.6
MMCA (Wei et al. 2020)	74.2	92.8	96.4	54.8	81.4	87.8	487.4
CAAN (Zhang et al. 2020b)	70.1	91.6	97.2	52.8	79.0	87.9	478.6
DPRNN (Chen and Luo 2020b)	70.2	91.6	95.8	55.5	81.3	88.2	482.6
PFAN (Wang et al. 2019a)	70.0	91.8	95.0	50.4	78.7	86.1	472.0
VSRN (Li et al. 2019)	71.3	90.6	96.0	54.7	81.8	88.2	482.6
IMRAM (Chen et al. 2020)	74.1	93.0	96.6	53.9	79.4	87.2	484.2
GSMN (Liu et al. 2020)	76.4	94.3	97.3	57.4	82.3	89.0	496.8
SGRAF(Diao et al. 2021)	77.8	94.1	97.4	58.5	83.0	88.8	499.6
CMCAN (ours)	**79.5**	**95.6**	**97.6**	**60.9**	**84.3**	**89.9**	**507.8**

Table 1: Comparisons with state-of-the-art methods on Flickr30K. The bests are in bold.

dicates the sigmoid function, and $\mathrm{LayerNorm}$ denotes the layer normalization operation. Note that the key idea here is that the matching confidence is inferred from how much semantic similarity between visual context of regions and the full text can be contained in the overall semantic similarity of image-text, since it indicates the relative extent to whether the region is really described from the global perspective of image-text.

Cross-Modal Relevance Measuring

To distinguish the matching confidence of region-word pairs in matching and filter out local semantic similarities contributed by the region-word pairs that locally matched but the regions are not really referred to in the global semantic of text, *i.e.*, unreliable matched region-word pairs, in overall cross-modal relevance measurement, we first multiply each region-queried semantic similarity $\boldsymbol{s}_i^v$ by the corresponding c_i. Thus, we can collect global semantic similarity and the scaled local similarities together as:

$$S_v = [\boldsymbol{s}^{glo}, c_1\boldsymbol{s}_1^v, \cdots, c_N\boldsymbol{s}_N^v] \tag{13}$$

Meanwhile, the global similarity $\boldsymbol{s}^{glo}$ and word-queried semantic similarities $\boldsymbol{s}_1^u, \boldsymbol{s}_2^u, \cdots, \boldsymbol{s}_M^u$ are collected together as $S_u = [\boldsymbol{s}^{glo}, \boldsymbol{s}_1^u, \cdots, \boldsymbol{s}_M^u]$.

We implement multi-layer self-attentional reasoning on the collected S_v and S_u, separately, in order to obtain modality specific enhanced global alignments:

$$S^{l+1} = \mathrm{ReLU}\left(W_r^l \cdot \mathrm{softmax}\left(W_q^l S^l \cdot \left(W_k^l S^l\right)^\top\right) \cdot S^l\right) \tag{14}$$

where $W_q^l \in \mathbb{R}^{P \times P}$ and $W_k^l \in \mathbb{R}^{P \times P}$ are parameter matices to transform attention query and key in the l^{th} layer respectively, and $W_r^l \in \mathbb{R}^{P \times P}$ is a parameter matrix to map the attended features to the next $l+1^{\text{th}}$ layer. Note that S_v^l and S_u^l are denoted as S^l in Eq.14.

Further, we concatenate the reasoned vision-enhanced global semantic similarity $\boldsymbol{s}_v^{glo_L}$ and language-enhanced global semantic similarity $\boldsymbol{s}_u^{glo_L}$ in the last L^{th} layer, and then feed the concatenated vision-language enhanced global similarity into a fully connected layer activated by the sigmoid function to measure the overall cross-modal relevance r between image I and text T:

$$r(I,T) = \sigma\left(\boldsymbol{w}_s\left([\boldsymbol{s}_v^{glo_L} : \boldsymbol{s}_u^{glo_L}]\right)\right) \tag{15}$$

where $\boldsymbol{w}_s \in \mathbb{R}^{1 \times 2P}$ is the learnable parameters to map the concatenated similarity vector to a scalar relevance score.

Objective Function

To cluster matched image-text pairs and enforce unmatched ones away from each other in the shared embedding space, the ranking objectives are widely employed in matching. Following (Faghri et al. 2018), we adopt the bi-directional triplet loss for end-to-end training, with focusing on the hard negatives within a minibatch for computational efficiency:

$$\begin{aligned}\mathcal{L}(I,T) = &\left[\lambda - r(I,T) + r\left(I, T_h^-\right)\right]_+ \\ &+ \left[\lambda - r(I,T) + r\left(I_h^-, T\right)\right]_+\end{aligned} \tag{16}$$

where λ is a margin constraint, $[x]_+ = \max(x, 0)$ and $r(\cdot)$ is the corss-modal semantic relevance measurement defined by Eq.15. Given a positive pair (I,T), $I_h^- = \arg\max_{I^- \neq I} r(I^-, T)$, and $T_h^- = \arg\max_{T^- \neq T} r(I, T^-)$ are the hardest negatives within the training minibatch.

Experiments

Datasets and Evalution Metrics

We evaluate our method on Flickr30K (Young et al. 2014) and MSCOCO (Lin et al. 2014) datasets. Flickr30K contains $31,000$ images and each image is captioned by 5 descriptions. Following dataset splits in (Lee et al. 2018), we use $29,000$ images for training, $1,000$ images for validation, and $1,000$ images for testing. MSCOCO contains $133,287$ images and each image is annotated with 5 sentences. We use $123,287$ images for training, $5,000$ images for validation, and $5,000$ images for testing, and the results are reported by both averaging over 5 folds of $1,000$ test images

Method	Text Retrieval			Image Retrieval			
	R@1	R@5	R@10	R@1	R@5	R@10	R@sum
CAMP (Wang et al. 2019b)	72.3	94.8	98.3	58.5	87.9	95.0	506.8
SCAN (Lee et al. 2018)	72.7	94.8	98.4	58.8	88.4	94.8	507.9
SGM (Wang et al. 2020)	73.4	93.8	97.8	57.5	87.3	94.3	504.1
MMCA (Wei et al. 2020)	74.8	95.6	97.7	61.6	89.8	95.2	514.7
CAAN (Zhang et al. 2020b)	75.5	95.4	98.5	61.3	89.7	95.2	515.6
DPRNN (Chen and Luo 2020b)	75.3	95.8	98.6	62.5	89.7	95.1	517.0
PFAN (Wang et al. 2019a)	76.5	96.3	99.0	61.6	89.6	95.2	518.2
VSRN (Li et al. 2019)	76.2	94.8	98.2	62.8	89.7	95.1	516.8
IMRAM (Chen et al. 2020)	76.7	95.6	98.5	61.7	89.1	95.0	516.6
GSMN (Liu et al. 2020)	78.4	96.4	98.6	63.3	90.1	95.7	522.5
SGRAF(Diao et al. 2021)	79.6	96.2	98.5	63.2	90.7	96.1	524.3
CMCAN (ours)	**81.2**	**96.8**	**98.7**	**65.4**	**91.0**	**96.2**	**529.3**

Table 2: Comparisons with state-of-the-art methods on MSCOCO 1K test images. The bests are in bold.

and testing on the full $5,000$ test images. As common in information retrieval, we measure the performance by R@K (recall at K) defined as the percentage of queries that are correctly matched in the closest K queried instances. R@1, R@5, R@10 are adopted as metrics. The higher R@K indicates better performance. To show overall matching performance, we sum up all recall values as R@sum at both image-to-text and text-to-image directions.

Implementation Details

We utilize the Faster R-CNN detector to extract $N = 36$ region proposals in each image then obtain a 2048-dimensional feature for each region. We set the word embedding dimension as 300. The dimension of vision-language shared embedding space D is set as 1024 and the dimension of distance-based similarity vectors P is 256. In region extended semantic representing, we extract $K = 3$ nearest detected regions in each of the top, bottom, left, and right scopes. For the region whose scopes are incomplete in the edges of the image, we use the region itself to supplement the lack, and further randomly discard one region in scopes to reduce the visual context distortion caused by the supplementation. The layer number L of the self-attentional mechanism for relevance measuring is 3. The Adam optimizer with 0.0002 as the initial learning rate is employed for model optimization. The learning rate is decayed by 10 times after 40 epochs in training on Flickr30K, and after 20 epochs in training on MSCOCO. The margin λ in triplet loss function is empirically set as 0.2. Source codes will be released.[1]

Comparisons with State-of-the-art Methods

We compare our proposed CMCAN with recent state-of-the-art methods on Flickr30K and MSCOCO datasets (For fair comparison, the feature extraction backbone of all methods is the same, *i.e.*, that for image is Faster R-CNN, and that for text is Bi-GRU). The experimental results are cited directly from respective papers. Comparison results are shown in Table 1 and Table 2 for Flickr30K and MSCOCO 1K, respectively. Note that our proposed CMCAN can achieve performance improvements on all metrics, compared to the state-of-the-art methods. On the Flickr30K test set, CMCAN outperforms other methods with R@1=79.5% for text retrieval and R@1=60.9% for image retrieval, obtaining performance improvements of 1.7% and 2.4%, respectively. On the MSCOCO 1K test set, our proposed CMCAN achieves the performance with R@1=81.2% for text retrieval and R@1=65.4% for image retrieval, which is a remarkable improvement. Our proposed CMCAN can outperform state-of-the-art methods by a large margin of 8.2% and 5.0% in terms of the overall performance R@sum on Flickr30K and MSCOCO, respectively. As shown in Table 3, CMCAN outperforms state-of-the-art models on almost all evaluation metrics in testing on the MSCOCO 5K test set. R@1=61.5% for text retrieval and R@1=44.0% for image retrieval, getting over 3.7% and 2.1% improvements, respectively. The consistently remarkable performance of CMCAN demonstrates its effectiveness and robustness.

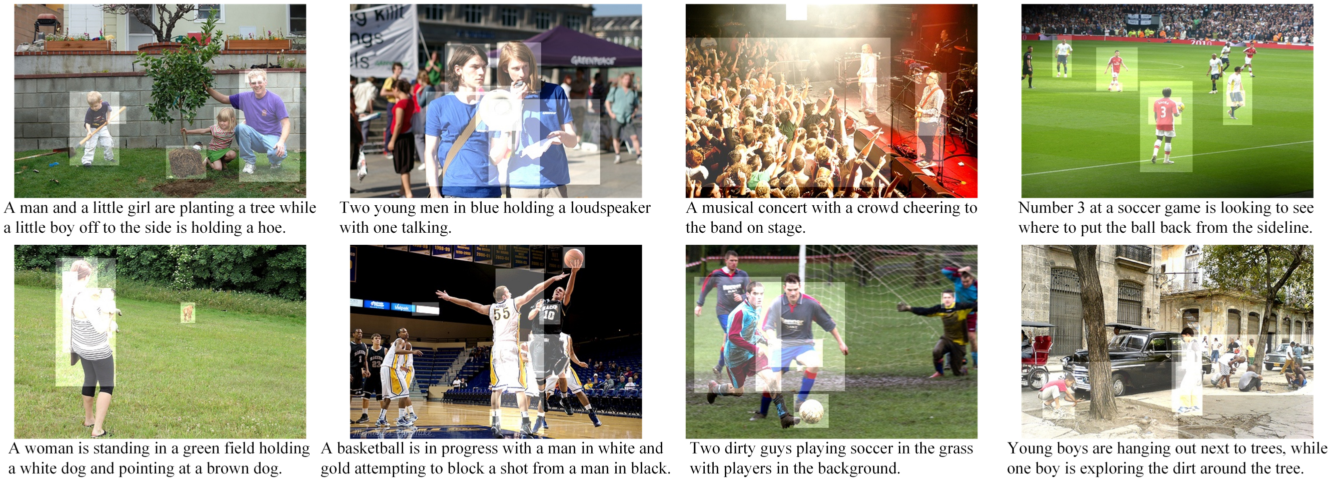
Method	Text Ret.		Image Ret.	
	R@1	R@10	R@1	R@10
CAMP (Wang et al. 2019b)	50.1	89.7	39.0	80.2
SCAN (Lee et al. 2018)	50.4	90.0	38.6	80.4
CAAN (Zhang et al. 2020b)	52.5	90.9	41.2	**82.9**
VSRN (Li et al. 2019)	53.0	89.4	40.5	81.1
IMRAM (Chen et al. 2020)	53.7	91.0	39.7	79.8
MMCA (Wei et al. 2020)	54.0	90.7	38.7	80.8
SGRAF (Diao et al. 2021)	57.8	91.6	41.9	81.3
CMCAN (ours)	**61.5**	**92.9**	**44.0**	82.6

Table 3: Comparisons on MSCOCO 5K test images.

Ablation Study

To show the effectiveness of the matching confidence in cross-modal relevance measurement, we enumerate the per-

[1] https://github.com/CrossmodalGroup/CMCAN



A man and a little girl are planting a tree while a little boy off to the side is holding a hoe.

Two young men in blue holding a loudspeaker with one talking.

A musical concert with a crowd cheering to the band on stage.

Number 3 at a soccer game is looking to see where to put the ball back from the sideline.

A woman is standing in a green field holding a white dog and pointing at a brown dog.

A basketball is in progress with a man in white and gold attempting to block a shot from a man in black.

Two dirty guys playing soccer in the grass with players in the background.

Young boys are hanging out next to trees, while one boy is exploring the dirt around the tree.

Figure 3: Visualization of the matching confidence. Brighter regions receive higher confidence w.r.t. the text, *i.e.*, the consistency degree with the global perspective of image-text. Results show CMCAN can accurately locate the really described regions.

Method	Text Ret.			Image Ret.			
	R@1	R@5	R@10	R@1	R@5	R@10	R@sum
without confidence	75.9	93.6	97.2	58.4	82.3	86.6	494.0
with confidence	77.5	94.3	96.9	58.8	82.9	88.9	499.3
CMCAN	79.5	95.6	97.6	60.9	84.3	89.9	507.8

Table 4: Ablation on Flickr30K. "without confidence" indicates the cross-modal relevance measuring without confidence, and "with confidence" is the opposite. CMCAN averages the relevance scores of two trained models in inference.

formance of the relevance measuring with and without matching confidence on Flickr30K in Table 4. The cross-modal relevance measurement with matching confidence outperforms that without the confidence on almost all metrics in both image retrieval and text retrieval directions. Specifically, the relevance measurement with matching confidence obtains improvements of 1.6% on R@1 and 0.7% on R@5 for text retrieval, 2.3% on R@10 for image retrieval, and 5.3% on the overall performance R@sum. CMCAN, which averages the cross-modal relevance scores of two trained models, outperforms the relevance measurement without confidence by 13.8% and measurement with confidence by 8.5% on the overall performance R@sum.

Qualitative Analysis

To verify the effectiveness of CMCAN, we visualize the learned matching confidence in Figure 3 which shows the relatively highest confidence in each image for brevity. The confidence is able to highlight the image regions that is really semantically consistent with the text to be matched, and guides to focus on the key scene in image-text matching. We also show the top-3 retrieval results given both image quries and text quries in Figure 4. It can be seen that images with similar contents are distinguished, since the inferred matching confidence can capture subtle visual clues.

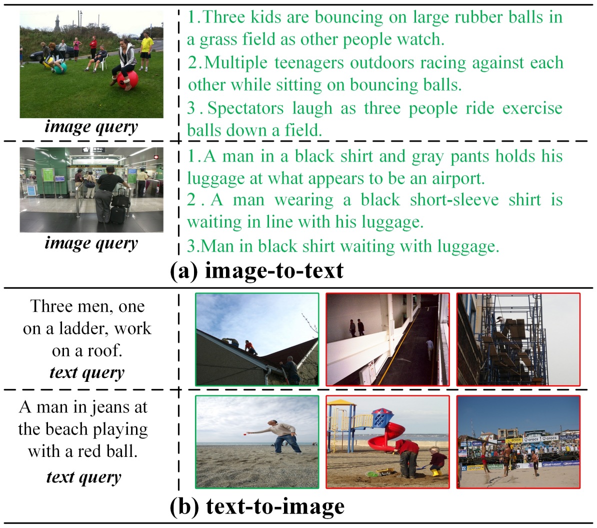


image query

1.Three kids are bouncing on large rubber balls in a grass field as other people watch.
2.Multiple teenagers outdoors racing against each other while sitting on bouncing balls.
3. Spectators laugh as three people ride exercise balls down a field.

image query

1.A man in a black shirt and gray pants holds his luggage at what appears to be an airport.
2. A man wearing a black short-sleeve shirt is waiting in line with his luggage.
3.Man in black shirt waiting with luggage.

(a) image-to-text

Three men, one on a ladder, work on a roof.
text query

A man in jeans at the beach playing with a red ball.
text query

(b) text-to-image

Figure 4: Case study, where the green texts or boxes denote the same with the ground-truth, and the red are not.

Conclusion

In this paper, we present a novel Cross-Modal Confidence-Aware Network for image-text matching, which infers the confidence of matched region-word pairs from the global perspective, enabling the model to be aware of whether the local matched pair is really described to refine the image-text relevance measurement. Moreover, bridging with the full text, we propose a delicately designed matching confidence measuring method via the whole image and the visual context of image regions. Extensive experiments are conducted to demonstrate the proposed method can significantly outperform state-of-the-art. Future works include employing the confidence-aware network into other multi-modal tasks, such as image captioning and visual question answering.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China under Grant U19A2057, Science Fund for Creative Research Groups under Grant 62121002, and Fundamental Research Funds for the Central Universities under Grants WK3480000008 and WK3480000010.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Bahdanau, D.; Cho, K. H.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12655–12663.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15789–15798.

Chen, T.; and Luo, J. 2020a. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10583–10590.

Chen, T.; and Luo, J. 2020b. Expressing Objects Just Like Words: Recurrent Visual Embedding for Image-Text Matching. In *AAAI*, 10583–10590.

Cui, H.; Zhu, L.; Li, J.; Yang, Y.; and Nie, L. 2019. Scalable deep hashing for large-scale social image retrieval. *IEEE Transactions on image processing*, 29: 1271–1284.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity Reasoning and Filtration for Image-Text Matching. In *AAAI*.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Gu, J.; Cai, J.; Joty, S. R.; Niu, L.; and Wang, G. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, Z.; Luo, Y.; Lin, J.; Yan, Y.; and Chen, J. 2019. Multi-Level Visual-Semantic Alignments with Relation-Wise Dual Attention Network for Image and Text Matching. In *IJCAI*, 789–795.

Huang, Y.; Wang, W.; and Wang, L. 2017. Instance-Aware Image and Sentence Matching With Selective Multimodal LSTM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, Y.; Wu, Q.; Song, C.; and Wang, L. 2018. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6163–6171.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–216.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4654–4662.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, C.; Mao, Z.; Zhang, T.; Xie, H.; Wang, B.; and Zhang, Y. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10921–10930.

Liu, Y.; Guo, Y.; Bakker, E. M.; and Lew, M. S. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, 4107–4116.

Nam, H.; Ha, J.-W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 299–307.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, volume 1, 2.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020. Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval. In *WACV*, 1497–1506.

Wang, Y.; Yang, H.; Qian, X.; Ma, L.; Lu, J.; Li, B.; and Fan, X. 2019a. Position Focused Attention Network for Image-Text Matching. In *IJCAI*, 3792–3798.

Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; and Shao, J. 2019b. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5764–5773.

Wehrmann, J.; Kolling, C.; and Barros, R. C. 2020. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12313–12320.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *CVPR*, 10941–10950.

Wu, Y.; Wang, S.; Song, G.; and Huang, Q. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2088–2096.

Xu, N.; Liu, A.-A.; Wong, Y.; Nie, W.; Su, Y.; and Kankanhalli, M. 2020. Scene graph inference via multi-scale context modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3): 1031–1041.

Xu, N.; Zhang, H.; Liu, A.-A.; Nie, W.; Su, Y.; Nie, J.; and Zhang, Y. 2019. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. *IEEE Transactions on Multimedia*, 22(5): 1372–1383.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yan, S.; Yu, L.; and Xie, Y. 2021. Discrete-continuous Action Space Policy Gradient-based Attention for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8096–8105.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yu, J.; Zhang, W.; Lu, Y.; Qin, Z.; Hu, Y.; Tan, J.; and Wu, Q. 2020. Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. *IEEE Transactions on Multimedia*, 22(12): 3196–3209.

Zhang, C.; Yang, Z.; He, X.; and Deng, L. 2020a. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3): 478–493.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020b. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3536–3545.

Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; and Zhang, H. 2020. Deep collaborative multi-view hashing for large-scale image search. *IEEE Transactions on Image Processing*, 29: 4643–4655.