Multi modal representation learning and cross-modal semantic matching

Wang, X.

5

Visual Representation Contextualization Based on Constrastive Learning

This chapter is based on the following publication:

Wang, X., Du, Y., Verberne, S. Verbeek, F.J. Improving Weakly Supervised Phrase Grounding via Visual Representation Contextualization with Contrastive Learning. Applied Intelligence. (under review)

CHAPTER SUMMARY

This chapter addresses RQ5 and RQ6.

RQ5: What is the effect of the attention mechanism to eliminate the different modal representations produced in the common embedding space?
RQ6: How to employ the correspondence between images and text as supervision instead of the matching annotations to address the limited data issue?

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 correspondences. We observe that the current studies are insufficient to model the complicated interactions between visual components (i.e., visual regions) and between visual and textual components (i.e., phrases). Therefore, this chapter 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 object proposals and benefits the visual representation in terms of trading off the richness of visual components and the computational efficiency. VSA aims to capture the correlation among the object proposals and generate the representation of each proposal by incorporating the visual information of the others. In order 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 difference of visual representations in the VSA module. Compared to the state-of-the-art methods, the proposed approach improves the performance by 3.88% point and 1.24% point on $R@1$, and by 2.23% point and 0.26% point on Pt_Acc, when trained on the MS COCO and Flickr30K Entities training set, respectively. We have made our code available for follow-up research.

Tasks combining cross-modal (visual-and-language) compatibility have attracted a lot of attention and contributed to the advancement of artificial intelligence in recent years. Examples of cross-modal tasks are image caption generation [147], visual question answering (VQA) [148], visual reasoning [149, 150], and phrase grounding [23]. Phrase grounding aims to localize the objects in images and at the same time, based on paired images and captions, maps them to the phrases in captions. Phrase grounding requires a model to understand the fine-grained correspondence between images and language. A large part of previous works plummer2017phrase,fukui2016multimodal,wang2018learning are based on supervised learning, i.e., with supervision of the correspondence between visual regions and phrases. However, the availability of this kind of labelled data is limited due to significant manual efforts in collecting the annotations for region-phrase correspondences.

To address the issue of limited availability of data, researchers have proposed a few weakly supervised phrase grounding methods, which only employ the correspondence between images and text as supervision instead of the matching annotations of visual regions and phrases. The attention mechanism has becoming 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 different regions in an image or different words in a caption [151] based on self-attention mechanism. The other seeks to mine the cross-modal interactions between textual words and visual regions based on inter-modality compatibility [152]. That is, most of the previous methods only consider the correlations either in inter-modality or in intra-modality.

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 [35], has shown promising results on a variety of applications. Gupta et al. [18] 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 chapter, inspired by the advancements of contrastive learning [18] and phrase grounding [17], we introduce a new approach, called VRC-PG, to improve weakly supervised phrase grounding with visual representation contextualization (VRC). In our method, the inter- and intra-modality interactions are modeled for inferring the compatibility between phrases and visual regions. Here, we also call the phrase and visual region as the textual component and 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 object detectors. This benefits the visual representation contextualization in terms of trading off the richness of 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., textual phrase, in a cross-modal common vector space, guided by the attention of a phrase to visual object proposals. In addition, we present a mixed contrastive loss function including two terms: one is to improve 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 contributions are three-fold: (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 Flickr30K Entities dataset and achieves the state-of-the-art performance, improving by 1.24% point and 3.88% point $Recall@1$ on the Flickr30K Entities test set when trained on the Flickr30K Entities training set and MS COCO, respectively.

5.1. Related Work

5.1.1. Phrase Grounding

The existing works are based on two different supervision processes, fully supervised learning and weakly supervised learning. Plummer et al. [23] proposed a global image-sentence canonical correlation analysis (CCA) model to analyze the region-phrase correspondence in the combined image-text embedding space, and achieved a state-of-the-art result for this task on the Flickr30K Entities dataset. Wang, and Sigal [153] used graphs to formulate more complex, non-sequential dependencies among object region proposals and phrase candidates. Most of these methods employ the annotations of region-phrase correspondences and are implemented under the supervised learning framework. Because manual labelling is expensive, also some other research has used the approach of weakly supervised learning. Plummer et al. [154] presented a weakly supervised learning method that modeled the appearance, object size and position of visual objects to localize phrases in images. Akbari et al. [36] proposed a multi-level multi-modal model to explicitly learn a non-linear mapping of the visual and textual modalities in a common semantic space, and do so at different granularities for each modality. Recently, the attention mechanism has been introduced to reconstruct the representation of vision and text guided by inter- or intra-modality. The result is a cross-modal attention mechanism with a fully supervised or weakly supervised learning framework. Chen et.al. [155] proposed a novel knowledge-aided network which was optimized by reconstructing input information of queries and region proposals extracted by a region proposal network (RPN). These existing methods lack the ability to model image-caption paired supervision. This is essential for grounding phrases in the images based on weak supervision from caption-image pairs. In this chapter, we propose the VRC-PG approach and model the fine-grained interactions in the inter- and intra-modality by jointly considering the visual self-attention mechanism and cross-modal attention mechanism.

5.1.2. NON-MAXIMUM SUPPRESSION (NMS)

NMS [156] has been an important technique for computer vision tasks, such as object detection [30, 58] and edge extraction [157]. In object detection, NMS is a post-processing step adopted by a number of modern object detectors, which removes duplicate bounding boxes based on detection confidence. A major issue with NMS is that it sets the score for neighboring detection to zero. Thus, if an object is actually present in an overlap region with an IoU greater than the threshold it would be missed and this would lead to a drop in average precision.

To alleviate this problem, Bodla et al. [158] 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 [159] proposed a bounding box regression Kullback-Leibler loss for learning bounding box transformation and localization variance together. As a downstream task of object detection, language grounding methods have used NMS to align the language with the proposals. Chen et al. [160] used NMS to yield expression-aware region proposals to improve the performance language grounding. In our work, we use Soft-NMS to replace the NMS module in Faster R-CNN 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 benefiting the weakly supervised phrase ground task.

5.1.3. CONTRASTIVE LEARNING IN CROSS-MODAL TASKS

Constrastive learning was first used as a powerful scheme for self-supervised representation learning [35, 161, 162, 163]. Until now, it has been explored to enforce consistency of different modal representations under different augmentations by contrasting positive pairs with negative ones. Zhang et al. [164] 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 images and text. Dai and Lin [13] proposed a method that encouraged the distinctiveness of positive pairs, while maintaining the overall quality of the generated captions. Gupta et al. [18] 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. [165] proposed a framework combining a self-attention mechanism with contrastive feature construction so as to effectively summarize common information

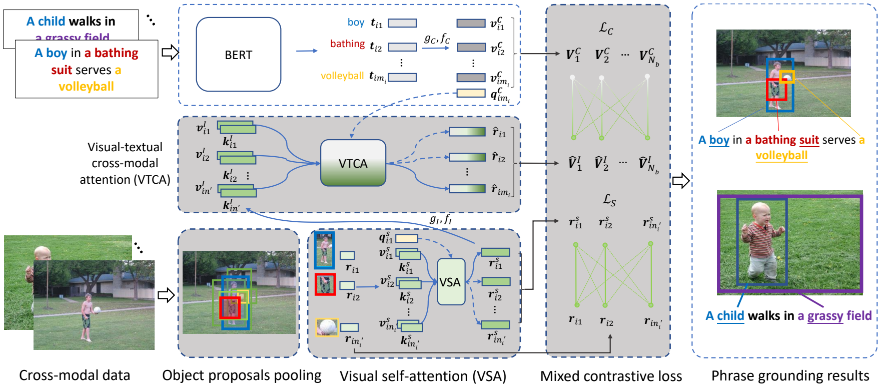


Figure 5.1: The framework of VRC-PG. The visual representation contextualization is comprised of three parts: 1) object proposals pooling, where thick bounding boxes (red, blue and yellow) are the output boxes and thin bounding boxes (green) are non-maximally suppressed, 2) visual self-attention, and 3) visual-textual cross-modal attention. The proposed model is trained with the contrastive learning paradigm by introducing our 4) mixed contrastive loss.

from each image group while capturing discriminative information between visual regions and phrases. CDMLMR [166] integrates the quadruplet ranking loss and semi-supervised constrastive loss for modeling cross-modal semantic similarity in a unified multi-task learning architecture. In our work, we learn our model with the contrastive learning paradigm and build a mixed contrastive loss function, which consists of two terms: one is control the difference of the visual representations induced by the VSA module, and the other is to improve the cross-modal compatibility in terms of the topics of images and captions.

5.2. Methodology

5.2.1. Overview

We are given a set of pairs, each consisting of an image and its caption. Formally, we have data $\mathscr{D}_i = \{(I_i, C_i)\}_{i=1}^N$, 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 $\mathscr{B}_i = \{b_{i1}, b_{i2}, \cdots, b_{in_i}\}$. The visual regions can be represented with the box location $\boldsymbol{B}_i = (\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2}, \cdots, \boldsymbol{b}_{in_i})$, confidence score $\boldsymbol{S}_i = (s_{i1}, s_{i2}, \cdots, s_{in_i})$, visual features $\boldsymbol{R}_i = (\boldsymbol{r}_{i1}, \boldsymbol{r}_{i2}, \cdots, \boldsymbol{r}_{in_i})$, and category predictions $\boldsymbol{L}_i = (l_{i1}, l_{i2}, \cdots, l_{in_i})$. Regarding the textual modality, each caption C_i can be considered a sequence of m_i tokens $T_i = (t_{i1}, t_{i2}, \cdots, t_{im_i})$ and transformed to the token representation $\boldsymbol{T}_i = (\boldsymbol{t}_{i1}, \boldsymbol{t}_{i2}, \cdots, \boldsymbol{t}_{im_i})$ using the BERT-base model [28]. A phrase consists of one or multiple tokens of captions. In this manner, the training data can be described by $\mathscr{D}_i = \{(\boldsymbol{B}_i, \boldsymbol{S}_i, \boldsymbol{R}_i, \boldsymbol{L}_i), \boldsymbol{T}_i\}_{i=1}^N$.

In this chapter, we present a novel approach called VRC-PG to the task of weakly supervised phrase grounding. As shown in Fig. 5.1, our VRC-PG approach includes four main parts: (1) object proposals pooling module, (2) visual self-attention module, (3) visual-textual cross-modal attention module and (4) mixed contrastive loss function. The proposed approach models visual representation contextualization by jointly considering the interactions in both the unimodal data and the cross-modal data, and trains the model with a contrastive learning paradigm under the weak supervision of the correspondence between images and text.

5.2.2. VISUAL REPRESENTATION CONTEXTUALIZATION MODEL

Feature extraction

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

We use the BERT-base model [28] to extract the text modal representation with caption as input.

$$\boldsymbol{t}_{ij} = BERT(C_i), \tag{5.1}$$

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

We utilize the Faster R-CNN [30] model trained on the Visual Genome dataset [22] to extract and represent the objects:

$$(\{\boldsymbol{b}_{ij}\}, \{s_{ij}\}, \{\boldsymbol{r}_{ij}\}, \{l_{ij}\}) = FasterRCNN(I_i), \tag{5.2}$$

where $\boldsymbol{b}_{ij} \in \mathbb{R}^4$ and $\boldsymbol{r}_{ij} \in \mathbb{R}^{d_r}$, s_{ij} is the maximum classification score among all 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.

Object Proposals Pooling (OPP)

As weakly supervised phrase grounding is performed without phrase grounding annotations, its quality depends on the performance of object box proposals extracted with Faster R-CNN. In order to keep more effective object box proposals, we replace NMS used in Faster R-CNN by Soft-NMS [158]. The advantage of Soft-NMS is to keep more proposals for an object. However, it will cause the mapping accuracy to be lower if two objects overlap between 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 boxes 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 from the weakly supervised phrase ground 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 produced the OPP module. Then, We update the confidence scores of all the bounding box b_{ij} by

$$s_{ij} = \begin{cases} s_{ij}, & IoU(\boldsymbol{b}_{ij}, \boldsymbol{b}_{iz}) < \theta, j \in 1, \ldots, n_i; \\ 0, & IoU(\boldsymbol{b}_{ij}, \boldsymbol{b}_{iz}) \geq \theta, j \in 1, \ldots, n_i. \end{cases} \tag{5.3}$$

Here, θ is a threshold to decide which object box should be directly excluded in each iteration of object proposals pooling. Based on the above process, we can choose more bounding boxes based on Eq. 5.3 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.

Visual Self-attention (VSA)

In general, the visual components, i.e., the visual object region proposals comprised in an image, have spatial and semantic correlation with each other. We introduce a visual self-attention module to model the context of visual object regions and build their representations. The general attention mechanism can be formulated accordingly as follows:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\text{sim}(\boldsymbol{Q}, \boldsymbol{K})) \cdot \boldsymbol{V}, \tag{5.4}$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ and Attention$(\cdot,\cdot,\cdot)$ refer to the query, key, value and output, respectively; and sim$(\cdot,\cdot)$ denotes a certain function to measure the corresponding of 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, implemented with a fully-connected layer as follows:

$$\begin{cases} \boldsymbol{q}_{ij}^s, \boldsymbol{k}_{ij}^s = f_I^s(\boldsymbol{r}_{ij}), \; j = 1, \cdots, n'_i; \\ \boldsymbol{v}_{ij}^s = g_I^s(\boldsymbol{r}_{ij}), \; j = 1, \cdots, n'_i. \end{cases} \tag{5.5}$$

Where, $\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{ij}^s$ and $\boldsymbol{v}_{r_{ij}}^s \in \mathbb{R}^{d_s}$ refer to the vector of query, key and value, respectively. The soft weight of self-attention from $\boldsymbol{r}_{ij}$ to $\boldsymbol{r}_{iu}$ can be measured by the corresponding between them defined as follows:

$$a_s(\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{iu}^s) = \frac{e^{\boldsymbol{q}_{ij}^s \cdot \boldsymbol{k}_{iu}^s / \sqrt{d_s}}}{\sum_w e^{\boldsymbol{q}_{ij}^s \cdot \boldsymbol{k}_{iw}^s / \sqrt{d_s}}}. \tag{5.6}$$

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

$$\boldsymbol{r}_{ij}^s = \sum_u a_s(\boldsymbol{q}_{ij}^s, \boldsymbol{k}_{iu}^s) \boldsymbol{v}_{iu}^s, \tag{5.7}$$

where $\boldsymbol{r}_{ij}^s$ denotes the contextualized visual representation for the object region $\boldsymbol{r}_{ij}$ that incorporates the global information of the i-th image.

Visual-textual Cross-modal Attention (VTCA)

To build an adaptive correspondence between the cross-modal components (i.e., object region 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 semantically related components in the visual modality for a given textual component. First, we transform the representation of 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 actor 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 $\boldsymbol{t}_{ij}$ are generated by the functions $f_C(\cdot)$ and $g_C(\cdot)$, respectively, and the key and value for the visual region proposal o_{ij} are obtained by $f_I(\cdot)$ and $g_I(\cdot)$, respectively, as follows:

$$\begin{cases} \boldsymbol{q}^C_{ij} = f_C(\boldsymbol{t}_{ij}),\ j = 1,\cdots,m_i; \\ \boldsymbol{k}^I_{ij} = f_I(\boldsymbol{r}^s_{ij}),\ j = 1,\cdots,n'_i; \\ \boldsymbol{v}^C_{ij} = g_C(\boldsymbol{t}_{ij}),\ j = 1,\cdots,m_i; \\ \boldsymbol{v}^I_{ij} = g_I(\boldsymbol{r}^s_{ij}),\ j = 1,\cdots,n'_i, \end{cases} \tag{5.8}$$

where $\boldsymbol{t}_{ij}$ refers to the representation of token t_{ij} generated by BERT, $\boldsymbol{r}^s_{ij}$ is the contextualized visual representation obtained with Eq. 5.7 and $\boldsymbol{q}^C_{ij}, \boldsymbol{k}^I_{ij}, \boldsymbol{v}^C_{ij}$ and $\boldsymbol{v}^I_{ij} \in \mathbb{R}^{d_c}$. 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., $\boldsymbol{q}^C_{ij}$, based on the attention mechanism [167] , the cross-modal attention [18] is defined as follows:

$$a_c(\boldsymbol{q}^C_{ij}, \boldsymbol{k}^I_{iu}) = \frac{e^{\boldsymbol{q}^C_{ij} \cdot \boldsymbol{k}^I_{iu} / \sqrt{d_c}}}{\sum_{w=1}^{n'_i} e^{\boldsymbol{q}^C_{ij} \cdot \boldsymbol{k}^I_{iw} / \sqrt{d_c}}}, \tag{5.9}$$

$$\hat{\boldsymbol{r}}_{ij} = \sum_{u=1}^{n'_i} a_c(\boldsymbol{q}^C_{ij}, \boldsymbol{k}^I_{iu}) \boldsymbol{v}^I_{iu}, \tag{5.10}$$

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

5.2.3. MIXED CONTRASTIVE LOSS FUNCTION

For a mini-batch of size N_b in the learning process, we have N_b captions and images represented with $\boldsymbol{V}^C_i$ and $\hat{\boldsymbol{V}}^I_j$. Here, the textual representation $\boldsymbol{V}^C_i = [\boldsymbol{v}^C_{i1}, \boldsymbol{v}^C_{i2}, \cdots, \boldsymbol{v}^C_{im_i}]$ and visual representation $\hat{\boldsymbol{V}}^I_j = [\hat{\boldsymbol{r}}_{i1}, \hat{\boldsymbol{r}}_{i2}, \cdots, \hat{\boldsymbol{r}}_{im_i}]$ obtained from VTCA, we measure the similarity of two cross-modal samples as follows:

$$S(\boldsymbol{V}_i^C, \hat{\boldsymbol{V}}_j^I) = \frac{e^{\mathrm{tr}(\boldsymbol{V}_i^{C\top} \cdot \hat{\boldsymbol{V}}_j^I)}}{\sum_{k=1}^{N_b} e^{\mathrm{tr}(\boldsymbol{V}_i^{C\top} \cdot \hat{\boldsymbol{V}}_k^I)}}, \tag{5.11}$$

where tr(·) and the superscript ⊤ denote the trace and transposition of a square matrix. Eq. 5.11 uses a softmax operator to normalize the similarity to sum 1.

For contrastive learning, in each mini-batch, an image and its matching caption are denoted a positive sample pair (i.e., $i = j$) and non-matching image-caption pairs are negative sample pairs (i.e., $i \neq j$). Based on the similarity measured by Eq. 5.11, we provide a contrastive loss function at the granularity of images and captions:

$$\mathscr{L}_C = -\frac{1}{N_b} \sum_{i=1}^{N_b} log(S(\boldsymbol{V}_i^C, \hat{\boldsymbol{V}}_i^I))/\mathscr{T}, \tag{5.12}$$

where $\mathscr{T}$ is a temperature hyper-parameter. The loss in Eq. 5.12 seems to only work on the positive pairs and do not involve the negative pairs. Actually, to maximize the similarity $S(\cdot,\cdot)$ in Eq. 5.12 for the positive pair will lead to the suppression of the similarity for the negative pairs due to the sum-to-one normalization in Eq. 5.11, which is just a manner of the 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:

$$\mathscr{L}_S = -\frac{1}{N_b} \sum_{i=1}^{N_b} \left(\frac{1}{n_i'} \sum_{j=1}^{n_i'} log \left(\frac{e^{(\boldsymbol{r}_{ij} \cdot \boldsymbol{r}_{ij}^s)}}{\sum_{u=1}^{n_i'} e^{(\boldsymbol{r}_{ij} \cdot \boldsymbol{r}_{iu}^s)}} \right) \right). \tag{5.13}$$

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

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

$$\mathscr{L} = \alpha \mathscr{L}_C + \mathscr{L}_S, \tag{5.14}$$

where α is a hyper-parameter to control the balance of both terms.

5.3. Experimental Results

In this section we first describe the datasets followed by the implementation details.

5.3.1. Datasets and Metrics

Datasets

The experiments are conducted on the Flickr30K Entities dataset and MS COCO 2014 dataset.

- Flickr30K Entities contains 31,873 images and 5 captions per image. Following Gupta et al. [18], we split the Flickr30K Entities in a training set with 29,783 images, a validation set with 1,000 images and a test set with 1,000 images. The Flickr30K Entites dataset provides the correspondence of phrases and visual object regions. Thus, the Flickr30K Entities validation set and test set are employed to validate the proposed model and test its performance, respectively, in this work.
- The MS COCO 2014 dataset contains 118,287 training images and 5,000 validation images, where each image is provided with 5 human-annotated captions. The MS COCO 2014 dataset does not contain the links between image regions and sentence phrases. We thus train our model on the MS COCO 2014 training set, validate and test on the Flickr30K Entities validation and test sets, respectively. In the training process, we randomly select one caption from 5 captions of each example as the textual segment.

Metrics

We use two standard metrics for this task:

- $Recall@K$ ($R@K$) for $K = 1, 5$ and 10 measures the percentage of phrases for which $IoU > 0.5$ between the top K predicted bounding boxes and the ground truth boxes.
- Pt_Acc refers to pointing accuracy and is commonly used to evaluate weakly supervised phrase grounding models [18]. *Pt_Acc* is the proportion of phrases for which center point of the predicted bounding box falls in the ground truth box. Unlike *R@K*, pointing accuracy 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 for computing pointing accuracy.

5.3.2. IMPLEMENTATION DETAILS

Visual Feature Representation

We extract visual region proposals from each image using Faster R-CNN with a backbone ResNet-101 [30] based on the bottom-up attention method [11], which was trained on the Visual Genome dataset. The region proposals contain the bounding boxes, visual features and Faster R-CNN's confidence scores (after Soft-NMS thresholding). We choose 50 regions of interest (RoI) based on confidence scores and obtain 2048-dimensional visual representations (i.e., $d_r = 2048$). By the VSA module we will reduce the dimension of visual representations from 2048d to 768d (i.e., $d_s = 768$).

Textual Feature Representation

We follow the setting of the BERT model used in the work of Gupta et al. [18] for the generation of the textual representation, where a pre-trained BERT model [28]

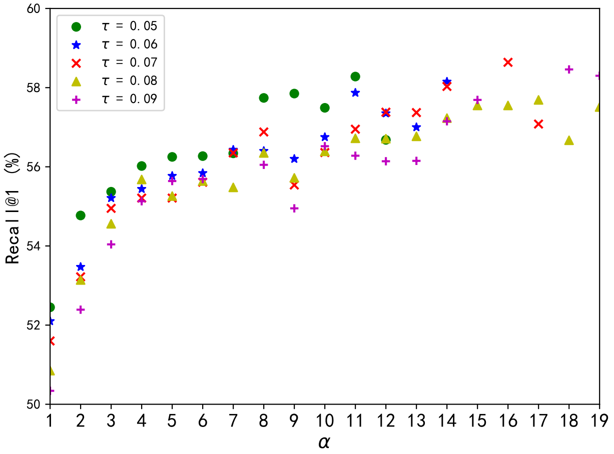


Figure 5.2: The hyper-parameters temperature τ and the loss function weight α optimized for Recall@1 on the validation set of Flickr30K.

is employed. 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 the VTCA is set to is 384, i.e., $d_c = 384$.

Parameter Tuning

The hyper-parameters are determined with a grid searching on the Flickr30K Entities validation set. The threshold θ in Eq. 5.3 is set to 0.5, a same value as used in the evaluation of models in terms of the $R@K$ metrics. In our research, we perform grid search for determining the parameters. Fig. 5.2 shows the optimization result of the hyperparameters α from Eq. 5.14 and temperature $\mathscr{T}$ in Eq. 5.12. We train our model for 10 epochs with a batch size of 30 using an SGD optimizer with momentum 0.9 and a learning rate of 10^{-5}. We select the final checkpoints on the basis of the model's best performance in terms of $R@1$ on the Flickr30K Entities validation set. Based on the validation results, we set $\alpha = 16$ and $\mathscr{T} = 0.07$.

5.3.3. Quantitative Results

Table 5.1 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 [15] by 1.24% point and 0.26% point 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% point and 2.23% point in terms of $R@1$ and Pt_Acc, respectively, compared to the state-of-the-art work [18]. For the other cases, we observe that our approach is superior to the compared methods as a whole.

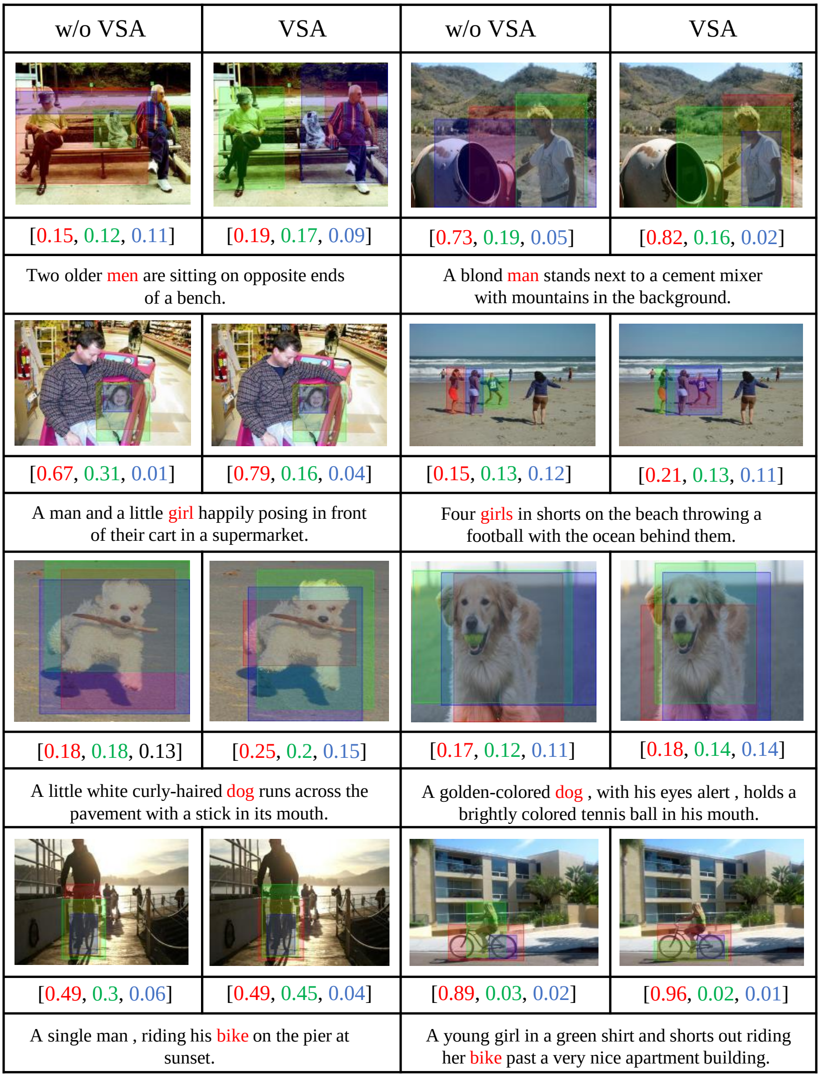
Table 5.1: The comparison of the results (%) of our approach with the state-of-the-art on the Flickr30K Entities test set. The models have been trained on Flickr30K Entities and MS COCO.

Methods	Training data	R@1	R@5	R@10	Pt_Acc
GroundeR [168]	Flickr30K	28.94	-	-	-
KAC Net [155]		38.71	-	-	-
InfoGround [18]		47.88	76.63	**82.91**	74.94
Wang et al. [14]		53.10	-	-	-
Liu et al. [15]		59.27	-	-	78.60
VRC-PG (ours)	Flickr30K	**60.51**	**78.77**	81.50	**78.86**
Fang et al. [65]	MS COCO	-	-	-	29.00
Akbari et al. [36]		-	-	-	69.19
Align2Ground [151]		-	-	-	71.00
InfoGround [18]		51.67	77.69	83.25	76.74
VRC-PG (ours)	MS COCO	**55.55**	**79.23**	**84.12**	**78.97**

In terms of $R@10$, our model obtains a lower performance (-1.41%) than InfoGround [18] when trained on the Flickr30K Entities training set. We analyzed this difference and found that our approach without the OPP module gets an $R@10$ of 83.86% which improves the performance of InfoGround by 0.95% point. The reason is that after the OPP module, we keep a smaller object proposals set as input to the next module than without the OPP module. The main contribution of the InfoGround model is that it uses the language model to generate a context-preserving negative caption set; the authors show that this improves the results in comparison to randomly sampling negatives from the training data. In our approach, we do not employ this negative caption set. In order to verify this, we re-train our model employing this negative caption set used in InfoGround [18]. Our proposed model with these negative captions results in 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 with negative captions achieves 59.47% and 79.34% in terms of $R@1$ and Pt_Acc, respectively. Both of them demonstrate that our approach achieves much higher performances than InfoGround when employing the same settings of negative captions.

5.3.4. ABLATION STUDY

In Table 5.2, we report the quantitative performance of 8 different design choices, i.e., c1-c8, within our proposed model on 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 in the model by Gupta et al. [18]. The introduction of VSA improves Pt_Acc from 62.43% to 64.26%, but results in a drop of $R@1$ from 32.12% to 29.64% (c1 vs. baseline). Our OPP module, as shown in Table 5.2, brings a performance gain of 3.24% in terms of $R@1$, but a 1.46% lower Pt_Acc (c2 vs. baseline).



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.	

Figure 5.3: Attention scores achieved in Eq. 5.9 of 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 by red, green and blue).

Table 5.2: Benefits of the different modules in our approach. All models are trained on the Flickr30K Entities training set and the results (%) are reported for the Flickr30K Entities validation set.

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 $\mathscr{L}_{\mathscr{S}}$	48.90	76.60
c5	-	✓	✓w/o $\mathscr{L}_{\mathscr{S}}$	52.71	78.31
c6	-	✓	✓	53.20	78.27
c7	✓	-	✓w/o $\mathscr{L}_{\mathscr{S}}$	55.64	77.58
c8	✓	✓	✓w/o $\mathscr{L}_{\mathscr{S}}$	57.90	77.24
VRC-PG	✓	✓	✓	58.64	77.03

When we use these two modules together, the $R@1$ 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 kind of modules can work in coordination with each other. We replace the InfoNCE loss in the baseline by our contrastive loss function (without $\mathscr{L}_S$), 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 $\mathscr{L}_S$, we can obtain a better result on $R@1$ and 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. 5.3, we visualize a few examples of 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 sentence than without VSA. For example, for the top-right example in the figure, we find that the setting with VSA gives a score (0.82) of attention 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 background.

5.3.5. Qualitative Results

In Fig. 5.4, we illustrate the qualitative results of visual grounding of phrases 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 phrases from the caption in the image. In Fig. 5.5, we show the attention scores obtained by Eq. 5.9 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. 5.5(a)). It is visible in the image that this region contains a head with white hair and exhibits a kind of visual appearance of ‘old’. Regions 29 and 3 are about the topic of scenes, and we can observe that the corresponding cells are high-

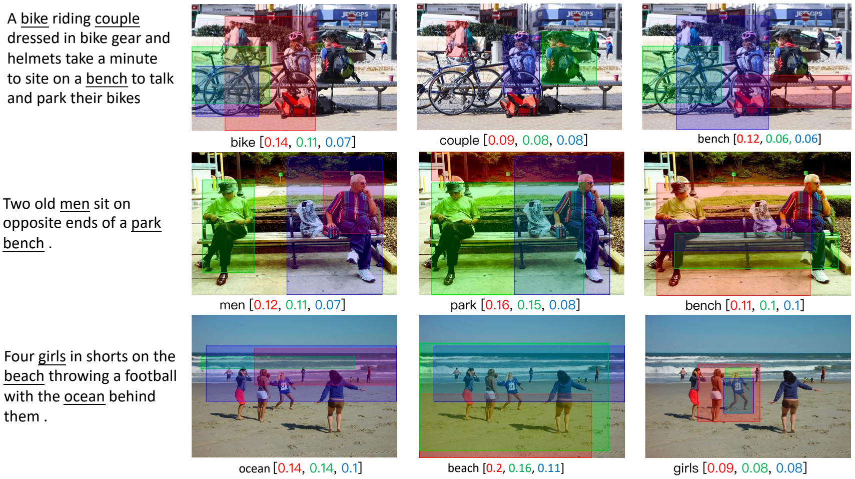
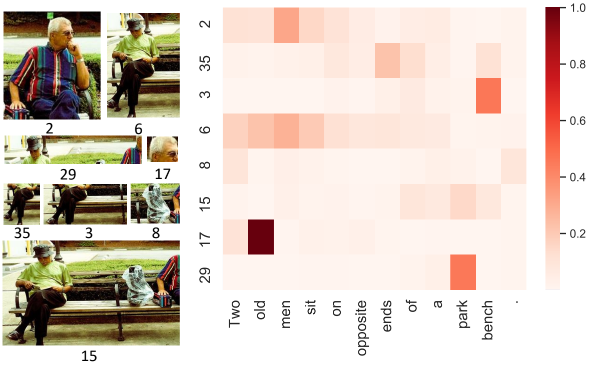

A bike riding couple dressed in bike gear and helmets take a minute to site on a bench to talk and park their bikes
bike [0.14, 0.11, 0.07]
couple [0.09, 0.08, 0.08]
bench [0.12, 0.06, 0.06]
Two old men sit on opposite ends of a park bench .
men [0.12, 0.11, 0.07]
park [0.16, 0.15, 0.08]
bench [0.11, 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]

Figure 5.4: 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 Eq. 5.9.



(a) Visual object proposals

(b) Attention weight map

Figure 5.5: Cross-modal attention scores achieved by Eq. 5.9 between visual object proposals and words. The darker cell color indicates that more attention is paid to the corresponding visual object proposals for a word query.

lighted in the attention weight map when the query of phrase is ‘park’ and ‘bench’. Regions 2 and 6 both relate to ‘men’, and they are really paid much attention to for the query of phrase ‘men’ as shown in the attention weight map.

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