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

Zhuowan Li¹, Quan Tran², Long Mai², Zhe Lin², and Alan Yuille¹
¹Johns Hopkins University ²Adobe 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.

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-
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.

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.\footnote{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).

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
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 \( D = \langle \hat{w}_1, ..., \hat{w}_l \rangle \) to describe the target image group in the 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), pre-trained on ImageNet \([10]\). The input of our model are the target features \( \Phi_t = [\phi_1^t, ..., \phi_{n_t}^t] \) and the reference features \( \Phi_r = [\phi_1^r, ..., \phi_{n_r}^r] \), 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:

\[
\phi_t' = \frac{1}{n_t} \sum_{i \in 1..n_t} \phi_t^i, \quad \phi_r' = \frac{1}{n_r} \sum_{i \in 1..n_r} \phi_r^i
\]

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:

\[
\begin{align*}
\hat{w}_t &= \text{softmax}(h_t) \\
h_t &= \text{LSTM}(h_{t-1}, \hat{w}_{t-1}) \\
\end{align*}
\]

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 grid 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 \(\Phi'\).

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

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 \(\Phi'_r\):

\[
\Phi'_t = F_{st}(\Phi_t) \quad \Phi'_r = F_{sr}(\Phi_r)
\]

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 \(\phi'\) between all the images. We do that by applying the same self-attention mechanism described above to all the images.

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

Then the joint information is “removed” from the group features \(\phi'_t\) and \(\phi'_r\) by subtraction to generate the contrastive/residual feature \(\phi''_t\) and \(\phi''_r\).

\[
\phi''_t = \phi'_t - \phi'_c \quad \phi''_r = \phi'_r - \phi'_c
\]

The contrastive features \(\phi''_t\) and \(\phi''_r\) 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\(^\text{[22]}\) 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\(^\text{[37]}\), CIDER\(^\text{[49]}\), METEOR\(^\text{[4]}\) and ROUGE\(^\text{[26]}\). 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
<table>
<thead>
<tr>
<th>Model</th>
<th>WordAcc</th>
<th>CIDER</th>
<th>WER</th>
<th>BLEU1</th>
<th>BLEU2</th>
<th>METEOR</th>
<th>ROUGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tgt0 + Ref15</td>
<td>24.4709</td>
<td>1.0399</td>
<td>0.0614</td>
<td>0.2341</td>
<td>0.3965</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tgt1 + Ref15</td>
<td>28.7479</td>
<td>1.3447</td>
<td>0.1292</td>
<td>0.2938</td>
<td>0.4415</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tgt3 + Ref15</td>
<td>34.6574</td>
<td>1.7641</td>
<td>0.2098</td>
<td>0.3698</td>
<td>0.5048</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tgt5 + Ref0</td>
<td>31.8061</td>
<td>1.6767</td>
<td>0.2095</td>
<td>0.3475</td>
<td>0.4552</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tgt5 + Ref15</td>
<td>40.6113</td>
<td>2.1561</td>
<td>0.2796</td>
<td>0.4332</td>
<td>0.5572</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Performance with varying the number of target and reference images. (evaluated on 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, images 2 and 3 are representative. For group (b) yoga on beach, image 5 is representative. Images with more distinguishable features become the representative images of a group and get higher attention weights.
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.

**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.

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


