GroupCap: Group-based Image Captioning with Structured Relevance and Diversity Constraints

Fuhai Chen[1], Rongrong Ji[1]*, Xiaoshuai Sun[2], Yongjian Wu[3], Jinsong Su[4]
[1]Fujian Key Laboratory of Sensing and Computing for Smart City, School of Information Science and Engineering, Xiamen University, 361005, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, 150006, China
[3]Tencent YouTu Lab, [4]Xiamen University, 361005, China
cfh3c@stu.xmu.edu.cn, {rrji,jssu}@xmu.edu.cn, xiaoshuaisun@hit.edu.cn, littlekenwu@tencent.com

Abstract

*Most image captioning models focus on one-line (single image) captioning, where the correlations like relevance and diversity among group images (*e.g.*, within the same album or event) are simply neglected, resulting in less accurate and diverse captions. Recent works mainly consider imposing the diversity during the online inference only, which neglect the correlation among visual structures in offline training. In this paper, we propose a novel group-based image captioning scheme (termed GroupCap), which jointly models the structured relevance and diversity among group images towards an optimal collaborative captioning. In particular, we first propose a visual tree parser (VP-Tree) to construct the structured semantic correlations within individual images. Then, the relevance and diversity among images are well modeled by exploiting the correlations among their tree structures. Finally, such correlations are modeled as constraints and sent into the LSTM-based captioning generator. We adopt an end-to-end formulation to train the visual tree parser, the structured relevance and diversity constraints, as well as the LSTM based captioning model jointly. To facilitate quantitative evaluation, we further release two group captioning datasets derived from the MS-COCO benchmark, serving as the first of their kind. Quantitative results show that the proposed GroupCap model outperforms the state-of-the-art and alternative approaches.*

1. Introduction

Automatic description of an image, *a.k.a.* image captioning, has recently attracted extensive research attention [1, 2, 3, 4, 5]. Typically, these methods train the image captioning models under a one-line paradigm, without regarding the correlations (*i.e.*, relevance and diversity) among group images. However, in many real-world applications like captioning photo albums or events, the images are not suitable to be captioned alone. In such situations, it would benefit the generated results by capturing the relevance and diversity among these group images as shown in Fig.1.

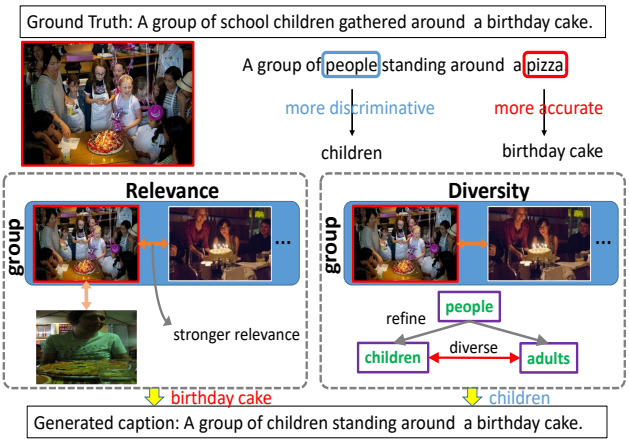

Figure 1. In one-line captioning, existing methods tend to generate less accurate and less discriminative captions compared to the ground truth (the top caption is generated by the state-of-the-art [5]). We focus on capturing the relevance and diversity among a group of images to reinforce and diversify the image captions (the bottom one, which is generated by the proposed GroupCap).

As far as we know, there is no existing work in the literature that addresses the task of group-based image captioning. On the one hand, there is no related work addressing the issue of modeling relevance. To this end, one should the model image relevance by maximizing the visual similarity of the inner-group images comparing to that of the inter-group images. On the other hand, there are two works [6, 7] that refer to modeling the image diversity, both of which however only focus on online inference. Sadovnik *et al.* [6] proposed a context-aware scheme to capture the particular items (entities, relations, and *etc.*) of the target image to diversify its description from the other inner-

*Corresponding Author.

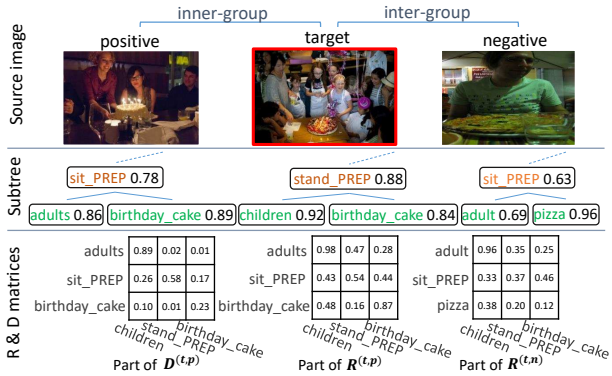

Figure 2. The examples of VP-Trees and structured relevance/diversity matrices among an image triplet. The second row presents the subtrees of the VP-Trees, where the value in each node denotes the probability of the corresponding entity/relation ("_PREP" means the preposition following verb). The third row presents the relevance and diversity sub-matrices ($\boldsymbol{R}$ and $\boldsymbol{D}$, where t, p, and n denote the target, the positive, and the negative images, respectively), where the value of each element in the matrix denotes the relevance/diversity score.

group images. However, the scheme in [6] only conducts a coarse-grained online inference by using a simple template matching for discriminative caption generation. Upon the off-the-shelf LSTM model, Vedantam *et al.* [7] proposed a fine-grained context-aware scheme to generate the discriminative caption given the specific distractor image, which exploits pairwise (*i.e.*, target-distractor) textual contents in online inference. However, the scheme in [7] only considers the diversity between words in the corresponding positions of the pairwise captions, while ignoring the structured alignment of semantic items, as well as the their visual correlation among images.

In this paper, we argue that the fundamental issue of group-based captioning among group images lies in modeling their relevance and diversity from the visual perspective in an offline manner. On the one hand, the visual structured correlation can accurately model the fine-grained diversity among inner-group images in the offline training period, which is different from the existing methods of coarse-grained template matching [6] or the rough alignment on words [7]. On the other hand, learning such visual structured correlation offline can better capture and accurately interpret relevance among the inter-group images, which is left unexploited in all existing works [4, 5, 8].

Driven by the above insights, we propose a novel group-based image captioning model, termed as *GroupCap*, based on offline learning with structured relevance and diversity constraints. Our main innovations lie in two aspects. Firstly, we introduce a visual parsing tree to extract the structured semantic relations in each image, and the examples are given in the second row of Fig.2. Secondly, we model the structured relevance and diversity upon the VP-Trees of group images, which are formulated as constraints to the unified image captioning model, the examples of which are given in the third row of Fig.2. In particular, taking an image triplet (including the target, the positive and the negative images) as the input for training, we firstly parse key entities and their relations of each image and organize them into a tree structure, which is trained by the supervision of the textual parsing trees. Then, based on parsing trees of these images, we design a structured relevance constraint among the image triplets by maximizing the similarity of the structured trees between the inner-group images, relative to that between the inter-group images. To measure the similarity among parsing trees, we further present an algorithm to align and compare between pairwise tree nodes, leading to an adaptive yet efficient calculation of structured relevance and diversity between image pairs. Finally, we embed such structured constraints into the decoder (an LSTM-based captioning model) for the caption generation. Note that, the parsing tree, the structured constraints, and the captioning model are integrated into an end-to-end joint training. In the online inference, we parse each image into a tree structure, which is fed into the LSTM-based decoder for the final caption generation.

The contributions of this paper are as follows: (1) We investigate a new problem, termed group based image captioning. (2) We are the first to model both relevance and diversity among image contents in the group based image captioning. (3) We propose an end-to-end offline training scheme towards generating very distinguished captioning among group images. (4) We release two group-based image captioning datasets to facilitate the subsequent research. Quantitative comparisons to the state-of-the-art and alternative schemes demonstrate our merits.

2. Related Work

Most existing methods for image captioning are based on Convolutional Neural Network + Recurrent Neural Network (CNN-RNN) [1, 2, 9, 10], where the visual features are extracted from CNN, and then fed into RNN to output word sequences as captions. The recent advances mainly focus on revising the above CNN-RNN architecture. For example, You *et al.* [4] proposed a semantic attention model to select semantic concepts detected from the image, which were embedded into the caption generation procedure. Lu *et al.* [11] introduced an adaptive attention encoder-decoder model, which relies on visual signals to decide when to compensate the language model. Liu *et al.* [12] proposed a semantically regularised CNN-RNN model to solve the vanishing gradients during backpropagation. Recently, Gan *et al.* [5] utilized a semantic compositional network to compose the semantic meaning of individual tags for image captioning. However, all above methods are based on one-line scheme that operates for individual images, without considering the correlations among group images to reinforce and diversify each other.

Recent works in image captioning also pay attention to

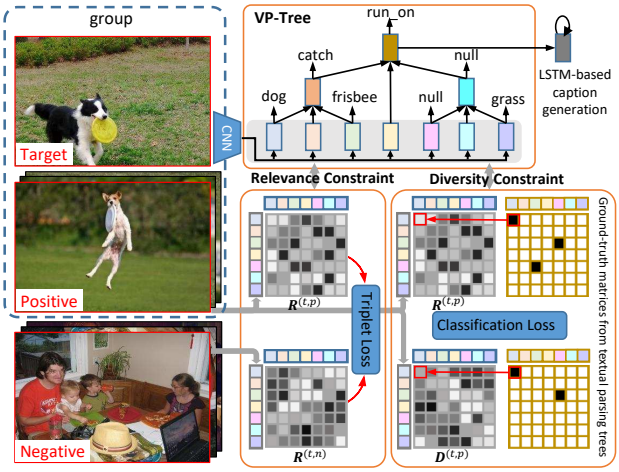

Figure 3. The framework overview of the proposed *GroupCap* model for group-based image captioning. The deep visual features of the given triplet images (the target (t), the positive (p), and the negative (n) images) are first extracted from a pre-trained CNN model, which are then sent to train a visual parsing tree (VP-Tree) model. Meanwhile, the structured relevance and diversity constraints are modeled based on these VP-Trees by minimizing the triplet loss and classification loss among the relevance ($\boldsymbol{R}$) and diversity ($\boldsymbol{D}$) matrices. Finally, the VP-Tree model, the structured relevance and diversity constraints, and the LSTM based captioning model are jointly trained in an end-to-end formulation.

exploiting the discriminability of caption generations, such as personalized image captioning [13, 14], stylistic image captioning [15], and context-aware discriminative image captioning [6, 7]. Specifically, context-aware schemes [6, 7] were proposed to capture the diversity among images during the online captioning. In addition to our advance described in Sec.1, we are dedicated to making the captioning more discriminative, *i.e.*, to describe specific concepts and their structured correlations, rather than only capturing their diversities as done in [6, 7]. For example, people tend to specify the general concepts, *e.g.*, they can easily name the object with *policeman*, *player*, or *fire man*, rather than generally naming *man*. This can be also verified by the annotating schemes that are commonly used in the datasets like COCO [16].

In terms of context-aware image captioning, our work is also related to the Referring Expression Generation (REG). REG aims to uniquely compose an expression for a specified object in comparison to other objects in an image, which can be regarded as a task of intra image contextual image captioning [17, 18, 19, 20, 21, 22, 23]. For example, Mao *et al.* [18] proposed to add a Maximum Mutual Information (MMI) constraint, which encourages the generated expression to describe the target object unambiguously relative to the others in the image. Luo *et al.* [23] proposed a *generate-and-rerank* pipeline to identify the target regions for unambiguous captions. Different from the above REG task, our image captioning scheme is based on the correlation (both relevance and diversity) among group images.

3. GroupCap

The framework of proposed group-based image captioning (GroupCap) scheme is presented in Fig.3, which aims at embedding both relevance and diversity among group images into the caption generation. It consists of four stages, *i.e.*, deep visual feature extraction, visual tree parsing, structured relevance and diversity modeling, and encoder-decoder based caption generating. In particular, we first employ a pre-trained CNN model to extract visual features from every given image. We then train a visual parsing tree model to extract visual entities and their relations for these images, as detailed in Sec.3.1. Then, we present our scheme to quantify tree-based correlations to model image-to-image correlations (both relevance and diversity) in Sec.3.2. Finally, we depict the joint training of the entire model in Sec.3.3.

3.1. Visual Parsing Tree

Visual parsing tree model (VP-Tree) is first designed in [24] to extract semantic entities and model their relations from an image. It is a fixed-structure of a three-layer complete binary tree, where each node represents a semantic item, *i.e.*, an entity or a relation (specifically, a subject, an object, a sub-relation or a main relation). We advance the VP-Tree by changing binary tree to ternary tree and adding the mapping for the relation from visual feature to the node feature as shown in Fig.3, where the information of relation can be strengthened. Given an image-caption pair, a deep visual feature G is first extracted from the last fully-connected layer of a pretrained CNN [25]. Then, the deep visual feature is mapped to the feature representations of entities/relations in the corresponding nodes, which is named as *Semantic mapping*. According to the structure of the tree, we combine the features of entities or relations and map them to the higher-level feature representations of relations. The operation is named as *Combination*. Meanwhile, the feature of each node is mapped to a category (entity or relation) space. And we name it as *Classification*. It's noted that the caption is parsed into a textual parsing tree denoted as $\boldsymbol{T}^t$ by the standard textual parser [26], which is employed as the supervision for the corresponding entity/relation classification during training. Finally, the whole VP-Tree can be generated with the parameters of the three operations.

To construct our VP-Tree, we first define it formally as $\mathbb{T}^v = \{\boldsymbol{h}^l_{j^l} \in \mathbb{R}^{d_n} | l \in \{1,2,3\}, j^1 \in \{1,\dots,7\}, j^2 \in \{1,2\}, j^3 \in \{1\}\}$, where $\boldsymbol{h}$, l, and d_n denote the node feature, the tree layer, and the dimensionality of the node feature, respectively. j^l denotes the index of node in the l-th layer. For example, $\boldsymbol{h}^1_{j^1=1}$, $\boldsymbol{h}^1_{j^1=7}$, and $\boldsymbol{h}^2_{j^2=2}$ represents the features of the first subject node (*dog*), the second object node (*grass*), and the second sub-relation node (*null*, *i.e.*, no specific relation) respectively in the VP-Tree as shown

in Fig.3. During *Semantic mapping*, the deep visual feature G can be mapped into the feature $\boldsymbol{h}^1_{j^1}$ as following:

$$\boldsymbol{h}^1_{j^1} = F^{sem}(G; \boldsymbol{W}^{sem}_{j^1}), \tag{1}$$

where F^{sem} is the linear mapping function, which transforms the visual feature to the semantic items (entities and relations). $\boldsymbol{W}^{sem}_{j^1}$ denotes the parameter for the j^1-th node in the first/leaf layer. We then combine the features of the children nodes in the lower layer and feed them into their parent nodes in the higher layer. Formally, we obtain the features of the parent nodes in the second/middle layer and the third/root layer respectively by:

$$\boldsymbol{h}^2_{j^2=1} = F^{com}([\boldsymbol{h}^1_{j^1=1}, \boldsymbol{h}^1_{j^1=2}, \boldsymbol{h}^1_{j^1=3}]; \boldsymbol{W}^{com}_{j^2=1}), \tag{2}$$

$$\boldsymbol{h}^2_{j^2=2} = F^{com}([\boldsymbol{h}^1_{j^1=5}, \boldsymbol{h}^1_{j^1=6}, \boldsymbol{h}^1_{j^1=7}]; \boldsymbol{W}^{com}_{j^2=2}), \tag{3}$$

$$\boldsymbol{h}^3_{j^3=1} = F^{com}([\boldsymbol{h}^2_{j^2=1}, \boldsymbol{h}^1_{j^1=4}, \boldsymbol{h}^2_{j^2=2}]; \boldsymbol{W}^{com}_{j^3=1}), \tag{4}$$

where $[\cdot,\cdot,\cdot]$ denotes the concatenation operation. F^{com} and $\boldsymbol{W}^{com}$ denote the linear mapping function and its parameter. Finally, each node is classified into the entity/relation category by the Softmax classifier as following:

$$\boldsymbol{y}^n_{j^1} = \sigma\big(F^{cat}(\boldsymbol{h}^1_{j^1}; \boldsymbol{W}^{cat}_{<e>})\big), j^1 \in \{1,3,5,7\}, \tag{5}$$

$$\boldsymbol{y}^n_{j^{2,3}} = \sigma\big(F^{cat}(\boldsymbol{h}^{(3)}_{j^{2,3}}; \boldsymbol{W}^{cat}_{<r>})\big), j^2 \in \{1,2\}, j^3 \in \{1\}, \tag{6}$$

where $\boldsymbol{y}^n$ denotes the predicted probability vectors of the entity or relation categories according to the entity dictionary or relation dictionary [24]. F^{cat} is a linear mapping function. $\boldsymbol{W}^{cat}_{<e>}$ and $\boldsymbol{W}^{cat}_{<r>}$ denote the mapping parameters for the entity and the relation categories, respectively.

The parameter set of VP-Tree can be denoted as $\boldsymbol{\Theta}^T = \{\boldsymbol{W}^{sem}, \boldsymbol{W}^{com}, \boldsymbol{W}^{cat}_{<e>}, \boldsymbol{W}^{cat}_{<r>}\}$. We minimize the loss of the category classification to optimize the whole tree model. And the offline training for VP-Tree is integrated into the overall training, as detailed in Subsec.3.3.

3.2. Structured Relevance and Diversity Constraint

We build a structured relevance and diversity constraints into the proposed GroupCap model, as illustrated in Fig.3. Our main ideas are: 1) The inner-group similarity is expected to be larger than that of inter-group, which is reflected during the training of VP-Tree model; 2) The diversity in two corresponding nodes of VP-Trees is expected to be classified accurately (*i.e.*, whether the nodes are diverse).

Given an image triplet, *i.e.*, the target (the i-th), the positive (the j-th), and the negative (the k-th) images, we estimate the relevance and diversity of pairwise images based on the features matrices of their leaf nodes, *i.e.*, $\boldsymbol{T}^v_i$, $\boldsymbol{T}^v_j$, and $\boldsymbol{T}^v_k \in \mathbb{R}^{K \times d_n}$ ($K = 7$ here, which denotes the number of the leaf nodes). Taking the target and the positive images for example, we have:

$$\boldsymbol{R}^{(i,j)} = \phi(\boldsymbol{T}^v_i \boldsymbol{U}_R (\boldsymbol{T}^v_j)^T), \tag{7}$$

$$\boldsymbol{D}^{(i,j)} = \phi(\boldsymbol{T}^v_i \boldsymbol{U}_D (\boldsymbol{T}^v_j)^T), \tag{8}$$

where $\boldsymbol{R}^{(i,j)}$ and $\boldsymbol{D}^{(i,j)}$ denote the $K \times K$ relevance and the $K \times K$ diversity matrices that align the visual tree nodes between the i-th and the j-th images. $\boldsymbol{U}_R$ and $\boldsymbol{U}_D$ are $d_n \times d_n$ factor matrices, which are the parameters of the relevance and diversity matrices, respectively. ϕ denotes a Sigmoid function. Then, we compare the similarities between inner-group and inter-group images as:

$$d^R(\boldsymbol{T}^v_i, \boldsymbol{T}^v_j, \boldsymbol{T}^v_k; \boldsymbol{U}_R, \boldsymbol{\Theta}^T) = \sum_{p,q}^{K} (\boldsymbol{R}^{(i,k)}_{p,q} - \boldsymbol{R}^{(i,j)}_{p,q}), \tag{9}$$

where $\boldsymbol{R}^{i,j}_{p,q}$ denotes the relevance score between the p-th and the q-th nodes of the i-th and the j-th images. $\boldsymbol{\Theta}^T$ denotes parameters of VP-Tree. Suppose there are N image triplets, we employ the triplet loss to maximize the inner-group similarity and minimize the inter-group similarity, leading to:

$$\mathcal{L}^R(\boldsymbol{U}_R, \boldsymbol{\Theta}^T) = \frac{1}{N} \sum_{<i,j,k>}^{N} \max(d^R(\boldsymbol{T}^v_i, \boldsymbol{T}^v_j, \boldsymbol{T}^v_k; \boldsymbol{U}_R, \boldsymbol{\Theta}^T), \tau), \tag{10}$$

where τ denotes the predefined margin of the triplet loss. To align the relevance of every two nodes in two VP-Trees, we introduce an alignment-wise logistic regression to compute the classification loss. Taking the i-th and the j-th images in the same group as example, we have:

$$\mathcal{L}^R_c(\boldsymbol{U}_R, \boldsymbol{\Theta}^T) = -\frac{1}{K^2} \sum_{p,q=1}^{K} \log P(\boldsymbol{y}^R_{p,q} = 1 | \boldsymbol{T}^v_i, \boldsymbol{T}^v_j; \boldsymbol{U}_R), \tag{11}$$

where $\boldsymbol{y}^R_{p,q}$ denotes the estimation of the relevance between the p-th and the q-th nodes in $\boldsymbol{R}^{(i,j)}_{p,q}$. If they are relevant, $\boldsymbol{y}^R_{p,q} = 1$, otherwise $\boldsymbol{y}^R_{p,q} = 0$. Similarly, to align diversity of every two nodes in two VP-Trees (specially for the node pairs in inner-group images, *i.e.*, the target and the positive images), we also adopt alignment-wise logistic regression to compute the classification loss:

$$\mathcal{L}^D_c(\boldsymbol{U}_D, \boldsymbol{\Theta}^T) = -\frac{1}{K^2} \sum_{p,q=1}^{K} \log P(\boldsymbol{y}^D_{p,q} = 1 | \boldsymbol{T}^v_i, \boldsymbol{T}^v_j; \boldsymbol{U}_D), \tag{12}$$

where $\boldsymbol{y}^D_{p,q}$ denotes the estimation of the diversity between the p-th and the q-th nodes in $\boldsymbol{D}^{(i,j)}_{p,q}$. To get the ground-truth relevance and diversity of two nodes, we use the textual parsing trees to decide if two nodes of different trees are relevant or diverse, the details of which will be provided in Sec.4. It's noted that the relevance and diversity are embedded as constraints only in the training period to refine the VP-Tree model.

3.3. Joint Learning

The training data for each image consist of deep visual feature G and caption words sequence $\{\boldsymbol{y}_t\}$. Our goal is to jointly learn all the visual parser parameters $\boldsymbol{\Theta}^T$, relevance

and diversity constraint parameters $\mathbf{\Theta}^C = \{\boldsymbol{U}_R, \boldsymbol{U}_D\}$, together with the LSTM parameters $\mathbf{\Theta}^L$ by minimizing a loss function over the training set. Given the deep visual feature set $\mathcal{S} = \{G_i | i = 1, 2, 3\}$ of an image triplet, the joint loss of the GroupCap model is defined as:

$$\begin{aligned}\mathcal{L}(\mathbf{\Theta}^T, \mathbf{\Theta}^C, \mathbf{\Theta}^L) = \sum_{i}^{|\mathcal{S}|} \Big(\sum_{t}^{T} \log P(\boldsymbol{y}_t^i | \boldsymbol{y}_{0:t-1}^i, G_i; \mathbf{\Theta}^L) \\ + \sum_{j}^{K} \log P(\boldsymbol{y}_j^{n,i} | G_i; \mathbf{\Theta}^T) \Big) \\ + \mathcal{L}^R(\boldsymbol{U}_R, \mathbf{\Theta}^T) + \mathcal{L}_c^R(\boldsymbol{U}_R, \mathbf{\Theta}^T) + \mathcal{L}_c^D(\boldsymbol{U}_D, \mathbf{\Theta}^T),\end{aligned} \tag{13}$$

where T and K denote the length of the sequence output and the number of tree nodes, respectively. $\boldsymbol{y}_t^i$ and $\boldsymbol{y}_j^{n,i}$ denote the word output in the t-th state and the entity/relation categories in the j-th node for the i-th sample, respectively.

We pre-train the VP-Tree model separately at the first time and the VP-Tree model with the caption generation model at the second time. Then, we use Adam algorithm [27] with learning rate 1×10^{-4} to optimize Eq.13, where the gradient is back-propagated over the caption generation model, the visual tree parser, and the structured relevance/diversity constraint. To avoid overfitting, we employ a dropout operation with a ratio of 0.5. Finally, the iteration ends until the cost of the final word prediction converges.

4. Experiments

In this section, we perform extensive experiments to evaluate the proposed GroupCap model. We first describe the datasets and experimental settings. Next, we quantitatively compare the results of our proposed model to the state-of-the-art methods on image captioning. Finally, we qualitatively analyze our merits in details.

Preprocessing on Textual Parsing Trees. Due to the irrelevant words and noise configurations generated by Stanford Parser [26], we whiten the source sentences by using the pos-tag tool and the lemmatizer tool in NTLK [28] simultaneously. After that, we convert the dynamic parsing tree to a fixed-structured, three-layer binary tree, which only contains nouns (or noun pair, adjective-noun pair), verbs, coverbs, prepositions, and conjunctions. Only nouns are regarded as entities and used as leaf nodes in the subsequent training. We select the frequent words and manually merge words with similar meaning to obtain the entity dictionary and the relation dictionary with size 748 and 246, respectively. For the judgment of relevance, we leave the noun pair and the adjective-noun pair out of leaf nodes. For the judgment of diversity, we keep two kinds of leaf nodes: leaf nodes with and without the noun pair and the adjective-noun pair[1]. Assuming there are K nodes in the fixed-structured tree, there would be K^2 alignments between two trees, each of which reflects whether each alignment in the relevance alignment matrix is relevant, as well as whether each alignment in the diversity alignment matrix is diverse.

[1]The judgment of diversity needs the coarse-categories and fine-categories simultaneously. For examples, as Fig.1 shows, the condition of diversity between *children* and *adults* is that they both belong to *people*.

Table 1. Performance comparisons to the state-of-the-art methods and baselines on FG-dataset. "-g" means using the grouped data (FG-dataset) from MS-COCO. The numbers in bold face are the best known results and (-) indicates unknown scores. All values are in %.

Methods	B1	B2	B3	B4	M
BRNN [2]	62.5	45.0	32.1	23.0	19.5
LRCN [29]	62.8	44.2	30.4	21.0	-
Google NIC [1]	66.6	45.1	30.4	20.3	-
Toronto [3]	71.8	50.4	35.7	25.0	23.0
ATT [4]	70.9	53.7	40.2	30.4	24.3
SCA-CNN [8]	71.9	54.8	41.1	31.1	25.0
StructCap [24]	72.6	56.3	43.0	32.9	25.4
SCN [5]	72.8	56.6	43.3	33.0	25.7
NIC-g	68.1	46.3	31.5	21.4	21.8
StructCap-g	73.1	56.8	43.1	32.8	25.7
SCN-g	73.4	57.0	43.4	33.0	25.7
GroupCap-T	73.4	57.0	43.3	32.9	25.8
GroupCap-T-SRC	73.7	57.3	43.5	33.0	25.9
GroupCap-T-SDC	73.6	57.2	43.2	32.8	25.8
GroupCap (w/o ensemble)	73.9	57.4	43.5	33.0	26.0
GroupCap (w/ ensemble)	**74.4**	**58.1**	**44.3**	**33.8**	**26.2**

Datasets and Evaluation Protocols. MS-COCO is a widely-used dataset for image captioning. There are over 123,000 images in MS-COCO, which has been split publicly into training, testing and validating sets[2]. We build two group captioning datasets[3] from MS-COCO to evaluate the performance of our models, where the images in the training set are grouped into two kinds of groups:

1) Frequency-based Group Captioning Dataset (FG-dataset). This dataset evaluates the accuracy and discriminability of the generated captions. To construct this dataset, we firstly filter and collect the top-784 entities and the top-246 relations with high frequencies in the textual parsing tree. Then, we combine the entities and the relations, and then keep the top-39,766 semantic combinations with high frequencies in the textual parsing tree. Finally, we divide the MS-COCO image-caption pairs into 39,766 image groups corresponding to the semantic combinations. We get the FG-dataset with totally 1,432,076 training images among 39,766 training groups, 5,000 valuation images, and 5,000 testing images (The valuation and testing sets are the same as MS-COCO). Note that all groups are unable to cover all semantic combinations, and any each group is unable to cover all the semantic items. However, by such high-frequency based sampling, the dataset is adequate to evaluate the performance of group-based models on accuracy and discriminability. To form the triplet, we randomly select the positive and negative images from the same and different

[2]https://github.com/karpathy/neuraltalk

[3]Datasets are available at mac.xmu.edu.cn/Data_cvpr18.html

Table 2. Performance comparisons to the state-of-the-art methods and baselines on ES-dataset. "-g" means using the grouped data (ES-dataset) from MS-COCO. The numbers in bold face are the best known results and (-) indicates unknown scores. All values are in %.

Methods	B1	B2	B3	B4	M
StructCap-g [24]	72.0	55.7	41.4	31.3	25.7
SCN-g [5]	72.1	55.2	41.6	**31.9**	25.7
GroupCap-T	72.3	56.0	41.9	31.1	25.6
GroupCap-T-SRC	72.6	56.3	42.3	31.5	25.8
GroupCap-T-SDC	72.3	56.0	42.1	31.2	25.7
GroupCap	**72.9**	**56.5**	**42.5**	31.6	**25.9**

groups respectively for each target image.

2) Entity-specific Group Captioning Dataset (ES-dataset). This dataset evaluates the accuracy and discriminability of the specific entities in the generated captions. We firstly collect the specific entities that are frequently missed in the generated captions by the state-of-the-art method [5]. We then filter and keep the images with the specific entity[4] from FG-dataset. Finally, we get the ES-dataset with totally $449,190$ training images among $28,937$ training groups, 5,000 valuation images, and 5,000 testing images (The valuation and testing sets are the same as MS-COCO). We further select the target and positive images in the *Missing Subset* and out of *Missing Subset*, respectively. The negative image is selected out of the group in the same way as that of FG-dataset.

Quantitative performance of all methods are evaluated by using microsoft COCO caption evaluation tool[5], including BLEU, METEOR, ROUGE-L[16]. We also evaluate our model by using accuracy and recall.

Baselines and State-of-the-Arts. We compare the proposed GroupCap with six baselines: 1) StructCap: A structured semantic embedding model for image captioning based on binary VP-Tree model [24]. 2) GroupCap-T: The mutated version of GroupCap without structured relevance and diversity constraints. 3) GroupCap-T-SRC: The mutated version of GroupCap, where the structured diversity constraint is removed. We compare our model to GroupCap-T-SRC to evaluate the effectiveness of the diversity constraint. 4) GroupCap-T-SDC: The mutated version of GroupCap, where the structured relevance constraint is removed. We compare our model to GroupCap-T-SDC to evaluate the effectiveness of the relevance constraint. We also compare the state-of-the-art methods, *i.e.*, ATT [4], SCA-CNN [8] and SCN [5].

Performance on FG-dataset. We compare GroupCap to the state-of-the-art and baseline methods on FG-dataset as shown in Tab.1. GroupCap, especially the one with ensemble scheme (4 models), achieves the best performance under all metrics, which reflects that considering the relevance and diversity can reinforce and diversify the caption generation (The generated results will be further analyzed in the part of *Evaluation for Caption Generation*). Moreover, GroupCap outperforms GroupCap-T-SRC and GroupCap-T-SDC, which reveals that our relevance and diversity constraints do contribute to the overall performance. In addition, GroupCap-T outperforms NIC-g, which indicates the effectiveness of our VP-Tree.

[4]We choose the specific entity in each group with high frequency in the ground-truth captions. And the images with this entity that the state-of-the-art method missed (We call it *Missing Subset* in each group) must be 40%-60% of all the images with this entity to guarantee the sample balance.

[5]https://github.com/tylin/coco-caption

Table 3. The performance comparisons of parsing models (Par. M.) on FG-dataset. "CNN-F", "CNN-sVP" and "CNN-gVP" denote the fully-connected layer, the original VP-Tree [24], and our VP-Tree respectively with the output of CNN. "Acc.(E.)" and "Acc.(R.)" denote the metrics of the correct classification (top-3) on entities and relations, respectively. All values are in %.

	Caption Generation				Classification	
Par. M.	B4	M	R	C	Acc.(E.)	Acc.(R.)
CNN-F	32.5	25.0	53.2	98.3	-	-
CNN-sVP	32.8	25.7	54.2	100.6	72.1	70.5
CNN-gVP	**32.9**	**25.8**	**54.5**	**101.9**	**74.7**	**73.0**

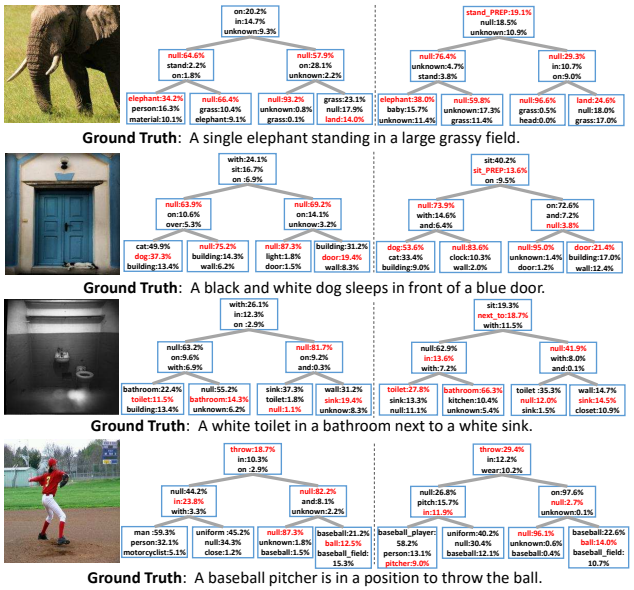

Figure 4. VP-Trees constructed by the proposed GroupCap compared to StructCap. The three columns are the source images, the generated VP-Trees by the model [24], the generated VP-Trees by our model, respectively. Red font means the correct semantic items according to the textual parsing trees.

Performance on ES-dataset. Tab.2 shows the comparing between GroupCap and state-of-the-art methods on ES-dataset. GroupCap achieves the best performance under all metrics. Since the specific entities are collected where the state-of-the-art method (SCN) fails, the superior performance reflects that GroupCap can refine the generated caption on the accuracy and discriminability. GroupCap also outperforms all the baselines on ES-dataset, such as GroupCap-T-SRC and GroupCap-T-SDC, which indicates

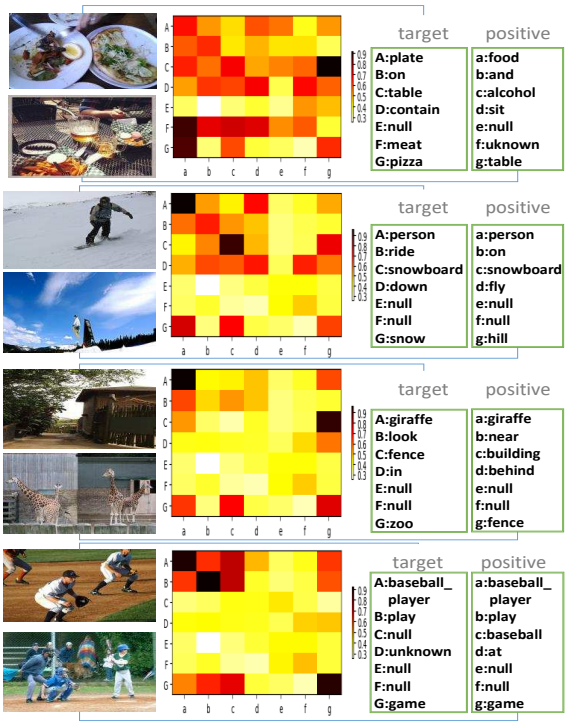

Figure 5. Visualization of the relevance matrices. Darker color means more relevant between two semantic items (entities or relations).

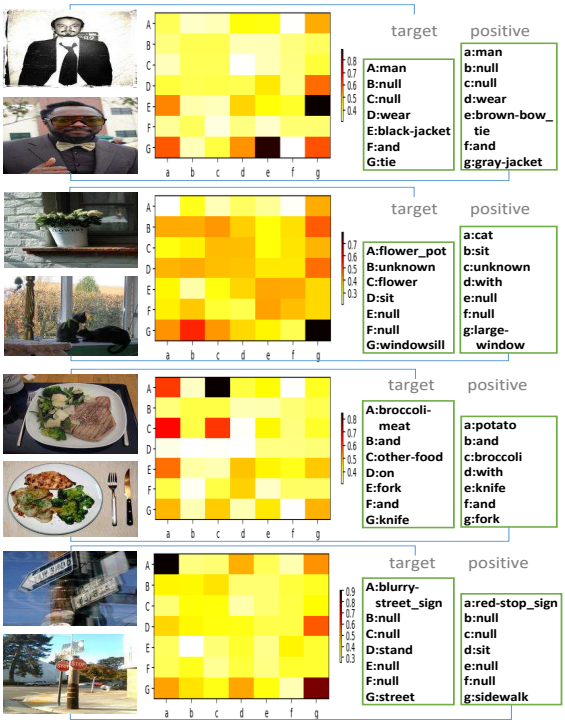

Figure 6. Visualization of the diversity matrices. Darker color means more diverse between two semantic items (entities or relations).

that considering relevance and diversity do contribute to the overall performance.

Evaluation for VP-Tree Construction. To evaluate the advanced VP-Tree, we compare it with the visual parser proposed in [24].[6] The qualitative results are shown in Fig.4. The proposed VP-Trees are more identical to the ground-truth captions compared to the ones generated by sVP. Also, the semantic items (entities or relations) in each VP-Tree generated by gVP are more accurate, which is due to more information of the relation from images.

Further, we estimate the performances of caption generation and classification (classify each node into an entity/relation category) by using sVP and gVP models in Tab.3. CNN-gVP outperforms the CNN-F and CNN-sVP, which indicates the effectiveness and superiority of our proposed gVP model. Additionally, the entity/relation classification of CNN-gVP is more accurate than that of CNN-sVP, which further manifests the rationality of the proposed gVP structure.

Evaluation for Relevance. We propose relevance constraint to make the generated captions more accurate. To evaluate the effect of this relevance constraint, we collect the specific entities where the state-of-the-art method (SCN) fails (as described in the part of *Datasets and Evaluation Protocols*), based on which we find whether GroupCap can predict such difficult cases and refine the generated captions. We estimate the occurrence recall (*Occ. Recall*) of specific entities (*S. E.*) in the generated captions in Tab.4, where the occurrence recall of a specific entity can be computed as:

$$Occ.\ Recall = \frac{No.\ of\ Correct\ Occ.\ of\ S.\ E.}{No.\ of\ Ground\ Truths\ with\ S.\ E.}, \quad (14)$$

where we count *Ground Truths with S. E.* when *S. E.* simultaneously occurs in all the ground-truth captions of an image. From Tab.4, we find that GroupCap can describe the captions more accurately compared to the StructCap [24] and SCN [5], which validates the effect of the structured relevance constraint.

We further explore the relevance captured by the proposed model. We visualize the relevance matrices of some examples with high relevance scores in Fig.5. The value of each element in the relevance matrix means the confidence score of the relevance between two corresponding semantic items (entities and relations) in two VP-Trees. We can find that the alignments of relevant semantic items appear darker color, and the gradation distribution of color is generally consistent to node alignments of the textual parsing trees (green boxes Fig.5). It reflects that the structured relevance among semantic items is well captured by the proposed relevance constraint.

Evaluation for Diversity. We further evaluate the quality of diversity captured by the proposed model. We visualize the diversity matrices of some examples with high

[6]For distinction, we call our VP-Tree and the original visual parser [24] as gVP and sVP, respectively

Table 4. Occurrence recalls (Values in %) of specific entities (with top-10 frequencies in ES-dataset) in generated captions.

	road	*table*	*person*	*food*	*ball*	*ski*	*phone*	*water*	*dog*	*track*
StructCap	8.0	81.8	100.0	70.0	20.0	75.0	73.3	57.1	76.9	90.0
SCN	6.0	84.8	75.0	70.0	20.0	75.0	73.3	42.9	76.0	70.0
GroupCap	14.0	90.9	100.0	80.0	30.0	87.5	80.0	85.7	77.7	90.0

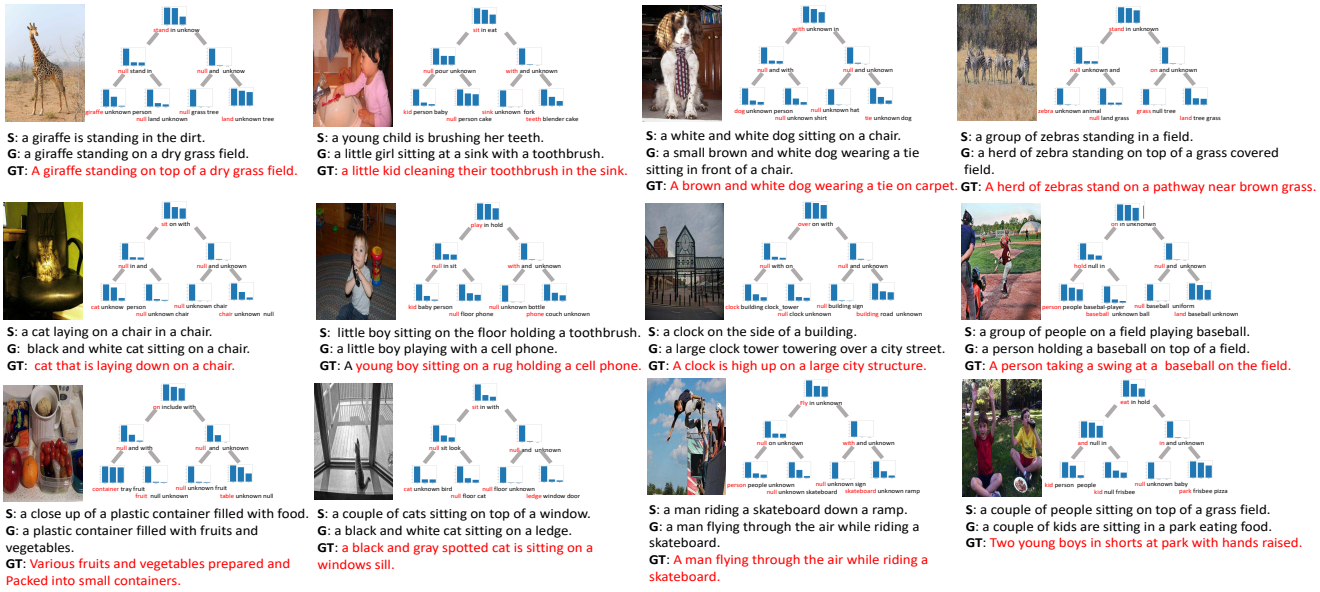

Figure 7. Visualization of the constructed VP-Trees and generated captions on FG-dataset. We compare the captions generated by the proposed GroupCap model to that of the state-of-the-art model (SCN) [5]. "S", "G", and "GT" denote SCN, GroupCap, ground truth, respectively.

diversity scores/probabilities in Fig.6. Different from relevance matrix, the value of each element in the diversity matrix represents the confidence score on whether two corresponding semantic items in two VP-Trees are first relevant and then diverse. We can find that the alignment of the diverse semantic items appears darker color, *e.g.*, the alignment of *black-jacket* and *gray-jacket* in the first example is with higher score compared to others. Moreover, the gradation distribution of color is generally consistent to node alignments of the textual parsing trees (green boxes in Fig.6), which reflects that the proposed diversity constraint can well capture the diversity among semantic items.

Evaluation for Caption Generation. Finally, we qualitatively evaluate our proposed GroupCap model in Fig.7. As we can see, the generated captions by GroupCap are more accurate and more discriminative compared to the state-of-the-art, which are also consistent with the ground truths. Moreover, the VP-Tree is mostly consistent with the image and the generated caption, which reveals the effect of the proposed joint training in Subsec.3.3.

5. Conclusion

In this paper, we propose a novel group-based image captioning model (GroupCap) by modeling relevance and diversity among group images for discriminative caption generation. Specifically, we first propose a visual parsing (VP) model to extract visual semantic items (entities and relations) and model their correlations, forming a tree structure. Then, we model the structured relevance and diversity among images via comparing between such tree structures. Finally, we embed the VP-Tree into the LSTM-based captioning model for the caption generation. In offline optimization, we further give an end-to-end formulation, which jointly trains the visual tree parser, the structured relevance and diversity constraints, and the LSTM based captioning model. Two group captioning datasets derived from MS-COCO are further released, serving as the first of its kind. Extensive experimental evaluations show that our model achieves state-of-the-art performance under several standard evaluation metrics.

6. Acknowledgements

This work is supported by the National Key R&D Program (No.2017YFC0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125).

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1, 2, 5

[2] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1, 2, 5

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015. 1, 5

[4] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016. 1, 2, 5, 6

[5] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[6] Amir Sadovnik, Yi-I Chiu, Noah Snavely, Shimon Edelman, and Tsuhan Chen. Image description with a goal: Building efficient discriminating expressions for images. In *CVPR*, 2012. 1, 2, 3

[7] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *CVPR*, 2017. 1, 2, 3

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*. IEEE, 2017. 2, 5, 6

[9] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 2

[10] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *ICCV*, pages 2533–2541, 2015. 2

[11] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[12] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. Semantic regularisation for recurrent image annotation. In *CVPR*, 2017. 2

[13] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *CVPR*, 2017. 3

[14] Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. Diverse image captioning via grouptalk. In *IJCAI*, 2016. 3

[15] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *CVPR*, 2017. 3

[16] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 6

[17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 3

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 3

[19] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 3

[20] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 3

[21] Amir Sadovnik, Andrew Gallagher, and Tsuhan Chen. Not everybody's special: Using neighbors in referring expressions with uncertain attributes. In *CVPR Workshops*, pages 269–276, 2013. 3

[22] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017. 3

[23] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *CVPR*, 2017. 3

[24] Fuhai Chen, Rongrong Ji, Jinsong Su, Yongjian Wu, and Yunsheng Wu. Structcap: Structured semantic embedding for image captioning. In *ACM MM*, pages 46–54, 2017. 3, 4, 5, 6, 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[26] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, pages 129–136, 2011. 3, 5

[27] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[28] Steven Bird. Nltk: the natural language toolkit. In *COLING/ACL*, pages 69–72, 2006. 5

[29] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634, 2015. 5