Semantic Relation-aware Difference Representation Learning for Change Captioning

Yunbin Tu[1]*, Tingting Yao[3], Liang Li[2], Jiedong Lou[3], Shengxiang Gao[1]†, Zhengtao Yu[1], Chenggang Yan[3]

[1]Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences
[3] Intelligent Information Processing Laboratory, Hangzhou Dianzi University
{tuyunbin1995,gaoshengxiang.yn}@foxmail.com
liang.li@ict.ac.cn, ztyu@hotmail.com

Abstract

Change captioning is to describe the difference in a pair of images with a natural language sentence. In this task, the distractors, such as the illumination or viewpoint change, bring the huge challenges about learning the difference representation. In this paper, we propose a semantic relation-aware difference representation learning network to explicitly learn the difference representation in the existence of distractors. Specifically, we introduce a self-semantic relation embedding block to explore the underlying changed objects and design a cross-semantic relation measuring block to localize the real change and learn the discriminative difference representation. Besides, relying on the POS of words, we devise an attention-based visual switch to dynamically use visual information for caption generation. Extensive experiments show that our method achieves the state-of-the-art performances on CLEVR-Change and Spot-the-Diff datasets [1].

1 Introduction

Change Captioning aims to describe a semantic change between a pair of "before" and "after" images, which has many practical applications such as facility monitoring (Sakurada and Okatani, 2015), medical imaging (Patriarche and Erickson, 2004), and aerial photography (Gueguen and Hamid, 2015).

The previous work (Jhamtani and Berg-Kirkpatrick, 2018) introduced this task with an ideal assumption that there is a semantic change between a completely-aligned image pair. However, there is always illumination change in a dynamic world, and same or similar scenes are prone to shoot under different viewpoints. Compared to semantic changes, both illumination and viewpoint changes are irrelevant distractors, so realistic change captioning requires a model: 1) distinguishing semantic changes (e.g., an object has moved) from distractors (e.g., a viewpoint change) and 2) conveying the detected change in a logically and grammatically accurate sentence. To this end, recent works (Park et al., 2019; Shi et al., 2020) focused on addressing change captioning in the presence of distractors.

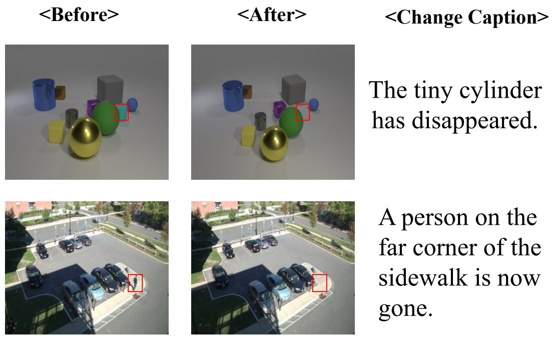
<Before>
<After>
<Change Caption>
The tiny cylinder has disappeared.
A person on the far corner of the sidewalk is now gone.

Figure 1: Two examples of change captioning with and without a viewpoint change.

Despite the progress, there are still two limitations for their approaches. First, the semantic difference was modeled only relying on the semantic features of objects, while ignoring their self-semantic relations. Hence, the feature difference is hard to capture the tiny change. As shown in Figure 1, compared with many unchanged objects, the dropped object is tiny and easy to ignore. Differently, if one of the objects has changed, especially number or position change (*e.g.*, "add", "drop", or "move"), the semantic relations surrounding it would change as well, which would be beneficial to explore the underlying objects that have changed. Second, due to the existing of irrelevant distractors, the model would capture the semantic difference

* This work was done when Yunbin Tu visited VIPL research group, CAS and was supervised by Prof. Liang Li.

† Corresponding author

[1]The code of this paper has been made publicly available at https://github.com/tuyunbin/SRDRL

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 63–73
August 1–6, 2021. ©2021 Association for Computational Linguistics

with noises and thus learn a wrong difference representation. However, both distractors are irrelevant to the semantics of image contents. Therefore, the cross-semantic relation between the captured semantic difference and the image pair is beneficial to judge whether the semantic change has actually happened, and further learn the difference representation in the “before” and “after” images.

Besides, during caption generation, previous works exploited visual information to generate each word, which is unnecessary or even misleading (Lu et al., 2017; Song et al., 2017). As words with different part-of-speech (POS) information not only play different grammatical roles in a sentence (Wang et al., 2019), but also have different relationships with the visual information in an image. As shown in the first example of Figure 1, some words (e.g., “tiny”, “cylinder” and “disappeared”) belong to adjective, noun and verb words, which denote the size, category, and state of the visual object, while the word (i.e.,“the”) is a determiner word which does not have corresponding canonical visual signals. Thus, it is useful to introduce the POS of words for switching visual information during change caption generation.

In this paper, we propose a Semantic Relation-aware Difference Representation Learning (SRDRL) network to localize the semantic change in the presence of distractors, and introduce an Attention-based Visual Switch (AVS) to dynamically decide when to use visual information during change caption generation. Specifically, first, a Self-Semantic Relation Embedding block (SSRE) builds semantic relations of objects for each image in the “before”/“after” pair via the self-attention mechanism. The built relations are embedded into image features for computing a relation-embedded feature difference. Second, a Cross-Semantic Relation Measuring block (CSRM) leverages the obtained difference to query the underlying “candidate change” in the each image. Further, CSRM uses the difference to generate an attention gate measuring its cross-semantic relations with respect to each image. Subsequently, the attention gate is applied to the candidate change to distinguish semantic change from the viewpoint/illumination change. Third, the change localizer is introduced to learn the accurate difference representation in the image pair under the guidance of a prior knowledge (the above distinguished information).

Finally, according to POS information of words, an Attention-based Visual Switch (AVS) is devised and incorporated into the caption generator to dynamically control visual information when predicting the next word. Extensive experiments show that our approach outperforms the state-of-the-art change captioning models with a large margin.

In summary, the contributions of this work have threefold: (1) We propose SRDRL that explicitly learns the semantic difference representation in the image pair by embedding self-semantic relations into object features of each image and further measuring the cross-semantic relations between the image pair and their difference. (2) Both SSRE and CSRM blocks are designed to help the change localizer to accurately focus on the changed objects. (3) An AVS is customized to dynamically utilize visual information for caption generation based on the POS information of words.

2 Related Work

Different from conventional image (Liu et al., 2020, 2019; Li et al., 2020; Yan et al., 2019, 2020a, 2021) or video captioning (Deng et al., 2021; Zhang et al., 2017; Tu et al., 2017, 2020; Yan et al., 2020b), change captioning addresses two-image captioning, especially to describe their difference. Jhamtani *et al.* (Jhamtani and Berg-Kirkpatrick, 2018) is the first work for change captioning. However, it is built upon an ideal situation by assuming there are no distractors (illumination/viewpoint change) between a pair of images. To make this task more close to our dynamic world, Park *et al.* and Shi *et al.* (Park et al., 2019; Shi et al., 2020) both aimed to address change captioning in the existence of distractors. On one hand, Park *et al.* directly concatenated the coarse feature difference with the image pair to operate spatial attention to localize the change. However, due to the existing of distractors, when the captured feature difference is not what the model really expects, the spatial attention module could be misled to give fallacious results. On the other hand, Shi *et al.* first exploited a cross-attention mechanism to search the most similar patches between the image pair and they are regarded as the unchanged representation. Then, they subtracted them from the original image to get the difference representation. However, as our aforementioned, the changed object is tiny and easy to ignore, so it is insufficient to capture the difference representation only at feature level.

Different from the above state-of-the-art meth-

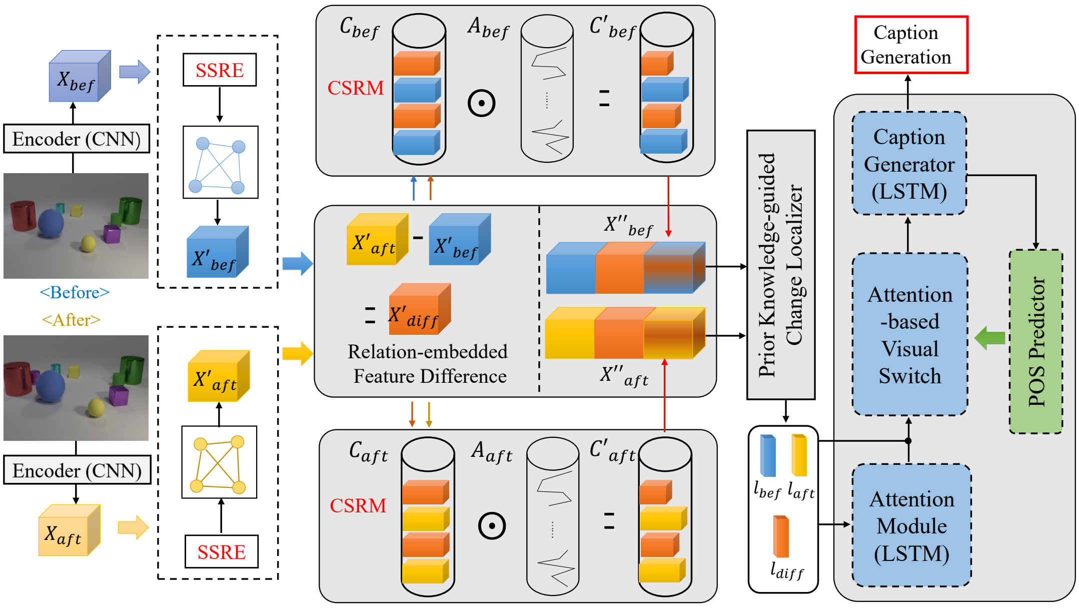

Figure 2: The architecture of the proposed semantic relation-aware difference representation learning (SRDRL) network and an attention-based visual switch (AVS). The SRDRL consists of a self-semantic relation embedding block (SSRE), a cross-semantic relation measuring block (CSRM) and a prior knowledge-guided change localizer. The AVS is incorporated into the caption generator and guided by a POS predictor.

ods, we first use SSRE to improve the fine-grained representation ability of object features by embedding the self-semantic relations among them. Then, we exploit CSRM to distinguish the actual semantic change from irrelevant distractors via measuring cross-semantic relations between the captured candidate difference and the original images. Finally, we use POS information to devise an attention-based visual switch that dynamically determines not only when to use visual information, but also which to use (*e.g.*, “before” and “after”). Compared to the aforementioned methods, our method not only can learn discriminative difference representation, but also can describe it using an accurate natural language sentence.

3 Methodology

We present a semantic relation-aware difference representation learning (SRDRL) network for change localization and devise an attention-based Visual Switch (AVS) under the guidance of POS information for caption generation. When a pair of “before” and “after” images are given (denoted as I_{bef} and I_{aft}), our SRDRL first detects *what* (position, number, attribute, or nothing) has changed in a scene and further decides *where* to localize on both I_{bef} and I_{aft}. Then, during caption generation, the AVS is able to dynamically decide *when* to use visual information and *which* to use (*e.g.*, “before” and “after”).

3.1 Semantic Relation-aware Difference Representation Learning Network

3.1.1 Self-Semantic Relation Embedding

Formally, given a pair of I_{bef} and I_{aft}, we first use pre-trained CNN model to extract object-level features and denote them as X_{bef} and X_{aft}, where $X_i \in \mathbb{R}^{C\times H\times W}$; C, H, W indicate the number of channels, height, and width. However, These original object features are independent, and there exist semantic relations among them (Huang et al., 2020; Wu et al., 2019; Yin et al., 2020). Inspired by the self-attention (Vaswani et al., 2017) using in machine translation, the self-semantic relation embedding block (SSRE) relies on it to implicitly model the semantic relations among objects in each image. Specifically, we first reshape X_i to $X_i \in \mathbb{R}^{N\times C}$ $(N = HW)$, where $i \in (bef, aft)$. Then, given (*key*, *value*), SSRE exploits the scaled dot-product attention on queries Q by:

$$SSRE(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (1)$$

In our case, the queries, keys and values are all projections of the object features of X_i:

$$(Q,K,V) = \left(X_iW^Q, X_iW^K, X_iW^V\right). \quad (2)$$

Though the SSRE, the semantic relations are embedded in the original object features; both X_{bef} and X_{aft} can be updated to X'_{bef} and X'_{aft}. Finally, we subtract X'_{bef} from X'_{aft} to capture the semantic difference X'_{diff} in the both object feature and relation aspects.

3.1.2 Cross-Semantic Relation Measuring

Due to the existing of distractors, the resulting X'_{diff} would include some irrelevant information, which would be noises for the accurate difference representation learning on both X'_{bef} and X'_{aft}. Thus, we propose a cross-semantic relation measuring block (CSRM) to distinguish the semantic change from the irrelevant illumination or viewpoint change by measuring the cross-semantic relation between the X'_{diff} and X'_{bef} (X'_{aft}). Concretely, the CSRM utilizes the X'_{diff} to first query the possible "candidate change" C_{bef} on the X'_{bef}, and then generates an "attention gate" A_{bef} measuring its semantic relations with respect to X'_{bef}. These are defined by using two separate non-linear transformations:

$$\begin{aligned} C_{bef} &= \phi\left(X'_{diff}W_q^i + X'_{bef}W_v^i + b^i\right), \\ A_{bef} &= \sigma\left(X'_{diff}W_q^g + X'_{bef}W_v^g + b^g\right), \end{aligned} \tag{3}$$

where $W_q^i, W_v^i, W_q^g, W_v^g \in \mathbb{R}^{C\times C}, b^i, b^g \in \mathbb{R}^C$, and C is the dimension of X'_{diff} and X'_{bef} ; σ and ϕ denote the sigmoid and tanh function. The value in the "attention gate" indicates the semantic relevance between the "candidate change" and the "before". Thus, the more information in the "candidate change" passes through the "attention gate", the more X'_{diff} is relevant to X'_{bef}.

Next, the CSRM applies the A_{bef} to the C_{bef} to filter all the underlying change information and focus on only the information about semantic change via element-wise multiplication:

$$C'_{bef} = A_{bef} \odot C_{bef}. \tag{4}$$

Besides, the information about semantic change C'_{aft} is computed via the similar operation between the X'_{diff} and X'_{aft} :

$$\begin{aligned} C_{aft} &= \phi\left(X'_{diff}U_q^i + X'_{aft}U_v^i + z^i\right), \\ A_{aft} &= \sigma\left(X'_{diff}U_q^g + X'_{aft}U_v^g + z^g\right), \\ C'_{aft} &= A_{aft} \odot C_{aft}. \end{aligned} \tag{5}$$

3.1.3 Prior Knowledge-guided Change Localizer

After obtaining the C'_{bef} and C'_{aft}, we use them as the prior knowledge to guide the change localizer to learn the difference representation. Specifically, the change localizer first predicts two separate attention maps under the guidance of C'_{bef} and C'_{aft}, respectively:

$$\begin{aligned} X''_{\text{bef}} &= [X'_{\text{bef}}\,;X'_{\text{diff}}\,;C'_{\text{bef}}\,], \\ a_{\text{bef}} &= \sigma\left(\text{conv}_2\left(\text{ReLU}\left(\text{conv}_1\left(X''_{\text{bef}}\right)\right)\right)\right), \\ X''_{\text{aft}} &= [X'_{\text{aft}}\,;X'_{\text{diff}}\,;C'_{\text{aft}}\,], \\ a_{\text{aft}} &= \sigma\left(\text{conv}_2\left(\text{ReLU}\left(\text{conv}_1\left(X''_{\text{aft}}\right)\right)\right)\right), \end{aligned} \tag{6}$$

where $[;]$, conv, and σ indicate concatenation, convolutional layer, and element-wise sigmoid, respectively. After that, the difference representation features l_{bef} and l_{aft} are attended to by applying a_{bef} and a_{aft} to the input image features X'_{bef} and X'_{aft} :

$$\begin{aligned} l_{\text{bef}} &= \textstyle\sum_{H,W} a_{\text{bef}} \odot X'_{\text{bef}}, l_{\text{bef}} \in \mathbb{R}^C, \\ l_{\text{aft}} &= \textstyle\sum_{H,W} a_{\text{aft}} \odot X'_{\text{aft}}, l_{\text{aft}} \in \mathbb{R}^C. \end{aligned} \tag{7}$$

3.2 Change Caption Generation

3.2.1 POS Predictor

Inspired by POS used in machine translation (Yin et al., 2019), we dynamically predict POS tags [1] of target words based on the previous hidden states $h_c^{(t-1)}$ of the caption generator. The predicted tags help the captioning model use visual information in a dynamic way.

Specifically, at time t, $h_c^{(t-1)}$ is first fed into a single hidden layer with the ReLU activation function:

$$d_t^p = \text{ReLU}\left(W_p^{(1)}h_c^{(t-1)} + b_p^{(1)}\right), \tag{8}$$

where $W_p^{(1)} \in \mathbb{R}^{M\times M}$ and $b_p^{(1)} \in \mathbb{R}^M$, and M is the dimension of the hidden state in caption generator. Then, a POS tag probability is predicted by a linear transformation with a softmax function:

$$w_t^p = \text{softmax}\left(W_p^{(2)}d_t^p + b_p^{(2)}\right), \tag{9}$$

where $W_p^{(2)} \in \mathbb{R}^{M\times n}$ and $b_p^{(2)} \in \mathbb{R}^n$, and n is the number of POS tag. After obtaining w_t^p, we represent the POS tag of the target word w_t using a semantic representation p_t:

$$p_t = E_p w_t^p, \tag{10}$$

where $E_p \in \mathbb{R}^{n\times N}$ is a POS embedding matrix and N is the dimension of the POS representation.

[1] The POS tags of words in ground truth are processed by Stanford Log-linear Part-Of-Speech Tagger (Toutanova et al., 2003).

3.2.2 Attention-based Visual Switch

Visual Attention. We first use a visual attention module to select a candidate feature from l_{bef} , l_{aft}, or l_{diff} (l_{aft} - l_{bef}), which could be relevant to the target word:

$$l_{\text{dyn}}^{(t)} = \sum_i \alpha_i^{(t)} l_i, \tag{11}$$

where $i \in$ (bef, diff, aft). $\alpha_i^{(t)}$ are current visual attention weights and they are computed by an attention LSTM$_a$:

$$\begin{aligned} v &= \text{ReLU}\left(W_{d_1}\left[l_{\text{bef}}\, ; l_{\text{diff}}\, ; l_{\text{aft}}\right] + b_{d_1}\right) \\ u^{(t)} &= \left[v; h_c^{(t-1)}\right] \\ h_a^{(t)} &= \text{LSTM}_a\left(h_a^{(t)} \mid u^{(t)}, h_a^{(0:t-1)}\right) \\ \alpha_i^{(t)} &\sim \text{Softmax}\left(W_{d_2} h_a^{(t)} + b_{d_2}\right) \end{aligned} \tag{12}$$

where W_{d_1}, b_{d_1}, W_{d_2}, and b_{d_2} are learnable parameters. $h_a^{(*)}$ and $h_c^{(*)}$ are hidden states of the attention module LSTM$_a$ and the caption generator LSTM$_c$, respectively.

Visual Switch. Then, we exploit a visual switch to decide whether to rely on visual information to predict the next word based on the predicted POS information p_t. At time step t, the visual switch β_t is defined as:

$$\begin{aligned} m_t &= \left[p_t; h_c^{t-1}; l_{dyn}^{(t)}\right], \\ \beta_t &= \sigma(W_{s2}(\text{ReLU}(W_{s1} m_t))), \end{aligned} \tag{13}$$

where σ is the sigmoid function and W_{s*} are the learnable parameters. The range of β_t is [0,1] and the value of it indicates how much visual information to use when predicting the target word. Then, we apply this switch to attended visual feature $l_{\text{dyn}}^{(t)}$ to control the use of visual information:

$$L_{dyn}^{(t)} = \beta_t \odot l_{dyn}^{(t)}. \tag{14}$$

3.2.3 Caption generator

After the proper visual information is obtained, we use it and the previous word w_{t-1} (ground-truth word during training, predicted word during inference) to the caption generator LSTM$_c$ to predict a series of distributions over the next word:

$$\begin{aligned} c^{(t)} &= \left[E\left[w_{t-1}\right]; L_{\text{dyn}}^{(t)}\right], \\ h_c^{(t)} &= \text{LSTM}_c\left(h_c^{(t)} \mid c^{(t)}, h_c^{(0:t-1)}\right), \\ w_t &\sim \text{Softmax}\left(W_c h_c^{(t)} + b_c\right), \end{aligned} \tag{15}$$

where E is a word embedding matrix; W_c and b_c are learnable parameters.

3.3 Joint Training

We jointly train the POS predictor and the caption generator end-to-end by maximizing the likelihood of the observed POS and word sequence. For the POS predictor, given the target ground-truth POS tags $(w_1^p, \ldots, w_m^p)$, we minimize its negative log-likelihood loss:

$$L_{pos}(\theta_p) = -\sum_{t=1}^{m} \log p\left(w_t^p \mid w_{<t}^p; \theta_p\right), \tag{16}$$

where θ_p are the parameters of the POS predictor and m is the length of the POS tag.

For the caption generator, given the target ground-truth caption words $(w_1^c, \ldots, w_m^c)$, we minimize its negative log-likelihood loss:

$$L_{cap}(\theta_c) = -\sum_{t=1}^{m} \log p\left(w_t^c \mid w_{<t}^c; \theta_c\right), \tag{17}$$

where θ_c are the parameters of the caption generator and m is the length of the caption. Thus, the final loss function is optimized as follows:

$$L(\theta) = L_{pos} + L_{cap} \tag{18}$$

4 Experiments

4.1 Datasets

CLEVR-Change. This dataset (Park et al., 2019) is a large scale dataset with a set of basic geometry objects, which consists of 79,606 image pairs and 493,735 captions. The change types consist of five cases, *i.e.*, “Color”, “Texture”, “Add”, “Drop”, and ‘’Move”. We use the official split with image pairs of 67,660 for training, 3, 976 for validation and 7,970 for testing.

Spot-the-Diff. This dataset (Jhamtani and Berg-Kirkpatrick, 2018) contains 13,192 real image pairs which are well aligned image pairs, with one or more changes between the images (but no distractors). Similar to (Park et al., 2019), we only evaluate our model in a single change setting and split it into training, validation, and test sets with a ratio of 8:1:1.

4.2 Evaluation Metrics

We use five standard metrics to evaluate the quality of generated sentences, *i.e.*, BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). We get all the results in this paper according to the Microsoft COCO evaluation server (Chen et al., 2015).

Table 1: Ablation studies on CLEVR-Change in terms of total performance, where B-4, M, R, C, and S are short for BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE, respectively.

	Total				
Method	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Baseline	53.1	37.3	70.6	115.6	31.2
SSRE	54.2	39.2	72.2	120.1	32.0
CSRM	53.7	38.5	71.6	118.0	32.0
SRDRL	54.8	40.1	73.2	121.0	32.6
AVS	53.2	38.5	71.3	115.7	31.6
SRDRL+AVS	**54.9**	**40.2**	**73.3**	**122.2**	**32.9**

Table 2: Ablation studies on CLEVR-Change in terms of different settings.

	Scene Change					None-scene Change				
Method	B-4	M	R	C	S	B-4	M	R	C	S
Baseline	50.9	33.0	65.3	100.9	27.7	62.0	50.0	75.9	116.1	34.7
SSRE	51.7	35.0	67.7	111.2	29.3	62.0	51.2	76.8	115.6	34.8
CSRM	51.8	34.6	67.3	106.5	29.4	61.4	49.9	75.9	115.5	34.7
SRDRL	52.0	35.8	68.9	112.3	30.3	62.1	**52.0**	**77.5**	116.3	**34.9**
AVS	50.9	34.2	66.5	103.6	28.8	60.3	50.5	76.1	113.5	34.4
SRDRL+AVS	**52.7**	**36.4**	**69.7**	**114.2**	**30.8**	**62.2**	51.3	76.9	**117.0**	**34.9**

4.3 Implementation Details

To extract image features, we use ResNet-101 (He et al., 2016) pre-trained on the Imagenet dataset (Russakovsky et al., 2015). We use features from the convolutional layer with dimensionality of 1024 × 14 × 14. The hidden size is set to 512 and the number of attention heads in SSRE is set to 4. The words are represented by trainable 300D word embedding features. POS tags are divided into 16 categories. In the training phase, on CLEVR-Change and Spot-the-Diff, we respectively set the mini-batch size as 128 and 96. We use Adam optimizer (Kingma and Ba, 2014) with the learning rate of 1 $\times 10^{-3}$ and 5 $\times 10^{-4}$, respectively. At inference, greedy decoding strategy is used to generate target captions. Both training and inference are implemented with PyTorch (Paszke et al., 2019) on a TITAN Xp GPU.

4.4 Ablation studies

In order to figure out the contribution of each module, we carry out the following ablation studies on CLEVR-Change: (1) Baseline which is based on DDUA (Park et al., 2019); (2) SSRE which only embeds the self-semantic relations of objects into their representations; (3) CSRM which only measures the cross-semantic relations between the captured candidate difference and the original images, and the learned discriminative difference representation is used as a prior knowledge to guide the change localizer; (4) SRDRL which is the combination of (2) and (3); (5) AVS which only relies on the POS information to determine when to use visual information and which of them should be used; (6) SRDRL+AVS which is the combination of (4) and (5).

The Evaluation on Total Performance. We frist study the total performance of each block of the proposed method under the whole dataset, including scene change and none-scene change. Experimental results are shown in Table 1. We can observe that each module of the proposed method improves the total performance of the baseline. Moreover, the best performance is achieved when putting them together, which indicates each block not only plays its unique role, but also can be a supplementary role for the others. This global statistical performance validates the generalization ability of the proposed method, that is, it not only can explicitly judge whether there is a semantic change between a pair of unaligned images, but also can describe the change using an accurate sentence.

The Evaluation on Scene Change and None-scene Change. The experimental results are shown

Table 3: Comparing with state-of-the-art methods on CLEVR-Change in Total Performance. RL is short for reinforcement learning training strategies.

		Total				
Method	RL	B-4	M	R	C	S
Capt-Dual (Park et al., 2019)	×	43.5	32.7	-	108.5	23.4
DDUA (Park et al., 2019)	×	47.3	33.9	-	112.3	24.5
M-VAM (Shi et al., 2020)	×	50.3	37.0	69.7	114.9	30.5
M-VAM+RAF (Shi et al., 2020)	✓	51.3	37.8	70.4	115.8	30.7
SRDRL+AVS	×	**54.9**	**40.2**	**73.3**	**122.2**	**32.9**

Table 4: Comparing with state-of-the-art methods on CLEVR-Change in terms of two settings.

		Scene Change				None-scene Change			
Method	RL	B-4	M	C	S	B-4	M	C	S
Capt-Dual (Park et al., 2019)	×	38.5	28.5	89.8	18.2	56.3	44.0	108.9	28.7
DDUA (Park et al., 2019)	×	42.9	29.7	94.6	19.9	59.8	45.2	110.8	29.1
M-VAM+RAF (Shi et al., 2020)	✓	-	-	-	-	-	**66.4**	**122.6**	33.4
SRDRL+AVS	×	**52.7**	**36.4**	**114.2**	**30.8**	**62.2**	51.3	117.0	**34.9**

in Table 2, in terms of scene change, we can observe that 1) SSRE, CSRM and AVS all achieve improvements over the baseline; 2) compared with SSRE, the improvement is relatively small when respectively using CSRM and AVS; 3) better performances are achieved when using two kinds of combinations (SRDRL and SRDRL+AVS). These indicate 1) the effectiveness of our proposed SRDRL and its single block, as well as the AVS; 2) the priority of this task is to capture the semantic difference in the image pair. The reason is that only if the semantic difference is captured sufficiently, can the following specific change localization and caption generation do well on itself part.

Besides, we can observe that although each single block can improve the baseline in the case of scene change, but they are worse than the baseline in one or more metrics in the case of none-scene change. Our conjecture is that the robustness of single block is relatively weak, so it would sometimes misidentify the illumination or viewpoint change as the actual semantic change. When observing the performance of two kinds of combinations (SRDRL and SRDRL+AVS), both of them improve the baseline in all metrics, which indicates the robustness of our overall model is strong.

4.5 Performance Comparison

4.5.1 Results on CLEVR-Change

In this dataset, we compare with four state-of-the-art methods, Capt-Dual (Park et al., 2019), DUDA (Park et al., 2019), M-VAM (Shi et al., 2020) and M-VAM+RAF (Shi et al., 2020), in four dimensions: 1) the total performance of scene change and none-scene change; 2) only scene change; 3) only none-scene change; 4) specific type of scene change. The comparison results are shown in Table 3, Table 4, and Table 5, respectively.

From Table 3, in terms of total performance, we can clearly observe that our method achieves significant improvements over them in all evaluation metrics, in particular with an increase of 34.3% and 7.2% in SPICE, respectively. From Table 4, under two kinds of settings, we can observe that our method outperforms DDUA with a large margin. Furthermore, since the M-VAM+RAF did not report the results on scene change, we only compare with them in the setting of none-change. We can observe that it outperforms us in METEOR and CIDEr. This superiority could derive from the reinforcement learning strategy. However, this strategy will remarkably increase training time and computation complexity. Moreover, as reported in Table 3, our total performance is much better than them, which is evaluated under the both scene change and none-scene change. Hence, compared to them, our method is more robust due to the discriminative difference representation learning.

Table 5 is the detailed breakdown of the evaluation based on five change types: “Color” (C), “Texture” (T), “Add” (A), “Drop” (D), and “Move” (M). Specifically, compare to all SOTA methods,

Table 5: A Detailed breakdown of Change Captioning evaluation on CLEVR-Change by different change types: “Color” (C), “Texture” (T), “Add” (A), “Drop” (D), and “Move” (M).

Method	RL	Metrics	C	T	A	D	M
Capt-Dual (Park et al., 2019)	×	CIDEr	115.8	82.7	85.7	103.0	52.6
DDUA (Park et al., 2019)	×	CIDEr	120.4	86.7	108.3	103.4	56.4
M-VAM+RAF (Shi et al., 2020)	✓	CIDEr	122.1	98.7	**126.3**	115.8	82.0
SRDRL+AVS	×	CIDEr	**136.1**	**122.7**	121.0	**126.0**	78.9
Capt-Dual (Park et al., 2019)	×	METEOR	32.1	26.7	29.5	31.7	22.4
DDUA (Park et al., 2019)	×	METEOR	32.8	27.3	33.4	31.4	23.5
M-VAM+RAF (Shi et al., 2020)	✓	METEOR	35.8	32.3	37.8	36.2	27.9
SRDRL+AVS	×	METEOR	**39.0**	**35.6**	**38.9**	**38.0**	**30.1**
Capt-Dual (Park et al., 2019)	×	SPICE	19.8	17.6	16.9	21.9	14.7
DDUA (Park et al., 2019)	×	SPICE	21.2	18.3	22.4	22.2	15.4
M-VAM+RAF (Shi et al., 2020)	✓	SPICE	28.0	26.7	30.8	**32.3**	22.5
SRDRL+AVS	×	SPICE	**32.4**	**30.9**	**33.0**	**32.4**	**25.4**

Table 6: Comparing with state-of-the-art methods on Spot-the-Diff.

Method	RL	M	R	C	S
DDLA	×	12.0	28.6	32.8	-
DDUA	×	11.8	29.1	32.5	-
SDCM	×	12.7	29.7	36.3	-
FCC	×	12.9	29.9	36.8	-
static rel-att	×	13.0	28.3	34.0	-
dynamic rel-att	×	12.2	31.4	35.3	-
M-VAM	×	12.4	31.3	38.1	14.0
M-VAM+RAF	✓	12.9	**33.2**	**42.5**	17.1
SRDRL+AVS	×	**13.0**	31.0	35.3	**18.0**

our method significantly raises the CIDEr scores in “Color” and “Texture” types, which indicates our method can better distinguish the attribute change of objects from an illumination change. Besides, for the number or position change of objects (“Add”, “Drop”, and “Move”), our method all outperforms them in most of metrics. Especially for SPICE, compared to them, our method has 64.9% and 12.9% improvements for “Move” case, respectively, which also shows our method can better localize the object movement from the viewpoint change. In particular, the most challenging change types are “Texture” and “Move” in this dataset, because they are most often confused with the illumination or viewpoint changes (Park et al., 2019). The relative experiments show that our method is more robust than SOTAs, and this benefits from the fact that the CSRM block helps attend to the actually semantic change by measuring the cross-semantic relations of the image pair and their difference.

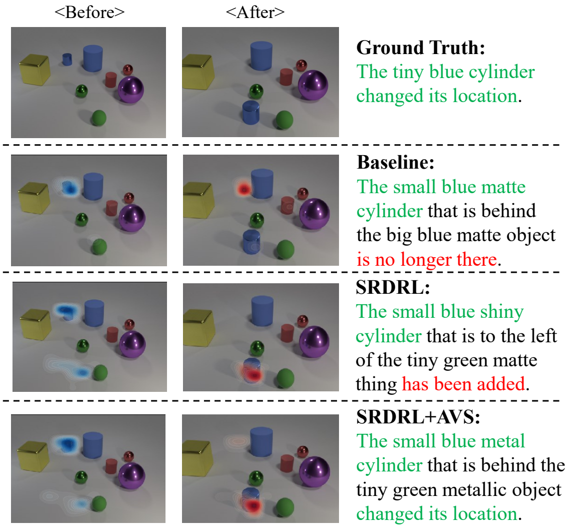
<Before>
<After>
Ground Truth:
The tiny blue cylinder changed its location.
Baseline:
The small blue matte cylinder that is behind the big blue matte object is no longer there.
SRDRL:
The small blue shiny cylinder that is to the left of the tiny green matte thing has been added.
SRDRL+AVS:
The small blue metal cylinder that is behind the tiny green metallic object changed its location.

Figure 3: A comparative example about “Move” case from the test set of CLEVR-Change, which involves the caption generated by the baseline, SRDRL, and SRDRL+AVS. We visualize the localization results on “before” (blue) and “after” (red).

4.5.2 Results on Spot-the-Diff

To validate the generalization ability of the proposed method, we conduct the experiments on a recent published Spot-the-Diff dataset, where the image pairs are mostly well aligned and their is no viewpoint change. We compare with eight SOTA methods and most of them cannot consider handling viewpoint changes: DDLA (Jhamtani and Berg-Kirkpatrick, 2018), DDUA (Park et al., 2019), SDCM (Oluwasanmi et al., 2019a), FCC (Oluwasanmi et al., 2019b), static rel-att / dynamic rel-att (Tan et al., 2019), and M-VAM / M-VAM+RAF (Shi et al., 2020).

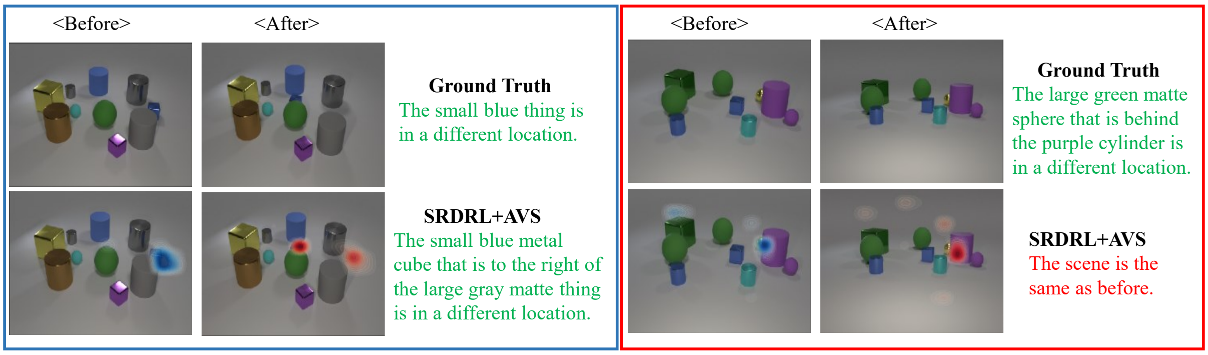
<Before>
<After>
Ground Truth
The small blue thing is in a different location.
SRDRL+AVS
The small blue metal cube that is to the right of the large gray matte thing is in a different location.
<Before>
<After>
Ground Truth
The large green matte sphere that is behind the purple cylinder is in a different location.
SRDRL+AVS
The scene is the same as before.

Figure 4: Qualitative examples of SRDRL+AVS. The left is a successful case that SRDRL+AVS localizes the accurate changed object and generates a correct sentence to describe the change. The right is a failure case that a slight movement of the object is not detected.

The results are reported in Table 6. We can observe that our method achieves the best performance in terms of METEOR and SPICE. Especially for SPICE which is recently designed for evaluating the image captioning task, our method achives 28.6% and 5.3% improvements over the current SOTA method M-VAM and M-VAM+RAF. Hence, compared to the above methods, the generated captions by our method are more in line with standards of human caption evaluation. This superiority results from that the SSRE block can capture the relation-embedded feature difference so as to better explore those tiny changed objects.

4.6 Qualitative Analysis

Figure 3 shows a comparative example about "Move" from the CLEVR-Change dataset, which includes the change captions generated by humans, baseline, SRDRL, and SRDRL+AVS. We also visualize the results of change detection. The baseline is implemented based on DDUA (Park et al., 2019). We can clearly observe that it localizes a wrong region on the "after" and thus misidentifies "Move" as "Drop". By contrast, both proposed methods (SRDRL and SRDRL+AVS) can accurately localize the moved object on both "before" and "after" images, which validates the effectiveness of the proposed SRDRL. Moreover, it is interesting to note that, for the proposed methods, although the results of change localization are accurate, only using SRDRL generates a wrong caption, which indicates the POS tags of target words indeed guide and regularize the change caption generation.

Figure 4 illustrates two examples with viewpoint changes on CLEVR-Change dataset. The left example is a success in which SRDRL+AVA can distinguish the small blue changed cube from the irrelevant viewpoint change. This benefits from that SRDRL can learn discriminative difference representation and overcome viewpoint changes. The right example shows a failure, where SRDRL+AVA judges there is no difference. Our conjecture is that the movement of this sphere is very slight and thus confused with the viewpoint change. Hence, we will improve our method to learn more fine-grained difference representation in the future work.

5 Conclusion

In this paper, we propose a semantic relation-aware difference representation learning network (SRDRL) and attention-based visual switch (AVS) to address change captioning in the presence of distractors, where SRDRL can explicitly learn the difference representation in the image pair and AVS can aid the caption generator to convey the localized change in a logically and grammatically accurate sentence. Extensive experiments conducted on both CLEVR-Change and Spot-the-Diff datasets show that the proposed method achieves state-of-the-art results.

Acknowledgements

The work was supported by National Natural Science Foundation of China (Grant Nos. 61761026, 61972186, 61732005 and 61762056), National Key Research and Development Plan (Grant Nos. 2019QY1802, 2019QY1801, and 2019QY1800), Yunnan high-tech industry development project (Grant No. 201606), Yunnan provincial major science and technology special plan projects (Grant No. 202002AD080001-5), and Yunnan Basic Research Project (Grant Nos. 202001AS070014, 2018FB104).

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *ECCV*, pages 382–398.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, pages 65–72.

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

Jincan Deng, Liang Li, Beichen Zhang, Shuhui Wang, Zhengjun Zha, and Qingming Huang. 2021. Syntax-guided hierarchical attention network for video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

L. Gueguen and R. Hamid. 2015. Large-scale damage detection using satellite imagery. In *CVPR*, pages 1321–1328.

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

Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned dual channel graph convolutional network for visual question answering. In *ACL*, pages 7166–7176.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In *EMNLP*, pages 4024–4034.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liang Li, Shijie Yang, Li Su, Shuhui Wang, Chenggang Yan, Zheng-jun Zha, and Qingming Huang. 2020. Diverter-guider recurrent network for diverse poems generation from image. In *ACM MM*, pages 3875–3883.

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

Xuejing Liu, Liang Li, Shuhui Wang, Zheng-Jun Zha, Dechao Meng, and Qingming Huang. 2019. Adaptive reconstruction network for weakly supervised referring expression grounding. In *ICCV*, pages 2611–2620.

Zhenhuan Liu, Jincan Deng, Liang Li, Shaofei Cai, Qianqian Xu, Shuhui Wang, and Qingming Huang. 2020. Ir-gan: Image manipulation with linguistic instruction by increment reasoning. In *ACM MM*, pages 322–330.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 3242–3250.

Ariyo Oluwasanmi, Muhammad Umar Aftab, Eatedal Alabdulkreem, Bulbula Kumeda, Edward Y Baagyere, and Zhiquang Qin. 2019a. Captionnet: Automatic end-to-end siamese difference captioning model with attention. *IEEE Access*, 7:106773–106783.

Ariyo Oluwasanmi, Enoch Frimpong, Muhammad Umar Aftab, Edward Y Baagyere, Zhiguang Qin, and Kifayat Ullah. 2019b. Fully convolutional captionnet: Siamese difference captioning attention model. *IEEE Access*, 7:175929–175939.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *ICCV*, pages 4623–4632.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

J. Patriarche and B. Erickson. 2004. A review of the automated detection of change in serial imaging studies of the brain. *Journal of Digital Imaging*, 17:158–174.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *In IJCV,*, pages 211–252.

Ken Sakurada and Takayuki Okatani. 2015. Change detection from a street image pair using cnn features and superpixel segmentation. In *BMVC*.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq R. Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encode for change captioning. In *ECCV*, pages 574–590.

Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. 2017. Hierarchical lstm with adjusted temporal attention for video captioning. In *IJCAI*, pages 2737–2743.

Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883, Florence, Italy. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*.

Yunbin Tu, Xishan Zhang, Bingtao Liu, and Chenggang Yan. 2017. Video description with spatial-temporal attention. In *ACM MM*, pages 1014–1022.

Yunbin Tu, Chang Zhou, Junjun Guo, Shengxiang Gao, and Zhengtao Yu. 2020. Enhancing the alignment between target words and corresponding frames for video captioning. *Pattern Recognition*, page 107702.

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 *NeurIPS*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, Jingwen Wang, and Wei Liu. 2019. Controllable video captioning with pos sequence guidance based on gated fusion network. In *ICCV*, pages 2641–2650.

Aming Wu, Linchao Zhu, Yahong Han, and Yi Yang. 2019. Connective cognition network for directional visual commonsense reasoning. In *NeurIPS*, pages 5669–5679.

Chenggang Yan, Biao Gong, Yuxuan Wei, and Yue Gao. 2020a. Deep multi-view enhancement hashing for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. 2021. Task-adaptive attention for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chenggang Yan, Liang Li, Chunjie Zhang, Bingtao Liu, Yongdong Zhang, and Qionghai Dai. 2019. Cross-modality bridging and knowledge transferring for image understanding. *IEEE Transactions on Multimedia*, 21(10):2675–2685.

Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2020b. STAT: spatial-temporal attention mechanism for video captioning. *IEEE Trans. Multimedia*, 22(1):229–241.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *ACL*, pages 3025–3035.

Yongjing Yin, Jinsong Su, Huating Wen, Jiali Zeng, Yang Liu, and Yidong Chen. 2019. Pos tag-enhanced coarse-to-fine attention for neural machine translation. *ACM transactions on Asian language information processing*, 18(4):46.1–46.14.

Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. 2017. Task-driven dynamic fusion: Reducing ambiguity in video description. In *CVPR*, pages 3713–3721.