Expressing Visual Relationships via Language

Hao Tan[1], **Franck Dernoncourt**[2], **Zhe Lin**[2], **Trung Bui**[2], **Mohit Bansal**[1]
[1]UNC Chapel Hill [2]Adobe Research
{haotan, mbansal}@cs.unc.edu, {dernonco, zlin, bui}@adobe.com

Abstract

Describing images with text is a fundamental problem in vision-language research. Current studies in this domain mostly focus on single image captioning. However, in various real applications (e.g., image editing, difference interpretation, and retrieval), generating relational captions for two images, can also be very useful. This important problem has not been explored mostly due to lack of datasets and effective models. To push forward the research in this direction, we first introduce a new language-guided image editing dataset that contains a large number of real image pairs with corresponding editing instructions. We then propose a new relational speaker model based on an encoder-decoder architecture with static relational attention and sequential multi-head attention. We also extend the model with dynamic relational attention, which calculates visual alignment while decoding. Our models are evaluated on our newly collected and two public datasets consisting of image pairs annotated with relationship sentences. Experimental results, based on both automatic and human evaluation, demonstrate that our model outperforms all baselines and existing methods on all the datasets.[1]

1 Introduction

Generating captions to describe natural images is a fundamental research problem at the intersection of computer vision and natural language processing. Single image captioning (Mori et al., 1999; Farhadi et al., 2010; Kulkarni et al., 2011) has many practical applications such as text-based image search, photo curation, assisting of visually-impaired people, image understanding in social media, etc. This task has drawn significant attention in the research community with numerous studies (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018), and recent state of the art methods have achieved promising results on large captioning datasets, such as MS COCO (Lin et al., 2014). Besides single image captioning, the community has also explored other visual captioning problems such as video captioning (Venugopalan et al., 2015; Xu et al., 2016), and referring expressions (Kazemzadeh et al., 2014; Yu et al., 2017). However, the problem of two-image captioning, especially the task of describing the relationships and differences between two images, is still under-explored. In this paper, we focus on advancing research in this challenging problem by introducing a new dataset and proposing novel neural relational-speaker models.[2]

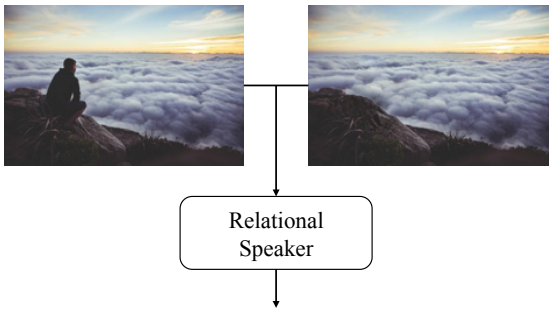

Figure 1: An example result of our method showing the input image pair from our Image Editing Request dataset, and the output instruction predicted by our relational speaker model trained on the dataset.

To the best of our knowledge, Jhamtani and Berg-Kirkpatrick (2018) is the only public dataset aimed at generating natural language descriptions for two real images. This dataset is about 'spotting the difference', and hence focuses more on describing exhaustive differences by learning align-

[1]Our data and code are publicly available at: https://github.com/airsplay/VisualRelationships

[2]We will release the full data and code upon publication.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1873–1883
Florence, Italy, July 28 - August 2, 2019.

ments between multiple text descriptions and multiple image regions; hence the differences are intended to be explicitly identifiable by subtracting two images. There are many other tasks that require more diverse, detailed and implicit relationships between two images. Interpreting image editing effects with instructions is a suitable task for this purpose, because it has requirements of exploiting visual transformations and it is widely used in real life, such as explanation of complex image editing effects for laypersons or visually-impaired users, image edit or tutorial retrieval, and language-guided image editing systems. We first build a new language-guided image editing dataset with high quality annotations by (1) crawling image pairs from real image editing request websites, (2) annotating editing instructions via Amazon Mechanical Turk, and (3) refining the annotations through experts.

Next, we propose a new neural speaker model for generating sentences that describe the visual relationship between a pair of images. Our model is general and not dependent on any specific dataset. Starting from an attentive encoder-decoder baseline, we first develop a model enhanced with two attention-based neural components, a static relational attention and a sequential multi-head attention, to address these two challenges, respectively. We further extend it by designing a dynamic relational attention module to combine the advantages of these two components, which finds the relationship between two images while decoding. The computation of dynamic relational attention is mathematically equivalent to attention over all visual "relationships". Thus, our method provides a direct way to model visual relationships in language.

To show the effectiveness of our models, we evaluate them on three datasets: our new dataset, the "Spot-the-Diff" dataset (Jhamtani and Berg-Kirkpatrick, 2018), and the two-image visual reasoning NLVR2 dataset (Suhr et al., 2019) (adapted for our task). We train models separately on each dataset with the same hyper-parameters and evaluate them on the same test set across all methods. Experimental results demonstrate that our model outperforms all the baselines and existing methods. The main contributions of our paper are: (1) We create a novel human language guided image editing dataset to boost the study in describing visual relationships; (2) We design novel relational-speaker models, including a dynamic relational attention module, to handle the problem of two-image captioning by focusing on all their visual relationships; (3) Our method is evaluated on several datasets and achieves the state-of-the-art.

2 Datasets

We present the collection process and statistics of our Image Editing Request dataset and briefly introduce two public datasets (viz., Spot-the-Diff and NLVR2). All three datasets are used to study the task of two-image captioning and evaluating our relational-speaker models. Examples from these three datasets are shown in Fig. 2.

2.1 Image Editing Request Dataset

Each instance in our dataset consists of an image pair (i.e., a source image and a target image) and a corresponding editing instruction which correctly and comprehensively describes the transformation from the source image to the target image. Our collected Image Editing Request dataset will be publicly released along with the scripts to unify it with the other two datasets.

2.1.1 Collection Process

To create a high-quality, diverse dataset, we follow a three-step pipeline: image pairs collection, editing instructions annotation, and post-processing by experts (i.e., cleaning and test set annotations labeling).

Images Pairs Collection We first crawl the editing image pairs (i.e., a source image and a target image) from posts on Reddit (Photoshop request subreddit)[3] and Zhopped[4]. Posts generally start with an original image and an editing specification. Other users would send their modified images by replying to the posts. We collect original images and modified images as source images and target images, respectively.

Editing Instruction Annotation The texts in the original Reddit and Zhopped posts are too noisy to be used as image editing instructions. To address this problem, we collect the image editing instructions on MTurk using an interactive interface that allows the MTurk annotators to either write an image editing instruction corresponding to a displayed image pair, or flag it as invalid (e.g., if the two images have nothing in common).

[3]https://www.reddit.com/r/photoshoprequest
[4]http://zhopped.com

Figure 2: Examples from three datasets: our Image Editing Request, Spot-the-Diff, and NLVR2. Each example involves two natural images and an associated sentence describing their relationship. The task of generating NLVR2 captions is converted from its original classification task.

	B-1	B-2	B-3	B-4	Rouge-L
Ours	52	34	21	13	45
Spot-the-Diff	41	25	15	8	31
MS COCO	38	22	15	8	34

Table 1: Human agreement on our datasets, compared with Spot-the-Diff and MS COCO (captions=3). B-1 to B-4 are BLEU-1 to BLEU-4. Our dataset has the highest human agreement.

Post-Processing by Experts Mturk annotators are not always experts in image editing. To ensure the quality of the dataset, we hire an image editing expert to label each image editing instruction of the dataset as one of the following four options: 1. correct instruction, 2. incomplete instruction, 3. implicit request, 4. other type of errors. Only the data instances labeled with "correct instruction" are selected to compose our dataset, and are used in training or evaluating our neural speaker model.

Moreover, two additional experts are required to write two more editing instructions (one instruction per expert) for each image pair in the validation and test sets. This process enables the dataset to be a multi-reference one, which allows various automatic evaluation metrics, such as BLEU, CIDEr, and ROUGE to more accurately evaluate the quality of generated sentences.

2.1.2 Dataset Statistics

The Image Editing Request dataset that we have collected and annotated currently contains 3,939 image pairs (3061 in training, 383 in validation, 495 in test) with 5,695 human-annotated instructions in total. Each image pair in the training set has one instruction, and each image pair in the validation and test sets has three instructions, written by three different annotators. Instructions have an average length of 7.5 words (standard deviation: 4.8). After removing the words with less than three occurrences, the dataset has a vocabulary of 786 words. The human agreement of our dataset is shown in Table 1. The word frequencies in our dataset are visualized in Fig. 3. Most of the images in our dataset are realistic. Since the task is image editing, target images may have some artifacts (see Image Editing Request examples in Fig. 2 and Fig. 5).

2.2 Existing Public Datasets

To show the generalization of our speaker model, we also train and evaluate our model on two public datasets, Spot-the-Diff (Jhamtani and Berg-Kirkpatrick, 2018) and NLVR2 (Suhr et al., 2019). Instances in these two datasets are each composed of two natural images and a human written sentence describing the relationship between the two

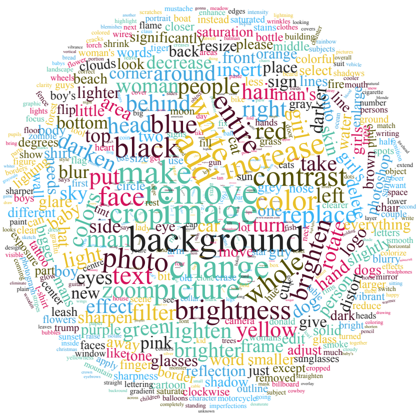

Figure 3: Word cloud showing the vocabulary frequencies of our Image Editing Request dataset.

images. To the best of our knowledge, these are the only two public datasets with a reasonable amount of data that are suitable for our task. We next briefly introduce these two datasets.

Spot-the-Diff This dataset is designed to help generate a set of instructions that can comprehensively describe all visual differences. Thus, the dataset contains images from video-surveillance footage, in which differences can be easily found. This is because all the differences could be effectively captured by subtractions between two images, as shown in Fig. 2. The dataset contains 13,192 image pairs, and an average of 1.86 captions are collected for each image pair. The dataset is split into training, validation, and test sets with a ratio of 8:1:1.

NLVR2 The original task of Cornell Natural Language for Visual Reasoning (NLVR2) dataset is visual sentence classification, see Fig. 2 for an example. Given two related images and a natural language statement as inputs, a learned model needs to determine whether the statement correctly describes the visual contents. We convert this classification task to a generation task by taking only the image pairs with correct descriptions. After conversion, the amount of data is 51,020, which is almost half of the original dataset with a size of 107,296. We also preserve the training, validation, and test split in the original dataset.

3 Relational Speaker Models

In this section, we aim to design a general speaker model that describes the relationship between two images. Due to the different kinds of visual relationships, the meanings of images vary in different tasks: "before" and "after" in Spot-the-Diff, "left" and "right" in NLVR2, "source" and "target" in our Image Editing Request dataset. We use the nomenclature of "source" and "target" for simplification, but our model is general and not designed for any specific dataset. Formally, the model generates a sentence $\{w_1, w_2, ..., w_T\}$ describing the relationship between the source image I^{SRC} and the target image I^{TRG}. $\{w_t\}_{t=1}^{T}$ are the word tokens with a total length of T. I^{SRC} and I^{TRG} are natural images in their raw RGB pixels. In the rest of this section, we first introduce our basic attentive encoder-decoder model, and show how we gradually improve it to fit the task better.

3.1 Basic Model

Our basic model (Fig. 4(a)) is similar to the baseline model in Jhamtani and Berg-Kirkpatrick (2018), which is adapted from the attentive encoder-decoder model for single image captioning (Xu et al., 2015). We use ResNet-101 (He et al., 2016) as the feature extractor to encode the source image I^{SRC} and the target image I^{TRG}. The feature maps of size $N \times N \times 2048$ are extracted, where N is the height or width of the feature map. Each feature in the feature map represents a part of the image. Feature maps are then flattened to two $N^2 \times 2048$ feature sequences f^{SRC} and f^{TRG}, which are further concatenated to a single feature sequence f.

$$f^{\text{SRC}} = \text{ResNet}\left(I^{\text{SRC}}\right) \quad (1)$$

$$f^{\text{TRG}} = \text{ResNet}\left(I^{\text{TRG}}\right) \quad (2)$$

$$f = \left[f_1^{\text{SRC}}, \ldots, f_{N^2}^{\text{SRC}}, f_1^{\text{TRG}}, , \ldots, f_{N^2}^{\text{TRG}}\right] \quad (3)$$

At each decoding step t, the LSTM cell takes the embedding of the previous word w_{t-1} as an input. The word w_{t-1} either comes from the ground truth (in training) or takes the token with maximal probability (in evaluating). The attention module then attends to the feature sequence f with the hidden output h_t as a query. Inside the attention module, it first computes the alignment scores $\alpha_{t,i}$ between the query h_t and each f_i. Next, the feature sequence f is aggregated with a weighted average (with a weight of α) to form the image context $\hat{f}$. Lastly, the context $\hat{f}_t$ and the hidden vector h_t are merged into an attentive hidden vector $\hat{h}_t$ with a

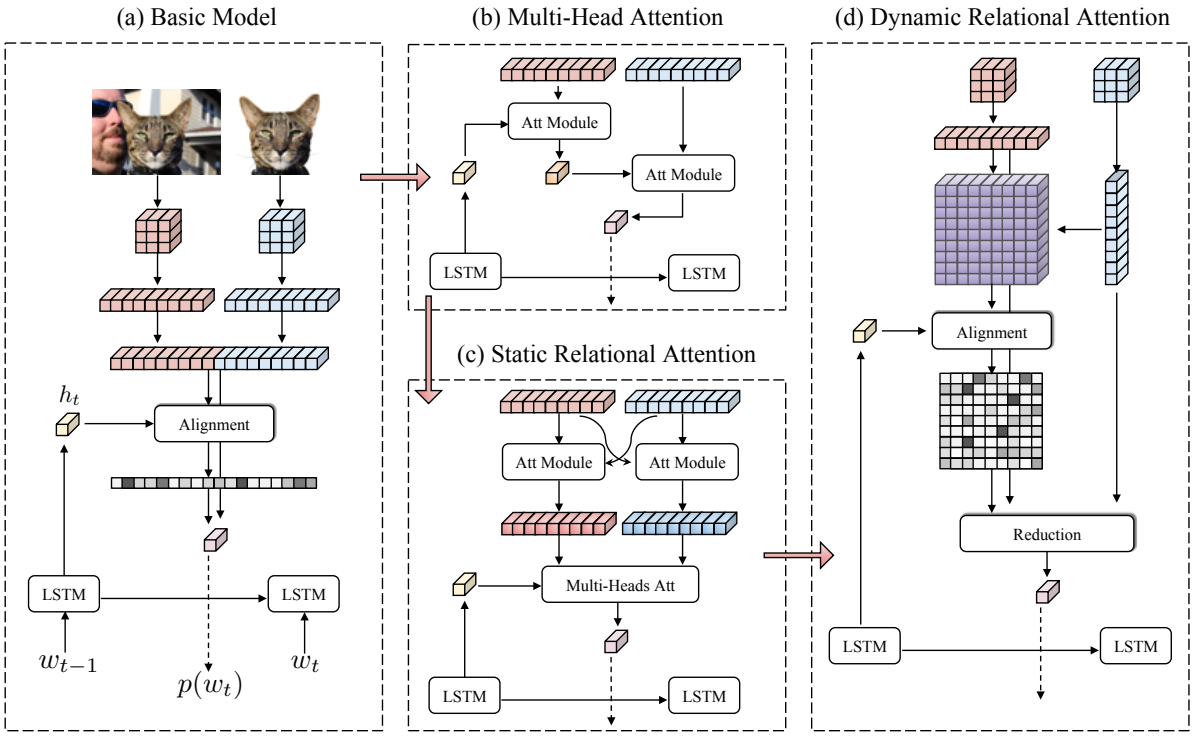

Figure 4: The evolution diagram of our models to describe the visual relationships. One decoding step at t is shown. The linear layers are omitted for clarity. The basic model (a) is an attentive encoder-decoder model, which is enhanced by the multi-head attention (b) and static relational attention (c). Our best model (d) dynamically computes the relational scores in decoding to avoid losing relationship information.

fully-connected layer:

$$\tilde{w}_{t-1} = \text{embedding}\,(w_{t-1}) \quad (4)$$

$$h_t, c_t = \text{LSTM}\,(\tilde{w}_{t-1}, h_{t-1}, c_{t-1}) \quad (5)$$

$$\alpha_{t,i} = \text{softmax}_i\left(h_t^\top W_{\text{IMG}} f_i\right) \quad (6)$$

$$\hat{f}_t = \sum_i \alpha_{t,i} f_i \quad (7)$$

$$\hat{h}_t = \tanh(W_1[\hat{f}_t; h_t] + b_1) \quad (8)$$

The probability of generating the k-th word token at time step t is softmax over a linear transformation of the attentive hidden $\hat{h}_t$. The loss $\mathcal{L}_t$ is the negative log likelihood of the ground truth word token w_t^*:

$$p_t(w_{t,k}) = \text{softmax}_k\left(W_{\text{w}}\,\hat{h}_t + b_{\text{w}}\right) \quad (9)$$

$$\mathcal{L}_t = -\log p_t(w_t^*) \quad (10)$$

3.2 Sequential Multi-Head Attention

One weakness of the basic model is that the plain attention module simply takes the concatenated image feature f as the input, which does not differentiate between the two images. We thus consider applying a multi-head attention module (Vaswani et al., 2017) to handle this (Fig. 4(b)). Instead of using the simultaneous multi-head attention [5] in Transformer (Vaswani et al., 2017), we implement the multi-head attention in a sequential way. This way, when the model is attending to the target image, the contextual information retrieved from the source image is available and can therefore perform better at differentiation or relationship learning.

In detail, the source attention head first attends to the flattened source image feature f^{SRC}. The attention module is built in the same way as in Sec. 3.1, except that it now only attends to the source image:

$$\alpha_{t,i}^{\text{SRC}} = \text{softmax}_i(h_t^\top W_{\text{SRC}} f_i^{\text{SRC}}) \quad (11)$$

$$\hat{f}_t^{\text{SRC}} = \sum_i \alpha_{t,i}^{\text{SRC}} f_i^{\text{SRC}} \quad (12)$$

$$\hat{h}_t^{\text{SRC}} = \tanh(W_2[\hat{f}_t^{\text{SRC}}; h_t] + b_2) \quad (13)$$

The target attention head then takes the output of the source attention $\hat{h}_t^{\text{SRC}}$ as a query to retrieve appropriate information from the target fea-

[5] We also tried the original multi-head attention but it is empirically weaker than our sequential multi-head attention.

ture f^{TRG}:

$$\alpha_{t,j}^{\text{TRG}} = \text{softmax}_j(\hat{h}_t^{\text{SRC}\top} W_{\text{TRG}} f_j^{\text{TRG}}) \quad (14)$$

$$\hat{f}_t^{\text{TRG}} = \sum_j \alpha_{t,j}^{\text{TRG}} f_j^{\text{TRG}} \quad (15)$$

$$\hat{h}_t^{\text{TRG}} = \tanh(W_3[\hat{f}_t^{\text{TRG}}; \hat{h}_t^{\text{SRC}}] + b_3) \quad (16)$$

In place of $\hat{h}_t$, the output of the target head $\hat{h}_t^{\text{TRG}}$ is used to predict the next word.[6]

3.3 Static Relational Attention

Although the sequential multi-head attention model can learn to differentiate the two images, visual relationships are not explicitly examined. We thus allow the model to statically (i.e., not in decoding) compute the relational score between source and target feature sequences and reduce the scores into two relationship-aware feature sequences. We apply a bi-directional relational attention (Fig. 4(c)) for this purpose: one from the source to the target, and one from the target to the source. For each feature in the source feature sequence, the source-to-target attention computes its alignment with the features in the target feature sequences. The source feature, the attended target feature, and the difference between them are then merged together with a fully-connected layer:

$$\alpha_{i,j}^{\text{S}\to\text{T}} = \text{softmax}_j((W_{\text{S}} f_i^{\text{SRC}})^\top (W_{\text{T}} f_j^{\text{TRG}})) \quad (17)$$

$$\hat{f}_i^{\text{S}\to\text{T}} = \sum_j \alpha_{i,j}^{\text{S}\to\text{T}} f_j^{\text{TRG}} \quad (18)$$

$$\hat{f}_i^{\text{S}} = \tanh(W_4[f_i^{\text{SRC}}; \hat{f}_i^{\text{S}\to\text{T}}] + b_4) \quad (19)$$

We decompose the attention weight into two small matrices W_{S} and W_{T} so as to reduce the number of parameters, because the dimension of the image feature is usually large. The target-to-source cross-attention is built in an opposite way: it takes each target feature f_j^{TRG} as a query, attends to the source feature sequence, and get the attentive feature $\hat{f}_j^{\text{T}}$. We then use these two bidirectional attentive sequences $\hat{f}_i^{\text{S}}$ and $\hat{f}_j^{\text{T}}$ in the multi-head attention module (shown in previous subsection) at each decoding step.

3.4 Dynamic Relational Attention

The static relational attention module compresses pairwise relationships (of size N^4) into two relationship-aware feature sequences (of size $2 \times N^2$). The compression saves computational resources but has potential drawback in information loss as discussed in Bahdanau et al. (2015) and Xu et al. (2015). In order to avoid losing information, we modify the static relational attention module to its dynamic version, which calculates the relational scores while decoding (Fig. 4(d)).

At each decoding step t, the dynamic relational attention calculates the alignment score $a_{t,i,j}$ between three vectors: a source feature f_i^{SRC}, a target feature f_j^{TRG}, and the hidden state h_t. Since the dot-product used in previous attention modules does not have a direct extension for three vectors, we extend the dot product and use it to compute the three-vector alignment score.

$$\text{dot}(x, y) = \sum_d x_d\, y_d = x^\top y \quad (20)$$

$$\text{dot}^*(x, y, z) = \sum_d x_d\, y_d z_d = (x \odot y)^\top z \quad (21)$$

$$a_{t,i,j} = \text{dot}^*(W_{\text{SK}} f_i^{\text{SRC}}, W_{\text{TK}} f_j^{\text{TRG}}, W_{\text{HK}} h_t) \quad (22)$$

$$= (W_{\text{SK}} f_i^{\text{SRC}} \odot W_{\text{TK}} f_j^{\text{TRG}})^\top W_{\text{HK}} h_t \quad (23)$$

where $\odot$ is the element-wise multiplication.

The alignment scores (of size N^4) are normalized by softmax. And the attention information is fused to the attentive hidden vector $\hat{f}_t^{\text{D}}$ as previous.

$$\alpha_{t,i,j} = \text{softmax}_{i,j}\,(a_{t,i,j}) \quad (24)$$

$$\hat{f}_t^{\text{SRC-D}} = \sum_{i,j} \alpha_{t,i,j} f_i^{\text{SRC}} \quad (25)$$

$$\hat{f}_t^{\text{TRG-D}} = \sum_{i,j} \alpha_{t,i,j} f_j^{\text{TRG}} \quad (26)$$

$$\hat{f}_t^{\text{D}} = \tanh(W_5[\hat{f}_t^{\text{SRC-D}}; \hat{f}_t^{\text{TRG-D}}; h_t] + b_5) \quad (27)$$

$$= \tanh(W_{5\text{S}} \hat{f}_t^{\text{SRC-D}} + W_{5\text{T}} \hat{f}_t^{\text{TRG-D}} + W_{5\text{H}} h_t + b_5) \quad (28)$$

where $W_{5\text{S}}, W_{5\text{T}}, W_{5\text{H}}$ are sub-matrices of W_5 and $W_5 = [W_{5\text{S}}, W_{5\text{T}}, W_{5\text{H}}]$.

According to Eqn. 23 and Eqn. 28, we find an analog in conventional attention layers with following specifications:

- Query: h_t
- Key: $W_{\text{SK}} f_i^{\text{SRC}} \odot W_{\text{TK}} f_j^{\text{TRG}}$
- Value: $W_{5\text{S}} f_i^{\text{SRC}} + W_{5\text{T}} f_j^{\text{TRG}}$

The key $W_{\text{SK}} f_i^{\text{SRC}} \odot W_{\text{TK}} f_j^{\text{TRG}}$ and the value $W_{5\text{S}} f_i^{\text{SRC}} + W_{5\text{T}} f_j^{\text{TRG}}$ can be considered as representations of the visual relationships between f_i^{SRC}

[6] We tried to exchange the order of two heads or have two orders concurrently. We didn't see any significant difference in results between them.

Method	BLEU-4	CIDEr	METEOR	ROUGE-L
Our Dataset (Image Editing Request)				
basic model	5.04	21.58	11.58	34.66
+multi-head att	6.13	22.82	11.76	35.13
+static rel-att	5.76	20.70	12.59	35.46
-static +dynamic rel-att	**6.72**	**26.36**	**12.80**	**37.25**
Spot-the-Diff				
CAPT(Jhamtani and Berg-Kirkpatrick, 2018)	7.30	26.30	10.50	25.60
DDLA(Jhamtani and Berg-Kirkpatrick, 2018)	**8.50**	32.80	12.00	28.60
basic model	5.68	22.20	10.98	24.21
+multi-head att	7.52	31.39	11.64	26.96
+static rel-att	8.31	33.98	**12.95**	28.26
-static +dynamic rel-att	8.09	**35.25**	12.20	**31.38**
NLVR2				
basic model	5.04	43.39	10.82	22.19
+multi-head att	**5.11**	44.80	10.72	22.60
+static rel-att	4.95	45.67	**10.89**	22.69
-static +dynamic rel-att	5.00	**46.41**	10.37	**22.94**

Table 2: Automatic metric of test results on three datasets. Best results of the main metric are marked in bold font. Our full model is the best on all three datasets with the main metric.

and f_j^{TRG}. It is a direct attention to the visual relationship between the source and target images, hence is suitable for the task of generating relationship descriptions.

4 Results

To evaluate the performance of our relational speaker models (Sec. 3), we trained them on all three datasets (Sec. 2). We evaluate our models based on both automatic metrics as well as pairwise human evaluation. We also show our generated examples for each dataset.

4.1 Experimental Setup

We use the same hyperparameters when applying our model to the three datasets. Dimensions of hidden vectors are 512. The model is optimized by Adam with a learning rate of $1e-4$. We add dropout layers of rate 0.5 everywhere to avoid over-fitting. When generating instructions for evaluation, we use maximum-decoding: the word w_t generated at time step t is $\arg\max_k p(w_{t,k})$. For the Spot-the-Diff dataset, we take the "Single sentence decoding" experiment as in Jhamtani and Berg-Kirkpatrick (2018). We also try to mix the three datasets but we do not see any improvement. We also try different ways to mix the three datasets but we do not see improvement. We first train a unified model on the union of these datasets. The metrics drop a lot because the tasks and language domains (e.g., the word dictionary and lengths of sentences) are different from each other. We next only share the visual components to overcome the disagreement in language. However, the image domain are still quite different from each other (as shown in Fig. 2). Thus, we finally separately train three models on the three datasets with minimal cross-dataset modifications.

4.2 Metric-Based Evaluation

As shown in Table 2, we compare the performance of our models on all three datasets with various automated metrics. Results on the test sets are reported. Following the setup in Jhamtani and Berg-Kirkpatrick (2018), we takes CIDEr (Vedantam et al., 2015) as the main metric in evaluating the Spot-the-Diff and NLVR2 datasets. However, CIDEr is known as its problem in up-weighting unimportant details (Kilickaya et al., 2017; Liu et al., 2017b). In our dataset, we find that instructions generated from a small set of short phrases could get a high CIDEr score. We thus change the main metric of our dataset to METEOR (Banerjee and Lavie, 2005), which is manually verified to be aligned with human judgment on the validation set in our dataset. To avoid over-fitting, the model is

	Basic	Full	Both Good	Both Not
Ours(IEdit)	11	**24**	5	60
Spot-the-Diff	22	**37**	6	35
NLVR2	24	**37**	17	22

Table 3: Human evaluation on 100 examples. Image pair and two captions generated by our basic model and full model are shown to the user. The user chooses one from 'Basic' model wins, 'Full' model wins, 'Both Good', or 'Both Not'. Better model marked in bold font.

early-stopped based on the main metric on validation set. We also report the BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores.

The results on various datasets shows the gradual improvement made by our novel neural components, which are designed to better describe the relationship between 2 images. Our full model has a significant improvement in result over baseline. The improvement on the NLVR2 dataset is limited because the comparison of two images was not forced to be considered when generating instructions.

4.3 Human Evaluation and Qualitative Analysis

We conduct a pairwise human evaluation on our generated sentences, which is used in Celikyilmaz et al. (2018) and Pasunuru and Bansal (2017). Agarwala (2018) also shows that the pairwise comparison is better than scoring sentences individually. We randomly select 100 examples from the test set in each dataset and generate captions via our full speaker model. We ask users to choose a better instruction between the captions generated by our full model and the basic model, or alternatively indicate that the two captions are equal in quality. The Image Editing Request dataset is specifically annotated by the image editing expert. The winning rate of our full model (dynamic relation attention) versus the basic model is shown in Table 3. Our full model outperforms the basic model significantly. We also show positive and negative examples generated by our full model in Fig. 5. In our Image Editing Request corpus, the model was able to detect and describe the editing actions but it failed in handling the arbitrary complex editing actions. We keep these hard examples in our dataset to match real-world requirements and allow follow-up future works to pursue the remaining challenges in this task. Our model is designed for non-localized relationships thus we do not explicitly model the pixel-level differences; however, we still find that the model could learn these differences in the Spot-the-Diff dataset. Since the descriptions in Spot-the-Diff is relatively simple, the errors mostly come from wrong entities or undetected differences as shown in Fig. 5. Our model is also sensitive to the image contents as shown in the NLVR2 dataset.

5 Related Work

In order to learn a robust captioning system, public datasets have been released for diverse tasks including single image captioning (Lin et al., 2014; Plummer et al., 2015; Krishna et al., 2017), video captioning (Xu et al., 2016), referring expressions (Kazemzadeh et al., 2014; Mao et al., 2016), and visual question answering (Antol et al., 2015; Zhu et al., 2016; Johnson et al., 2017). In terms of model progress, recent years witnessed strong research progress in generating natural language sentences to describe visual contents, such as Vinyals et al. (2015); Xu et al. (2015); Ranzato et al. (2016); Anderson et al. (2018) in single image captioning, Venugopalan et al. (2015); Pan et al. (2016); Pasunuru and Bansal (2017) in video captioning, Mao et al. (2016); Liu et al. (2017a); Yu et al. (2017); Luo and Shakhnarovich (2017) in referring expressions, Jain et al. (2017); Li et al. (2018); Misra et al. (2018) in visual question generation, and Andreas and Klein (2016); Cohn-Gordon et al. (2018); Luo et al. (2018); Vedantam et al. (2017) in other setups.

Single image captioning is the most relevant problem to the two-images captioning. Vinyals et al. (2015) created a powerful encoder-decoder (i.e., CNN to LSTM) framework in solving the captioning problem. Xu et al. (2015) further equipped it with an attention module to handle the memorylessness of fixed-size vectors. Ranzato et al. (2016) used reinforcement learning to eliminate exposure bias. Recently, Anderson et al. (2018) brought the information from object detection system to further boost the performance.

Our model is built based on the attentive encoder-decoder model (Xu et al., 2015), which is the same choice in Jhamtani and Berg-Kirkpatrick (2018). We apply the RL training with self-critical (Rennie et al., 2017) but do not see significant improvement, possibly because of the relatively small data amount compared to MS COCO. We also observe that the detection system in An-

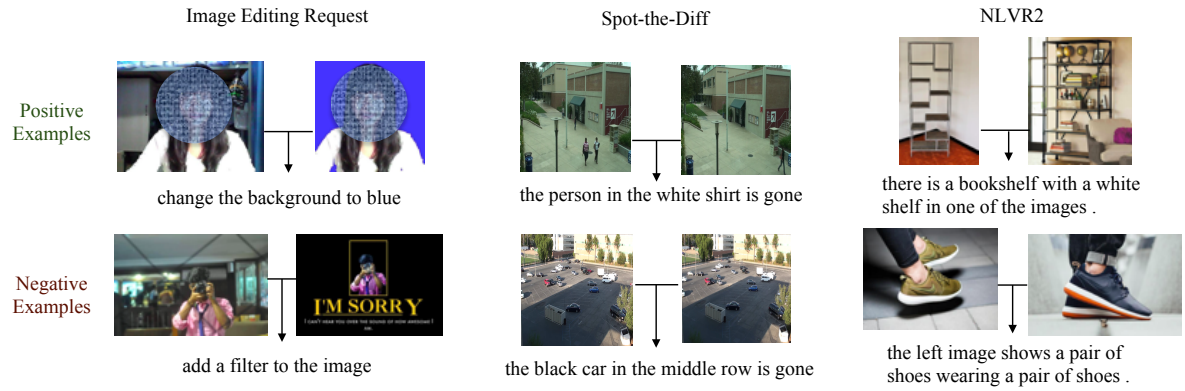

Figure 5: Examples of positive and negative results of our model from the three datasets. Selfies are blurred.

derson et al. (2018) has a high probability to fail in the three datasets, e.g., the detection system can not detect the small cars and people in spot-the-diff dataset. The DDLA (Difference Description with Latent Alignment) method proposed in Jhamtani and Berg-Kirkpatrick (2018) learns the alignment between descriptions and visual differences. It relies on the nature of the particular dataset and thus could not be easily transferred to other dataset where the visual relationship is not obvious. The two-images captioning could also be considered as a two key-frames video captioning problem, and our sequential multi-heads attention is a modified version of the seq-to-seq model (Venugopalan et al., 2015). Some existing work (Chen et al., 2018; Wang et al., 2018; Manjunatha et al., 2018) also learns how to modify images. These datasets and methods focus on the image colorization and adjustment tasks, while our dataset aims to study the general image editing request task.

6 Conclusion

In this paper, we explored the task of describing the visual relationship between two images. We collected the Image Editing Request dataset, which contains image pairs and human annotated editing instructions. We designed novel relational speaker models and evaluate them on our collected and other public existing dataset. Based on automatic and human evaluations, our relational speaker model improves the ability to capture visual relationships. For future work, we are going to further explore the possibility to merge the three datasets by either learning a joint image representation or by transferring domain-specific knowledge. We are also aiming to enlarge our Image Editing Request dataset with newly-released posts on Reddit and Zhopped.

Acknowledgments

We thank the reviewers for their helpful comments and Nham Le for helping with the initial data collection. This work was supported by Adobe, ARO-YIP Award #W911NF-18-1-0336, and faculty awards from Google, Facebook, and Salesforce. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Aseem Agarwala. 2018. Automatic photography with google clips.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675.

Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 439–443.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

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

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6485–6494.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. 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.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 199–209.

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

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6116–6124.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. 2017a. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864.

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

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

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

Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. 2018. Learning to color from language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 764–769.

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

Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens van der Maaten. 2018. Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, pages 1–9. Citeseer.

Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Ramakanth Pasunuru and Mohit Bansal. 2017. Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 979–985.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.

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

Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th annual meeting on association for computational linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

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

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

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542.

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

Hai Wang, Jason D Williams, and SingBing Kang. 2018. Learning to globally edit images with textual description. *arXiv preprint arXiv:1810.05786*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296.

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

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.