Exploring Semantic Relationships for Image Captioning without Parallel Data

Fenglin Liu[1], Meng Gao[2], Tianhao Zhang[3] and Yuexian Zou[1*],
*Corresponding Author
[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]School of ICE, Beijing University of Posts and Telecommunications, Beijing, China
[3]Beijing Foreign Studies University, Beijing, China
{fenglinliu98, zouyx}@pku.edu.cn, gaomeng@bupt.edu.cn, wszth666@bfsu.edu.cn

Abstract—Recently, image captioning has aroused great interest in both academic and industrial worlds. Most existing systems are built upon large-scale datasets consisting of image-sentence pairs, which, however, are time-consuming to construct. In addition, even for the most advanced image captioning systems, it is still difficult to realize deep image understanding. In this work, we achieve unpaired image captioning by bridging the vision and the language domains with high-level semantic information. The motivation stems from the fact that the semantic concepts with the same modality can be extracted from both images and descriptions. To further improve the quality of captions generated by the model, we propose the *Semantic Relationship Explorer*, which explores the relationships between semantic concepts for better understanding of the image. Extensive experiments on MSCOCO dataset show that we can generate desirable captions without paired datasets. Furthermore, the proposed approach boosts five strong baselines under the paired setting, where the most significant improvement in CIDEr score reaches 8%, demonstrating that it is effective and generalizes well to a wide range of models.

Index Terms—vision and language, unpaired training data, image captioning, deep neural network

I. INTRODUCTION

Image captioning has drawn remarkable attention in both natural language processing and computer vision. The task, which combines image understanding and language generation, is tough yet practical. Above all, it has various kinds of applications such as human-robot interaction [1], text-based image retrieval [2] and helping visually impaired people see [3], among others. The deep neural networks, particularly those based on the encoder-decoder framework [4]–[11], have achieved great success in advancing the state-of-the-art.

Despite the impressive achievements of the deep learning frameworks, they unduly rely on large-scale paired data, which is not easy to access in the real world. Especially, when it comes to the Non-English caption systems, image-caption pairs are labor-intensive to obtain. In recent years, unsupervised encoder-decoder models have been proposed for neural machine translation [12]–[14]. Typically, the source language and the target language are mapped into a common latent space, where sentences with the same semantic meaning are well aligned, thus the unsupervised translation can be carried out . For the task of image captioning, due to the great disparities between the vision and the language domains, and
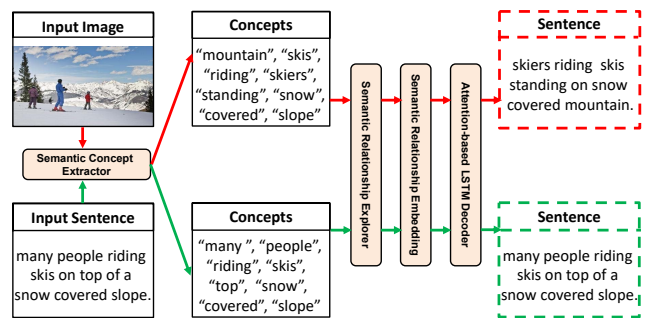

Figure 1: Illustration of our proposed framework. Our framework consists of a Semantic Concept Extractor, a Semantic Relationship Explorer, a *Semantic Relationship Embedding* and an Attention-based Sentence Decoder.

distinct features of the modalities, unsupervised learning is even more challenging. Although unpaired image-to-sentence translation has been explored in the literature, the training of neural models on unpaired datasets has not yet been conducted at a semantic level.

Recently, there have been a few works [15]–[17] trying to reduce the dependence of image captioning models on paired data. [15] proposed a method to generate captions in a central language (Chinese) and then translate them into a target language (English), without requirement for image-caption pairs for the target language. However, a paired image captioning corpus for the source language and a paired corpus for translation are still indispensable. [16] connected the visual and the textual modalities with visual objects (e.g., *dog*, *mirror*) and an adversarial manner. It achieved unsupervised image captioning without any image-caption pairs. Nevertheless, their method relied on the recognized visual objects to decide whether the generated captions are image-related, which necessitated complex models and schemes to obtain higher-quality image captions, e.g. image reconstruction [16], sentence reconstruction [18] and adversarial learning [19]. Moreover, the visual object based approach does not consider attributive words (e.g. attributes (*small*), relations (*standing*) and color (*white*)), which could help improve the image representation. [17] used the scene graph to bridge

the gap between the visual and the textual modalities. Since the scene graph constructed a series of semantic relationship information, the model achieved good results. However, in order to construct the semantic relationships, they needed to use Faster-RCNN [20] as the object detector, MOTIFS [21] as the relationship detector, and an additional classifier for attribute identification [11] for generating the image scene graph. In all, although existing methods for unsupervised image captioning have made appreciable progress, they are hard to implement and still far from real world applications.

In this paper, we bridge the gap between images and captions with semantic concepts, which explicitly represent high-level information [22]. To solve the challenges posed by different characteristics between visual and textual modalities, we propose an approach to explore the relationships between individual semantic concepts. The proposed framework (Figure 1) includes a *Semantic Concept Extractor*, a *Semantic Relationship Explorer*, a *Semantic Relationship Embedding* and an *Attention-based Sentence Decoder*. To begin with, our model extracts relevant semantic concepts from the images and the captions, respectively, which is the *Semantic Concept Extractor*. Then, the *Semantic Relationship Explorer*, which consists of an *"Attribute" Aggregator*, an *"Object" Aggregator* and a *"Relation" Aggregator*, explores effective semantic relationships among the extracted concepts from three aspects, i.e., "attribute", "object" and "relation". Meanwhile, the *Semantic Relationship Embedding* can extract principle features of the semantic relationships from different parts, which strengthens the expressive ability of relationship features. As a result, characteristic information of the semantic concepts is fully exploited and the representation of the image concept features is getting closer to the textual modality and, as input to the decoder, adopted effectively to generate meaningful sentences. Our main contributions are as follows:

- We propose a semantic based framework for unpaired image captioning. The framework can effectively and automatically use the semantic information to generate high quality image captions by bridging the gap between the vision and the language domains.
- To explore the relationships between individual semantic concepts, we propose a *Semantic Relationship Explorer*, which can extract both coarse-grained and fine-grained semantic relationships to help the model generate more accurate and meaningful captions.
- The proposed approach proves to be effective by the experiments on the MSCOCO image captioning dataset. We outperforms the latest unpaired image captioning systems. More encouragingly, our method also makes improvements of up to 8% in terms of CIDEr score for paired image captioning model, when equipping it with our approach.

II. RELATED WORK

A. Image Captioning.

Recently, many neural model systems have been proposed for image captioning [4]–[11]. The most advanced methods [8]–[11] rely on the encoder-decoder framework and combine the attention mechanism [4] to transform the image into coherent caption. However, the paired image-sentence data for training these models is expensive to collect. The first model under the unsupervised training setting, built by [15], aims to take the pivot language to connect source image and target caption. Although image and target language caption pairs are not used, their method relies on image-pivot pairs and a pivot-target parallel translation corpus. [16] and [17] align between the vision and the language domain in a adversarial way, so they don't need any paired image-sentence data. In contrast to the approaches mentioned above, our approach attempts to regard semantic concepts as a bridge and explore the semantic relationships among them, which has not been well studied yet.

B. Unsupervised Machine Translation.

Unsupervised image captioning is essentially similar to unsupervised machine translation. Nevertheless, due to the apparently different characteristics of the image and text modality, unpaired image-to-sentence translation is more challenging than unpaired sentence-to-sentence translation.

C. Exploring Semantic Relationships.

A most recent advancement [9], [11] attempts to use graph networks to explicitly explore visual relationships by encoding scene graphs, which model the spatial and semantic relationships of image regions based on visual features. However, they built an additional model to explicitly predict the relationships between visual objects with extra annotated data. Instead, our model associates the semantic concepts based on attention, and the relationships are implicitly modelled as weighted combinations and trained with the captioning model.

III. THE PROPOSED FRAMEWORK

As shown in Figure 2, the proposed framework consists of four main modules: (1) *Semantic Concept Extractor*: this module is able to extract the semantic concepts from images or sentences; (2) *Semantic Relationship Explorer* (Figure 3): since the extracted semantic concepts are independent and not associated with each other, e.g., the three words *riding*, *boy* and *bike*, this module can associate these words together as a phrase *boy riding bike* to represent a complete semantics; (3) *Semantic Relationship Embedding*: this module can strengthen the expressive ability of relationship features by extracting different features of the explored semantic relationships from different parts; (4) Through the above three steps, we are able to provide rich semantic relationship information which is beneficial for the fourth module, i.e., *Attention-based Sentence Decoder*, to generate the complete and coherent image captions under the unpaired setting. In the following parts of this section, we will describe these steps in detail.

A. Semantic Concept Extractor

In recent studies [6], [22]–[25], the semantic concepts are introduced to represent the explicit high-level information of the image [22]. In particular, it consists of a series of words,
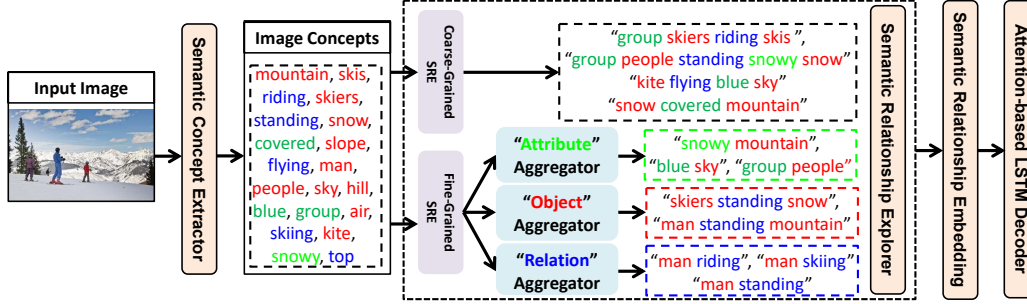
440

Figure 2: Illustration of the proposed framework. The semantic concepts are extracted by an extractor [23]. The color *Green* denotes the attribute words, the color *Red* denotes the object words and *Blue* color denotes the relation words. The *Semantic Relationship Explorer* consists of a *Coarse-Grained Semantic Relationship Explorer* and a *Fine-Grained Semantic Relationship Explorer*. The latter one includes three sub-modules: a *"Attribute" Aggregator*, an *"Object" Aggregator* and a *"Relation" Aggregator*, then adaptively explores the relationships among the extracted semantic concepts. The proposed approach could achieve deep semantic understanding of the image and in turn generate more complete and coherent captions.

including attributes (e.g., *young*, *black*), objects (e.g., *woman*, *shirt*) and relationships (e.g., *sitting*, *wearing*). In this case, these semantic concept words for an image can be generated by a *Semantic Concept Extractor* [6], [22]. In implementation, we adopt a weakly-supervised method of Multiple Instance Learning (MIL) [26] to build the *Semantic Concept Extractor*, following [23]. Technically, for each concept, MIL takes the "positive" and "negative" bags of bounding boxes as input sets, where every bag corresponds to an image. If the current concept is in the semantic label (e.g. description and scene graph) of the image, the bag is regarded to be positive, otherwise it is said to be negative. MIL iteratively selects the positive instances to train an extractor for the semantic concepts. Due to limited space, please refer to [23] for detailed explanation. Other semantic concept extracting approaches can also be used and may produce better results, which, however, are not the main focuses of this work.

To extract the semantic concepts of a sentence, we directly use the words, which appear in both the ground truth and the pre-defined semantic concepts vocab, as semantic concepts.

In addition, because semantic concepts are discrete word tokens, we adapt a word embedding to project them into vectors $C_g = \{c_1, c_2, \ldots, c_{N_g}\}, c_i \in \mathbb{R}^e$, where e is the dimension of word embedding. Then all words, words containing only "attribute", words containing only "object" and words containing only "relation" are respectively represented as global semantic concepts $C_g \in \mathbb{R}^{N_g \times e}$, attribute semantic concepts $C_a \in \mathbb{R}^{N_a \times e}$, object semantic concepts $C_o \in \mathbb{R}^{N_o \times e}$ and relation semantic concepts $C_r \in \mathbb{R}^{N_r \times e}$, where N_g, N_a, N_o and N_r represent the number of words per type.

B. Semantic Relationship Explorer

In this section, we will first introduce the method we use to extract the semantic relationships - MultiHead Attention, then we will introduce the proposed *Semantic Relationship Explorer*, Figure 3 shows a sketch.

1) MultiHead Attention: In order to extract the semantic relationships, we adopt multi-head attention [27], which can calculate the association weights between each pair of the given semantic concepts [28]. The following multi-head attention consists of N parallel heads and each head is represented as a scaled dot-product attention.

$$\text{Att}(Q,K,V) = \text{softmax}\left(\frac{QW^Q(KW^K)^T}{\sqrt{d_k}}\right)VW^V \quad (1)$$

where $Q \in \mathbb{R}^{l \times d_h}$, $K \in \mathbb{R}^{k \times d_h}$ and $V \in \mathbb{R}^{k \times d_h}$ represent respectively the query matrix, the key matrix and the value matrix. $W^Q, W^K, W^V \in \mathbb{R}^{d_h \times d_k}$ are learnable parameters and $d_k = d_h/n$, where n represents the number of heads.

Following the multi-head attention is a fully-connected feed-forward network, which is defined as follows:

$$\text{FFN}(x) = \text{ReLU}(xW_f + b_f)W_{ff} + b_{ff} \quad (2)$$

where W_f and W_{ff} denote matrices for linear transformation; b_f and b_{ff} represent the bias terms.

We use the attention-based method to make each semantic concept adaptively find the most relevant attributes, objects and relations, and establish related relationships, thus completing the exploration process of the semantic relationships.

2) Coarse-Grained Semantic Relationship Explorer: The *Coarse-Grained SRE* is a simple and intuitive way to explore the semantic relationships. The relationships are built via multi-head attention on the global semantic concepts C_g. Particularly, we use C_g as Query, Key and Value simultaneously. The result turns out to be a set of attended semantic concepts:

$$R_g = \text{FFN}(\text{MultiHead}(C_g, C_g, C_g)) \quad (3)$$

with *Coarse-Grained SRE*, all semantic concepts can attend to other semantic concepts related to them, making it easy to capture global relationships quickly. However, the semantic relationships extracted by *Coarse-Grained SRE* are coarse-grained, because for any kind of semantic concept as a Query, it is not necessary to treat all words as Key and Value, i.e., there is no need to attend to the most relevant among all words. For example, when an "attribute" word is used as
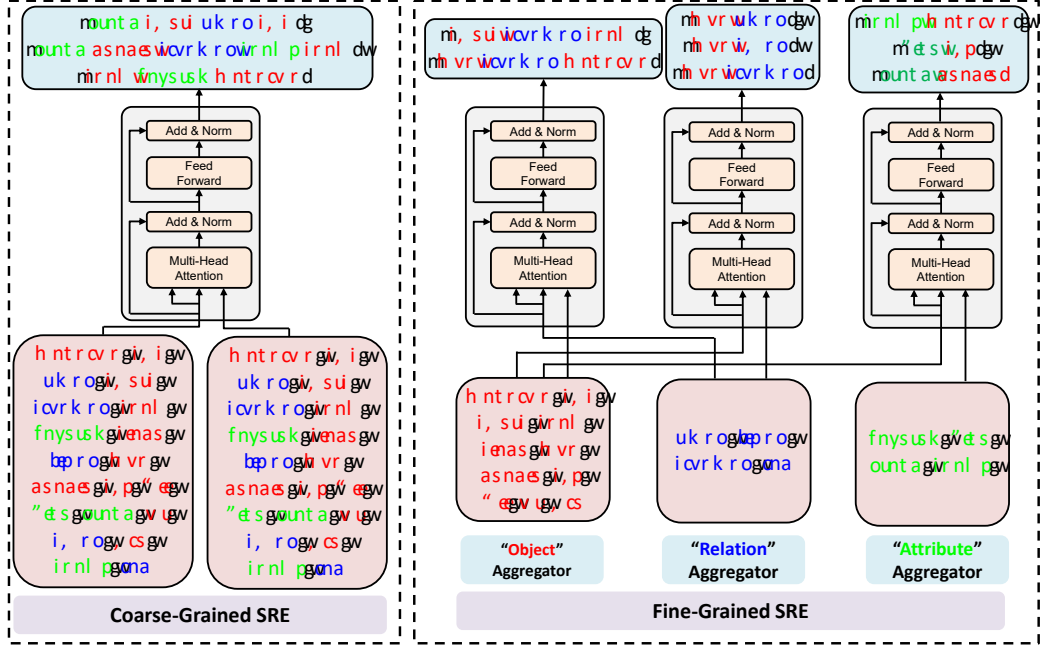
Figure 3: Illustration of the proposed Semantic Relationship Explorer. The color *Green* denotes the Attribute words, the color *Red* denotes the Object words and *Blue* color denotes the Relation words.

a Query, it does not need to pay attention to the "relation" words. This is because the "attribute" words are usually used to describe a specific object, while the "relation" words are usually used to describe the relationship among two or more objects, that is, the association between the "attribute" words and the "relation" words is not very strong. Therefore, the semantic relationships explored by *Coarse-Grained SRE* are not good due to the noise introduced by many unrelated words.

3) Fine-Grained Semantic Relationship Explorer: In order to solve the problems of the *Coarse-Grained SRE* and to extract more fine-grained and more precise semantic relationships, we propose the *Fine-Grained Semantic Relationship Explorer* (*Fine-Grained SRE*). Inspired by the proposed attribute embedding, object embedding and relationship embedding in [11], we propose *"Attribute" Aggregator*, *"Object" Aggregator* and *"Relation" Aggregator*, which use attribute words, object words and relation words, respectively, as Value and Key. In addition, we get finer and more accurate semantic relationships by using the words of the category most relevant to them as Query.

a) "Attribute" Aggregator: A specific object may have multiple attributes. Therefore, we have to focus on the attributes that are most relevant to it from a set of attribute semantic concepts. In this case, according to the attention theorem, the object semantic concepts C_o are the Queries, and the attribute semantic concepts C_a are the Keys and Values. Consequently, the result turns out to be a set of attended attribute semantic concepts:

$$R_a = \text{FFN}(\text{MultiHead}(C_o, C_a, C_a)) \quad (4)$$

Through this formula, we can get detailed object-attribute relationship information.

b) "Object" Aggregator: For a particular relation concept, we tend to focus on seeking for two objects that are related to each other to form the object-relationship-object, e.g., *women-wearing-shirt*. Therefore, for the *"Object" Aggregator*, the relation semantic concepts should focus on the object semantic concepts. The relation semantic concepts C_r serve as Queries, and the object semantic concepts C_o serve as Keys and Values, which can be defined as follows:

$$R_o = \text{FFN}(\text{MultiHead}(C_r, C_o, C_o)) \quad (5)$$

Thus, we can get detailed object-relation-object relationship information by the *"Object" Aggregator*.

c) "Relation" Aggregator: A specific object may have many relations with other objects, so we need to focus on the relation that is most relevant to it from a set of relation semantic concepts, e.g., *woman-eating, woman-holding, woman-wearing* and so on. Therefore, for *"Relation" Aggregator*, we consider what relationships they are attached to. Accordingly, the object semantic concepts C_o should pay attention to the relation semantic concepts C_r, which is defined as follows:

$$R_r = \text{FFN}(\text{MultiHead}(C_o, C_r, C_r)) \quad (6)$$

where we can get object-relation relationship information.

C. Semantic Relationship Embedding

For the semantic concepts of each image or sentence, through *Coarse-Grained SRE* and *Fine-Grained SRE*, we can get coarse-grained relationships $R_g \in \mathbb{R}^{N_g \times e}$ and fine-grained

relationships $\boldsymbol{R}_a \in \mathbb{R}^{N_a \times e}$, $\boldsymbol{R}_o \in \mathbb{R}^{N_o \times e}$ and $\boldsymbol{R}_r \in \mathbb{R}^{N_r \times e}$, respectively. These semantic relationships are of different quantity and contain different information. So in order to have the same number of features for each type of semantic relationships and to encourage the diversity of extracted main features for decoding effectively, we use a self-attention mechanism provided by [29]. This mechanism can be defined as follows:

$$A = \mathrm{softmax}(\boldsymbol{W}_{s_2}\tanh(\boldsymbol{W}_{s_1}\boldsymbol{R}_k{}^T)) \tag{7}$$

$$\boldsymbol{M}_k = A\boldsymbol{R}_k, k \in \{g, a, o, r\} \tag{8}$$

where $\boldsymbol{W}_{s_1} \in \mathbb{R}^{d_a \times e}$ and $\boldsymbol{W}_{s_2} \in \mathbb{R}^{r \times d_a}$ are learnable parameters; r stands for the number of extracted features in different parts. Through the formula, we can get different semantic relationship embeddings, i.e., $\boldsymbol{M}_k \in \mathbb{R}^{r \times e}(k \in \{g, a, o, r\})$, which share the same dimensions. For coarse-grained relationships embedding, $\boldsymbol{M}_c = \boldsymbol{M}_g$. We combine $\boldsymbol{M}_a$, $\boldsymbol{M}_o$ and $\boldsymbol{M}_r$ to obtain the fine-grained relationships embedding, $\boldsymbol{M}_f = [\boldsymbol{M}_a; \boldsymbol{M}_o; \boldsymbol{M}_r]$, where [;] stands for concatenation operation.

D. Attention-Based Decoder

The decoder part is essential in the way that it translates the extracted semantic relationships into their corresponding captions. The decoder is an attention-based LSTM [30], which takes the word embedding vector $\boldsymbol{w}_t^e$, concatenated with the averaged semantic concepts vector $\boldsymbol{c}_{avg} = \frac{1}{N_g}\sum_{i=1}^{N_g} \boldsymbol{c}_i$, as input, which is denoted as $\boldsymbol{x}_t = [\boldsymbol{w}_t^e; \boldsymbol{c}_{avg}]$.

At each time step, when the decoder generates captions, the semantic relationship information that needs attention varies with different time steps. Inspired by the attention mechanism in [30], we design an attention LSTM sentence decoder by injecting all of the coarse-grained or fine-grained semantic relationship embedding features ($\boldsymbol{M}_k, k \in \{c, f\}$) into a LSTM with attention mechanism. Specifically, at each time step t, according to the current hidden layer $\boldsymbol{h}_t \in \mathbb{R}^d$, the attention LSTM decoder firstly generates the attention distribution over the n (n is r for $\boldsymbol{M}_c$ and $3r$ for $\boldsymbol{M}_f$) parts of semantic relationship embedding features $\boldsymbol{M}_k, k \in \{c, f\}$:

$$\boldsymbol{\alpha}_t = \mathrm{softmax}(\boldsymbol{w}_\alpha\tanh(\boldsymbol{W}_M\boldsymbol{M}_k{}^T \oplus \boldsymbol{W}_h\boldsymbol{h}_t)) \tag{9}$$

where $\boldsymbol{W}_M \in \mathbb{R}^{d_a \times e}$, $\boldsymbol{W}_h \in \mathbb{R}^{d_a \times d}$ and $\boldsymbol{w}_\alpha \in \mathbb{R}^{d_a}$ are learnable parameters; $\oplus$ denotes the matrix-vector addition, which is calculated by adding the vector to each column of the matrix. $\boldsymbol{\alpha}_t \in \mathbb{R}^n$ is the attentive weight of $\boldsymbol{M}_k$ and the attentive relationship vector $\boldsymbol{r}_t \in \mathbb{R}^e$ can be defined as follows:

$$\boldsymbol{r}_t = \boldsymbol{\alpha}_t\boldsymbol{M}_k \tag{10}$$

With a further combination with $\boldsymbol{h}_t$, the attentive relationship can predict the current output word:

$$y_t \sim \boldsymbol{p}_t = \mathrm{softmax}(\boldsymbol{W}_{p,r}\boldsymbol{r}_t + \boldsymbol{W}_{p,h}\boldsymbol{h}_t) \tag{11}$$

where $\boldsymbol{W}_{p,r} \in \mathbb{R}^{|D| \times e}, \boldsymbol{W}_{p,h} \in \mathbb{R}^{|D| \times d}$, and each value of $\boldsymbol{p}_t \in \mathbb{R}^{|D|}$ denotes the probability suggesting how likely the corresponding word in vocabulary D is the current output word.

The approach mentioned above encourages the model to exploit all the available semantic relationship information. Thanks to the abundant and enriched information extracted by the *Semantic Relationship Explorer*, the source semantic concept information turns into a deep image understanding.

IV. IMPLEMENTATION DETAILS

In this section, we will introduce how to train our model and how to use our model to generate image captions under the unpaired settings.

A. Training Strategy

a) Training in Language Domain: Since there is no parallel data, we train the full model in the language domain, and use the method of reconstructing the original sentence to train the model, that is:

Sentence → Semantic Concepts → Semantic Relationships → Semantic Relationship Embedding → Decoder → Sentence

The *Semantic Relationship Explorer* explores the Semantic Relationships $\boldsymbol{R}^S$ from the Semantic Concepts $\boldsymbol{C}^S$, which are extracted from the sentence $\boldsymbol{S}$, and obtain Semantic Relationship Embeddings $\boldsymbol{E}^S$ through a self-attention mechanism. Finally, the original sentence $\boldsymbol{S}$ is reconstructed by an *Attention-based Sentence Decoder*.

b) Training Objectives: Firstly, we train the full model by minimizing the Cross-Entropy (XE) Loss [6], [30], [32]–[35]. Given a target ground truth sequence $y_{1:T}^*$ and a captioning model with parameters θ, the goal is to minimize the following cross entropy loss:

$$L_{CE}(\theta) = -\sum_{t=1}^{T}\log(p_\theta(y_t^*|y_{1:t-1}^*)) \tag{12}$$

Recently, reinforcement learning methods have also been widely used for captioning model training [8], [9], [11], [36]. Consequently, we further employ a reinforcement learning (RL) loss to improve the performance of our model. In this case, the cross-entropy loss method is used to pre-train the model, after which our goal is to minimize the negative expected score as:

$$L_{RL}(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})] \tag{13}$$

where r is the score function (e.g., CIDEr). Following the Self-Critical Sequence Training [36] (SCST), the gradient of $L_{RL}(\theta)$ can be approximated by

$$\nabla_\theta L_{RL}(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T}))\nabla_\theta log p_\theta(y_{1:T}^s) \tag{14}$$

where $r(y_{1:T}^s)$ is the score of a sampled caption $y_{1:T}^s$ and $r(\hat{y}_{1:T})$ suggests the baseline score of a caption which is generated by the current model using greedy decode. Through this gradient, sampled captions with higher CIDEr scores are more likely to be generated by the model because their corresponding probabilities are increased. Through this approach, we complete CIDEr optimization.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Language Pivoting [15]	46.2	24.0	11.2	5.4	13.2	-	17.7	-
Adversarial+Reconstruction [16]	58.9	40.3	27.0	18.6	17.9	43.1	54.9	11.1
Graph-Align† [17]	66.7	48.0	31.5	21.1	20.5	46.2	68.7	14.5
Graph-Align [17]	67.1	47.8	32.3	21.5	20.9	47.2	69.5	15.0
Coarse-Grained SRE	50.3	36.6	24.7	16.5	14.3	33.4	37.2	10.6
Fine-Grained SRE	57.5	38.9	27.1	19.7	17.4	41.9	49.7	13.3
Graph-Align [17] + Fine-Grained SRE	**67.8**	**48.7**	**33.6**	**21.8**	**22.1**	**48.4**	**75.7**	**16.1**

Table I: Performance on the MSCOCO Karpathy test split [31] under the unpaired setting. All values are reported in percentage (%). Higher is better in all columns. † denotes the results of the re-implemented Graph-Align for fair comparison. The proposed *Fine-Grained SRE* has a better performance than the *Coarse-Grained SRE*. Additionally, it is clear that the *Fine-Grained SRE* approach also boosts the performance of Graph-Align, indicating that our approach learns very effective semantic relationship information even for scene-graphs.

B. Inference Strategy

We use the full model to train the model in the language domain to generate image captions. The inference process is defined as follows:

Image → Semantic Concepts → Semantic Relationships → Semantic Relationship Embedding → Decoder → Sentence

Given an image I, we can extract the Semantic Concepts C^I from I through the *Semantic Concept Extractor*. Then, similar to sentence reconstruction process, we use the *Semantic Relationship Explorer* to extract the Semantic Relationship R^I, which embedded to *Semantic Relationship Embedding*s E^I. At last, E^I will use the decoder to generate captions.

V. Experiments

In this section, we will firstly describe a benchmark dataset for image captioning as well as some widely-used metrics and experimental settings. Then we will present the proposed model.

A. Datasets, Metrics and Settings

Several datasets consist of images-caption pairs. We use the popular Microsoft COCO [38] dataset to evaluate our reported results. This dataset contains 123,287 images and each image is paired with 5 sentences. To make a fair comparison, we use the widely-used splits in the work of Karpathy and Li [31] to report our results. There are 5,000 images each in validation set and test set, and 566,435 sentences for training full model.

We test the model performance with MSCOCO captioning evaluation toolkit [38]. It reports the widely-used automatic evaluation metrics SPICE, CIDEr, BLEU, METEOR and ROUGE. SPICE [39] is based on scene graph matching and CIDEr [40] is based on n-gram matching. These two metrics are specifically designed to evaluate image captioning systems. They both incorporate the consensus of a set of references for an example. BLEU [41] and METEOR [42], [43] are originally designed for machine translation evaluation, while ROUGE [44] is proposed for automatic evaluation of the extracted text summarization. A related research suggests that SPICE correlates the best with human judgments and does especially well in judging detailedness, where the other metrics present

negative correlations; CIDEr and METEOR follows with no conspicuous superiority, then comes ROUGE-L, and BLEU-4, in that order [39], [40]. Among them, SPICE and CIDEr are specifically designed to evaluate image captioning systems and will be the main considered metrics.

Following convention, we replace caption words appearing less than 5 times in the training set with the commonly unknown word token UNK. This has resulted in 9,487 words for MSCOCO. The size of the word embedding e is 512 and the hidden size d of the LSTM is 512. N, the number of heads in multi-head attention is set to 8 and d_{ff}, the feed-forward network dimension is set to 2048. Besides, the hidden layer size d_a is 350. The relationship embedding has 30 rows (the r), which is similar to [29]. Following [17], we train the *Semantic Concept Extractor* on the Visual Genome [45] dataset. Only top 20 semantic concepts are selected for each image. We also share the semantic concept embedding and the input word embedding in implementation. During the training process, firstly, we train the model with the cross-entropy loss for 15 epochs. Then, in reinforcement learning, we train the entire model with the batch size of 50 and use Adam [46] for parameter optimization. For full model, the learning rate is 4e-4. The β_1, β_2 are respectively set to 0.8 and 0.999. During sampling for MSCOCO dataset, we apply beam-search with a beam size of 5.

B. Results

1) Under the Unpaired Setting: In this section, in order to show the advantages of our method, firstly we compare our approach with the current unpaired image captioning methods, including the recently proposed Language Pivoting [15], Adversarial+Reconstruction [16] and Graph-Align [17], which are the most advanced on the MSCOCO dataset in comparable settings. As shown in Table I, in terms of SPICE, our proposed *Fine-Grained SRE* is competitive with all baselines, which use complex models and schemes. This is reasonable because both semantic concepts and semantic relationships are high-level and understandings of image [22]. Additionally, we boost the performance of Graph-Align, which indicates vision domain and language domain can be connected by our approach effectively. As expected, the *Fine-Grained SRE* has a better

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
ATT-FCN [5]								
Baseline	70.9	53.7	40.2	30.4	24.3	53.3	95.6	18.6
+ Proposal	**72.3**	**55.6**	**41.6**	**32.0**	**25.9**	**54.8**	**100.2**	**19.2**
LSTM-A2 [6]								
Baseline	73.3	56.5	42.7	32.2	25.3	53.9	99.1	18.3
+ Proposal	**77.2**	**58.4**	**44.6**	**33.1**	**25.9**	**54.2**	**103.5**	**18.7**
LSTM-A3 [6]								
Baseline	73.5	56.6	42.9	32.4	25.5	53.9	99.8	18.5
+ Proposal	**76.8**	**57.9**	**44.2**	**33.2**	**26.7**	**54.5**	**104.1**	**19.1**
LSTM-A4 [6]								
Baseline	72.1	55.5	41.7	31.4	24.9	53.2	95.7	17.8
+ Proposal	**72.8**	**56.4**	**43.4**	**32.1**	**26.6**	**54.6**	**101.4**	**18.4**
LSTM-A5 [6]								
Baseline	73.4	56.7	43.0	32.6	25.4	54.0	100.2	18.6
+ Proposal	**75.2**	**57.9**	**44.2**	**34.0**	**26.7**	**54.8**	**108.0**	**19.7**

Table II: Evaluation of representative systems under the paired setting. The proposed approach can further improve the already strong baselines in all metrics, which demonstrates the generalization ability of our approach to a wide range of existing systems. The significant improvements come from the explored semantic relationships rather than the simple incorporation of the semantic concepts, which indicates its effectiveness in exploring semantic relationships.

effect than the *Coarse-Grained SRE*. This indicates that the *Fine-Grained SRE* helps improve the quality of the extracted semantic relationships by extracting relevant relationships in more precise ranges.

2) Under the Paired Setting: We have done experiments to further illustrate the advantages of our method. Under the paired setting, we use five strong baselines without exploring the semantic relationships for semantic concepts. We replace the source semantic concept information of the baselines with extracted semantic relationship information, which is extracted by *Fine-Grained Semantic Relationship Explorer*. Since our focus is to provide semantic relationship information, we preserve the original settings for all baselines.

ATT-FCN [5] attends semantic concepts in all decoding steps. LSTM-A [6] uses a series of models (LSTM-A2,3,4,5) to incorporate the semantic concepts. We will study the effect of our method on them. LSTM-A3 takes semantic concepts as input at the first decoding step and visual features at the second step. LSTM-A4 presents textual concepts to the decoder, leaving visual features for the following steps; Contrary to LSTM-A3 and LSTM-A4, LSTM-A2 and LSTM-A5 reverse this order by firstly presenting visible features, respectively.

As shown in Table II, *Fine-Grained SRE* successfully promotes all baseline systems. It has brought improvements up to 8% and 6% in terms of CIDEr and SPICE, respectively, which further demonstrates its wide-range generalizing ability and indicates its effectiveness in exploring semantic relationships, which is less prone to the variations of model structures, hyperparameters (e.g., learning rate and batch-size), and learning paradigm.

To conclude, we not only performs better than existing methods under the unpaired setting, but manage to greatly improve the baselines under the paired setting.

VI. ANALYSIS

In this section, we will first analyze the contribution of each component in the proposed method. Then examples and error analysis under the unpaired setting are given to demonstrate the strengths and areas for improvement of our model. The following analysis are conducted on the the proposed *Fine-Grained SRE*.

A. Quantitative Analysis

The generic captioning performance can be affected by the *Semantic Relationship Embedding* and *Attention-based Sentence Decoder* we adopted. Therefore, we first assume that ground truth semantic concepts are fixed and then analyze them. The results are shown in Table III. In addition, because of our main contribution is proposing the *Semantic Relationship Explorer*, we select the Graph-Align to conduct a series of experiments under real unpaired situations to evaluate the contribution of sub-modules. The results are shown in Table IV. Meanwhile, we have also listed results of SPICE sub-categories to make it easier to analyze the quality and differences of the captions.

a) Contribution of Semantic Relationship Embedding and Attention-based Sentence Decoder: As shown in Table III, the comparison of the three models indicates the upper limit of the performance of attention mechanism (*Att*) and embedding mechanism (*Emb*). It is demonstrated that the *Attention-based Sentence Decoder* (*GT-SRE-Att*) outperforms the non-attention model (*GT-SRE*) in all metrics. This proves that weighing over features can improve the global dependency of features. The performance of model is greatly improved when it is combined with *Semantic Relationship Embedding* (*GT-SRE-Emb-Att*). This indicates that *Semantic Relationship Embedding* can

445

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
GT-ATT [22]	82.3	71.7	54.2	44.9	30.7	59.7	143.9	24.1
GT-SRE	84.7	72.9	57.8	46.4	32.0	61.5	154.1	26.2
GT-SRE-Att	90.6	80.4	65.9	57.2	35.5	66.9	161.5	27.4
GT-SRE-Emb-Att (Full Model)	**92.4**	**86.9**	**74.4**	**63.1**	**36.7**	**69.1**	**174.4**	**28.3**

Table III: Performance of the proposed models using the ground-truth semantic concepts on MSCOCO. The attention mechanism (*Att*) and the embedding mechanism (*Emb*) can promote the baseline in all metrics. Additionally, we report the performance of a representative system ATT [22]. As we can see, our approach outperforms the ATT substantially in all metrics, which further demonstrates the effectiveness of our approach.

Methods	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE						
					All	Objects	Attributes	Relations	Color	Count	Size
Graph-Align	21.1	20.5	46.2	68.7	14.5	28.8	3.8	3.8	1.7	0.7	1.9
Graph-Align w/ "Attribute"	**21.5**	20.9	**47.0**	**71.2**	15.0	29.2	**5.6**	3.4	**3.1**	**2.6**	**2.3**
Graph-Align w/ "Object"	21.3	**21.4**	46.7	70.1	**15.6**	**30.7**	4.2	4.3	2.4	1.5	2.0
Graph-Align w/ "Relation"	20.7	20.4	46.4	69.8	14.7	29.0	3.6	**4.5**	2.0	1.1	1.9
Full Model	21.8	22.1	48.4	75.7	16.1	31.0	5.8	4.9	3.4	3.7	2.5

Table IV: Results of incremental analysis of our proposed approach under the unpaired setting. For a better understanding of the differences, we further list the breakdown of SPICE F-scores. *"w/ Attribute"*, *"w/ Object"* and *"w/ Relation"* stand for the baseline further equipped with the *"Attribute" Aggregator*, *"Object" Aggregator* and *"Relation" Aggregator*, respectively. The bold numbers are best numbers before applying full model. We can see that the *"w/ Attribute"* has a higher attributes, colors, count and size scores than the other baselines. The *"w/ Object"* promotes the baseline in objects scores and the *"w/ Relation"* reaches better scores in relations. As we can see, incorporating all the sub-modules (i.e., *Full Model*) directly on the baseline leads to overall improvements.

further extract multiple features from semantic relationships, which strengthens its representing ability.

Furthermore, our full model performs better than *GT-ATT*. The latter directly uses semantic concepts at all steps instead of exploring the semantic relationships for semantic concepts. This proves the necessity to build relationships between semantic concepts.

b) Effect of "Attribute" Aggregator.: As shown in Table IV, if we incorporate our model with the *"Attribute" Aggregator* and aggregate correlated attributes in each specific object, it will have strengthened attribute relationship representations. Consequently, the object united with the strengths of attribute collocations can produce a balanced improvement. This is especially true in *attributes*, *color* and *count* score. It can also promote the decoder to generate accurate captions.

c) Effect of "Object" Aggregator.: Just as we have expected, *"Object" Aggregator* performs well in connecting related objects in the semantic concepts according to a relation word, which is proved by the increased scores in *objects*. The object collocations use the "relation" word as a pivot to extract constrained collocations, which distills more detailed relationship information between objects, promoting the decoder to generate comprehensive captions.

d) Effect of "Relation" Aggregator.: As is shown in Table IV, the base model, which directly incorporates *"Relation" Aggregator*, have improved *relations*. This indicates that the introduction of relation collocations has induced more powerful relation information to the decoder. The decoder therefore establishes more fine-grained relationships around a specific object.

e) Full Model.: As we expected, each component, with different functions, brings about improvements from different aspects to the model. Consequently, their advantages are united to produce abundant and enriched information. This achieves deep image understanding, brings an overall improvement (see Table IV) and results in generating comprehensive and detailed captions (see Figure V).

B. Qualitative Analysis

In the Figure V, we list examples to intuitively show the differences between models. As is shown, all models can generate fluent and descriptive sentences of input images. However, different models present different amount of semantic relationship information. The *w/ Object* has more objects but lacks details, e.g. color and number. The *w/ Attribute* generates more detailed attributes and color. The *w/ Relation* performs well in portraying the relations but worse in attributes. The proposed *Full Model*, compared with the models mentioned above, helps the baseline to maintain good balances as well as generates more complete and coherent captions.

The reason for this phenomenon is *"Object" Aggregator* can automatically extend its focus in a specific relation and find related objects to establish a relationship group. *"Attributes" Aggregator* learns attribute-word collocations from various visual attributes and rich semantic information. *"Relation" Aggregator* learns to automatically extend its focus in a specific object and look for related relations. Based on rich semantic relationship information, the decoder can retrieve
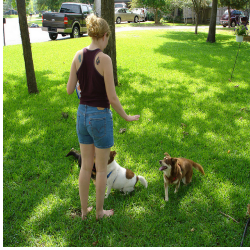
Comparison of Models				
Reference	a zebra grazing on grass in a green open field.	a woman is outside playing with her dogs in the grass.	a herd of wild animals walking across a large dry field.	a woman playing tennis on a tennis court while people watch.
Graph-Align	an animal standing on a field.	a woman standing on a field.	an animal standing on a field.	a woman standing on a field.
w/ "Attribute"	an animal standing on lush green grass field.	a beautiful woman standing on a grass field.	a herd of animals standing on a dry field.	a woman standing on a tennis court.
w/ "Object"	a zebra standing on a field with grass.	a woman in shirt and dogs playing on a grass field.	zebras and giraffes standing on a dry field.	a woman with a racket standing on a tennis court.
w/ "Relation"	an animal standing on a field eating grass.	a woman wearing shirt standing on a grass field with dogs nearby.	animals walking in a field.	a woman holding a racket standing on a tennis court.
Full Model	a zebra eating lush green grass in a grass field.	a woman wearing black shirt is playing with dogs on a green grass field.	giraffes and zebras standing on a dry grass field.	a woman wearing white shirt holding a tennis racket on the court.

Table V: Examples of the captions generated by different methods. The first line is the input image. The second and the third lines are the ground truth and the captions generated by the Graph-Align. The fourth, the fifth and the sixth lines are captions generated by adding *"Attribute" Aggregator*, *"Object" Aggregator* and *"Relation" Aggregator*, respectively. From the captions, we can find that the *"Attribute" Aggregator* helps the baseline to generate more detailed captions in attributes and colors for each object. The *"Object" Aggregator* results in more comprehensiveness in objects, and the *"Relation" Aggregator* helps the baseline to establish more reasonable relationships among objects. The *Full Model*, which incorporates all sub-modules, is able to generate more complete captions that are detailed in the objects, attributes, relations and colors.

Figure 4: Illustrations of the error analysis of the proposal.

information through attention mechanism. It will also decide on the most probable next word.

The qualitative analysis indicates that the semantic relationship information, extracted by *Semantic Relationship Explorer*, is much more aggregated than the source semantic concept information, ensuring the focus of current caption generation.

C. Error Analysis

In this section, we conduct error analysis using the proposed (full) model on the test set to provide insights on how the model may be improved. We find some generated captions that do not meet the standard. There are mainly three types of errors, i.e., location (45.9%), attribute (21.7%), and object (18.4%), with the left (14.0%) falling into other categories. In the Figure 4, we also give some examples.

Location error means that the model describes an irrelevant location relationship for an object (e.g., mislocating *banana*). Attribute error occurs when there are a lot of attributes, which can respectively establish logical relationships with a specific object. Therefore, the model mistakenly treats an incorrect relationship as grounded. In the given example, *field* can establish a relationship with either *dry* or *grass*, but only *grass field* is correct as shown in the picture. Object error occurs when there are incorrect extracted semantic concepts. The model treats the semantic concept as grounded in the image. In the given example, the incorrect semantic concept is *man*.

The challenging part is that these semantic relationships, which are extracted by the *Semantic Relationship Explorer*, are logically correct, otherwise the proposed model will choose other semantic relationships. A more powerful and accurate *Semantic Concept Extractor* may be needed to solve the problem, but it's unlikely to be completely avoided.

VII. CONCLUSIONS

In this work, we have proposed a framework for unpaired image captioning and a *Semantic Relationship Explorer* (SRE). Without any parallel data, SRE extracts abundant and rich

447

semantic relationship information for image captioning. Our method employs the semantic relationships to bridge the gap between vision and language domains. It is a powerful basis for the image description. We also use *Semantic Relationship Embedding* to get more powerful relationship features. Meanwhile the *Attention-based Sentence Decoder* is adopted to improve the global dependency of features. Our proposal is validated by experiments on the MSCOCO image captioning dataset. Remarkably, the proposed method outperforms the existing methods under the unpaired setting. In addition, it manages to boost five strong baselines to achieve a significant improvement under the paired setting. As shown in the quantitative and qualitative analysis, generated captions achieve a great balance among "object", "attribute" and "relation", which are more complete and coherent compared to other existing models.

VIII. ACKNOWLEDGMENTS

REFERENCES

[1] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *CVPR*, 2017.

[2] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *ECCV*, 2018.

[3] S. Wu, J. Wieland, O. Farivar, and J. Schiller, "Automatic alt-text: Computer-generated image descriptions for blind users on a social network service," in *CSCW*, 2017.

[4] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.

[6] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.

[7] F. Liu, X. Ren, Y. Liu, H. Wang, and X. Sun, "simnet: Stepwise image-topic merging network for generating detailed and comprehensive image captions," in *EMNLP*, 2018.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," in *CVPR*, 2018.

[9] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018.

[10] F. Liu, X. Ren, Y. Liu, K. Lei, and X. Sun, "Exploring and distilling cross-modal information for image captioning," in *IJCAI*, 2019.

[11] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR*, 2019.

[12] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *ICLR*, 2018.

[13] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," in *ICLR*, 2018.

[14] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based & neural unsupervised machine translation," in *EMNLP*, 2018.

[15] J. Gu, S. R. Joty, J. Cai, and G. Wang, "Unpaired image captioning by language pivoting," in *ECCV*, 2018.

[16] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *CVPR*, 2019.

[17] J. Gu, S. R. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," *arXiv preprint arXiv:1903.10658*, 2019.

[18] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *ICML*, 2008.

[19] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[20] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[21] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018.

[22] Q. Wu, C. Shen, L. Liu, A. R. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *CVPR*, 2016.

[23] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2015.

[24] P. Lu, L. Ji, W. Zhang, N. Duan, M. Zhou, and J. Wang, "R-VQA: learning visual relation facts with semantic attention for visual question answering," in *KDD*, 2018.

[25] D. Yu, J. Fu, T. Mei, and Y. Rui, "Multi-level attention networks for visual question answering," in *CVPR*, 2017.

[26] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *NIPS*, 2006.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[28] F. Liu, Y. Liu, X. Ren, X. He, K. Lei, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *NeurIPS*, 2019.

[29] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," in *ICLR*, 2017.

[30] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.

[31] A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[32] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *CVPR*, 2018.

[33] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017.

[34] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *CVPR*, 2017.

[35] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, "Skeleton key: Image captioning by skeleton-attribute decomposition," in *CVPR*, 2017.

[36] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.

[37] F. Liu, M. Gao, Y. Liu, and K. Lei, "Self-adaptive scaling for learnable residual structure," in *CoNLL*, 2019.

[38] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[39] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[40] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[41] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for automatic evaluation of machine translation," in *ACL*, 2002.

[42] C. Lin and E. H. Hovy, "Automatic evaluation of summaries using n-gram co-occurrence statistics," in *HLT-NAACL*, 2003.

[43] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *ACL*, 2005.

[44] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL*, 2004.

[45] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2014.