# Context-Aware Group Captioning via Self-Attention and Contrastive Features

Zhuowan Li[1]*, Quan Tran[2], Long Mai[2], Zhe Lin[2], and Alan Yuille[1]
[1]Johns Hopkins University [2]Adobe Research
{zli110, alan.yuille}@jhu.edu {qtran, malong, zlin}@adobe.com

## Abstract

*While image captioning has progressed rapidly, existing works focus mainly on describing single images. In this paper, we introduce a new task, context-aware group captioning, which aims to describe a group of target images in the context of another group of related reference images. Context-aware group captioning requires not only summarizing information from both the target and reference image group but also contrasting between them. To solve this problem, we propose a framework combining self-attention mechanism with contrastive feature construction to effectively summarize common information from each image group while capturing discriminative information between them. To build the dataset for this task, we propose to group the images and generate the group captions based on single image captions using scene graphs matching. Our datasets are constructed on top of the public Conceptual Captions dataset and our new Stock Captions dataset. Experiments on the two datasets show the effectiveness of our method on this new task.* [1]

## 1. Introduction

Generating natural language descriptions from images, the task commonly known as image captioning, has long been an important problem in computer vision research [3, 15, 29]. It requires a high level of understanding from both language and vision. Image captioning has attracted a lot of research attention in recent years thanks to the advances in joint language-vision understanding models [1, 19, 39, 54]. While image captioning has progressed rapidly, existing works mostly focus on describing individual images. There are practical scenarios in which captioning images in group is desirable. Examples include summarizing personal photo albums for social sharing or understanding web user intention from their viewed or clicked images. Moreover, it is often the case that the target image group to be captioned naturally belongs to a larger set that provides the context. For instance, in text-based image retrieval applications, given a group of user-interested images and other images returned by the search engine, we could predict the user hidden preferences by contrasting the two groups and suggest a new search query accordingly. Figure 1 shows an example of such scenario. Among all the images returned by search query `woman`, the user can indicate his/her interest in some of the images (in orange boxes). The objective is to recognize that the user wants `woman with cowboy hat` and suggest the query accordingly.


Figure 1. Context-ware group captioning. Given a group of target images (shown in orange boxes) and a group of reference images which provide the context (`woman`), the goal is to generate a language description (`woman with cowboy hat`) that best describes the target group while taking into account the context depicted by the reference group.

Inspired by these real-world applications, we propose the novel problem of *context-aware group captioning*: given a group of target images and a group of reference images, our goal is to generate a language description that best describes the target group in the context of the reference group. Compared to the conventional setting of single-image based captioning, our new problem poses two fundamental requirements. First, the captioning model needs to effectively summarize the common properties of the image groups. Second, the model needs to accurately describe the distinguish-

*This work has been done during the first author's internship at Adobe.
[1]Related Datasets and code are released at https://lizw14.github.io/project/groupcap.

ing content in the target images compared to the reference images.

To address those requirements, we develop a learning-based framework for context-aware image group captioning based on self-attention and contrastive feature construction. To obtain the feature that effectively summarizes the visual information from the image group, we develop a group-wise feature aggregation module based on self-attention. To effectively leverage the contrastive information between the target image group and the reference images, we model the context information as the aggregated feature from the whole set and subtract it from each image group feature to explicitly encourage the resulting feature to capture the differentiating properties between the target image group and the reference image group.

Training our models requires a large number of image groups with text descriptions and associated reference image sets. In this paper, we leverage large-scale image caption datasets to construct the training data. In particular, we build our annotations on top of the Conceptual Captions [40], a recently introduced large-scale image captioning dataset. We parse the single-image caption into scene graphs and use the shared scene graphs of image groups to generate the groups' ground-truth captions. In addition, we apply the same procedure on a large-scale image set collected from a photograph collection. This dataset contains a large number of images with compact and precise human-generated per-image descriptions. That results in our second dataset, *Stock Captions*, which we plan to contribute to the research community to encourage future research in this new problem.

Our main contributions in this paper are three-fold. First, we introduce the problem of context-aware group captioning. This novel image captioning setting can potentially be important for many real-world applications such as automatic query suggestion in image retrieval. Second, we present a learning-based approach which learns to aggregate image group visual feature for caption generation. Our framework combines the self-attention mechanism with contrastive feature construction to effectively encode the image group into a context-aware feature representation, which effectively summarizes relevant common information in the groups while capturing discriminative information between the target and context group. Third, we introduce two large-scale datasets specifically for the context-aware group captioning problem. Experiments on the two datasets demonstrate that our model consistently outperforms various baselines on the context-based image group captioning task.

## 2. Related Work

**Image captioning** has emerged as an important research topic with a rich literature in computer vision [3, 15, 29]. With the advances in deep neural networks, state-of-the-art image captioning approaches [1, 12, 17, 19, 36, 39, 50, 56] are based on the combination of convolutional neural networks [24] and recurrent neural networks [14] (CNN-RNN) architecture, where the visual features are extracted from the input image using CNNs which is then decoded by RNNs to generate the language caption describing the given image. Research in image captioning has progressed rapidly in recent years. Novel network architectures [1, 6, 32, 51], loss functions [7, 28, 30, 33, 39, 41], and advanced joint language-vision modeling techniques [18, 21, 32, 54, 55, 57] have been developed to enable more diverse and discriminative captioning results. Recent works have also proposed to leverage the contextual and contrastive information from additional images to help generating more distinctive caption for the target image [2, 5, 8, 48] or comparative descriptions between image pairs [38, 43, 44, 46]. Existing works, however, mostly focus on generating captions for a single image. Our work, on the other hand, focuses on the novel setting of context-based image group captioning which aims to describe a target image group while leveraging the context of a larger pool of reference images.

**Referring expression generation** [20, 34, 59] is a related problem to image captioning, which aims to generate natural language descriptions for a target object in an image. Contrastive modeling has been successfully applied in state-of-the-art referring expression generation methods to describe the target image region in contrast with other image regions. Yu *et al*. [58] use relative location and feature difference to discriminate the target object. Mao *et al*. [35] maximize the probability of generated expression describing a specific region over other regions by Maximum Mutual Information training. While referring expression generation considers the target region in contrast with each negative region respectively, our problem requires contrastive context modeling among and between image groups.

**Attention mechanism** has been successful in image captioning [6, 27, 32, 54, 57]. These works focused on applying visual attention to different spatial regions at each text generation time step. More recently, attention in transformer[47] and pretrained BERT[11] has been very successful in natural language processing tasks. [25, 31, 45] adapted the idea of BERT to vision and language tasks and showed improved performance on multiple sub-tasks. These works focus on learning attention between every word token. In our work, we apply attention over images and show its effectiveness for summarizing information in an image group.

Our setting is inspired by **query suggestion** [9, 16, 42, 53] in the context of document retrieval systems. Query suggestion aims to predict the expanded query given the previous query used by the users while taking into account

additional context such as search history [9, 16, 42] or user interaction (e.g. clicked and skipped documents) [53]. We are inspired by this task formulation and extend it to vision domain. Earlier works on query suggestion in image search focus on forming visual descriptors to help obtain better search results [60, 61] while the suggested text query is obtained solely from the current user query without taking visual content understanding into account. Our work, on the other hand, can potentially be applied to enable query suggestion from images. In this work, we focus on the image captioning aspect without relying on modeling user information and behavior as in existing query suggestion works, thus making it applicable beyond retrieval tasks.

## 3. Dataset

To train our models, we need a large-scale dataset where each data sample contains a group of target images with an associated ground-truth description and a larger group of reference images. The reference images need to be relevant to target images while containing a larger variety of visual contents and thus provides context for describing target images. The description should be both specific to the target group and conditioned on the reference group.

In this section, we first describe the intuition and method for dataset creation, then provide details of our proposed datasets built on the Conceptual Captions dataset and our proposed Stock Captions dataset.

### 3.1. Data Construction Method

We build our dataset on top of large-scale per-image captioning datasets by leveraging the shared scene graphs among images, motivated by [5]. The overall data generation process is shown in Figure 2.

Images with shared scene graphs compose an image group. More specifically, images with the same *(attribute)-object-relationship-(attribute)-object* are chosen to compose the target image group, while images with partially overlapping scene graphs with the target group are chosen as the reference image group. For example, as in Figure 2, images with the scene graph *woman in chair* are selected to form the target group, while images containing *woman* are selected to form the reference group paired with the target group. In this way, the reference group contains a larger variety of contents (woman in any places or poses) while the target group is more specific in terms of certain attributes (in chair).

In order to get the scene graphs for each image to support our grouping process, we use a pretrained language parser (improved upon [52]) to parse each ground-truth per-image caption into a scene graph. We choose to parse the scene graph from image captions instead of using the annotated scene graph in Visual Genome dataset [23] because our scene graph needs to focus on the most "salient" content in the image. Since Visual Genome is densely annotated without the information of which object is the main content of the image, scene graphs of small trivial objects may dominate the grouping process while the main content is ignored. This will produce very noisy data, potentially unsuitable for training our models. On the other hand, while parsing errors may introduce noise, scene graphs parsed out of image captions focus on the main objects because the caption usually describes the most important contents in an image.

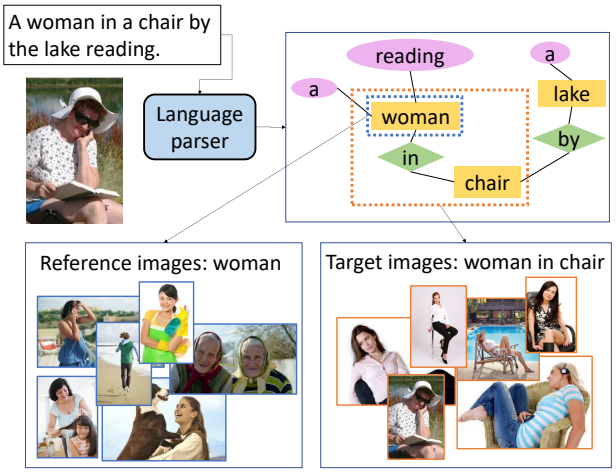


Figure 2. Dataset construction method. Our datasets are constructed from image collections with per-image descriptions. A pretrained language parser is used to parse each image caption into a scene graph. Then the images with shared scene graph are grouped to form the target group. Images with scene graphs that partially match the targets' form the reference group.

After getting the target and reference groups using scene graph matching, the shared scene graph among target images is flattened into text to serve as the ground truth group description. For example, in Figure 2, the ground-truth group caption is *woman in chair*. Other examples of ground-truth group captions include: *colorful bag on white background, girl in red, business team holding terrestrial globe, woman with cowboy hat, etc.*

To construct our datasets for group captioning, the per-image captioning datasets need to be large-scale to provide enough image groups. We build our group captioning datasets on top of two datasets: Conceptual Captions dataset [40], which is the largest existing public image captioning dataset, and Stock Captions dataset, which is our own large-scale per-image captioning dataset characterized by precise and compact descriptions. Details about construction on the two datasets are provided as follows.[2]

[2]For simplicity, in this paper, we call our newly constructed group captioning datasets by the same name as their parent datasets: Conceptual Captions, and Stock Captions.

### 3.2. Conceptual Captions

Conceptual Captions is a large-scale image captioning dataset containing 3.3 million image-caption pairs. (By the time we download the images through the urls provided, only 2.8 million are valid.) Because the captions are automatically collected from alt-text enabled images on the web, some of the captions are noisy and not natural. However, the high diversity of image contents and large number of images makes Conceptual a suitable choice for data generation using our method.

After sampling from 2.7 million images from Conceptual Captions, we obtain around 200k samples with 1.6 million images included. Each sample contains 5 target images and 15 reference images. The images with rare scene graphs that cannot be made into groups are not used. We manually cleaned the sampled data to remove samples that are not meaningful. For example, target group of *portrait or woman* and reference group of *woman* are not semantically different so they are removed. We also cleaned the vocabulary to remove rare words.

The 200k samples are split into test, validation and train splits, where these three splits share the same image pool. While the validation and train splits may contain samples of same group captions (because group captions are usually short), we make sure that captions in test split do not overlap with train split. More detailed statistics are provided in Table 1.

### 3.3. Stock Captions

While the Conceptual dataset excels in image diversity, we found that its captions are often long and sometime noisy. Motivated by the query suggestion application where the suggested search queries are usually short and compact, we propose to construct the dataset on a new image captioning dataset named Stock Captions. Stock Captions is a large-scale image captioning dataset collected in text-to-image retrieval setting. Stock Captions dataset is characterized by very precise, short and compact phrases. Many of the captions in this dataset are more attribute-like short image titles, e.g. "colorful bags", "happy couple on a beach", "Spaghetti with dried chilli and bacon", etc.

After grouping and filtering the 5.8 million raw images, we get 1.9 million images, grouped into 1.5 million data samples for the Stock Captions dataset. The dataset sampling and split details are similar to Conceptual.(Table 1).

| **Original Per-Image Captioning** | | |
|---|---|---|
| | Conceptual | Stock |
| Size | 2766614 | 5785034 |
| Avg Length | 9.43 | 4.12 |
| **Context-aware Group Captioning** | | |
| | Conceptual | Stock |
| Size | 199442 | 146339 |
| Train Split | 175896 | 117829 |
| Val Split | 10000 | 10000 |
| Test Split | 13546 | 18510 |
| # of images | 1634523 | 1941370 |
| Vocab Size | 5811 | 2437 |
| Avg Length | 3.74 | 2.96 |

Table 1. Statistics of Conceptual Captions and Stock Captions, in terms of original per-image captioning dataset and our group captioning dataset constructed on top of per-image captions.

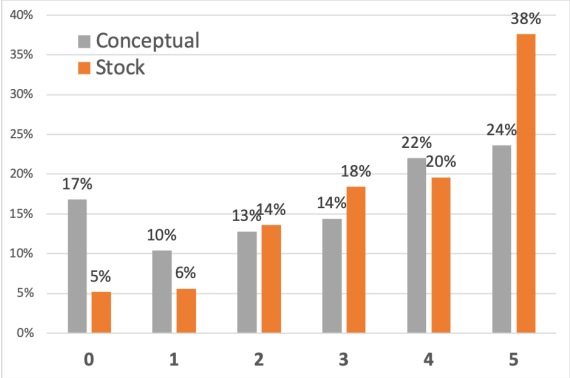


Figure 3. Distribution of human-given scores for our two constructed datasets. Dataset constructed on Stock Captions gets higher human scores.

### 3.4. User Study for Dataset Comparisons

To test the quality of our data and compare our two datasets, we conduct a user study by randomly selecting 500 data samples (250 from each dataset) and ask 25 users to give a 0-5 score for each sample.

To better compare the two datasets, we ask the users to give strict scores. A caption needs to be precise, discriminative and natural to be considered good. Many captions with the score of 0 and 1 are semantically good, but are unnatural. The distribution of scores is shown in Figure 3. As expected, in overall quality, the Stock Captions data scores significantly higher as it is based on compact and precise human-generated captions. However, several users do note that the captions in the Conceptual Captions dataset seems to be more specific, and "interesting".

## 4. Method

In this section, we explore methods to address the two main challenges in our proposed problem: a) **feature aggregation**, *i.e.* how to summarize the images within one image group, and (b) **group contrasting**, *i.e.*, how to figure out the difference between two image groups. By comparing different methods, our goal is not only finding the best

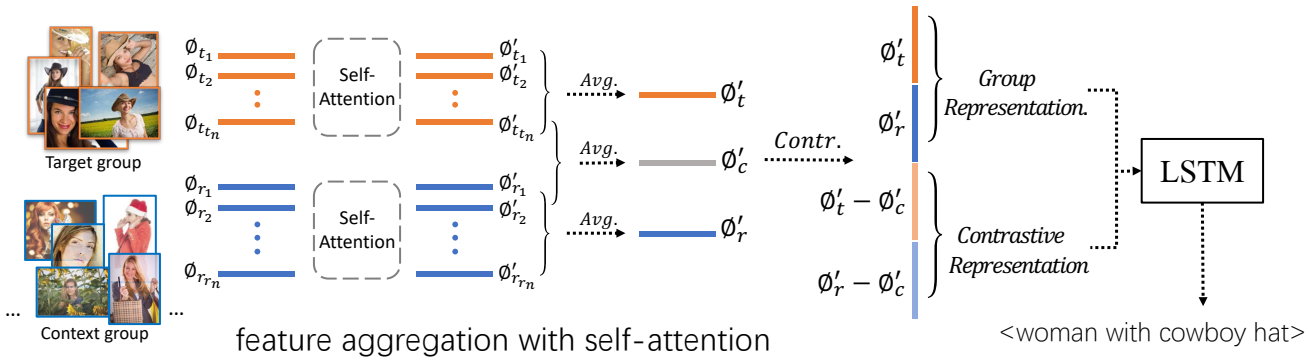


Figure 4. Context-aware group captioning with self-attention and contrastive features. Image features are aggregated with self-attention to get the group representation for each image group. Then the group representation is concatenated with contrastive representation to compose the input to LSTM decoder, which finally generates context-aware caption for the target image group.

performing models, but also drawing insights into the characteristics of the task, and hopefully, setting the focus for future exploration in this problem.

To begin the section, we first formalize the problem settings in Section 4.1. In the subsequent sub-sections, we describe our method explorations path starting with a simple baseline. We then gradually introduce more computationally specialized modules. For each module, we describe our intuition and back them up with quantitative results and visual illustrations.

### 4.1. Problem Setting

Given a group of $n_t$ target images and a group of $n_r$ reference images, our task is to generate a description $D = < \hat{w}_1, ..., \hat{w}_l >$ to describe the target image group in context of the reference group. Here $\hat{w}_i$ denotes the word in the sentence and $l$ is sentence length, which varies for each data sample. In our setting, $n_t = 5, n_r = 15$.

Each image is represented by a 2048-d feature extracted using the ResNet50 network [13] (after pool5 layer), pretrained on ImageNet [10]. The input of our model are the target features $\Phi_t = [\phi_t^1, ..., \phi_t^{n_t}]$ and the reference features $\Phi_r = [\phi_r^1, ..., \phi_r^{n_r}]$, where $\phi^i \in \mathbb{R}^{2048}$. We use $\Phi$ to denote a list of features, while a single feature is denoted as $\phi$.

While we believe that more detailed features (e.g. spatial features without mean-pooling, or object-level features) may improve performance, they increase the computational complexity, and by extension, the training time to an unacceptably high level in our initial testing. Thus, we simply use the mean-pooled feature vector.

### 4.2. Baseline: feature averaging and concatenation

From the problem setting above, one intuitive approach would be to summarize the target and reference features by averaging, and concatenating them to create the final feature for description generation. The process can be formalized as follows.

We compute the target group feature $\phi'_t$ and the reference group feature $\phi'_r$ by averaging the features in each group:

$$\phi'_t = \frac{1}{n_t} \sum_{i \in 1..n_t} \phi_{t_i} \qquad \phi'_r = \frac{1}{n_r} \sum_{i \in 1..n_r} \phi_{t_r}$$

Following standard captioning pipeline, we then use the concatenation of the two group features as input to LSTM to predict the context-aware descriptions. We use LSTM-RNN [14] to generate the caption in an auto-regressive manner. Denoting the output of the LSTM module at time step $t$ as $h_t$, we have the equations for decoding:

$$\begin{aligned} h_1 &= [\phi'_t, \phi'_r] \\ h_t &= \text{LSTM}(h_{t-1}, \hat{w}_{t-1}) \\ \hat{w}_t &\sim \text{softmax}(h_t). \end{aligned}$$

Finally, we follow standard beam search process to generate the captions. This decoding architecture is used in all of our subsequent model variants.

### 4.3. Feature aggregation with self attention

While the average-pooling method used for feature aggregation above is intuitive, it treats all image features equally. We note that many groups of images have prominent members that encapsulate the joint information of the whole groups (Figure 5). We argue that the group summarizing process could be improved if we can identify these prominent features/images. Motivated by that observation, we propose to use the transformer architecture [47] for this task. The transformer relies on a grid of attention between the elements of the set to learn a better representation. Intuitively, by learning the self-attention grid, the model can detect the prominent features as each element in the set can "vote" for the importance of the other elements through the

attention mechanism. In the subsequent analysis, we show that, in our task, the self-attention gird indeed puts a lot more weights to the prominent images. The core computations of our transformer-based architecture can be summarized as follows.[3]

The first step is calculating the contextualized features using self-attention mechanism. Given the input features $\Phi$; three different sets of features: queries $Q$, keys $K$ and values $V$ are calculated using a linear transformation:

$$Q = W^Q \Phi + b^Q$$
$$K = W^K \Phi + b^K$$
$$V = W^V \Phi + b^V$$

Then the self-attention grid is calculated by a scaled dot product between $Q$ and $K$ (the scaling factor $d$ is the dimension of the vectors in $Q$ and $K$). The self-attention layer uses this attention grid and the value matrix $V$ to compute its outputs.[4]

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

The self-attention output is then coupled with the residual signal to create the contextualized features $\Phi'$.

$$V' = V + \text{Attention}(Q, K, V)$$
$$\Phi' = V' + \max\left(0, V'W_1 + b_1\right) W_2 + b_2$$

From this point, we denote the process of transforming from the original features set $\Phi$ to the contextualized feature set $\Phi'$ as $\Phi' = F(\Phi)$. With this formulation, we have the contextualized set of features $\Phi'_t$ and $\Phi'_r$:

$$\Phi'_t = F_{st}(\Phi_t) \qquad \Phi'_r = F_{sr}(\Phi_r)$$

We tried both sharing and not-sharing weights of $F_{st}$ and $F_{sr}$, and found that sharing weights lead to slightly better performance. This is intuitive as the task of grouping target images are not different from the task of grouping reference images, and thus, the grouping model can share the same set of weights.

In our experiments, the self-attention architecture provides a significant boost in performance compared to the average-pooling variant.

### 4.4. Group contrasting with contrastive features

The second major challenge in our proposed problem is the image group contrasting. With the aforementioned self-attention mechanism, we have good representations for the target and reference groups. The most intuitive way to learn the difference between the two features is either concatenation (which is implemented in our baseline) or feature subtraction.

We argue that, to learn the difference between two groups of images, we first need to capture their similarity. Our hypothesis is that, when we identify the similarity between all the images, we can "remove" this similarity portion from the two features to deduce more discriminative representation. This process is formalized as follows.

The first step is learning the common information $\phi'_c$ between all the images. We do that by applying the same self-attention mechanism described above to all the images.

$$\Phi'_c = F_a([\Phi_t; \Phi_r])$$
$$\phi'_c = \frac{1}{n_t + n_r} \sum \Phi'_c$$

Then the joint information is "removed" from the group features $\phi'_t$ and $\phi'_r$ by subtraction to generate the contrastive/residual feature $\phi^d_t$ and $\phi^d_r$.

$$\phi^d_t = \phi'_t - \phi'_c \qquad \phi^d_r = \phi'_r - \phi'_c$$

The contrastive features $\phi^d_t$ and $\phi^d_r$ are concatenated with the group features $\phi'_t$ and $\phi'_r$, which are then fed into LSTM-RNN to generate captions. In our subsequent analysis, we show that the contrastive features indeed focus on the difference between two image groups.

## 5. Experiments

In this section, we first describe our evaluation results on the two datasets. Then we provide quantitative analysis and visualization to expose the effectiveness of different components of our model.

### 5.1. Group Captioning Performance

We evaluate our context-aware group captioning method on both Conceptual Captions and Stock Captions datasets. The same hyper-parameters are used for all experiments on each dataset. On the Stock Captions dataset, we use batch size 512 and initial learning rate $1 \times 10^{-4}$. On the Conceptual Captions dataset, we use batch size 512 and learning rate $5 \times 10^{-5}$. We train the model for 100 epochs with Adam optimizer[22] on both datasets.

We measure the captioning performance on the test splits in both datasets using a variety of captioning metrics. Specifically, we consider the standard metrics widely used in image captioning literature, including BLEU[37], CIDER[49], METEOR[4] and ROUGE[26]. In addition, since group descriptions are often short and compact, we put more emphasis on single word accuracy compared to traditional image captioning. We thus consider two additional metrics, Word-by-word accuracy(WordAcc), word

[3]We only describe the core computation steps of the self-attention due to space constraint and to improve clarity. More details can be found in the original paper [47]. We also release our implementation if accepted.

[4]We don't use the multi-head attention in this work, as in our preliminary experiments, the multi-head attention provides no performance gain compared to a single head.

| | WordAcc | CIDER | WER | BLEU1 | BLEU2 | METEOR | ROUGE |
|---|---|---|---|---|---|---|---|
| Conceptual | | | | | | | |
| Per-Img. Caption | 5.4638 | 0.4671 | 2.6587 | 0.1267 | 0.0272 | 0.0868 | 0.1466 |
| Average | 36.7329 | 1.9591 | 1.6859 | 0.4932 | 0.2782 | 0.3956 | 0.4964 |
| SA | 37.9916 | 2.1446 | 1.6423 | 0.5175 | 0.3103 | 0.4224 | 0.5203 |
| Average+Contrast | 37.8450 | 2.0315 | 1.6534 | 0.5007 | 0.2935 | 0.4057 | 0.5027 |
| **SA+Contrast** | **39.4496** | **2.2917** | **1.5806** | **0.5380** | **0.3313** | **0.4405** | **0.5352** |
| Stock | | | | | | | |
| Per-Img. Caption | 5.8931 | 0.3889 | 1.8021 | 0.1445 | 0.0359 | 0.0975 | 0.1620 |
| Average | 37.9428 | 1.9034 | 1.1430 | 0.5334 | 0.2429 | 0.4042 | 0.5318 |
| SA | 39.2410 | 2.1023 | 1.0829 | 0.5537 | 0.2696 | 0.4243 | 0.5515 |
| Average+Contrast | 39.1985 | 2.0278 | 1.0956 | 0.5397 | 0.2632 | 0.4139 | 0.5375 |
| **SA+Contrast** | **40.6113** | **2.1561** | **1.0529** | **0.5601** | **0.2796** | **0.4332** | **0.5572** |

Table 2. Group captioning performance on the Conceptual Captions and Stock Captions dataset.

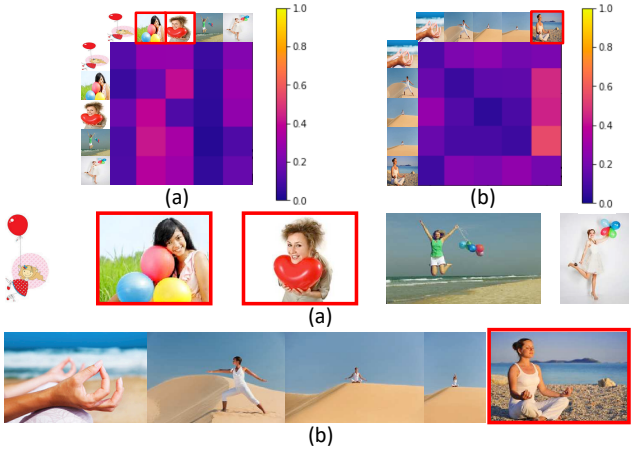


Figure 5. Visualization of $5 \times 5$ self-attention weight matrix for target image group. Each row sums up to 1. For group (a) `woman with balloon`, image 2 and image 3 are representative. For group (b) `yoga on beach`, image5 is representative. Images with more distinguishable features become the representative images of a group and get higher attention weights.

| Model | WordAcc | CIDER | BLEU2 | METEOR | ROUGE |
|---|---|---|---|---|---|
| Tgt0 + Ref15 | 24.4709 | 1.0399 | 0.0614 | 0.2341 | 0.3965 |
| Tgt1 + Ref15 | 28.7479 | 1.3447 | 0.1292 | 0.2938 | 0.4415 |
| Tgt3 + Ref15 | 34.6574 | 1.7641 | 0.2098 | 0.3698 | 0.5048 |
| Tgt5 + Ref0 | 31.8061 | 1.6767 | 0.2095 | 0.3475 | 0.4552 |
| **Tgt5 + Ref15** | **40.6113** | **2.1561** | **0.2796** | **0.4332** | **0.5572** |

Table 3. Performance with varying the number of target and reference images. (evaluated on Stock Captions dataset)

error rate(WER), that specifically assess word-based accuracy[5]. We also note that as some group descriptions may contain as few as two words, we do not consider BLEU3 and BLEU4 scores which evaluates tri-grams and 4-grams.

The captioning performance on the testing set of Conceptual Captions and Stock Captions datasets are reported in Table 2. To compare with a simple baseline, we caption each image individually and summarize them using our dataset building method. The result (**Per-Img. Caption**) shows that the group captioning problem cannot be solved by simply summarizing per-image captions. More details are shown in supplementary materials. Compared to aggregating features by averaging (**Average**, as in Section 4.2), self-attention (**SA**) is more effective in computing group representation and leads to significant performance improvement. On top of feature aggregation, contrastive feature is critical for the model to generate context-aware caption which emphasizes the difference of target image group on context of reference group. Applying contrastive features (**Contrast**) to either feature aggregation methods leads to performance boost (**Average+Contrast, SA+Contrast**). To this end, our full model, which combines self-attention for group aggregation and contrastive feature for group comparing performs best, achieving 39.4% WordAcc on Conceptual Captions and 40.6% on Stock Captions.

## 5.2. Discussion

**Effectiveness of self-attention on feature aggregation.** To better understand the effectiveness of self-attention, in Figure 5, we visualize the $5 \times 5$ self-attention weight matrix between 5 target images. The i-th row of the attention matrix represents the attention weights from i-th image to each of the 5 images, which sum up to 1. In (a), images with larger and centered balloons (Image2 and Image3) gets higher attention. In (b), image5 where the woman doing yoga is larger and easier to recognize gets higher attention. In both examples, images with more recognizable features get higher attention weights and thus contribute more to the aggregated group representation.

**Importance of multiple target and reference images.** To investigate the effectiveness of giving multiple images in each group, we vary the number of target and reference images. Results are shown in Table 3. Fewer target or reference images results in performance decline, which indi-

[5]Here we consider position-specific word accuracy. For example, prediction *woman with straw hat* with ground truth *woman with cowboy hat* has accuracy 75%, while prediction *woman with hat* has accuracy 50%.

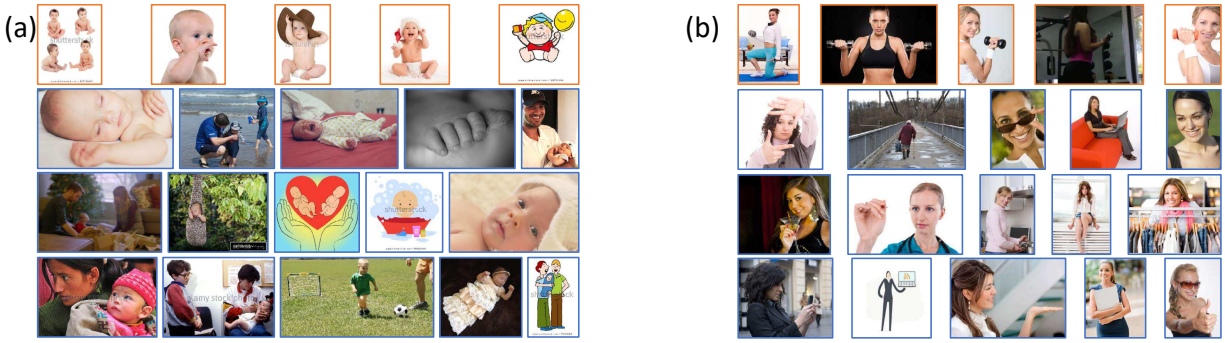


Figure 6. Qualitative prediction examples on Conceptual Captions (a) and Stock Captions (b) datasets. In each example, images in first row (in orange boxes) are target images while second to fourth rows (in blue boxes) are reference images. Our model can effectively summarize relevant information in the image groups during captioning. Our model also effectively takes discriminative information between the target and reference group into account during captioning to predict accurate group captioning results.

| Contrastive + Group | Group | Contrastive |
|---|---|---|
| woman with cowboy hat | woman | country with cowboy straw hat |
| white girl | girl | white rule white and... |
| woman in boxing glove | woman | is go in boxing... |

Table 4. Analysis of contrastive representation. Column `Contrastive + Group` is the prediction of our full model. Column `Group` and column `Contrastive` are the predictions when only the group or only the contrastive representation is fed into the decoder respectively. Blue text denotes the common part while red text denotes the contrastive part.

cates that a larger number of images is more informational for the model to get better descriptions. We also qualitatively study the importance of the reference image group. Examples are shown in Figure 6. The examples indicate that when not giving reference group the predictions tend to be more generic and less discriminative.

**Contrastive representation versus group representation.** Table 4 shows example descriptions when only the group representations or only the contrastive representations are fed into LSTM decoder. Although the model does not treat the features independently and removing the features might break the grammar structure of the caption, looking at the lexicons returned by the two variants, we can clearly observe the focus of two features. When the decoder uses only the group representations, the predictions emphasize the common part of two image groups. On the other hand, when the decoder only uses the contrastive representations, the predictions emphasize the difference between two image groups. This reveals that the group representation encodes similarity information, while the contrastive representation encodes discriminative information.

**Robustness to noise images.** To investigate the model's robustness to noise in the image group, we tried adding random unrelated images to the target group. Figure 7 shows performances of models trained and tested with different number (0-4) of noise images on Conceptual Captions dataset. Training with more noise increases robustness of the model but hinder performance when tested with no noise. The model shows robustness to small noise. Qualitatively, when testing with small (1 or 2) noise (trained with 0 noise), the caption loses details, e.g. woman in red dress becomes woman in dress. The generated caption is broken when the noise is severe, which is reasonable.

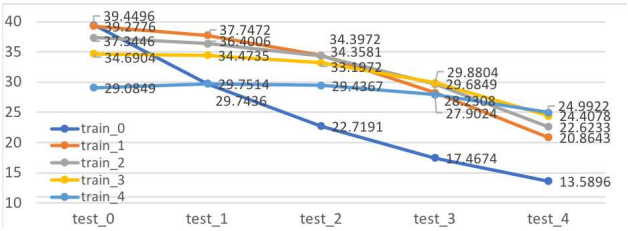


Figure 7. Performance change on Conceptual Captions dataset when trained and tested with 0-4 random images in the target group. Training with more noise increases robustness of the model but hinder performance when tested with no noise.

## 6. Conclusion

In this paper, we present the novel context-aware group captioning task, where the objective is to describe a target image group in contrast to a reference image group. To explore this problem, we introduce two large scale datasets, Conceptual Captions and Stock Captions respectively, both of which will be released for future research. We also propose a framework with self-attention for grouping the images and contrastive representation for capturing discriminative features. We show the effectiveness of our proposed model both quantitatively and qualitatively on our datasets. We also thoroughly analyze the behavior of our models to provide insights into this new problem.

## References


[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 2

[2] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016. 2

[3] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018. 1, 2

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[5] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1345–1353, 2018. 2, 3

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017. 2

[7] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 2

[8] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017. 2

[9] Mostafa Dehghani, Sascha Rothe, Enrique Alfonseca, and Pascal Fleury. Learning to attend, copy, and generate for session-based query suggestion. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1747–1756. ACM, 2017. 2, 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 5

[15] MD Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):118, 2019. 1, 2

[16] Jyun-Yu Jiang and Wei Wang. Rin: Reformulation inference network for context-aware query suggestion. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 197–206. ACM, 2018. 2, 3

[17] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018. 2

[18] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2

[19] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 1, 2

[20] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[21] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 2

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 3

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[27] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[28] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017. 2

[29] Xiaoxiao Liu, Qingyang Xu, and Ning Wang. A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3):445–470, 2019. 1, 2

[30] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 338–354, 2018. 2

[31] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[32] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2

[33] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018. 2

[34] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111, 2017. 2

[35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2

[36] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011. 2

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 6

[38] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4624–4633, 2019. 2

[39] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 1, 2

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 2, 3

[41] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 2

[42] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM, 2015. 2, 3

[43] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017. 2

[44] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 2

[45] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019. 2

[46] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689*, 2019. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 5, 6

[48] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260, 2017. 2

[49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[50] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[51] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997. ACM, 2016. 2

[52] Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. *arXiv preprint arXiv:1803.09189*, 2018. 3

[53] Bin Wu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Query suggestion with feedback memory network. In *Proceedings of the 2018 World Wide Web Conference*, pages 1563–1571. International World Wide Web Conferences Steering Committee, 2018. 2, 3

[54] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 1, 2

[55] Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2193–2202, 2017. 2

[56] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018. 2

[57] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[58] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 2

[59] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290, 2017. 2

[60] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang. Visual query suggestion. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 15–24. ACM, 2009. 3

[61] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Trans. Multimedia Comput. Commun. Appl.*, 6(3), August 2010. 3