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

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

1. Details of Datasets

There are six types of captions: subject-relation-object, adjective-object, noun-object, attribute-object-relation-object, object-relation-attribute-object, attribute-object-relation-attribute-object. Table 1 shows the number of samples in each type of captions.

	Conceptual	Stock
Sub-Rel-Obj	46810	40620
Adj-Obj	24890	33650
NN-Obj	18466	32774
Att-Sub-Rel-Obj	55124	19170
Sub-Rel-Att-Obj	30944	16683
Att-Sub-Rel-Att-Obj	23208	3442
Total	199442	146339

Table 1. Statistics of each caption type on Conceptual Captions and Stock Captions.

2. Experiments

2.1. Varying the Number of Reference Images

In Table 3 of the main paper, we give experiment results of varying the number of target and reference images. Here in Table 2 we give more detailed results of varying the number of reference images. As shown in the table, the performance improves when more reference images are given. We also notice that while the differences between giving 0, 5 or 10 references images are large, the gap between 10 and 15 reference images are not significant. So we use 15 reference images in the overall experiment setting.

	WordAcc	CIDER	WER	BLEU1	BLEU2	METEOR	ROUGE
Tgt5 + Ref0	31.8061	1.6767	1.2539	0.4600	0.2095	0.3475	0.4552
Tgt5 + Ref5	37.1283	1.9536	1.1413	0.5219	0.2503	0.3987	0.5185
Tgt5 + Ref10	39.4072	2.076	1.0923	0.5451	0.2684	0.4201	0.5424
Tgt5 + Ref15	40.6113	2.1561	1.0529	0.5601	0.2796	0.4332	0.5572

Table 2. Performance change when varying the number of reference images on Stock Captions dataset.

2.2. Variations of Contrastive Representation

In this subsection we show the experimental results of model variations we tried for contrasting the two image groups. The results of variation models are shown in Table 3.

*This work has been done during the first author's internship at Adobe.

2.2.1 The cross-attention models

Given the effectiveness of attention on grouping images, we tried applying attention to contrast two image groups. We investigate three different variants:

AttenAll: Applying self-attention between all the target and reference images simultaneously (we use two different fully-connected layers to differentiate target and reference). This variant decreases the performance over self-attention only. We hypothesize that treating two distinct relations : intra-group relations (which the model must focus on the similarity) and inter-group relations (which the model must focus on the difference) might not be the ideal solution. Thus, we develop the second variant, which treat these two relations separately: Cross attention (**CA**)

CA: In **CA**, we tried applying self-attention within each image group first and then cross-attention between two image groups. When doing cross-attention, we apply a mask to the self-attention kernel to remove attention connections within each image group and only keep connections between groups. This leads to slight improvement over **AttenAll**, but the performance is still behind the Self-attention only variant.

NCA: Going a step further, we experiment with the negative cross attention mechanism (**NCA**), which is to negate the reference image features before computing attention. The intuition is, by negating one group of features, two feature vectors that are close in the feature space will become distant. Thus, we want to force the to focus on the difference between the features, instead of the similarities. Negative cross attention improves the performance over **CA** but does not lead to consistent improvement of self-attention only.

From the experimental results, we hypothesize that the self-attention kernel is only effective in similarity detection, not in extracting the difference, even with the negative trick. However, if we consider two feature groups as two mathematical sets, and if we can detect the common elements between the two sets, we can just "remove" them from both sets and get the "difference" of the two sets. This intuition leads us to the development of the contrastive representation models. Our formulation in the main paper is the translation of this intuition in neural network language.

2.2.2 Variants of the contrastive representation model

We also tried different variants of contrastive representation. In the method part, we derive the contrastive representation by concatenating the difference of target and reference features with their joint information, i.e., $\phi^d = [\phi^d_t; \phi^d_r] = [\phi'_t - \phi'_c; \phi'_r - \phi'_c]$). Besides this variant, we also tried computing contrastive representation by taking difference of target and reference features, i.e., $\phi^d = \phi'_t - \phi'_r$ (**SA+Contrast1**) or taking difference between target features and joint features, i.e., $\phi^d = \phi^d_t = \phi'_t - \phi'_c$ (**SA+Contrast2**). Both methods improve performance over self-attention (**SA**) but the results are lower than our best method (**SA+Contrast**), which indicates the contribution of term ϕ^d_r and the advantage to minus the joint information of all images instead of minus reference features.

3. Comparison with Single Image Captioning

In this section, we describe the difference between our group captioning task and existing individual image captioning task. Captioning each image individually and then summarizing the per-image captions can not solve our task.

Figure 1 shows one example from Conceptual Captions and one from Stock Captions. The individual image captions are generated using existing image captioning models[1]. In each figure, the 20 captions on the right corresponds to the 20 images on the left in order, where the first 5 are targets and the other 15 are references.

In (a), while the image group is characterized by `man in black suit`, the individual captions focus on *man in dark*, *man with a gun*, *portrait of a man*, *man working on a laptop*, etc, thus summarizing them by finding the most frequent phrase will lead to `portrait of a young man`, which is not a good caption for the image group. In (b), while the image group features for woman in cowboy hat, individual captions focus on other aspects including *with a cup of tea*(this is an error of the captioning model), *beautiful*, *in the field* or *lying on bed*. Only one per-image caption notices that the woman is *in a hat*. Therefore, if we are summarizing the target per-image captions to get group caption, we will get result `young woman` or `beautiful woman`, which miss out the most important feature of the image group (`woman in cowboy hat`).

While individual captions might be able to describe each image discriminatively, they does not necessarily include the common properties of the image group, because the common property of the group might not be the significant and distinguishing feature for each image. Therefore, captioning images as a group can capture the information that individual image

[1]For Conceptual Captions, we use the winning model of Conceptual Captions Challenge Workshop in CVPR2019 to generate captions for each image (https://github.com/ruotianluo/GoogleConceptualCaptioning). More details of the model can be found at https://ttic.uchicago.edu/~rluo/files/ConceptualWorkshopSlides.pdf. For Stock Captions, we use the Show, Attend and Tell [1] captioning model and finetune it on Stock Captions

	WordAcc	CIDER	WER	BLEU1	BLEU2	METEOR	ROUGE
Conceptual							
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**
AttenAll	36.1231	2.0727	1.6851	0.5044	0.2976	0.4089	0.5059
SA+CA	36.2892	2.1282	1.6697	0.5041	0.3094	0.4145	0.5062
SA+NCA	37.6046	2.2109	1.6344	0.5155	0.3183	0.4237	0.5165
SA+Contrast1	38.2574	2.1499	1.6332	0.5213	0.3106	0.4228	0.5203
SA+Contrast2	38.5916	2.1821	1.6230	0.5218	0.3156	0.4261	0.5229
Stock							
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**
AttenAll	38.9215	2.0271	1.0904	0.5451	0.2578	0.4166	0.5428
SA+CA	38.6316	2.0414	1.0894	0.5440	0.2579	0.4139	0.5417
SA+NCA	39.3278	2.0833	1.0704	0.5490	0.2664	0.4207	0.5459
SA+Contrast1	39.9114	2.1006	1.0699	0.5553	0.2731	0.4271	0.5523
SA+Contrast2	40.2068	2.1115	1.0620	0.5537	0.2725	0.4262	0.5516

Table 3. Group captioning performance on the Conceptual Captions and Stock Captions dataset.

captioning tend to miss out and thus lead to more informative group captions. Therefore, captioning the group as a whole is different from processing each image individually and then summarizing them. This also explains why merging the image features in early stage using self-attention before generating text descriptions is beneficial.

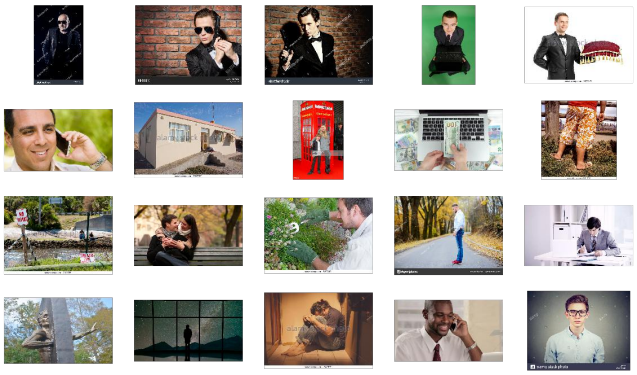
4. Analysis of contrastive representation

In Table 4 of the main paper, we show example results of the captions generated using only group representation or using only contrastive representation. Here in Figure 2 we show the images of these examples in Table 4. We also provide more examples to illustrate the function of group representation and contrastive representation. The first 3 examples are from Conceptual Captions dataset while the last 3 examples are from Stock Captions. Each example contains of 20 images (four rows), where the first row is target group and the second to fourth rows are reference group.

As shown, the common information in both image groups is encoded in the group representation, while the difference between two image groups is captured by the contrastive representation. The first four examples are good cases while the last two examples are failure cases. In failure case `woman in red glove`, the contrastive representation fails to capture the red information. In failure case `girl wearing white dress`, the color white is encoded in the contrastive representation, but its relationship with the girl is wrong in the prediction.

5. More Examples

Figure 3 and Figure 4 show more good examples on Conceptual Captions and Stock Captions respectively. Figure 5 and Figure 6 show failure cases on the two datasets respectively. Similar as above, in each example, the first row is target group while the other rows are reference group. Analysis for the failure cases (Figure 5, Figure 6) can be found in the captions of each figure.



Ground Truth Group Caption: man in black suit
Our Prediction: man with black suit

Individual Captions

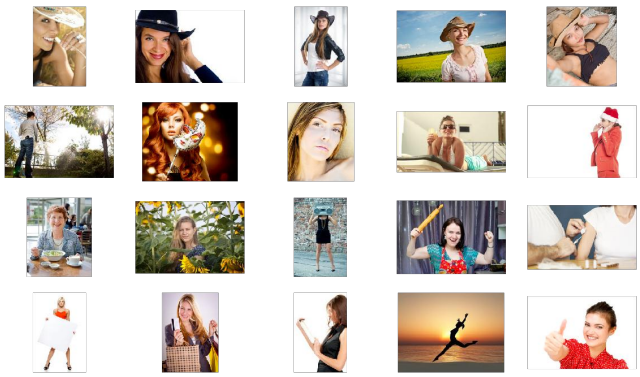
portrait of a man in the dark
portrait of a young man with a gun
portrait of a young man
businessman working on a laptop
man in a suit with a red umbrella

man with a smart watch
modern house built in the style
actors arrive at the premiere
a hand holding a mobile phone
a young couple standing in front of a tree

a road sign on a flooded road
couple in the city park
a bride and groom at their wedding
a man walking along a road
man working on a laptop

a close up of the statue
a silhouette of a man standing in front of a starry sky
a man looking at a painting
doctor working at the hospital
portrait of a young man

(a)



Ground Truth Group Caption: woman in cowboy hat
Our Prediction: woman with cowboy hat

young woman with a cup of tea.
portrait of a beautiful woman.
girl in a hat.
beautiful girl in the field.
young woman lying on bed.

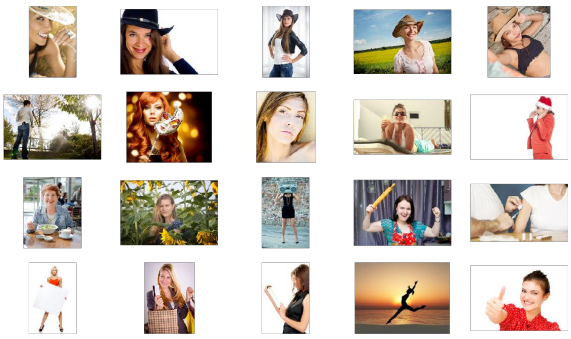
woman in the park.
young woman eating a cake.
portrait of a beautiful girl.
young woman with a laptop in the gym.
young woman in red dress with red hat.

young woman eating salad in a cafe.
beautiful girl with flowers.
young woman in the park.
young woman with shopping bags.
wedding rings on a white background.

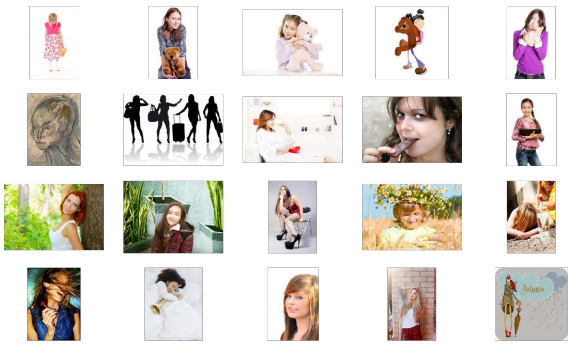
young woman holding a blank card.
young woman holding a blank card.
young woman with a laptop.
sunset on the beach.
young woman holding a heart.

(b)

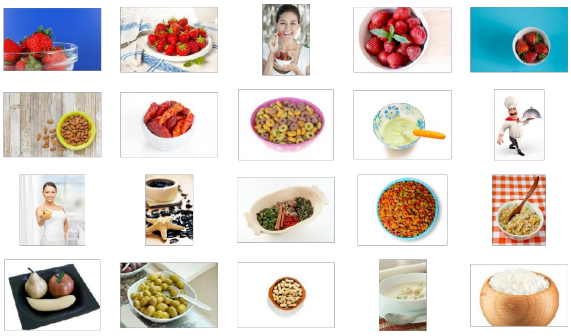
Figure 1. An example on Conceptual Captions dataset to show that the group captioning cannot be easily solved by captioning each image individually. The 20 model-generated captions on the right corresponds to the 20 images on the left in order, where the first 5 are targets and the other 15 are references. In (a), if we are summarizing the 5 target captions on context of reference captions, `portrait of a man`, which is the most frequent phrase, might be the result, which is not a good description as `man in black suit`. In (b), if we are summarizing the individual captions to get the group caption, `young woman` might be the result, which is not as good as `woman in cowboy hat`. The information needed for group captioning may be missed out in individual captions because the common feature of the group might not be important for individual images. Therefore, captioning images as a group can be more informative. We also perform a limited user study, and most users note that it is almost impossible for them to come up with a summarizing phrase given the individual captions.



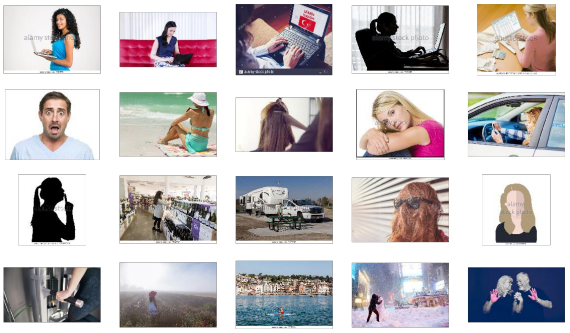
Ground Truth: woman in cowboy hat
Our Prediction: woman with cowboy hat
Group: woman
Contrastive: country with cowboy straw hat



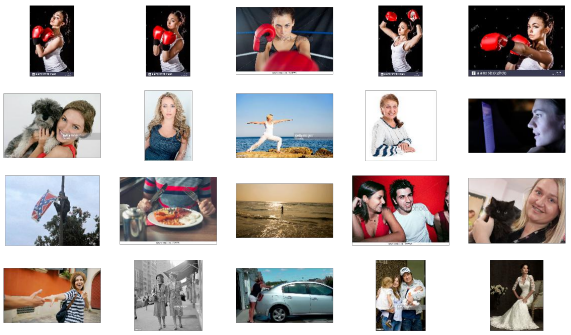
Ground Truth: girl holding teddy bear
Our Prediction: girl with toy bear
Group: girl
Contrastive: gingerbread bank with bear toy



Ground Truth: bowl of strawberry
Our Prediction: bowl of strawberry
Group: bowl
Contrastive: strawberry playing with strawberry ...



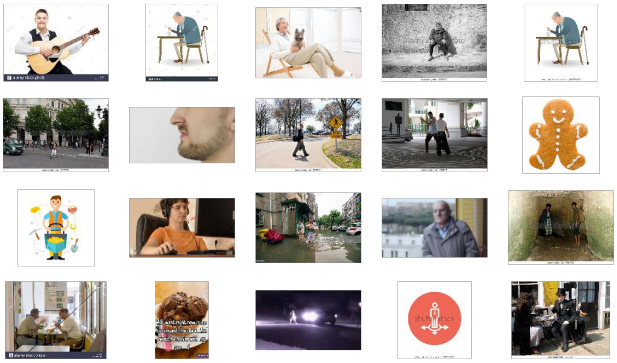
Ground Truth: woman using laptop computer
Our Prediction: woman using laptop
Group: woman
Contrastive: business people speaks working on digital laptop computer



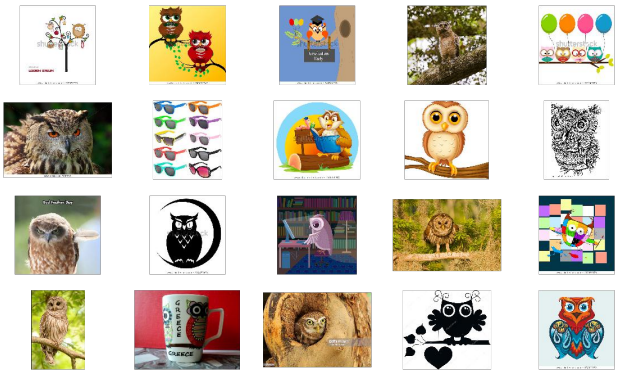
Ground Truth: woman in red glove
Our Prediction: woman in boxing glove
Group: woman
Contrastive: is go in boxing in boxing ...

Ground Truth: girl wearing white dress
Our Prediction: white girl
Group: girl
Contrastive: white rule white and white natural...

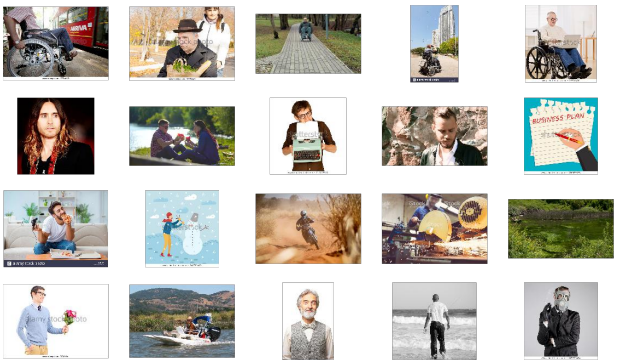
Figure 2. Examples of only using group representation or only using contrastive representation (Corresponding to Table 4 in the main paper). As shown, common information in both image groups (blue text) is encoded in the group representation, while the difference between two groups (red or orange text) is in contrastive representation. The first four examples are good cases while the last two examples are failure cases.



Ground Truth: man sitting on chair
Our Prediction: man sitting on chair

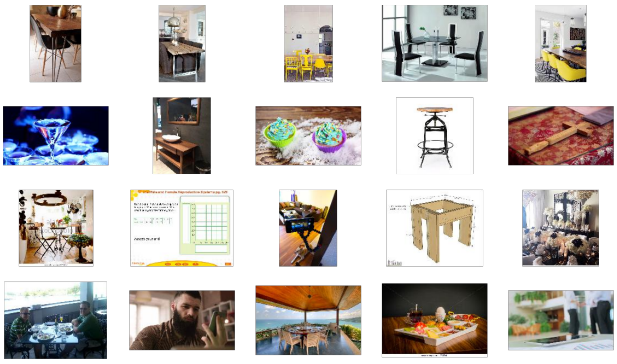


Ground Truth: owl on branch
Our Prediction: owl sitting on branch

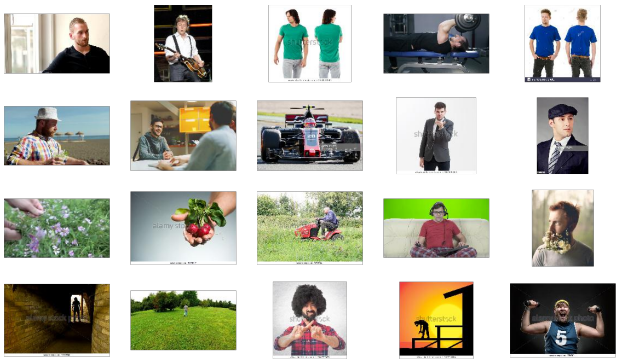


Ground Truth: disabled man in wheelchair
Our Prediction: man in wheelchair

Ground Truth: football player celebrates scoring with team mate
Our Prediction: football player celebrates scoring with mate

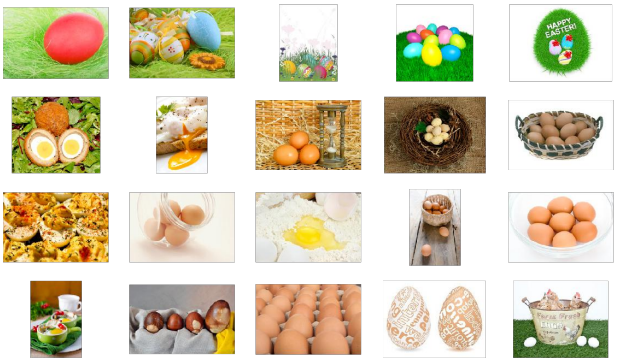


Ground Truth: dining table with chair
Our Prediction: dining table with chair

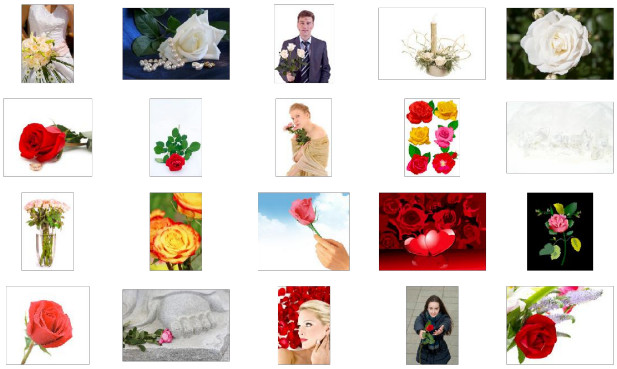


Ground Truth: man wearing shirt
Our Prediction: man with shirt

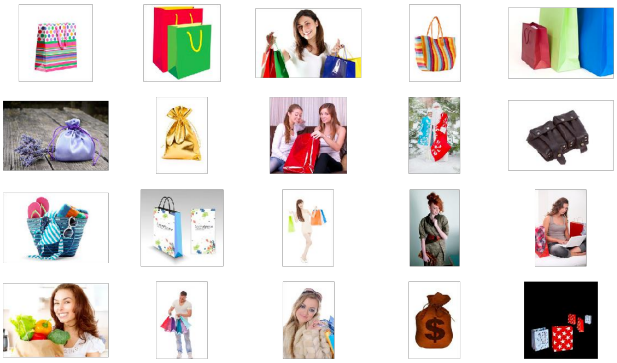
Figure 3. Good examples Conceptual Captions dataset.



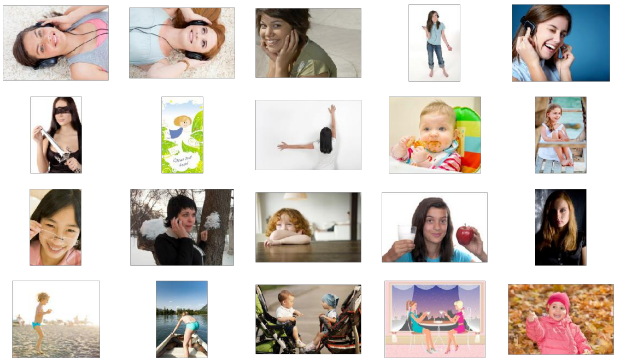
Ground Truth: easter eggs on grass
Our Prediction: colorful eggs on grass



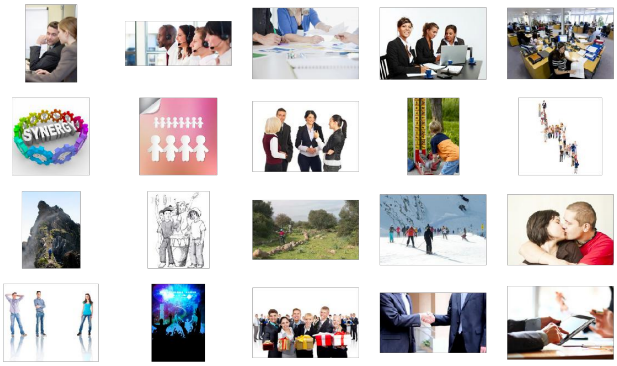
Ground Truth: white rose
Our Prediction: white rose



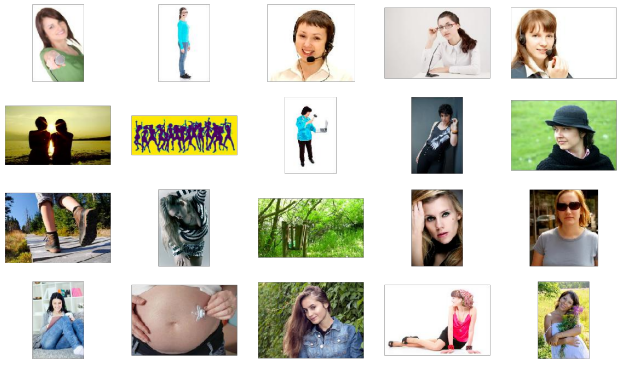
Ground Truth: colorful bag on white background
Our Prediction: colorful bag on white background



Ground Truth: teen girl listening to music
Our Prediction: girl listening to music

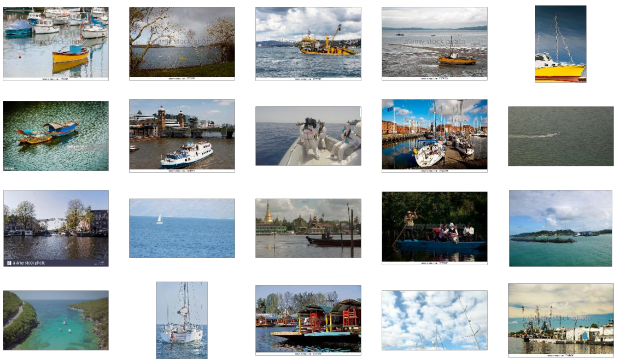


Ground Truth: people working in office
Our Prediction: business people in office

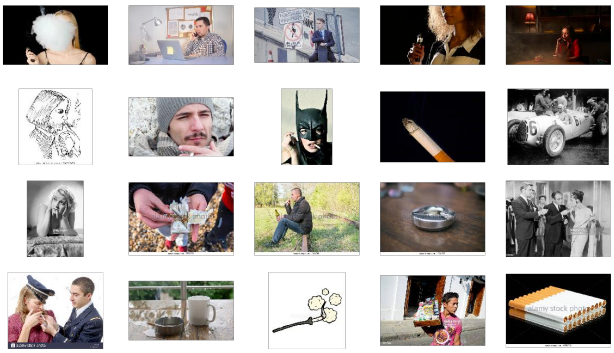


Ground Truth: woman with microphone
Our Prediction: woman with headset

Figure 4. Good examples on Stock Captions dataset.



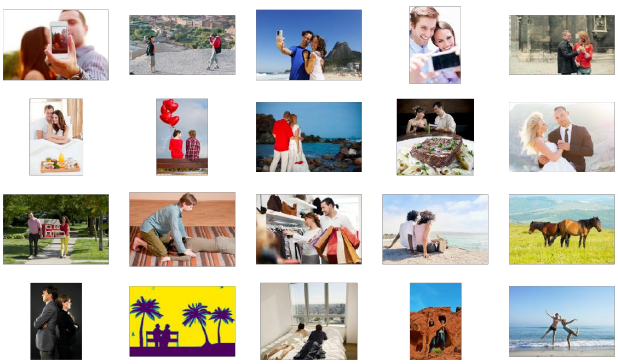
Ground Truth: yellow boat
Our Prediction: fishing boat
Group: boat
Contrastive: flaming colorful chart appears colorful hypnotic colorful stylized wavy stylized colorful



Ground Truth: electronic cigarette
Our Prediction: white smoke
Group: white
Contrastive: blue rhythm blue activist on electric on pov cigarette call

Figure 5. Failure cases on Conceptual Captions dataset. For the first example, the model predicts `fishing boat` instead of `yellow boat`, which is less discriminative. This may be because the model does not capture features of the small boat well. For the example on the right, the model prediction (`white smoke`) may be dominated by one dominant image in the target group.

Ground Truth: sick boy
Our Prediction: boy child
Group: boy
Contrastive: sick sitting in bed sick bed



Ground Truth: couple taking photo
Our Prediction: couple using phone
Group: couple
Contrastive: smile pretty credit mobile using mobile phone device

Figure 6. Failure cases on Stock Captions dataset. For the first example, the model prediction does not notice that the boy is `sick`. We further look into the model output when using only the group representation or contrastive representation, where the *sick* information is captured in the contrastive representation, but may not be strong enough to be decoded out in the prediction. For the second example, the model prediction is correct but not as good as groundtruth.

References

[1] 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. 2