



Universiteit
Leiden
The Netherlands

Improving weakly supervised phrase grounding via visual representation contextualization with contrastive learning

Wang, X.; Du, Y.; Verberne, S.; Verbeek, F.J.; Meng, X.; Wang, F.; ... ; Xie, X.

Citation

Wang, X., Du, Y., Verberne, S., & Verbeek, F. J. (2022). Improving weakly supervised phrase grounding via visual representation contextualization with contrastive learning. *Applied Intelligence*, 53(11), 14690-14702. doi:10.1007/s10489-022-04259-9

Version: Publisher's Version

License: [Licensed under Article 25fa Copyright Act/Law \(Amendment Taverne\)](#)

Downloaded from: <https://hdl.handle.net/1887/3640505>

Note: To cite this publication please use the final published version (if applicable).



Improving weakly supervised phrase grounding via visual representation contextualization with contrastive learning

Xue Wang^{1,2} · Youtian Du¹ · Suzan Verberne² · Fons J. Verbeek²

Accepted: 10 October 2022 / Published online: 2 November 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Weakly supervised phrase grounding aims to map the phrases in an image caption to the objects appearing in the image under the supervision of image-caption correspondence. We observe that the current studies are insufficient to model the complicated interactions between the visual components (i.e., the visual regions) and between the visual and textual components (i.e., the phrases). Therefore, this paper presents a novel weakly supervised learning approach to phrase grounding in which we systematically model the visual contextualized representation with three modules: (1) object proposals pooling (OPP), (2) visual self-attention (VSA) and (3) visual-textual cross-modal attention (VTCA). OPP alleviates the suppression of the object proposals and benefits the visual representation in terms of trading off the richness of the visual components and the computational efficiency. VSA aims to capture the correlation among the object proposals and generate a representation of each proposal by incorporating the visual information of the others. To measure the cross-modal compatibility in terms of topics, we introduce the VTCA module to represent the visual topic corresponding to each textual component in a cross-modal common vector space. In the training process, we build a mixed contrastive loss function by considering both the cross-modal compatibility and the differences in the visual representations in the VSA module. Compared with the state-of-the-art methods, the proposed approach improves the performance by 3.88% points and 1.24% points on $R@I$, and by 2.23% points and 0.26% points on $Pt.Acc$, when trained on the MS COCO and Flickr30K Entities training sets, respectively. We have made our code available for follow-up research.

Keywords Visual representation · Phrase grounding · Contrastive learning · Weakly supervised learning

1 Introduction

Tasks combining cross-modal (visual-and-language) compatibility have attracted much attention and have contributed

to the advancement of artificial intelligence in recent years. Examples of cross-modal tasks are image caption generation [1], visual question answering (VQA) [2], visual reasoning [3, 4] and phrase grounding [5]. Phrase grounding localizes the objects in images and at the same time, based on the paired images and captions, maps them to the phrases in the captions. Phrase grounding requires a model to understand the fine-grained correspondences between images and language. A large part of the previous works [6–8] are based on supervised learning, i.e., there is supervision of the correspondence between the visual regions and phrases. However, the availability of this kind of labeled data is limited due to the significant amount of manual effort required to collect the annotations for region-phrase correspondences.

To address the issue of limited data availability, researchers have proposed a few weakly supervised phrase grounding methods, which employ only the correspondence between images and text as supervision instead of the matching annotations of the visual regions and phrases. The

✉ Youtian Du
duyt@mail.xjtu.edu.cn

Xue Wang
nimowangxue1989@stu.xjtu.edu.cn

Suzan Verberne
s.verberne@liacs.leidenuniv.nl

Fons J. Verbeek
f.j.verbeek@liacs.leidenuniv.nl

¹ Faculty of Electronic and Information Engineering,
Xi'an Jiaotong University, Xi'an, 710049, China

² Leiden Institute of Advanced Computer Science, Leiden
University, Street, Leiden, 2333 CA, Leiden, The Netherlands

attention mechanism has become an important technique in solving the task of weakly supervised phrase grounding, and can generally be divided into two types. The first type models the intra-modality compatibility that infers the latent correlations between the different regions in an image or the different words in a caption [9] based on a self-attention mechanism. The other seeks to mine the cross-modal interactions between textual words and visual regions based on inter-modality compatibility [10]. That is, most of the previous methods only consider the inter-modality or intra-modality correlations.

Another issue of weakly supervised phrase grounding is how to choose loss functions to obtain a better learning result. Recently, contrastive learning, e.g., InfoNCE [11], has shown promising results on a variety of applications. Gupta et al. [12] proposed a novel contrastive learning approach to the task of weakly supervised phrase grounding, which improved the performance by employing the InfoNCE loss defined on the positive and negative samples.

In this paper, inspired by the advancements of contrastive learning [12] and phrase grounding [13], we introduce a new approach called VRC-PG to improve the weakly supervised phrase grounding with visual representation contextualization (VRC). In our method, the inter- and intra-modality interactions are modeled to infer the compatibility between the phrases and the visual regions. Here, we also call the phrase and visual region the textual component and the visual component, respectively. VRC-PG consists of three modules: object proposals pooling (OPP), visual self-attention (VSA) and visual-textual cross-modal attention (VTCA). In the visual representation contextualization, OPP is introduced to alleviate the suppression of object proposals (candidates) generated by the object detectors. This benefits visual representation contextualization in terms of trading off the richness of the visual components and the computational efficiency. VSA aims to capture the correlation between visual object proposals for each image and generate the representation of each candidate by incorporating the visual information of the other candidates. To measure the cross-modal compatibility at the level of topics, we subsequently introduce the VTCA module to distill the visual topic corresponding to each textual component, i.e., the textual phrase, in a cross-modal common vector space, guided by the attention of a phrase to the visual object proposals. In addition, we present a mixed contrastive loss function including two terms; one term is to improve the cross-modal compatibility in terms of topics of images and captions, and the other is to control the difference of the visual representations induced by the VSA module.

In summary, our research contributions are threefold: (1) We propose a novel approach to weakly supervised phrase grounding based on visual representation contextualization

under the weak supervision of image-caption correspondences without region-phrase matching annotations. Moreover, a mixed contrastive loss is introduced to improve the performance of our model. (2) We present an architecture of visual representation contextualization that consists of object proposals pooling (OPP), visual self-attention (VSA) and visual-textual cross-modal attention (VTCA). (3) The proposed model is evaluated on the Flickr30K Entities dataset and achieves a state-of-the-art performance, improving by 1.24% and 3.88% in terms of *Recall@1* on the Flickr30K Entities test set when trained on the Flickr30K Entities training set and MS COCO, respectively.

2 Related work

2.1 Phrase grounding

The existing works on phrase grounding are carried out mainly under two learning paradigms, namely, fully supervised learning and weakly supervised learning. Fully supervised learning methods employ the correspondence between visual regions and phrases in the training procedure. Plummer et al. [5] proposed a global image-sentence canonical correlation analysis model based on full supervision to measure the region-phrase correspondence in a combined image-text embedding space and achieved a state-of-the-art result on the Flickr30K Entities dataset. Bajaj et al. [14] utilized graphs to formulate more complex, non-sequential dependencies among object proposals and phrase candidates. However, it is time-consuming in practice to manually annotate data for the correspondences between phrases and visual regions. Thus, researchers have started to address the phrase grounding task under weak supervision. Plummer et al. [6] presented a weakly supervised learning method to localize the phrases in images by modeling the appearance, object size and position of the visual objects. Chen et al. [15] proposed a weak supervision novel knowledge-aided network, which was optimized by reconstructing the input information of queries and region proposals with the prediction labels extracted with a region proposal network (RPN). Akbari et al. [16] proposed a multi-level multimodal model to explicitly learn a nonlinear mapping of the visual and textual modalities in a common semantic space based on a weakly supervised learning process, and did so at different granularities for each modality. Most weakly supervised phrase grounding methods can be categorized into two classes; one class directly measures the similarity between the predicted textual labels of the visual region proposals and the phrases in the captions, the other class transforms both modalities into a common space and measures the similarity of a sentence and an image in such a space.

Recently, with the success of the attention mechanism in the cross-modal research field, researchers have employed the attention mechanism in phrase grounding to model the correspondence of visual regions and phrase problems. Vaswani et al. [17] introduced an attention mechanism in the Transformer model to mine the relation between the terms fed to the model. A number of works have been proposed to use the attention mechanism for an estimation of the similarity between the data from different modalities. Yu et al. [18] proposed a heterogeneous attention network (HAN) to build a cross-modal self-attention in the union of word features and bounding box features. To integrate more cross-modal information from the other modalities, Dong et al. [19] proposed a cross-modal graph attention strategy to generate a graph attention representation for each sample. To address the problem that image regions and words do not strictly match, Xu et al. [20] proposed an approach named cross-modal attention with semantic consistency (CASC) by jointly considering the local alignment and global semantic consistency. Overall, the main goal of the attention mechanism in cross-modal understanding is to reconstruct the representation of an example by aggregating the contextual information from both modalities.

In our work, we propose to build a visual representation contextualization architecture to enhance the performance of weakly supervised phrase grounding, which jointly considers the visual self-attention for the visual modality and the visual-textual cross-modal attention between both modalities. To optimize the proposed model, a mixed contrastive loss is defined in the visual space and cross-modal common space.

2.2 Non-maximum suppression (NMS)

NMS [21] has been an important technique for computer vision tasks, such as object detection [22, 23] and edge extraction [24]. In object detection, NMS is a post-processing step adopted by a number of modern object detectors, which can remove duplicate bounding boxes based on the confidence of detection. A major issue with NMS is that it sets the confidence of the neighboring detection results to zero. Thus, the object region proposals will be removed when their intersection over union (IoU) with the region proposal of the highest classification confidence is greater than a threshold, which will lead to a drop in the average precision. To alleviate this problem, Bodla et al. [25] presented the Soft-NMS algorithm to decrease the confidence scores as an increasing function of overlap instead of setting the score to zero as in NMS. Softer-NMS [26] proposed a bounding box regression Kullback-Leibler loss for learning the bounding box transformation and localization variance together. As a downstream task of object detection, language grounding

has been performed with NMS to align the language with the visual object proposals. Chen et al. [27] employed NMS to yield expression-aware region proposals to improve the performance of language grounding.

In our work, we use Soft-NMS to replace NMS in the generation of the regions of interest (RoIs) to keep more bounding box proposals and introduce an extra object proposals pooling module with NMS to adaptively choose those proposals with high confidence scores and to benefit the weakly supervised phrase grounding task.

2.3 Contrastive learning in cross-modal tasks

Contrastive learning was first used as a powerful scheme in self-supervised representation learning [11, 28–30]. Recently, contrastive learning has been introduced in cross-modal understanding tasks to enforce the consistency of representations from different modalities by leveraging contrasting positive example pairs and negative pairs. Zhang et al. [31] proposed a cross-modal model called XMC-GAN, which introduced an attentional self-modulation generator and a contrastive discriminator to maximize the cross-modal information between the images and text. Dai and Lin [32] proposed a method that encouraged the distinctiveness of positive pairs while maintaining the overall quality of the generated captions. Gupta et al. [12] built a weakly supervised phrase grounding model based on optimizing the lower bound of InfoNCE on mutual information (MI) with respect to parameters of a word-region attention model. Li et al. [33] proposed a framework combining a self-attention mechanism with contrastive feature construction to effectively summarize the common information from each image group while capturing the discriminative information between the visual regions and phrases. CDMLMR [34] integrates the quadruplet ranking loss and semi-supervised contrastive loss to model the cross-modal semantic similarity in a unified multi-task learning architecture.

Different from previous works, our work introduces a mixed contrastive loss for the learning of contextualized visual representations in the phrase grounding task. The mixed contrastive loss consists of two terms; one term improves the cross-modal compatibility in terms of the topics of images and captions, and the other controls the closeness of the visual representations before and after the VSA module.

3 Methodology

3.1 Overview

We are given a set of pairs, each consisting of an image and its caption. Formally, we have data $\mathcal{D}_i =$

$\{(I_i, C_i)\}_{i=1}^N$, where I_i and C_i denote the i -th image and its corresponding caption, respectively. In general, the content of an image, I_i , can be described by a set of n_i visual object regions enclosed with bounding boxes $\mathcal{B}_i = \{b_{i1}, b_{i2}, \dots, b_{in_i}\}$. The visual regions can be represented with the box location $\mathbf{B}_i = (b_{i1}, b_{i2}, \dots, b_{in_i})$, confidence score $\mathbf{S}_i = (s_{i1}, s_{i2}, \dots, s_{in_i})$, visual features $\mathbf{R}_i = (r_{i1}, r_{i2}, \dots, r_{in_i})$ and category predictions $\mathbf{L}_i = (l_{i1}, l_{i2}, \dots, l_{in_i})$. Regarding the textual modality, each caption C_i can be considered a sequence of m_i tokens $T_i = (t_{i1}, t_{i2}, \dots, t_{im_i})$, and transformed to the token representation $\mathbf{T}_i = (t_{i1}, t_{i2}, \dots, t_{im_i})$, using the BERT-base model [35]. A phrase consists of one or multiple tokens of captions. In this manner, the training data can be described by $\mathcal{D}_i = \{(\mathbf{B}_i, \mathbf{S}_i, \mathbf{R}_i, \mathbf{L}_i), \mathbf{T}_i\}_{i=1}^N$.

In this paper, we present a novel approach called VRC-PG for the task of weakly supervised phrase grounding. As shown in Fig. 1, our VRC-PG approach includes four main parts: (1) an object proposals pooling module, (2) a visual self-attention module, (3) a visual-textual cross-modal attention module and (4) a mixed contrastive loss function. The proposed approach models visual representation contextualization by jointly considering the interactions in both the unimodal data and cross-modal data and trains the model with a contrastive learning paradigm under the weak supervision of the correspondence between images and text.

3.2 Visual representation contextualization model

3.2.1 Feature extraction

The purpose of the visual representation contextualization model is to build the correspondence between the token representations $\mathbf{T}_i = (t_{i1}, t_{i2}, \dots, t_{im_i})$ and object candidate representations $\mathbf{R}_i = (r_{i1}, r_{i2}, \dots, r_{in_i})$ by measuring their attention.

We use the BERT-base model [35] to generate the textual representation, with captions as input.

$$t_{ij} = \text{BERT}(C_i), \quad (1)$$

where $t_{ij} \in \mathbb{R}^{d_t}$ is a dense vector representation.

We utilize the Faster R-CNN model [22] trained on the Visual Genome dataset [36] to extract and represent the visual objects from images:

$$(\{b_{ij}\}, \{s_{ij}\}, \{r_{ij}\}, \{l_{ij}\}) = \text{FasterRCNN}(I_i), \quad (2)$$

where $b_{ij} \in \mathbb{R}^4$ and $r_{ij} \in \mathbb{R}^{d_r}$ and s_{ij} is the maximum classification score among all the categories. In this work, we do not employ the predicted category labels, l_{ij} , generated by Faster R-CNN for each object region in our task.

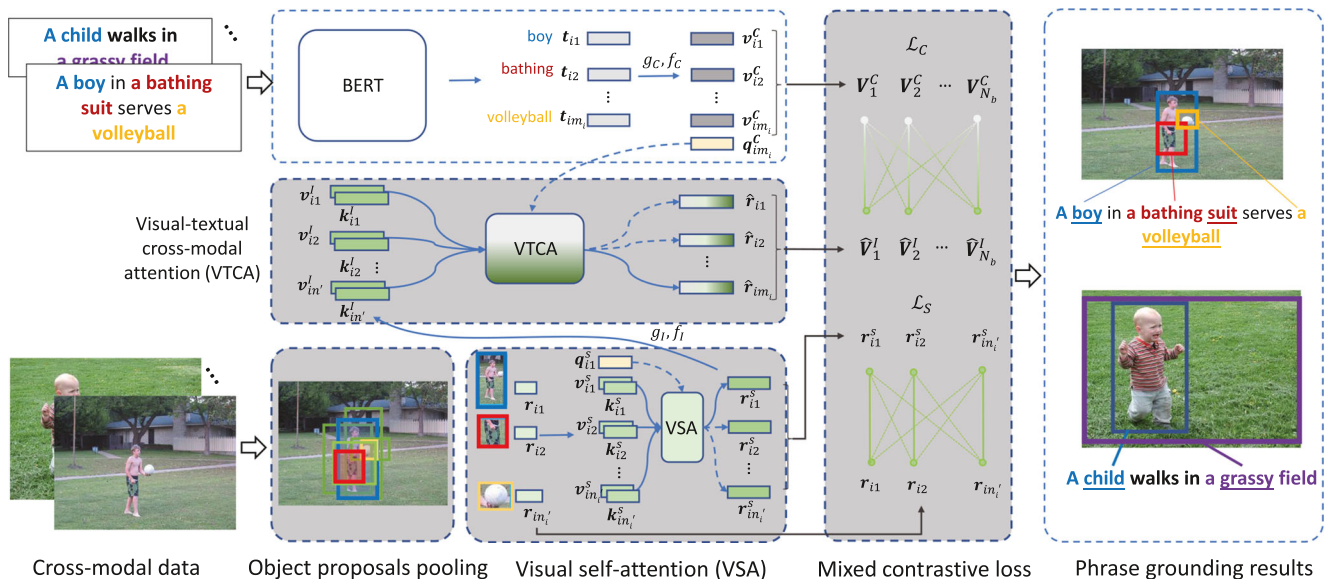


Fig. 1 The framework of VRC-PG. The visual representation contextualization comprises three parts: 1) object proposals pooling, where thick bounding boxes (red, blue and yellow) are the output boxes and thin bounding boxes (green) are non-maximally suppressed, 2)

visual self-attention, and 3) visual-textual cross-modal attention. The proposed model is trained with the contrastive learning paradigm by introducing our 4) mixed contrastive loss

3.2.2 Object proposals pooling (OPP)

As weakly supervised phrase grounding is performed without phrase grounding annotations, its quality depends on the accuracy of the object proposals extracted with Faster R-CNN. To keep more effective object proposals, we replace NMS used in Faster R-CNN with Soft-NMS [25]. The advantage of Soft-NMS is that it keeps more proposals for an object. However, it will cause a lower mapping accuracy if two objects overlap with each other. To alleviate this problem, we propose an object proposals pooling module based on NMS to further prune the detected objects and only keep the boxes with less than an IoU threshold θ in the training process. The OPP module can adaptively choose those proposals with high confidence scores, $\{s_{ij}\}$, and benefit the weakly supervised phrase grounding task.

For an image I_i , the pruning starts with a bounding box, b_{iz} , with the highest confidence score $s_{iz} = \max_j(s_{ij})$. b_{iz} is kept as one of the bounding boxes. Then, we update the confidence score of each bounding box b_{ij} by

$$s_{ij} = \begin{cases} s_{ij}, & \text{IoU}(\mathbf{b}_{ij}, \mathbf{b}_{iz}) < \theta, j \in 1, \dots, n_i; \\ 0, & \text{IoU}(\mathbf{b}_{ij}, \mathbf{b}_{iz}) \geq \theta, j \in 1, \dots, n_i. \end{cases} \quad (3)$$

Here, θ is a threshold to decide which object box should be directly excluded in each iteration of the object proposals pooling. Based on the above process, we can choose more bounding boxes based on (3) until all the confidence scores are updated to zero. Finally, the OPP module produces n'_i object proposals. In this module, we do not employ the category predictions generated by Faster R-CNN.

3.2.3 Visual self-attention (VSA)

In general, the visual components, i.e., the visual object proposals in an image, have spatial and semantic correlations with each other. In this paper, we introduce a visual self-attention module to model the context of the visual object regions and build their representations. A general attention mechanism [17] can be formulated as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\text{sim}(\mathbf{Q}, \mathbf{K})) \cdot \mathbf{V}, \quad (4)$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ and $\text{Attention}(\cdot, \cdot, \cdot)$ refer to the query, key, value and output, respectively, and $\text{sim}(\cdot, \cdot)$ denotes a certain function to measure the similarity of the queries and keys. In this work, the query (key) and value are obtained by the projection functions $f_I^s(\cdot)$ and $g_I^s(\cdot)$, respectively, and are implemented with a fully-connected layer as follows:

$$\begin{cases} \mathbf{q}_{ij}^s, \mathbf{k}_{ij}^s = f_I^s(\mathbf{r}_{ij}), & j = 1, \dots, n'_i; \\ \mathbf{v}_{ij}^s = g_I^s(\mathbf{r}_{ij}), & j = 1, \dots, n'_i, \end{cases} \quad (5)$$

where $\mathbf{q}_{ij}^s, \mathbf{k}_{ij}^s$ and $\mathbf{v}_{ij}^s \in \mathbb{R}^{d_s}$ refer to the vector of query, key and value, respectively. The soft weight of self-attention

from $\mathbf{r}_{ij}$ to $\mathbf{r}_{iu}$ can be measured by the correspondence between them defined as follows:

$$a_s(\mathbf{q}_{ij}^s, \mathbf{k}_{iu}^s) = \frac{e^{\mathbf{q}_{ij}^s \cdot \mathbf{k}_{iu}^s / \sqrt{d_s}}}{\sum_w e^{\mathbf{q}_{ij}^s \cdot \mathbf{k}_{iw}^s / \sqrt{d_s}}}. \quad (6)$$

Thus, the contextualized visual representation of an object region is obtained by considering the self-attention:

$$\mathbf{r}_{ij}^s = \sum_u a_s(\mathbf{q}_{ij}^s, \mathbf{k}_{iu}^s) \mathbf{v}_{iu}^s, \quad (7)$$

where $\mathbf{r}_{ij}^s$ denotes the contextualized visual representation for the object region $\mathbf{r}_{ij}$ that incorporates the global information of the i -th image. Figure 2 illustrates the structure of the VSA module.

3.2.4 Visual-textual cross-modal attention (VTCA)

To build an adaptive correspondence between the components of different modalities (i.e., the object proposals and tokens), we make a cross-modal alignment between the visual and textual components. Here, we introduce a visual-textual cross-modal attention module to find the visual components semantically related to a given textual component. First, we transform the representation of the textual components generated by BERT and the contextualized visual representation into a common space of dimensionality, d_c . In this module, we take the textual token as the query and measure the weight of attention to the visual components by computing the cross-modal correlation.

In the common space, the query and value for the token representation $\mathbf{t}_{ij}$ are generated by the functions $f_C(\cdot)$ and $g_C(\cdot)$, respectively, and the key and value for the visual region proposal $\mathbf{o}_{ij}$ are obtained by $f_I(\cdot)$ and $g_I(\cdot)$, respectively, as follows:

$$\begin{cases} \mathbf{q}_{ij}^C = f_C(\mathbf{t}_{ij}), & j = 1, \dots, m_i; \\ \mathbf{k}_{ij}^I = f_I(\mathbf{r}_{ij}^s), & j = 1, \dots, n'_i; \\ \mathbf{v}_{ij}^C = g_C(\mathbf{t}_{ij}), & j = 1, \dots, m_i; \\ \mathbf{v}_{ij}^I = g_I(\mathbf{r}_{ij}^s), & j = 1, \dots, n'_i, \end{cases} \quad (8)$$

where $\mathbf{t}_{ij}$ refers to the representation of token t_{ij} generated by BERT, $\mathbf{r}_{ij}^s$ is the contextualized visual representation obtained with (7) and $\mathbf{q}_{ij}^C, \mathbf{k}_{ij}^I, \mathbf{v}_{ij}^C$ and $\mathbf{v}_{ij}^I \in \mathbb{R}^{d_c}$. In this work, $f(\cdot)$ and $g(\cdot)$ are implemented with fully-connected layers.

Given the representation of a token obtained from BERT as a query, i.e., $\mathbf{q}_{ij}^C$, based on the attention

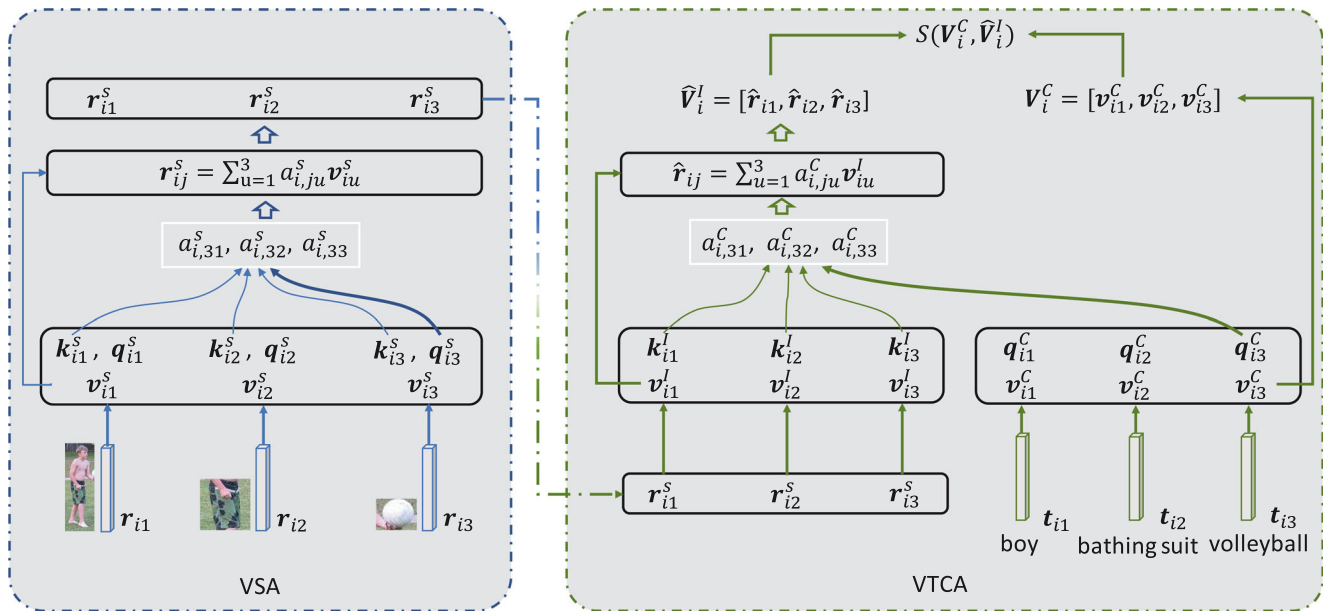


Fig. 2 The detailed structure of the VSA and VTCA modules. $a_{i,ju}^s$ and $a_{i,ju}^l$ denote the abbreviation of $a_s(q_{ij}^s, k_{iu}^s)$ in (6) and $a_c(q_{ij}^l, k_{iu}^l)$ in (9), respectively

mechanism [17], the cross-modal attention [12] is defined as follows:

$$a_c(q_{ij}^C, k_{iu}^l) = \frac{e^{q_{ij}^C \cdot k_{iu}^l / \sqrt{d_c}}}{\sum_{w=1}^{n'_i} e^{q_{ij}^C \cdot k_{iw}^l / \sqrt{d_c}}}, \quad (9)$$

$$\hat{r}_{ij} = \sum_{u=1}^{n'_i} a_c(q_{ij}^C, k_{iu}^l) v_{iu}^l, \quad (10)$$

where $\hat{r}_{ij}$ represents a visual topic correlated to the semantics of token t_{ij} by incorporating the textual token information with the cross-modal attention. Figure 2 illustrates the structure of the VTCA module.

3.3 Mixed contrastive loss function

For a mini-batch in the learning process, we have N_b pairs of captions and images represented with $V_i^C = [v_{i1}^C, v_{i2}^C, \dots, v_{im_i}^C]$ and $\hat{V}_j^l = [\hat{r}_{i1}, \hat{r}_{i2}, \dots, \hat{r}_{im_i}]$, respectively, based on VTCA. We measure the similarity between two samples of different modalities as follows:

$$S(v_i^C, \hat{v}_j^l) = \frac{e^{\text{tr}(V_i^{\text{CT}} \cdot \hat{V}_j^l)}}{\sum_{k=1}^{N_b} e^{\text{tr}(V_i^{\text{CT}} \cdot \hat{V}_k^l)}}, \quad (11)$$

where $\text{tr}(\cdot)$ and the superscript T denote the trace and transposition of a square matrix. (11) uses a softmax operator to normalize the similarity to sum 1.

In contrastive learning, an image and its matching caption construct a positive sample pair (i.e., $i = j$), and the non-matching image-caption pairs in a mini-batch are

considered negative sample pairs (i.e., $i \neq j$). Based on the similarity measured in (11), we introduce a contrastive loss function at the granularity of images and captions:

$$\mathcal{L}_C = -\frac{1}{N_b} \sum_{i=1}^{N_b} \log(S(v_i^C, \hat{v}_i^l)) / \mathcal{T}, \quad (12)$$

where $\mathcal{T}$ is a temperature hyperparameter. The loss in (12) seems to only work on the positive pairs and does not involve the negative pairs. Actually, to maximize the similarity $S(\cdot, \cdot)$ in (12) for the positive pair will lead to the suppression of the similarity for the negative pairs due to the sum-to-one normalization shown in (11), which is just a manner of contrastive learning.

In addition, we introduce a loss to force the outputs of the visual self-attention module to be close to its inputs. The visual self-attention loss is defined as follows:

$$\mathcal{L}_S = -\frac{1}{N_b} \sum_{i=1}^{N_b} \left(\frac{1}{n'_i} \sum_{j=1}^{n'_i} \log \left(\frac{e^{(r_{ij} \cdot r_{ij}^s)}}{\sum_{u=1}^{n'_i} e^{(r_{ij} \cdot r_{iu}^s)}} \right) \right). \quad (13)$$

Clearly, the visual self-attention loss is also a contrastive loss.

Finally, we build a mixed contrastive loss function in the form of

$$\mathcal{L} = \alpha \mathcal{L}_C + \mathcal{L}_S, \quad (14)$$

where α is a hyperparameter to control the balance of both terms.

4 Experimental results

4.1 Datasets and metrics

The experiments are conducted on the Flickr30K Entities dataset [5] and MS COCO 2014 dataset. The Flickr30K Entities contains 31,873 images and 5 captions per image. Following the work by Gupta et al. [12], we split the Flickr30K Entities dataset into a training set with 29,783 images, a validation set with 1,000 images and a test set with 1,000 images. The MS COCO 2014 dataset [1] contains 118,287 training images and 5,000 validation images, where each image is provided with 5 human-annotated captions. In the training process, we randomly select one caption from 5 captions of each example as the textual input.

We use two standard metrics, namely, *Recall@K* ($R@K$) and *Pt_Acc* [12], to evaluate the performance. *Recall@K* ($R@K$) for $K \in \{1, 5, 10\}$ measures the percentage of phrases for which $IoU > 0.5$ between the top K predicted bounding boxes and the ground truth boxes. Unlike *Recall@K*, *Pt_Acc* does not require identifying the IoU of the predicted object box. Generally, the center point of the selected bounding box is used as the prediction for each phrase to compute the point accuracy.

4.2 Implementation details

4.2.1 Visual feature representation

We extract the visual region proposals from images using Faster R-CNN with a backbone ResNet-101 [22] based on the bottom-up attention method [37], which was trained on the Visual Genome dataset. The region proposals contain the bounding boxes, visual features and confidence scores (after Soft-NMS thresholding). We choose 50 RoIs based on the confidence scores and obtain 2048-dimensional visual representations (i.e., $d_r = 2048$). By the VSA module, we will reduce the dimension of the visual representations from 2048 to 768 (i.e., $d_s = 768$).

4.2.2 Textual feature representation

We follow the setting of the BERT model proposed by Gupta et al. [12] and employ a pre-trained BERT [35] for the generation of textual representations. A 768-dimensional token representation, i.e., $d_t = 768$, is generated for a word, t_{ij} , in captions with the BERT model. The dimension of the common space generated by VTCA is set to 384, i.e., $d_c = 384$.

4.2.3 Parameter tuning

The hyperparameters are determined with a grid search on the Flickr30K Entities validation set. The threshold θ in (3) is set to 0.5, the same value used in the evaluation of models in terms of the $R@K$ metrics. We train our model for 10 epochs with a batch size of 30 using an SGD optimizer with a momentum of 0.9 and a learning rate of 10^{-5} .

Figure 3 shows the effect of the hyperparameter temperature $\mathcal{T}$ in (12) and α in (14) on the performance in terms of $R@1$, $R@5$, $R@10$ and *Pt_Acc* on the Flickr30K Entities validation set. From Fig. 3(a), (b) and (d), we can see that the model with $\mathcal{T} = 0.07$ achieves the best performance in terms of $R@1$, $R@5$ and *Pt_Acc*. In terms of $R@10$, the model with $\mathcal{T} = 0.07$ may also achieve a result close to the best one (e.g., $\alpha = 4$ and 9). It is noted that the best performances mentioned above are achieved at different values of α . In general, we consider that $R@1$ is a more important metric, and thus select the final checkpoint for the best performance in terms of $R@1$. Finally, we set $\alpha = 16$ and $\mathcal{T} = 0.07$.

4.3 Quantitative results

Table 1 presents the experimental results of the compared methods on the Flickr30K Entities test set. From this table, we observe that our proposed approach outperforms the state-of-the-art work [38] by 1.24% and 0.26% in terms of $R@1$ and *Pt_Acc*, respectively, with the model trained on the Flickr30K Entities training set. For the models trained on MS COCO, our approach improves the performance by 3.88% points and 2.23% points in terms of $R@1$ and *Pt_Acc*, respectively, compared with the state-of-the-art work [12]. For the other cases, we observe that our approach is superior to the compared methods as a whole.

In terms of $R@10$, our model obtains a lower performance (−1.41%) than InfoGround [12] when trained on the Flickr30K Entities training set. We analyzed this difference and found that our approach without the OPP module obtains an $R@10$ of 83.86%, improving the performance by 0.95% compared with InfoGround. The reason is that after the OPP module, we keep a smaller set of object proposals as input to the next module than without the OPP module. The main contribution of InfoGround is that it generates a context-preserving negative caption set based on a language model, which improves the results in comparison with randomly sampling negatives from the training data. In our approach, we do not employ this negative caption set. To verify this, we retrain our model by employing the negative caption set used in InfoGround [12]. Our proposed

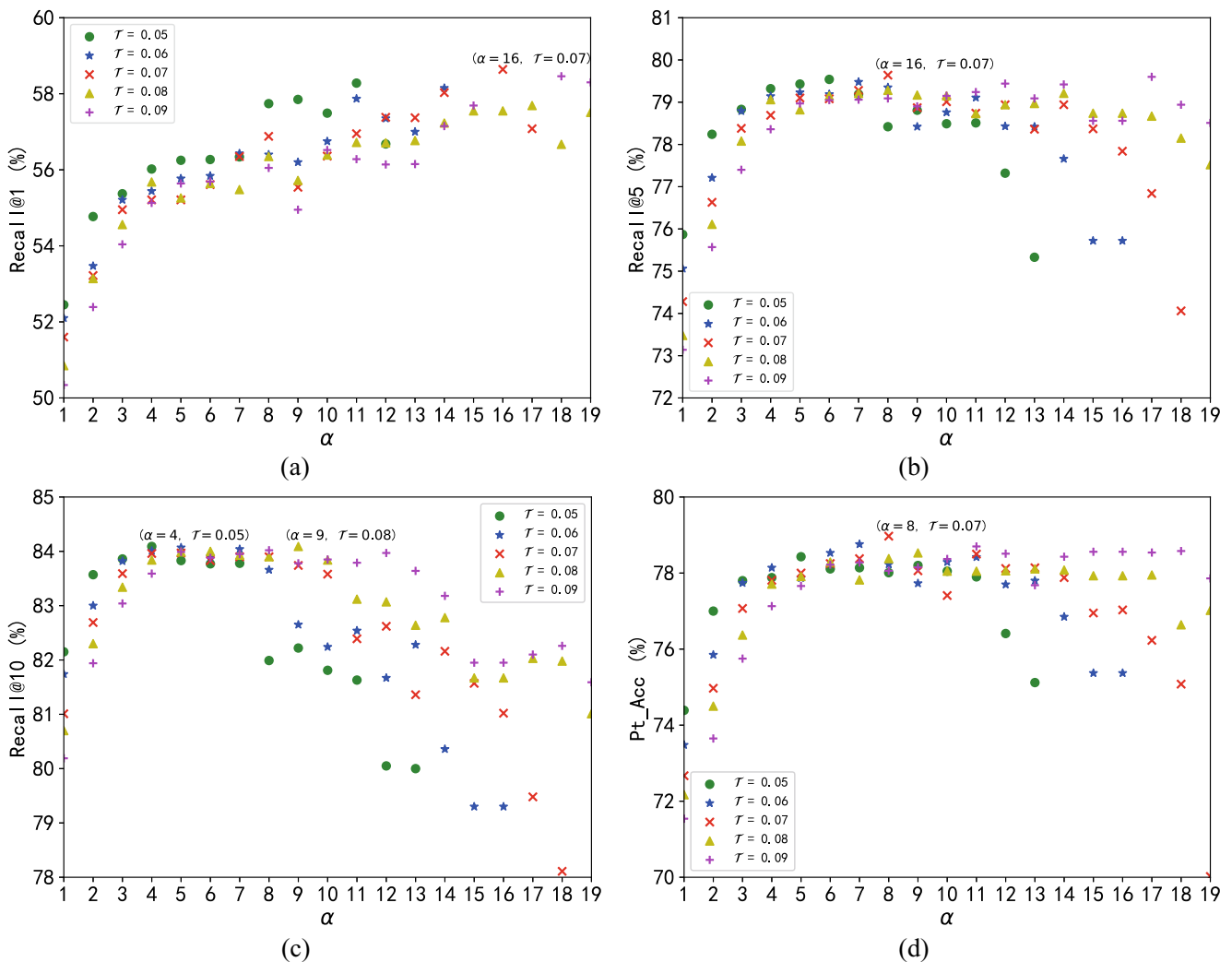


Fig. 3 The effect of the hyperparameter temperature, $\mathcal{T}$, and loss function weight, α , on (a) $R@1$, (b) $R@5$, (c) $R@10$ and (d) Pt_Acc on the validation set of the Flickr30K Entities dataset

model with these negative captions achieves 66.60% and 78.83% in terms of $R@1$ and Pt_Acc , respectively, with the model trained on the Flickr30K Entities training set. For the models trained on MS COCO, our approach achieves 59.47% and 79.34% in terms of $R@1$ and Pt_Acc , respectively, when the negative captions are employed. Both of them demonstrate that our approach achieves much higher performances than InfoGround when employing the same negative caption settings.

4.4 Ablation study

In Table 2, we report the quantitative performance of 8 different design choices, i.e., c1-c8, within our proposed model on the Flickr30K Entities validation set. In this

experiment, we take the design only consisting of the VTCA module as our baseline model, which is only supervised by image-caption pairs based on InfoNCE loss, similar to the model by Gupta et al. [12]. The introduction of VSA improves Pt_Acc from 62.43% to 64.26% but results in a drop in $R@1$ from 32.13% to 29.64% (c1 vs. baseline). Our OPP module, as shown in Table 2, brings a performance gain of 3.24% in terms of $R@1$ but a 1.46% lower Pt_Acc (c2 vs. baseline). When we use these two modules together, $R@1$ is improved from 32.13% to 39.21% and Pt_Acc from 62.43% to 63.61% (c3 vs. baseline). Thus, OPP is more positive for $R@1$ and VSA for Pt_Acc . If we want to simultaneously optimize both metrics, these two kinds of modules can work in coordination with each other. We replace the InfoNCE loss in the baseline with our contrastive

Table 1 The comparison of the results (%) of our approach with the state-of-the-art on the Flickr30K entities test set

Methods	Training data	R@1	R@5	R@10	Pt_Acc
GroundeR [39]	Flickr30K	28.94	—	—	—
KAC Net [15]		38.71	—	—	—
InfoGround [12]		47.88	76.63	82.91	74.94
Contrastive Distillation[40]		53.10	—	—	—
RIR [38]		59.27	—	—	78.60
VRC-PG (ours)	Flickr30K	60.51	78.77	81.50	78.86
MS Research [41]	MS COCO	—	—	—	29.00
MultiGrounding [16]		—	—	—	69.19
Align2Ground [9]		—	—	—	71.00
InfoGround [12]		51.67	77.69	83.25	76.74
VRC-PG (ours)		55.55	79.23	84.12	78.97

The models have been trained on Flickr30K entities and MS COCO

loss function (without $\mathcal{L}_S$) and achieve an improvement of 16.77% on $R@1$ and 14.17% on Pt_Acc (c4 vs. baseline). If we further add the visual self-attention loss $\mathcal{L}_S$, we can obtain a better result on $R@1$ and a close result on Pt_Acc (c6 vs. c5 and VRC-PG vs. c8). This shows that our contrastive loss is very useful in the phrase grounding task.

In Fig. 4, we visualize a few examples of phrase grounding for different model settings, i.e., with and without VSA, on the Flickr30K validation set. The figure indicates that the setting with VSA can lead to more attention being paid to the correct visual region corresponding to the phrase in the caption than without VSA. For example, for the top-right example in the figure, we find that the setting with VSA gives an attention score (0.82) to the bounding box (red) enclosing a man, while the setting without VSA

generates a lower attention score (0.73) for the region (red) covering the man and a large area of the background.

4.5 Qualitative results

In Fig. 5, we illustrate a few qualitative results of phrase grounding obtained by our approach on three image-caption pairs from the Flickr30K Entities test data. From this figure, it is evident that our model has the ability to localize the phrases from a caption in an image. In Fig. 6, we show the attention scores obtained by (9) from the VTCA module in our model. For example, for the word ‘old’, our approach generates a high attention to visual region No. 17 (cf. Fig. 6(a)). It is visible in the image that this region contains a head with white hair and exhibits a visual appearance of

Table 2 Benefits of the different modules in our approach

Methods	OPP	VSA	Loss	R@1	Pt_Acc
baseline	—	—	—	32.13	62.43
c1	—	✓	—	29.64	64.26
c2	✓	—	—	35.37	60.97
c3	✓	✓	—	39.21	63.61
c4	—	—	✓ w/o $\mathcal{L}_S$	48.90	76.60
c5	—	✓	✓ w/o $\mathcal{L}_S$	52.71	78.31
c6	—	✓	✓	53.20	78.27
c7	✓	—	✓ w/o $\mathcal{L}_S$	55.64	77.58
c8	✓	✓	✓ w/o $\mathcal{L}_S$	57.90	77.24
VRC-PG	✓	✓	✓	58.64	77.03

All models are trained on the Flickr30K Entities training set, and the results (%) are reported for the Flickr30K Entities validation set

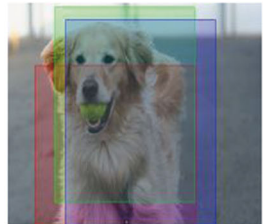
w/o VSA	VSA	w/o VSA	VSA
			
[0.15, 0.12, 0.11]	[0.19, 0.17, 0.09]	[0.73, 0.19, 0.05]	[0.82, 0.16, 0.02]
Two older men are sitting on opposite ends of a bench.		A blond man stands next to a cement mixer with mountains in the background.	
			
[0.67, 0.31, 0.01]	[0.79, 0.16, 0.04]	[0.15, 0.13, 0.12]	[0.21, 0.13, 0.11]
A man and a little girl happily posing in front of their cart in a supermarket.		Four girls in shorts on the beach throwing a football with the ocean behind them.	
			
[0.18, 0.18, 0.13]	[0.25, 0.2, 0.15]	[0.17, 0.12, 0.11]	[0.18, 0.14, 0.14]
A little white curly-haired dog runs across the pavement with a stick in its mouth.		A golden-colored dog , with his eyes alert, holds a brightly colored tennis ball in his mouth.	
			
[0.49, 0.3, 0.06]	[0.49, 0.45, 0.04]	[0.89, 0.03, 0.02]	[0.96, 0.02, 0.01]
A single man, riding his bike on the pier at sunset.		A young girl in a green shirt and shorts out riding her bike past a very nice apartment building.	

Fig. 4 The attention scores achieved in (9) of the region proposals on the Flickr30K Entities validation set for the setting without/with the visual self-attention module (i.e., w/o VSA and VSA). The visual

regions surrounded by bounding boxes refer to the object proposals with top-3 cross-modal attention scores (colored red, green and blue)

‘old’. Regions 29 and 3 are about the topic of scenes, and we can observe that the corresponding cells are highlighted in the attention weight map when the query phrase is ‘park’

and ‘bench’. Regions 2 and 6, both related to the semantics of a man, are paid much attention weight for the query phrase ‘men’, as shown in the attention weight map.

A bike riding couple dressed in bike gear and helmets take a minute to sit on a bench to talk and park their bikes



bike [0.14, 0.11, 0.07]



couple [0.09, 0.08, 0.08]



bench [0.12, 0.06, 0.06]

Two old men sit on opposite ends of a park bench.



men [0.12, 0.11, 0.07]



park [0.16, 0.15, 0.08]

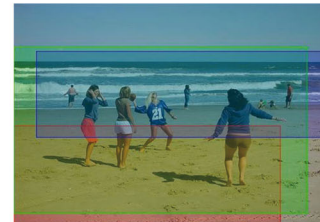


bench [0.11, 0.1, 0.1]

Four girls in shorts on the beach throwing a football with the ocean behind them.



ocean [0.14, 0.14, 0.1]



beach [0.2, 0.16, 0.11]



girls [0.09, 0.08, 0.08]

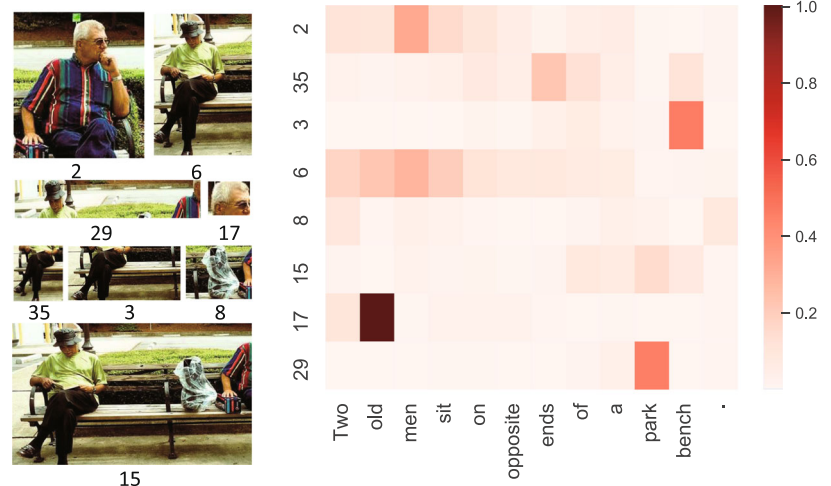
Fig. 5 Visualization of weakly supervised phrase grounding. In each image, for a given word query, we show the visual regions in the form of bounding boxes with top-3 cross-modal attention scores (colored by red, green and blue) achieved in (9)

5 Conclusion

In this work, we have proposed a novel weakly supervised approach to phrase grounding under the supervision of the correspondence between images and captions. Our key contribution lies in systematically learning contextualized visual representations with a mixed contrastive loss

function. In the visual representation contextualization, the three modules, OPP, VSA and VTCA, work in coordination with each other for representing local visual semantics by considering the unimodal and cross-modal contexts. In addition, we define a novel contrastive loss function on the intra- and inter-modal representations and clearly demonstrate that this leads to better results. Overall, we

Fig. 6 The cross-modal attention scores achieved by (9) between the visual object proposals and words. The darker cell color indicates that more attention is given to the corresponding visual object proposals for a word query



(a) Visual object proposals

(b) Attention weight map

report improvements of 3.88% points and 1.24% points on $R@1$, and 2.23% points and 0.26% points on Pt_Acc , with the models trained on the MS COCO and Flickr30K Entities training set, respectively, compared with the state-of-the-art methods. Our qualitative analysis, using a visualization of the attention between words and the image regions, also illustrates the capability of our model to learn the joint representations of images and text using the attention mechanism.

Acknowledgements This research is supported in part by National Key R&D Program of China (2018AAA0101501), China Scholarship Council (201906280464), and National Natural Science Foundation of China (61772415).

Funding This research is supported in part by National Key R&D Program of China (2018AAA0101501), China Scholarship Council (201906280464), and National Natural Science Foundation of China (61772415).

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests The authors declare that there is no conflict of interests regarding the publication of this article.

References

- Chen X, Fang H, Lin TY et al (2015) Microsoft COCO captions: data collection and evaluation server arXiv:150400325
- Antol S, Agrawal A, Lu J et al (2015) VQA: visual question answering. In: Proceedings of the IEEE international conference on computer vision, pp 2425–2433
- Suhr A, Zhou S, Zhang A et al (2018) A corpus for reasoning about natural language grounded in photographs. arXiv:181100491
- Zellers R, Bisk Y, Farhadi A et al (2019) From recognition to cognition: visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 6720–6731
- Plummer BA, Wang L, Cervantes CM et al (2015) Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, pp 2641–2649
- Plummer BA, Mallya A, Cervantes CM et al (2017) Phrase localization and visual relationship detection with comprehensive image-language cues. In: Proceedings of the IEEE international conference on computer vision, pp 1928–1937
- Fukui A, Park DH, Yang D et al (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:160601847
- Wang L, Li Y, Huang J et al (2018) Learning two-branch neural networks for image-text matching tasks. IEEE Trans Pattern Anal Mach Intell 41(2):394–407
- Datta S, Sikka K, Roy A et al (2019) Align2Ground: weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2601–2610
- Lai F, Xie N, Doran D et al (2019) Contextual grounding of natural language entities in images. arXiv:191102133
- Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv:180703748
- Gupta T, Vahdat A, Chechik G et al (2020) Contrastive learning for weakly supervised phrase grounding. In: Proceedings of the European conference on computer vision, Springer, pp 752–768
- Yu T, Hui T, Yu Z et al (2020) Cross-modal omni interaction modeling for phrase grounding. In: Proceedings of the 28th ACM international conference on multimedia, pp 1725–1734
- Bajaj M, Wang L, Sigal L (2019) G3raphGround: graph-based language grounding. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4281–4290
- Chen K, Gao J, Nevatia R (2018) Knowledge aided consistency for weakly supervised phrase grounding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4042–4050
- Akbari H, Karaman S, Bhargava S et al (2019) Multi-level multimodal common semantic space for image-phrase grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12,476–12,486
- Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. arXiv:170603762
- Yu T, Yang Y, Li Y et al (2021) Heterogeneous attention network for effective and efficient cross-modal retrieval. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval, pp 1146–1156
- Dong X, Zhang H, Dong X et al (2021) Iterative graph attention memory network for cross-modal retrieval. Knowl-Based Syst 226:107,138
- Xu X, Wang T, Yang Y et al (2020) Cross-modal attention with semantic consistence for image-text matching. IEEE Trans Neural Netw Learning Syst 31(12):5412–5425
- Neubeck A, Van Gool L (2006) Efficient non-maximum suppression. In: 18th international conference on pattern recognition (ICPR'06), pp 850–855
- Ren S, He K, Girshick R et al (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
- Redmon J, Divvala S, Girshick R et al (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
- Zitnick CL, Dollár P (2014) Edge boxes: locating object proposals from edges. In: European conference on computer vision. Springer, pp 391–405
- Bodla N, Singh B, Chellappa R et al (2017) Soft-NMS – improving object detection with one line of code. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 5561–5569
- He Y, Zhang X, Savvides M et al (2018) Softer-NMS: rethinking bounding box regression for accurate object detection, vol 2(3) arXiv:180908545
- Chen L, Ma W, Xiao J et al (2021) Ref-NMS: Breaking proposal bottlenecks in two-stage referring expression grounding. In: Proceedings of the AAAI conference on artificial intelligence, pp 1036–1044
- He K, Fan H, Wu Y et al (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9729–9738
- Chen T, Kornblith S, Norouzi M et al (2020) A simple framework for contrastive learning of visual representations. In:

- International Conference on Machine Learning, PMLR, pp 1597–1607
30. Wu Z, Xiong Y, Yu SX et al (2018) Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3733–3742
 31. Zhang H, Koh JY, Baldrige J et al (2021) Cross-modal contrastive learning for text-to-image generation. arXiv:210104702
 32. Dai B, Lin D (2017) Contrastive learning for image captioning. arXiv:171002534
 33. Li Z, Tran Q, Mai L et al (2020) Context-aware group captioning via self-attention and contrastive features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3440–3450
 34. Huang X, Peng Y (2017) Cross-modal deep metric learning with multi-task regularization. In: 2017 IEEE international conference on multimedia and expo (ICME). IEEE, pp 943–948
 35. Devlin J, Chang MW, Lee K et al (2019) BERT: pre-training Of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, vol 1, (Long and short papers), pp 4171–4186
 36. Krishna R, Zhu Y, Groth O et al (2017) Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis* 123(1):32–73
 37. Anderson P, He X, Buehler C et al (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, pp 6077–6086
 38. Liu Y, Wan B, Ma L et al (2021) Relation-aware instance refinement for weakly supervised visual grounding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5612–5621
 39. Rohrbach A, Rohrbach M, Hu R et al (2016) Grounding of textual phrases in images by reconstruction. In: European conference on computer vision. Springer, pp 817–834
 40. Wang L, Huang J, Li Y et al (2021) Improving weakly supervised visual grounding by contrastive knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14,090–14,100
 41. Fang H, Gupta S, Iandola F et al (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1473–1482

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.