Topic Scene Graph Generation by Attention Distillation from Caption

Wenbin Wang\textsuperscript{1,2}, Ruiping Wang\textsuperscript{1,2,3}, Xilin Chen\textsuperscript{1,2}

\textsuperscript{1}Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, 100049, China
\textsuperscript{3}Beijing Academy of Artificial Intelligence, Beijing, 100084, China

wenbin.wang@vipl.ict.ac.cn, \{wangruiping, xilchen\}@ict.ac.cn

Abstract

If an image tells a story, the image caption is the briefest narrator. Generally, a scene graph prefers to be an omniscient “generalist”, while the image caption is more willing to be a “specialist”, which outlines the gist. Lots of previous studies have found that a scene graph is not as practical as expected unless it can reduce the trivial contents and noises. In this respect, the image caption is a good tutor. To this end, we let the scene graph borrow the ability from the image caption so that it can be a specialist on the basis of remaining all-around, resulting in the so-called \textbf{Topic Scene Graph}. What an image caption pays attention to is distilled and passed to the scene graph for estimating the importance of partial objects, relationships, and events. Specifically, during the caption generation, the attention about individual objects in each time step is collected, pooled, and assembled to obtain the attention about relationships, which serves as weak supervision for regularizing the estimated importance scores of relationships. In addition, as this attention distillation process provides an opportunity for combining the generation of image caption and scene graph together, we further transform the scene graph into linguistic form with rich and free-form expressions by sharing a single generation model with image caption. Experiments show that attention distillation brings significant improvements in mining important relationships without strong supervision, and the topic scene graph shows great potential in subsequent applications.

1. Introduction

A picture is worth a thousand words. However, only a few person prefers to know all of the “thousand words”, while others would like to be informed the “topic words”. Therefore, the scene graph and the image caption are used for conveying the image contents out of different purposes.

Concretely, the scene graph \cite{20} consists of objects in an image and the relationships between pairs of objects. A series of studies have tried to generate the scene graph and realize its potential in advanced intelligence tasks, e.g., visual Q&A \cite{2, 47}, visual reasoning \cite{44}, and vision-and-language navigation (VLN) \cite{53}, etc. Nevertheless, as pointed in \cite{25, 35, 52}, the scene graph is helpful only if it is informative, while the current generated scene graph with such a lot of noises does not meet this standard. This is mainly due to the explosive combination possibility of two objects \cite{52, 65}, which brings the double-edge effect that the scene graph is comprehensive but the key information is overwhelmed by massive trivial details. It is necessary and practical to make the scene graph well-circumscribed between important and trivial contents. Fortunately, the image caption exactly shows this ability and is a good teacher from which a scene graph should learn.

In the context of scene graph generation, few researches devote endeavor to discovering the important relationships, which is a meaningful step for restricting the scale of the
scene graph when it is used for downstream tasks. The most popular approach is to keep the relationships with large products of the predicted subject, object, and predicate scores. However, this product measures the accuracy of prediction rather than the importance. Yang et al. [58] and Lv et al. [33] either use a light-weight relationship proposal network to extract some probably related pairs, or predict an attention score for each relationship, based on the perspective that annotated relationships are the important ones. This may be questionable because the mainstream scene graph datasets (e.g., Visual Genome [22]) are suffered from serious long-tailed problem [6, 46] and the annotated pairs (head pairs) are usually trivial ones [52]. To more precisely define what is important relationship, image caption is found helpful because a caption almost exactly reveals what humans think important [14], as shown in Figure 1. Consequently, Yu et al. [65] and Wang et al. [52] learn to mine the important relationships under the strong supervision from the important relationship annotation which is obtained under the guidance of captions. But this process need to transform the captions into triplets first and then align two groups of heterologous triplets, which is so expensive and complicates the scene graph generation.

In this work, we propose to let the scene graph learn the important relationships from the image caption in an economical way, resulting in the Topical Scene Graph. The importance of relationship is estimated by distilling the visual attention during the image caption generation, which is treated as the weak supervision. Specifically, most advanced image captioners are able to fix its gaze on the correct object regions. We apply an image captioner and collect the first-order attention information with respect to the object regions, which are used for assembling the second-order attention about the relationships. In this way, we actually transform humans’ attention into a new form, converting it from the concern about individual to that about relational events. The second-order attention is used as the weak supervision for guiding the estimation of the importance of the relationship. In this way, strong supervision is no longer necessary.

Furthermore, as the attention distillation process makes it possible to generate the image caption and scene graph simultaneously and both of them are the description of an image, why not generating them with a single model? It is noted that the most popular scene graph dataset, Visual Genome (VG) [22] of world scale contains more than 40,000 types of relationships which are originally extracted from humans’ language, while the traditional definition of scene graph treats the relationship recognition as predicate classification and makes most of the relationships filtered, which is harmful to the diversity of relationship description. What is worse, there exist huge interior differences in some certain predicate classes, e.g., the appearance of two relationship triplet instances for the predicate class riding, person-riding-horse and dog-riding-skateboard are totally different. It is difficult to clearly define the semantic boundaries between different predicates. Inspired by [21], we redefine the scene graph as the set of short relational sentences. In this way, a shared captioning module can be used for the so-called linguistic scene graph generation and image captioning at the same time.

2. Related Work

Scene Graph Generation (SGG) and Visual Relationship Detection (VRD) focus on understanding the relationships between objects. Early studies [10, 42] treat each distinct combination of object categories and relationship predicates as a distinct class. Lu et al. [32] formally define the VRD task and address the object and predicate classification separately. Recent state-of-the-art VRD works [8, 18, 26, 34, 38, 64, 66, 71, 72, 73] pay attention on the prediction of each relationship triplet independently. The scene graph which describes the image faithfully from a bird’s-eye view is proposed in [20]. After that, a batteries of studies contribute to generation of a high-quality scene graph. Message passing mechanism [55] has been proved effective and its variants are widely adopted in [27, 28].

The latest essential practice achieves more promising results through constructing reasonable context among objects and visual relationships [31, 39, 47, 51, 58, 70], or taking the advantage of external knowledge and commonsense [6, 13, 67, 69]. Besides, Zhang et al. [74] propose contrastive losses to resolve the related pair configuration ambiguity. Zareian et al. [68] creatively consider SGG as an edge role assignment problem. Tang et al. [46] diversify the predicted relationships through addressing the causal effect. Most of these works struggle to fit the VG dataset but always overlook the fact that the scene graph annotation suffers from serious long-tailed problem and the valuable relationships are usually overwhelmed by trivial ones. This problem is general because of the reporting bias [11] and should not be blamed on a particular dataset. A growing number of works are considering how to make the scene graph more practical. Liang et al. [29] prune the dominant and easy-to-predict relations while keeping the visually relevant relations in VG. Lv et al. [33] estimate the importance of relationships with an attention module, but actually they still think that the annotated relationships are semantically important, which may not be true. Yu et al. [65] and Wang et al. [52] provide annotations with relationships of humans interest under guidance of image caption and explicitly use them as supervision. However, only semantically important relationships are detected in [65], which is not enough for a comprehensive scene graph. In this work, we distill the attention from image caption as weak supervision rather than constructing the high-cost annotation of important re-
relationships, and reasonably estimate the importance of all relationships so that the number of remaining relationships are controllable.

**Image captioning and scene graph.** Compared to the scene graph, the image caption is usually treated as the final presentation to humans (for interaction). Restricted by the length, an image caption usually contains the most important contents in an image [14], but misses details. Some researchers propose the dense captioning task [19] which generates diverse but aimless region-level descriptions passively. Our proposed topic scene graph is naturally a structured representation of an image, and is especially able to actively estimate the importance of the image contents.

Early studies on image captioning are rule or template-based [45, 61]. Modern captioning models have achieved great progress benefiting from encoder-decoder framework [49], attention technique [4, 7, 15, 17, 24, 37, 48, 56, 60] and RL-based training objective [41]. Our work distills the attention from image caption. Besides, as the scene graph contains much semantic information, lots of works have tried to incorporate it into captioning models [5, 12, 25, 35, 57, 59, 62, 63]. Inspired by this direction, we propose to generate linguistic scene graph, and creatively let the scene graph benefit from image caption.

3. Approach

3.1. Overview

Given an image I, its scene graph G consists of a set of objects (nodes) O = \{o_i | i ∈ [1, n]\} with the assigned class labels C = \{c_i | i ∈ [1, n], c_i ∈ C\} and the corresponding bounding boxes B = \{b_i | i ∈ [1, n], b_i ∈ R^4\}, and a set of relationships (edges) R = \{r_k | k ∈ [1, m]\}. Conventionally, each relationship r_k is a triplet of the start node o_i, the end node o_j, and the relationship label x_{ij} ∈ R where R is the set of predicate types. These relationships are disordered. Thinking of the limitation of this representation, we redefine the relationship r_k as a relational caption in the form of word sequence \( (y_{kt})_{t=1}^{Y_k} \), where \( y_{kt} \in Y \) and \( Y \) denotes the vocabulary. \( t \) is the positional index of the word in the sequence and \( T_R \) is the sequence length. More importantly, these relationships are sorted according to their importance. Specifically, as depicted in Figure 1 (b), the detected objects are used for generating the relational captions and the image caption \((w_t)_{t=1}^{Y} \) (\( w_t ∈ Y \) and \( T_C \) is the caption length), during which the subjective interest (attention) \( α \) is collected from the image caption. The \( α \) is used for the estimation of importance scores of the relationships.

In the following parts, we will describe the generation of the image caption (Sec. 3.2) and the linguistic scene graph (Sec. 3.3) using a shared captioning module. Then we will elaborate on importance score estimation and attention distillation for obtaining the final topic scene graph (Sec. 3.4).

3.2. Captioning Module

In this work, we adopt two types of state-of-the-art captioning models, the Up-Down model [1] based on LSTM [16], and the Transformer [48]. The Up-Down model comprises of an attention LSTM layer and a language LSTM layer. Specifically, the detected objects are represented by their visual features \( V = [v_1, \ldots, v_n] ∈ R^{d_v × n} \) and bounding boxes \([b_1, \ldots, b_n] ∈ R^{4 × n}\). The object visual features are firstly transformed to \( V' = [v'_1, \ldots, v'_n] ∈ R^{d'v × n} \) with a lower dimension:

$$ v'_i = W_v v_i + b_v, \quad v'_i ∈ R^{d'_v}, \quad (1) $$

where \( W_v ∈ R^{d'_v × d_v} \) and \( b_v ∈ R^{d'_v} \) are trainable parameters. At each time step \( t \), the previous hidden state of the language LSTM \( h_{t-1}^l \) is concatenated with the mean-pooled image feature \( \bar{v} = \frac{1}{n} \sum_{i} v'_i \) and the previous word embedding \( e_{t-1} = W_e w_{t-1} \), and fed into the attention LSTM:

$$ h_t^l = LSTM_{Att}([h_{t-1}^2; \bar{v}; e_{t-1}], h_{t-1}^l), \quad h_t^l ∈ R^{d_h}, \quad (2) $$

where \([:\] denotes the concatenation and \( W_e ∈ R^{d_e × |V|} \) is the embedding matrix. The \( w_{t-1} \) stands for the \(|V|\) dim one-hot vector where the \( w_{t-1} \)-th element is 1 in practice. The attention about the objects \( α = [α_1, …, α_T_C] ∈ R^{n × T_C} \) is calculated as:

$$ z_{i,t} = w_a^T \tanh(W_{va} v'_i + W_{ha} h_t^l), \quad α_t = \text{softmax}(z_{i,t}), \quad (3) $$

where \( w_a ∈ R^{d_a × 1} \), \( W_{va} ∈ R^{d'_v × d_a} \), and \( W_{ha} ∈ R^{d_h × d_a} \) are trainable parameters. Finally, the attended image visual feature \( \bar{v}_t = \sum_{i=1}^{n} z_{i,t} v'_i \) and \( h_t^l \) are used as the input of the language LSTM, which predicts the conditional distribution over the possible word:

$$ h_t^2 = LSTM_{Lang}([\bar{v}_t; h_t^l], h_{t-1}^2), \quad h_t^2 ∈ R^{d_h}, \quad \quad p(w_t | w_{1:t-1}) = \text{softmax}(W_w h_t^2 + b_w), \quad (4) $$

with trainable parameters \( W_w ∈ R^{d_w × d'_h} \) and \( b_w ∈ R^{d'_w} \).

As for the Transformer model, it consists of an encoder and a decoder, both of which contain a stack of layers. We provide details in the Supp. and especially explain how to extract the attention on objects here. For captioning task, the transformed visual features \( V' \) are fed into the encoder and we get the output \( V^* ∈ R^{d_{tr} × n} \). For each decoder layer in the decoder, it contains a multi-head self-attention layer and a multi-head cross-attention layer. All the word embeddings \( E \) are fed into the self-attention layer to get the output \( E^* ∈ R^{d_{tr} × T_C} \). In each head \( j ∈ [1, H] \), the attention weights \( α_j ∈ R^{n × T_C} \) about objects are computed by:

$$ α_j = \text{softmax} \left( \frac{V'^T E^*}{\sqrt{d_{tr}}} \right). \quad (5) $$

We average the \( α_j \) across \( H \) heads and obtain the final \( α \).
3.3. Linguistic Scene Graph

In this part, we share the captioning module to make it applicable to relational captioning so that a linguistic scene graph is realizable. Following the general scene graph generation process, we build combinations of $N$ detected objects and obtain $O(N^2)$ object pairs. For the subject $o_i$ and the object $o_j$, we extract their union visual feature $v_{ij} \in \mathbb{R}^{d_u}$ which contains rich context information by applying ROI pooling [40] with the union box of $b_i$ and $b_j$. Besides, as the relative position between two objects is found as effective prior information, we follow [38] to build the geometry feature:

$$g_{ij} = \left[\frac{x_j-x_i}{\sqrt{w_i h_i}}, \frac{y_j-y_i}{\sqrt{w_i h_i}}, \sqrt{w_j h_j}, w_i, w_j, h_i, h_j, b_i \cap b_j\right],$$

where the $(x, y)$ is the center position and $w$ and $h$ denote width and height of a box. It is further projected to a 64-dim feature and concatenated with $v_{ij}$ to obtain the final union feature $v'_{ij} \in \mathbb{R}^{d_l}$:

$$v'_{ij} = W_u[v_{ij}; (W_g g_{ij} + b_g)] + b_u,$$

where $W_u \in \mathbb{R}^{d_l \times (d_u+64)}$, $b_u \in \mathbb{R}^{d_l}$, $W_g \in \mathbb{R}^{64 \times 6}$, and $b_g \in \mathbb{R}^{64}$ are trainable parameters.

Different from the image captioning which should pay attention to all objects, relational captioning focuses on two designated objects. Specifically, for the Up-Down model, only the $v'_i$, $v'_j$ and $v'_{ij}$ are used for decoding. For the Transformer, all object features are fed into the encoder to construct contextual information $V^*$, but only the $v'_i$, $v'^*_j$ and $v'_{ij}$ are fed into the decoder.

3.4. Topic Scene Graph

With the attention about objects provided by image captioning, we propose to assemble it to obtain attention about the relationships, which is used as the weak supervision to guide the estimation of the importance of the relationships.

Suppose that there are $m$ relationships. We first estimate the importance score for each relationship consisting of subject $o_i$ and object $o_j$. Specifically, as depicted in the top middle part in Figure 2, we concatenate the $v'_i$, $v'_j$, $v'_{ij}$ and the semantic embeddings of the subject and object categories, $e_i, e_j \in \mathbb{R}^{d_r+m}$, to form a query $q_{ij}$, and compute the key using the global feature $\Psi$:

$$q_{ij} = f([v'_i; v'_j; v'_{ij}; e_i; e_j]) \in \mathbb{R}^{d_s},$$

$$k = g(\Psi) \in \mathbb{R}^{d_s},$$

where $f$ and $g$ are two learnable linear transformation functions. The estimated importance score $s_{ij}$ is calculated as the inner product of the query and key and then normalized with softmax function:

$$s_{ij} = \frac{d_{ij}^T k}{\sqrt{d_s}},$$

$$\eta = \text{softmax}(s) \in \mathbb{R}^m.$$

On the other hand, we have the attention information $\alpha$ with respect to individual objects, which is used to assemble the attention $\beta$ with respect to relationships. As shown in the bottom left part (with gray background) in Figure 2, firstly, we gather the attention score for each object over multiple time steps with a pooling function $P$, resulting in
\( \gamma \in \mathbb{R}^n \). Then the second order attention for a relationship is assembled as:

\[
\delta_{ij} = \gamma_i + \gamma_j, \quad \beta = \text{softmax}(\delta) \in \mathbb{R}^m. \tag{12}
\]

Finally, the estimated score \( \eta \) is regularized with the induced second-order attention \( \beta \) via KL-divergence.

### 3.5. Optimization

The optimization process is divided into two stages. In the first stage, given a single ground truth image caption \((w_t)_{t=1}^{T_c}\) and \(m\) ground truth relational captions \([[(y_{kt})_{t=1}^{T_R}]_{k=1}^{m}]\), the captioning module is optimized with the traditional cross-entropy loss consisting of the image captioning part and the relational captioning part:

\[
L_{CE} = \sum_{t=1}^{T_c} - \log p(w_t|w_{1:t-1}) + \lambda \left( \frac{1}{m} \sum_{k=1}^{m} \sum_{t=1}^{T_R} - \log p \left( y_{kt}|y_{k1:k(t-1)} \right) \right), \tag{13}
\]

where \( \lambda \) is the balance parameter. In the second stage, the attention distillation module is optimized with the KL-divergence loss:

\[
L_{KL} = K L(\eta || \beta). \tag{14}
\]

Although the reinforce algorithm such as SCST [41] is widely used for further optimizing the captioning models, some researches [75] found that SCST actually does harm to the text-to-image grounding because it encourages the n-gram consistency rather than visual semantic alignment. As our framework has a high demand for superior grounding performance, we do not optimize the captioning module with SCST in this research and leave it to future works.

### 4. Experiments

#### 4.1. Datasets and Evaluation Metrics

**Datasets.** There is no existing dataset with both the image caption and relational caption. Inspired by [21], we refer to their data construction procedure to collect the relational caption, and further collect the image caption, as well as the important relationship annotation which is only for training the upper bound models and evaluation. Specifically, we use the 51,208 images in both VG and MSCOCO [30] datasets which have both relationship and image caption annotation. Firstly, we cleanse the VG dataset and keep a large scale vocabulary including 3,000 object categories and 800 attributes. The relationships about the objects beyond these categories are filtered. To obtain the important relationship annotation, we apply the Scene Graph Parser [43] to extract the relationships from the image caption, and align them with the annotated ones by matching their subject and object WordNet [36] synsets. Finally we convert the remaining relationships into sentences. It is worth mentioning that we do not make a category-wise vocabulary for relationships, but keep the relationships in their original free and open form. To further enrich the concepts, the attributes of each object are randomly selected to add to the sentences. We obtain 35,928 images for 29,928/1,000/5,000 splits for train/validation/test sets respectively and 11,437 vocabularies (including 3,000 object categories and 800 attributes).

**Evaluation metrics.** We use BLEU, METEOR, CIDEr-D, ROUGE-L, and SPICE for image captioning. For relational captioning, we refer to [19] and [21] and use the following metrics. (1) mean Average Precision (mAP): it uses METEOR score [9] with thresholds \{0, 0.05, 0.10, 0.15, 0.20, 0.25\} for language and IoU thresholds \{0.2, 0.3, 0.4, 0.5, 0.6\} for localization. Only the pair whose subject and object have IoUs greater than thresholds is a true positive sample. The mAP is obtained by averaging across all the combinations of language and localization thresholds. (2) image-level recall (Img-Lv.Recall): it ignores the localization and evaluate the recall of the bag of predicted relational captions.

Besides, in order to evaluate whether the important relationships are properly found, we refer to the metrics in traditional scene graph generation [52, 55], i.e., Recall@K where K is set to 20, 50, and 100. Under this metric, only the important relationships are regarded as ground truth and the top K relationships are evaluated, which means that the predicted relationships should be sorted. A relational caption is correct only if the following two conditions are satisfied: (1) both the subject and object have IoUs greater than 0.5, and (2) the METEOR score is greater than the thresholds above. We average the recall on different language thresholds. To evaluate the performance on discovering correct important object pairs, we derive the Recall@K metric which only requires the above first condition and does not consider the METEOR scores.

#### 4.2. Implementation Details

We firstly train the faster-RCNN [40] detector with the ResNeXt-101 [54] backbone on the objects of 3,000 categories of our dataset. During the scene graph training, the parameters of the object detector are frozen. More details are given in the Supp. Source codes are available at https://vipl.ict.ac.cn/resources/codes or http://www.kennethwong.tech/.

#### 4.3. Experiments on Linguistic Scene Graph

In our method, a shared captioning module is trained for image captioning (IC) and relational captioning (RC), which has never been explored before. We start with exploring the effectiveness of this practice. To this end, we
Table 1. Image captioning results. B1, B4, M, R, C, S denote BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D, and SPICE respectively. “-ICRC” denotes that the model is trained with image captions and relational captions.

<table>
<thead>
<tr>
<th>Model</th>
<th>λ</th>
<th>B1</th>
<th>B4</th>
<th>M</th>
<th>R</th>
<th>C</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>UD [21]</td>
<td>0.0</td>
<td>69.8</td>
<td>30.6</td>
<td>25.0</td>
<td>52.3</td>
<td>94.1</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>71.1</td>
<td>30.4</td>
<td>25.1</td>
<td>52.6</td>
<td>95.3</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>70.7</td>
<td>30.0</td>
<td>24.9</td>
<td>52.5</td>
<td>94.6</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>70.5</td>
<td>30.1</td>
<td>25.0</td>
<td>52.4</td>
<td>94.8</td>
<td>18.2</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>71.0</td>
<td>30.0</td>
<td>24.8</td>
<td>52.5</td>
<td>93.5</td>
<td>17.9</td>
</tr>
<tr>
<td>Transformer</td>
<td>0.7</td>
<td>68.8</td>
<td>26.8</td>
<td>23.5</td>
<td>50.4</td>
<td>85.6</td>
<td>17.3</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>0.7</td>
<td>70.3</td>
<td>28.6</td>
<td>24.4</td>
<td>51.7</td>
<td>91.5</td>
<td>18.0</td>
</tr>
</tbody>
</table>

adjust the balance parameter λ to control the proportion of the loss of the relational captioning in the final loss function. Results under more value settings are provided in Supp. The evaluation is divided into two parts. On the side of image captioning, the baselines are Up-Down (UD) and Transformer, which are trained only using the image captions. We compare the baselines to the UD-ICRC and Transformer-ICRC trained with the image captions and relational captions. From Table 1, we observe that mixed training actually brings benefit to the image captioning, but as the λ increases, this benefit will slightly drop. It suggests that mixed training is feasible despite the fact that the assembled relational captions will bring some noises.

On the side of relational captioning, we use TriLSTM [21], UD-RC, and Transformer-RC as baselines, which are trained only using the relational captions. The TriLSTM is re-implemented and trained on our dataset. The relational captions are sorted by the product of the probabilities of the generated words, i.e., likelihoods. As shown in Table 2, compared to the TriLSTM, both the UD-RC and UD-ICRC outperform it obviously. Comparing the UD-RC with UD-ICRC, we find that as the λ increases, the UD-ICRC roughly performs better on the image level metrics, and surpasses the UD-RC baseline when λ is greater than 0.7. However, the performance drops on the important relationship recall metrics. We think that it is because the increasing λ makes the model fit the relational captions data better, but the increasing sentence likelihood loses its discrimination and is less suitable for importance estimation. It also suggests that neither the sentence likelihood of the relational caption nor the score product of the traditional triplet are unstable for importance estimation. As for the Transformer, mixed training has little impact on the performances. With a comprehensive consideration on the performances of the two tasks, we set the λ as 0.7 and freeze the UD-ICRC / Transformer-ICRC models for generating the topic scene graph in the following experiments.

4.4. Experiments on Topic Scene Graph

As we know this is the first time to study the topic linguistic scene graph generation. We replace some key components to show the effectiveness of our proposed model (UD-ICRC-attn) and facilitate the ablation study. As we have the important relationship annotation, we train the upper bound models named as UD-ICRC-label and Transformer-ICRC-label under the supervision of the annotated important relationships with binary cross entropy loss. The results are shown in Table 3.

### Pooling function

The pooling function P is used for gathering attention information over multiple time steps for each individual object. We compare two functions: max pooling (MAX) and mean pooling (MEAN). Comparing the 4th row and the 5th row in the UD-ICRC-attn section, the max pooling function is much more effective than mean pooling. It is reasonable because we want to maximize the scores of the objects which are mentioned in the image captions, while the mean pooling reduces the attention scores and makes it hard to shed light on the key objects.

### Input features

We try to use different concatenated features for obtaining the query q when estimating the importance scores, including the union features (U), subject and object features (SO), subject, object, and union features (SOU), and subject, object, union features together with the semantic embeddings of subject and object categories (SOUS). By comparing the 1st~3rd rows and the 6th row in the UD-ICRC-attn section, and the rows in the UD-ICRC-label section, it is found that the SOU significantly improves the performances compared with U and SO, suggesting that these three types of features cannot be used independently, as the SO provides information about objects and the U provides relative spatial information. The semantic embeddings bring slight improvement, and it is not as obvious as that in the upper bound models.

### Masking non-noun words

When gathering the attention information, we explore whether all the words of a sentence should be considered or not. Different from considering all words, we try an alternative way that only the attention of noun words are collected since they are probably to be correctly grounded to the regions, and other words are masked. To this end, we apply the NLTK POS tagger [3] to filter out the non-noun words. Comparing the 5th row and 6th row in the UD-ICRC-attn section, it is interesting to find that masking the non-noun words does harm to the performance instead. This phenomenon may imply that the context plays a crucial role and the non-noun words would also contribute to the attention of the center nouns.

Overall, compared with the TriLSTM, UD-ICRC and Transformer-ICRC baselines, the application of our attention alignment module significantly improves the performances, and obviously reduces the gap between the baselines and the upper bound. The best configuration is to use the SOUS input features, max pooling function and collect the attention from all words. It’s noted that our method does not need the complicated collection of important relationship annotation, but can still provide the useful important
Table 2. Results of relational captioning (\%). “-RC” denotes that the model is only trained with the relational captions. “-ICRC” denotes that the model is trained with image captions and relational captions. R-ns means Recall-ns. Img-Lv. Recall means the image level recall.

<table>
<thead>
<tr>
<th>Model</th>
<th>λ</th>
<th>mAP</th>
<th>METEOR</th>
<th>Img-Lv. Recall</th>
<th>R@20</th>
<th>R-ns@20</th>
<th>R@50</th>
<th>R-ns@50</th>
<th>R@100</th>
<th>R-ns@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>UD-RC [1]</td>
<td>-</td>
<td>5.61</td>
<td>42.40</td>
<td>88.77</td>
<td>3.02</td>
<td>3.71</td>
<td>10.46</td>
<td>12.92</td>
<td>22.97</td>
<td>28.90</td>
</tr>
<tr>
<td>UD-ICRC</td>
<td>0.1</td>
<td>4.84</td>
<td>38.31</td>
<td>84.81</td>
<td>3.45</td>
<td>4.43</td>
<td>10.22</td>
<td>13.99</td>
<td>20.77</td>
<td>29.00</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>5.14</td>
<td>40.36</td>
<td>86.93</td>
<td>3.39</td>
<td>4.18</td>
<td>9.87</td>
<td>12.88</td>
<td>21.57</td>
<td>27.99</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>5.43</td>
<td>42.26</td>
<td>89.15</td>
<td>2.75</td>
<td>3.49</td>
<td>9.97</td>
<td>12.40</td>
<td>20.76</td>
<td>26.46</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>5.41</td>
<td>42.75</td>
<td>89.52</td>
<td>2.31</td>
<td>2.90</td>
<td>8.09</td>
<td>10.20</td>
<td>19.97</td>
<td>25.56</td>
</tr>
<tr>
<td>Transformer-RC [48]</td>
<td>-</td>
<td>5.26</td>
<td>41.62</td>
<td>88.65</td>
<td>2.11</td>
<td>2.73</td>
<td>6.83</td>
<td>9.12</td>
<td>16.36</td>
<td>21.91</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>-</td>
<td>5.15</td>
<td>41.63</td>
<td>88.64</td>
<td>2.05</td>
<td>2.70</td>
<td>6.86</td>
<td>9.19</td>
<td>16.21</td>
<td>21.91</td>
</tr>
</tbody>
</table>

Table 3. Results (%) comparison on discovering the important relationships. “Feat.” denotes different input features. “P” denotes the pooling function. “Mask” denotes masking the non-noun words (✓) or not (✗).

<table>
<thead>
<tr>
<th>Feat.</th>
<th>P</th>
<th>Mask</th>
<th>R@20</th>
<th>R-ns@20</th>
<th>R@50</th>
<th>R-ns@50</th>
<th>R@100</th>
<th>R-ns@100</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>TriLSTM</td>
<td>-</td>
<td>-</td>
<td>1.31</td>
<td>3.20</td>
<td>3.93</td>
<td>9.58</td>
<td>8.42</td>
<td>20.88</td>
<td>7.89</td>
</tr>
<tr>
<td>UD-ICRC</td>
<td>-</td>
<td>-</td>
<td>2.75</td>
<td>3.49</td>
<td>9.97</td>
<td>12.40</td>
<td>20.76</td>
<td>26.46</td>
<td>12.64</td>
</tr>
<tr>
<td>UD-ICRC-attn</td>
<td>U</td>
<td>MAX</td>
<td>7.27</td>
<td>10.53</td>
<td>17.12</td>
<td>24.10</td>
<td>30.44</td>
<td>42.22</td>
<td>21.95</td>
</tr>
<tr>
<td></td>
<td>SO</td>
<td>MAX</td>
<td>7.49</td>
<td>10.88</td>
<td>20.61</td>
<td>28.79</td>
<td>37.06</td>
<td>51.07</td>
<td>25.98</td>
</tr>
<tr>
<td></td>
<td>SOU</td>
<td>MAX</td>
<td>15.71</td>
<td>21.80</td>
<td>28.85</td>
<td>39.39</td>
<td>41.09</td>
<td>55.73</td>
<td>33.76</td>
</tr>
<tr>
<td></td>
<td>SOU</td>
<td>MEAN</td>
<td>2.74</td>
<td>4.53</td>
<td>8.76</td>
<td>13.71</td>
<td>19.27</td>
<td>28.05</td>
<td>12.84</td>
</tr>
<tr>
<td></td>
<td>SOU</td>
<td>MAX</td>
<td>10.72</td>
<td>15.43</td>
<td>21.59</td>
<td>30.26</td>
<td>34.43</td>
<td>47.34</td>
<td>26.63</td>
</tr>
<tr>
<td></td>
<td>SOU</td>
<td>MAX</td>
<td>15.46</td>
<td>21.81</td>
<td>29.55</td>
<td>40.72</td>
<td>41.14</td>
<td>55.68</td>
<td>34.06</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>U</td>
<td>-</td>
<td>13.04</td>
<td>17.35</td>
<td>25.25</td>
<td>33.28</td>
<td>36.72</td>
<td>49.22</td>
<td>29.14</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>SO</td>
<td>-</td>
<td>30.14</td>
<td>38.86</td>
<td>41.45</td>
<td>53.95</td>
<td>51.55</td>
<td>67.70</td>
<td>47.28</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>SOU</td>
<td>-</td>
<td>32.17</td>
<td>41.38</td>
<td>43.57</td>
<td>56.68</td>
<td>53.65</td>
<td>70.81</td>
<td>49.71</td>
</tr>
<tr>
<td>Transformer-ICRC</td>
<td>SOUS</td>
<td>-</td>
<td>34.39</td>
<td>45.13</td>
<td>46.03</td>
<td>60.97</td>
<td>54.60</td>
<td>72.44</td>
<td>52.26</td>
</tr>
<tr>
<td>Transformer-ICRC-atttn</td>
<td>SOUS</td>
<td>MAX</td>
<td>17.52</td>
<td>24.96</td>
<td>31.88</td>
<td>44.46</td>
<td>43.71</td>
<td>61.10</td>
<td>37.27</td>
</tr>
<tr>
<td>Transformer-ICRC-label</td>
<td>SOUS</td>
<td>-</td>
<td>25.79</td>
<td>34.68</td>
<td>39.06</td>
<td>53.02</td>
<td>48.76</td>
<td>66.43</td>
<td>44.62</td>
</tr>
</tbody>
</table>

relationships. We observe that the attention alignment module is more effective for the Transformer, which may imply that the attention in Transformer is more precise.

4.5. Qualitative Results

In Figure 3 (b), we visualize the attention about the objects of each word (the purple heat map) during captioning and the pooled attention over all words (the reddish brown heat map). It can be observed that although the caption may not be so precise, the objects are still correctly attended (the first sample, the woman and bus in the first sample, the court, person, and the tennis racket in the second sample). The max pooling function highlights the mentioned objects in the caption. It’s also found that an object can be activated by several words, which gives an explanation for the performance drop when masking the non-noun words. In Figure 3 (c), we draw the scores of relationships for sorting. All the relationships are firstly sorted according to the assembling attention scores \( \beta \) induced from image caption, and then their \( \eta \) scores and the sentence likelihoods are drawn, which are used for sorting by UD-ICRC-attn and UD-ICRC respectively. The line charts show that the predicted \( \eta \) scores have a similar trend with the \( \beta \) and therefore it can correctly rank the relationships according to their importance. However, the sentence likelihoods do not show this trend, suggesting that these scores (including the product scores used in traditional scene graph) are irrelevant to the importance of the relationships. In Figure 3 (d-e), we also compare the topic linguistic scene graph and the traditional scene graph from the motifs [70]. The topic linguistic scene graph focuses on the relationships of humans interest which are more important in the images. Besides, the scene graph of linguistic style allows the relationships to be expressed in a natural way with more suitable words, despite that the given detected object categories may be not so appropriate in the language context, e.g., in the first example, the photograph is expressed as station in the relationships.

4.6. Topic Scene Graph for Retrieval

As the topic scene graph provides relationships relevant to the major events in an image, it can be utilized for image retrieval.
Table 4. The image retrieval results using top 1 relationships. We use the recall at K (R@K, higher is better) and the median rank of the target image (Med, lower is better).

<table>
<thead>
<tr>
<th>Model</th>
<th>R@1</th>
<th>R@5</th>
<th>R@10</th>
<th>Med</th>
</tr>
</thead>
<tbody>
<tr>
<td>TriLSTM</td>
<td>1.73</td>
<td>7.47</td>
<td>12.83</td>
<td>135.33</td>
</tr>
<tr>
<td>UD-ICRC</td>
<td>5.67</td>
<td>20.40</td>
<td>31.73</td>
<td>27.33</td>
</tr>
<tr>
<td>UD-ICRC-attn</td>
<td>9.73</td>
<td>31.67</td>
<td>46.13</td>
<td>12.33</td>
</tr>
<tr>
<td>UD-ICRC-label</td>
<td>17.77</td>
<td>49.17</td>
<td>67.37</td>
<td>5.67</td>
</tr>
</tbody>
</table>

Figure 3. The qualitative results. (a) The objects with their bounding boxes and ids. (b) The attention about the objects during caption generation and the pooled attention are visualized. Darker colors indicate larger weights. (c) Importance scores of the relationships are drawn. Along the X-axis, the relationships are sorted by the $\beta$ scores in a descending order. All the lines are smoothed. (d-e) The scene graph from our method and motifs [70] consisting of top 5 relationships are shown.

Figure 4. Two important relationships given by topic scene graph of the left image are used to retrieve similar images respectively. The results are shown on the right column.

5. Conclusion

In this work, we propose to generate the scene graph jointly with the image caption so that it can not only understand the image comprehensively, but also balance the important and trivial contents. The attention information from the image caption provides guideline to emphasize the important relationships. In addition, we generate the scene graph together with the image caption using a shared captioning module, making it express in a more natural style. Experimental results show the advantages of the proposed method in both performance and its feasibility in mining the important relationships without strong supervision. Besides, the topic scene graph has shown its practicality for controllable and fine-grained retrieval.

Acknowledgements. This work is partially supported by National Key R&D Program of China (2020AAA0105200), Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, and CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009.
References


