

Finding It at Another Side: A Viewpoint-Adapted Matching Encoder for Change Captioning

Xiangxi Shi¹, Xu Yang¹, Jiuxiang Gu², Shafiq Joty¹, and Jianfei Cai^{1,3}

¹ Nanyang Technological University, 50 Nanyang Avenue, Singapore
 $\{\text{xxshi, srjoty, asjfcai}\}@ntu.edu.sg$, $s170018@e.ntu.edu.sg$

² Adobe Research, USA
 $jigu@adobe.com$

³ Monash University, Victoria 3800, Australia
 $jianfei.cai@monash.edu$

Abstract. Change Captioning is a task that aims to describe the difference between images with natural language. Most existing methods treat this problem as a difference judgment without the existence of distractors, such as viewpoint changes. However, in practice, viewpoint changes happen often and can overwhelm the semantic difference to be described. In this paper, we propose a novel visual encoder to explicitly distinguish viewpoint changes from semantic changes in the change captioning task. Moreover, we further simulate the attention preference of humans and propose a novel reinforcement learning process to fine-tune the attention directly with language evaluation rewards. Extensive experimental results show that our method outperforms the state-of-the-art approaches by a large margin in both Spot-the-Diff and CLEVR-Change datasets.

Keywords: Image Captioning, Change Captioning, Attention, Reinforcement learning

1 Introduction

The real world is a complex dynamic system. Changes are ubiquitous and significant in our day-to-day life. As humans, we can infer the underlying information from the detected changes in the dynamic task environments. For instance, a well-trained medical doctor can judge the development of a patient’s condition better by comparing the CT images captured in different time steps in addition to locating the lesion.

In recent years, a great deal of research has been devoted to detecting changes in various image content [5,27,1,7], which typically require a model to figure out the differences between the corresponding images. However, most of them focus on the changes in terms of object classes without paying much attention to various attribute changes (e.g., color and size). Moreover, *change detection* usually requires pixel-level densely labeled ground-truth data for training the model. Acquiring such labeled data is too time-consuming and labor-intensive,

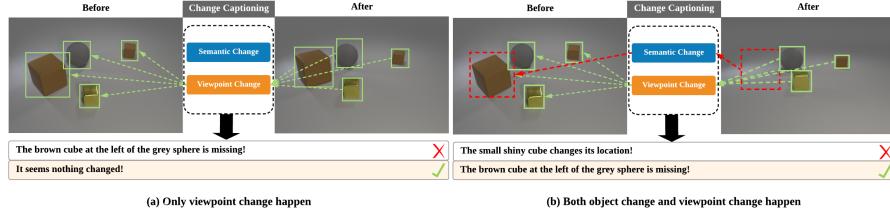


Fig. 1. Current change captioning methods can be influenced by the viewpoint change between image pairs, which results in generating wrong captions. To address this problem, we propose a novel framework that aims at explicitly separating viewpoint changes from semantic changes.

some of which, such as in medical image applications even require human experts for generating annotations.

In addition, current change detection methods usually assume that the differences occur in an ideal situation. However, in practice, changes such as illumination and viewpoint changes occur quite frequently, which makes the change detection task even more challenging.

As humans, we can describe the differences between images accurately, since we can identify the most significant changes in the scene and filter out tiny differences and irrelevant noise. Inspired by the research development from dense object detection to image captioning, which generates natural language descriptions (instead of object labels) to describe the salient information in an image, the *change captioning task* has recently been proposed to describe the salient differences between images [22,14,20,19]. Arguably, captions describing the changes are more accessible (hence preferred) for users, compared to the map-based labels (e.g., pixel-level binary maps as change labels).

Despite the progress, the existing change captioning methods cannot handle the viewpoint change properly [22]. As shown in Fig. 1, the viewpoint change in the images can overwhelm the actual object change leading to incorrect captions. Handling viewpoint changes is more challenging as it requires the model to be agnostic of the changes in viewpoints from different angles, while being sensitive to other salient changes.

Therefore, in this paper, following the prevailing architecture of a visual encoder plus a sentence decoder, we propose a novel viewpoint-agnostic image encoder, called Mirrored Viewpoint-Adapted Matching (M-VAM) encoder, for the change captioning task. Our main idea is to exhaustively measure the feature similarity across different regions in the two images so as to accurately predict the changed and unchanged regions in the feature space. The changed and unchanged regions are formulated as probability maps, which are used to synthesize the changed and unchanged features. We further propose a Reinforcement Attention Fine-tuning (RAF) process to allow the model to explore other caption choices by perturbing the probability maps.

Overall, the contributions of this work include:

- We propose a novel M-VAM encoder that explicitly distinguishes semantic changes from the viewpoint changes by predicting the changed and unchanged regions in the feature space.
- We propose a RAF module that helps the model focus on the semantic change regions so as to generate better change captions.
- Our model outperforms the state-of-the-art change captioning methods by a large margin in both Spot-the-Diff and CLEVR-Change datasets. Extensive experimental results also show that our method produces more robust prediction results for the cases with viewpoint changes.

2 Related Work

Image Captioning. Image captioning has been extensively studied in the past few years [9,26,28,3,11,12,30]. Farhadi *et al.* [9] proposed a template-based alignment method, which matches the visual triplets detected from the images with the language triplets extracted from the sentences. Unlike the template-based solutions, Vinyals *et al.* [26] proposed a Long Short-Term Memory (LSTM) network-based method. It takes the image features extracted from a pre-trained CNN as input and generates the image descriptions word by word. After that, Xu *et al.* [28] introduced an attention mechanism, which predicts a spatial mask over encoded features at each time step to select the relevant visual information automatically. Anderson *et al.* [3] attempted to mimic the human’s visual system with a top-down attention structure. Yang *et al.* [29] introduced a method to match the scene graphs extracted from the captions and images using Graph Convolutional Networks.

Another theme of improvement is to use reinforcement learning (RL) to address the training and testing mismatch (aka. exposure bias) problem [23,31]. Rennie *et al.* [24] proposed the Self-Critical Sequence Training (SCST) method, which optimizes the captioning model with sentence-level rewards. Gu *et al.* [10] further improved the SCST-based method and proposed a stack captioning model to generate the image description from coarse to fine. Chen *et al.* [6], introduced Recurrent Neural Network based discriminator to evaluate the naturalness of the generated sentences.

Captioning for Change Detection. Change detection tasks aim to identify the differences between a pair of images taken at different time steps to recognize the information behind them. It has a wide range of applications, such as scene understanding, biomedical, and aerial image analysis. Bruzzone *et al.* [5] proposed an unsupervised method to overcome the huge cost of human labeling. Wang *et al.* [27] used CNN to predict the changing regions with their proposed mixed-affinity matrix. Recently, Alcantarilla *et al.* [1] proposed a deconvolutional network combined with a multi-sensor fusion SLAM (Simultaneous Localization and Mapping) system to achieve street-view change detection in the real world.

Instead of detecting the difference between images, some studies have explored to describe the difference between images with natural language. Jham-

tani *et al.* [14] first proposed the change captioning task and published a dataset extracted from the VIRAT dataset [18]. They also proposed a language decoding model with an attention mechanism to match the hidden state and features of annotated difference clusters. Oluwasanmi *et al.* [20] introduced a fully convolutional siamese network and an attention-based language decoder with a modified update function for the memory cell in LSTM. To address the viewpoint change problem in the change captioning task, Park *et al.* [22] built a synthetic dataset with a viewpoint change between each image pair based on the CLEVR engine. They proposed a Dual Dynamic Attention Model (DUDA) to attend to the images (“before” and “after”) for a more accurate change object attention.

In contrast, in our work, we propose a M-VAM encoder to overcome the confusion raised by the viewpoint changes. Compared with Difference Description with Latent Alignment (DDLA) proposed by Jhamtani *et al.* [14], our method does not need ground-truth difference cluster masks to restrict the attention weights. Instead, our method can self-predict the difference (changed) region maps by feature-level matching without any explicit supervision. Both DUDA [22] and Fully Convolutional CaptionNet (FCC) [20] apply subtraction between image features at the same location in the given image pair, which has problems with large viewpoint changes and thus cannot reliably produce correct change captions. Different from them, our model overcomes the viewpoint change problem by matching objects in the given image pair without the restriction of being at the same feature location.

3 Method

Figure 2 gives an overview of our proposed change captioning framework. The entire framework basically follows the prevailing encoder-decoder architecture and consists of two main parts: a Mirrored Viewpoint-Adapted Matching encoder to distinguishing the semantic change from the viewpoint-change surrounding, and a top-down sentence decoder to generate the captions. During training, we use two types of supervisions: we first optimize our model with traditional cross-entropy (XE) loss with the model at the bottom branch. And then we finetune our model with reinforcement learning using both top and bottom branches, which share the parameters, to further explore the information of objects matched in the image pairs.

3.1 Mirrored Viewpoint-Adapted Matching Encoder

We first encode the given “before” image I_b and the “after” image I_a into the spatial image feature maps $f_b \in \mathbb{R}^{W \times H \times C}$ and $f_a \in \mathbb{R}^{W \times H \times C}$, respectively, with a pretrained CNN, where W , H , and C are the width, the height, and the channel size of the feature maps, respectively.

Fig. 3 shows our proposed M-VAM encoder. It has of two VAM cells with shared parameters. Each VAM cell takes f_b and f_a as input and outputs the changed feature and the unchanged feature. Here, different orders of the inputs

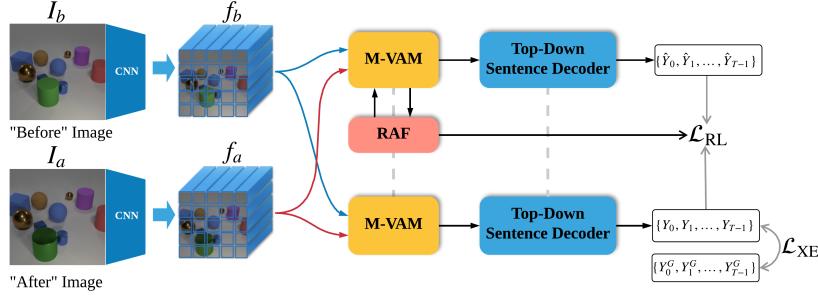


Fig. 2. Overview of the proposed change captioning framework. Our model mainly consists of one M-VAM encoder and one top-down sentence decoder. We use dashed lines to denote that the parameters of two models are shared with each other. Note that with a pre-trained CNN, the model first encodes the *before* and *after* images (I_b and I_a) into features (f_b and f_a) before feeding into M-VAM. The entire model is trained with a cross-entropy loss (\mathcal{L}_{XE}) first, followed by a reinforcement learning loss (\mathcal{L}_{RL}) based on our RAF process.

lead to different outputs. For example, when a VAM takes the input (f_b, f_a) , it treats f_b as the reference and computes the changed (or difference) feature $f_{d_{b \rightarrow a}}$ and the unchanged (or background) feature f_{bb} . Similarly, for input order (f_a, f_b) , the VAM computes the difference feature $f_{d_{a \rightarrow b}}$ and the background feature f_{ba} .

Viewpoint-Adapted Matching (VAM) Cell Given the extracted image features f_a and f_b , we aim to detect where the changes are in the feature space. The traditional way is to directly compute the difference of the two feature maps [22], which can hardly represent the semantic changes of the two images in the existence of a viewpoint change. To solve this problem, we propose to identify the changed and the unchanged regions by comparing the similarity between different patches in the two feature maps.

In particular, for any cell location (x, y) in f_b and any cell location (i, j) in f_a , we compute their feature-level similarity as follows

$$\mathbf{E}_{x,y}(i, j) = f_{b(x,y)}^T f_{a(i,j)} \quad (1)$$

With the obtained $\mathbf{E} \in \mathbb{R}^{WH \times WH}$, we then generate a synthesized feature map f_{sb} for image I_b from the other image feature f_a as

$$f_{sb(x,y)} = \sum_{i,j} f_{a(i,j)} \cdot \text{softmax}(\mathbf{E}_{x,y}(i, j)) \quad (2)$$

The softmax in Eq. (2) serves as a soft feature selection operation for $f_{a(i,j)}$ according to the similarity matrix \mathbf{E} . When an object in $f_{b(x,y)}$ can be found at

$f_{a(i,j)}$, the score of $E_{x,y}(i,j)$ should be higher than the score of other features. In this way, the most corresponding feature can be selected out to synthesize the feature of $f_{sb(x,y)}$, i.e., the objects in the *after* image can “move back” to the original position in the *before* image at the feature level.

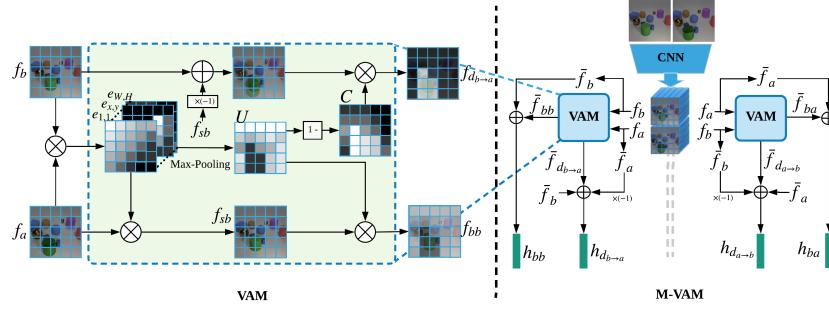


Fig. 3. Left: structure of a VAM Cell. Right: proposed M-VAM encoder, which consists of two VAM cells with shared parameters but with the inputs of f_a and f_b in different orders.

So far, our method just blindly searches for the most similar patch in f_a to reconstruct f_b without taking into account the fact that the most similar one in f_a might not be a correct match. Thus, we further compute an unchanged probability map $U_{x,y} \in [0, 1]^{WH}$ and a changed probability map $C_{x,y} \in [0, 1]^{WH}$ as follows.

$$U_{x,y} = \max_{i,j} (\sigma(a_u E_{x,y}(i,j) + b_u)) \quad (3)$$

$$C_{x,y} = 1 - U_{x,y} \quad (4)$$

where σ is the sigmoid function, a_u and b_u are the learned (scalar) parameters. Finally, as shown in Fig. 3 (left), with the obtained probability maps, the original and the synthesized feature maps f_b and f_{sb} , we compute the outputs of the VAM, i.e., the unchanged feature f_{bb} and the changed feature $f_{d_{b \rightarrow a}}$ as:

$$f_{bb} = f_{sb} \odot U \quad (5)$$

$$f_{d_{b \rightarrow a}} = (f_b - f_{sb}) \odot C \quad (6)$$

where \odot is the element-wise product.

M-VAM Using one VAM cell with the input order of (f_b, f_a) , we can now predict the unchanged and changed features (f_{bb} and $f_{d_{b \rightarrow a}}$), but all of them are with respect to the “before” image, I_b . For an ideal change captioner, it should catch all the changes in both of the images. Therefore, we introduce the mirrored

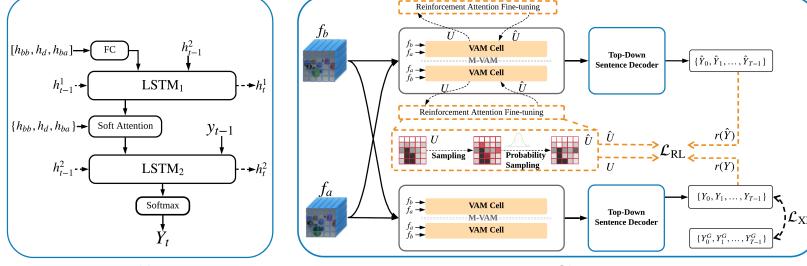


Fig. 4. (a): structure of the top-down decoder. (b): proposed RAF process.

encoder structure, as shown in Fig. 3 (right). Specifically, we first use the VAM cell to predict the changed and unchanged features w.r.t. I_b , and then the same VAM cell is used again to predict the output features w.r.t. I_a . Formally, the encoding process can be written as

$$f_{d_{b \rightarrow a}}, f_{bb} = \text{VAM}(f_b | f_a) \quad (7)$$

$$f_{d_{a \rightarrow b}}, f_{ba} = \text{VAM}(f_a | f_b) \quad (8)$$

Note that such a mirror structure also implicitly encourages the consistency of the matched patch pairs from the two opposite directions.

Finally, we treat the features obtained in Eq. (7) and Eq. (8) as the residual information and add them to the original image feature maps:

$$h_{bi} = \bar{f}_{bi} + \bar{f}_i, \quad (9)$$

$$h_{d_{i \rightarrow j}} = \bar{f}_{d_{i \rightarrow j}} + \bar{f}_i - \bar{f}_j, \quad (10)$$

where $i, j \in \{a, b\}$ and $i \neq j$, and \bar{f}_* is the resultant feature vector after applying an average pooling on f_* .

3.2 Sentence Decoder

As shown in Fig. 4(a), we adopt a top-down structure [3] as our sentence decoder. It consists of two LSTM networks and a soft attention module. Since we have two changed features generated from the M-VAM encoder, to obtain a consistent changed feature, we first apply a fully connected layer to fuse the two features

$$h_d = \mathbf{W}_1[h_{d_{a \rightarrow b}}, h_{d_{b \rightarrow a}}] + \mathbf{b}_1 \quad (11)$$

where \mathbf{W}_1 and \mathbf{b}_1 are learned parameters.

With the set of the encoded changed and unchanged features $[h_d, h_{bb}, h_{ba}]$, the sentence decoder operates as follows:

$$h_t^1 = \text{LSTM}_1([\mathbf{W}_2[h_d, h_{bb}, h_{ba}] + \mathbf{b}_2, h_{t-1}^2]) \quad (12)$$

$$a_{i,t} = \text{softmax}(\mathbf{W}_3 \tanh(h_i + h_t^1) + \mathbf{b}_3), \quad (13)$$

$$h_t^2 = \text{LSTM}_2([\sum a_{i,t} \times h_i, y_{t-1}]) \quad (14)$$

$$Y_t \sim \text{softmax}(\mathbf{W}_4 h_t^2 + \mathbf{b}_4) \quad (15)$$

where h_t^1 and h_t^2 are the hidden states of the first LSTM and the second LSTM, respectively, $a_{i,t}$ is the soft attention weight for the visual feature h_i at time step t , $i \in \{d, bb, ba\}$ refers to different input features, y_{t-1} is the previous word embedding, and Y_t is the output word drawn from the dictionary at time step t according to the softmax probability.

3.3 Learning Process

Cross-entropy Loss We first train our model by minimizing the traditional cross-entropy loss. It aims to maximize the probability of the ground truth caption words given the two images:

$$\mathcal{L}_{\text{XE}} = - \sum_{t=0}^{T-1} \log(p_\theta(Y_t^G | Y_{0:t-1}^G, I_a, I_b)) \quad (16)$$

where θ is the parameter of our model, and Y_t^G is the t -th ground-truth word.

Reinforcement Attention Fine-tuning (RAF) Training with the XE loss alone is usually insufficient, and the model suffers from the exposure bias problem [4]. The ground-truth captions of an image contain not only language style information but also information about preferential choices. When it comes to describing the change in a pair of images, a good captioning model needs to consider the relations between the objects and how the objects change. In other words, the model needs to learn what and when to describe. Besides, as humans, when we describe the changed objects, we often include some surrounding information. All these different expressions are related to different visual information in the images in a complicated way. To model such complicated relations, we propose to predict attention regions (i.e., focus areas to describe the changes) in a more human-like way that can help the model to generate more informative captions. Different from image captioning tasks, the images here are encoded only once to decode the sentence (i.e., before starting the decoding, our M-VAM encoder encodes an image pair into features, h_d, h_{bb}, h_{ba} , and keeps them unchanged during the entire decoding process as opposed to computing a different attention or context vector at each decoding step), for which the standard RL would have much larger influence on the decoder rather than on the encoder. Therefore, we propose a RAF process based on the attention perturbation to further optimize our encoder-decoder model.

As shown in Fig. 4(b), our RAF process generates two captions for the same image pair. The first caption $Y = \{Y_0, Y_1, \dots, Y_{T-1}\}$ is generated in the usual way using our model as shown in the bottom branch of Fig. 4(b), while the second caption $\hat{Y} = \{\hat{Y}_0, \hat{Y}_1, \dots, \hat{Y}_{T-1}\}$ is generated in a slightly different way according to a sampled probability map \hat{U} (and \hat{C}), as shown in the top branch. Here, \hat{U} is a slightly modified map of U by introducing some small noise, which will lead to the generation of a different caption. By comparing the two captions Y and \hat{Y} , and rewarding the better one with the reinforcement learning, our model can essentially *explore* for a better solution.

Specially, to generate \hat{U} , we first randomly select some elements to perturb from the unchanged probability map U with a sampling rate of γ . For the selected element $U_{i,j}$, we get its perturbed version by sampling from the following Gaussian distribution:

$$\hat{U}_{i,j} \sim \mathcal{N}(U_{i,j}, c) \quad (17)$$

where the standard deviation c is a small number (e.g., 0.1 in our experiments) that is considered as a hyperparameter. For the elements that are not selected for perturbation, $\hat{U}_{i,j} = U_{i,j}$. With \hat{U} , we can then calculate the corresponding changed map \hat{C} using Eq. (4) and generate the caption \hat{Y} accordingly.

After that, we calculate the CIDEr scores for the two captions and treat them as rewards, i.e., $r(Y) = \text{CIDEr}(Y)$ and $r(\hat{Y}) = \text{CIDEr}(\hat{Y})$. If $r(\hat{Y})$ is higher than $r(Y)$, it means that \hat{Y} describes the change of the objects better than Y (according to CIDEr). Consequently, the predicted probability map U should be close to the sampled probability map \hat{U} . On other other hand, if $r(\hat{Y})$ is smaller than $r(Y)$, it suggests that the sampled probability map is worse than the original one, thus should not be used to guide the learning. Based on this intuition, we define our reinforcement loss as follows.

$$\mathcal{L}_{\text{RL}} = \mathbb{1}^+(r(\hat{Y}) - r(Y)) \|\hat{U} - U\|_2^2, \quad (18)$$

where $\mathbb{1}^+(x)$ represents $\max(x, 0)$. The gradients from \mathcal{L}_{RL} will backward to the M-VAM module. The overall training loss becomes:

$$\mathcal{L}_{\text{XE+RL}} = \lambda \mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{XE}} \quad (19)$$

where λ is a hyper-parameter to control the relative weights of the two losses.

4 Experiment

4.1 Datasets and Metrics

We evaluate our method on two published datasets: Spot-the-Diff [14] and CLEVR-Change [22]. Spot-the-Diff contains 13,192 image pairs extracted from the ViRAT dataset [18], which were captured by stationary ground cameras. Each image pair can differ in multiple ways; on average, each image pair contains 1.86 reported differences. CLEVR-Change is a synthetic dataset with a set of basic geometry objects, which is generated based on the CLEVR engine. Different from Spot-the-Diff dataset, to address the viewpoint change problem, CLEVR-Change involves a viewpoint change between each image pair. The dataset contains 79,606 image pairs and 493,735 captions. The changes in the dataset can be categorized into six cases, which are Color, Texture, Movement, Add, Drop, and Distractor.

We evaluate our generated captions with five evaluation metrics: BLEU [21], METEOR [8], CIDEr [25], ROUGE_L [17], and SPICE [2]. For comparison in terms of BLEU scores, we mainly show the results for BLEU4 which reflects the matching ratio of 4-word sub-sequences between the generated caption and the ground truth caption.

4.2 Training Details

We use ResNet101 [13] trained on the ImageNet dataset [16] to extract the image features. The images are resized to 224×224 before extracting the features. The feature map size is $\mathbb{R}^{14 \times 14 \times 1024}$. During the training, we set the maximum length of captions as 20 words. The dimension of hidden states in the Top-Down decoder is set to 512. We apply a drop-out rate of 0.5 before the word probability prediction (Eq.(15)). We apply ADAM [15] optimizer for training.

When training with the cross-entropy loss, we train the model for 40 epochs with a learning rate of 0.0002. During the reinforcement attention fine-tuning (RAF) process, we set the learning rate to 0.00002. Due to the low loss value from Eq. (18), we set the hyperparameter λ in Eq. (19) as 1000.0 to put more weight to the reinforcement learning loss. In the RAF process, we set the random sampling rate γ as 0.025. As for the Gaussian distribution (Eq. (17)), we set c as 0.1.

4.3 Model Variations

To analyze the contribution of each component in our model, we consider the following two variants of our model.

- 1) **M-VAM:** This variant models the encoder-decoder captioning framework with the proposed M-VAM encoder introduced in Sec. 3.1 and the top-down decoder introduced in Sec. 3.2.
- 2) **M-VAM + RAF:** This variant fine-tunes M-VAM with the proposed RAF process.

4.4 Results

Quantitative Results on Spot-the-Diff. Quantitative results of different methods on Spot-the-Diff dataset are shown at Table 1. We compare with the four existing methods – FCC [20], SDCM [19], DDLA [14] and DUDA [22]. It can be seen that both of our models outperform existing methods in all evaluation measures. Compared with FCC, M-VAM provides improvements of 0.2, 1.4 and 1.3 in BLEU4, ROUGE_L and CIDEr, respectively. After using the RAF process, the scores are further improved from 10.1 to 11.1 in BLEU4, 31.3 to 33.2 in ROUGE_L, 38.1 to 42.5 in CIDEr, and 14.0 to 17.1 in SPICE. The big gains in CIDEr (+4.4) and SPICE (+3.1) for the RAF fine-tuning demonstrate its effectiveness in change captioning.

Quantitative Results on CLEVR-Change. In Table 2, we compare our results with DUDA [22] on the CLEVR-Change dataset. CLEVR-Change contains additional distractor image pairs, which may not be available in other scenarios. Therefore, we performed separate experiments with and without the distractors to show the efficacy of our method.

Table 1. Comparisons of the results of different methods on Spot-the-Diff dataset.

Model	BLEU4	ROUGE_L	METEOR	CIDEr	SPICE
DUDA [22]	8.1	28.3	11.5	34.0	-
FCC [20]	9.9	29.9	12.9	36.8	-
SDCM [19]	9.8	29.7	12.7	36.3	-
DDLA [14]	8.5	28.6	12.0	32.8	-
M-VAM	10.1	31.3	12.4	38.1	14.0
M-VAM + RAF	11.1	33.2	12.9	42.5	17.1

Table 2. Comparisons of the results of different methods on CLEVR-Change dataset.

Model	BLEU4	ROUGE_L	METEOR	CIDEr	SPICE
DUDA [22]	47.3	-	33.9	112.3	24.5
<i>With Distractor</i>					
M-VAM	50.3	69.7	37.0	114.9	30.5
M-VAM + RAF	51.3	70.4	37.8	115.8	30.7
<i>Without Distractor</i>					
M-VAM	45.5	69.5	35.5	117.4	29.4
M-VAM + RAF	50.1	71.0	36.9	119.1	31.2

We can see that our model tested on the entire dataset (with distractor) achieves state-of-the-art performance and outperforms DUDA by a wide margin. Our M-VAM pushes the BLEU4, METEOR, SPICE and CIDEr scores from 47.3, 33.9, 24.4 and 112.3 to 50.3, 69.7, 37.0, 30.7 and 114.9, respectively. The RAF process further improves the scores by about 1 point in all measures.

Our model also achieves impressive performance on the dataset without the distractors. The M-VAM achieves a CIDEr score of 117.4. RAF further improves the scores – from 45.5 to 50.1 in BLEU4 and from 117.4 to 119.1 in CIDEr. The gains are higher in this setting than the gains with the entire dataset. There is also a visible increase in SPICE compared with the models trained with the distractors (from 30.7 to 31.2). This is because, without the influence of the distractors, the model can distinguish the semantic change more accurately from the viewpoint change.

Table 3 shows the detailed breakdown of the evaluation according to the categories defined in [22]. In the table, “Metrics” represent the metric used to evaluate the results for a certain model. “C” represents the changes related to “Color” attributes. Similarly, “T”, “A”, “D”, “M” and “DI” represent “Texture”, “Add”, “Drop”, “Move” and “Distractor”, respectively.

Comparing our results with DUDA, we see that our method can generate more accurate captions for all kinds of changes. In particular, it obtains large improvements in CIDEr for “Move” and “Add” cases (from 108.2 to 126.3 and from 56.4 to 82.0, respectively). For SPICE, the “Drop” cases have the largest improvement (from 22.2 to 32.3) among the six cases. This indicates that our method can better distinguish the object movement from a viewpoint change

Table 3. The Detailed breakdown of evaluation on the CLEVR-Change dataset.

Model	Metrics	C	T	A	D	M	DI
DUDA [22]	CIDEr	120.4	86.7	108.2	103.0	56.4	110.8
M-VAM + RAF (w DI)	CIDEr	122.1	98.7	126.3	115.8	82.0	122.6
M-VAM + RAF (w/o DI)	CIDEr	135.4	108.3	130.5	113.7	107.8	-
DUDA [22]	METEOR	32.8	27.3	33.4	31.4	23.5	45.2
M-VAM + RAF(w DI)	METEOR	35.8	32.3	37.8	36.2	27.9	66.4
M-VAM + RAF (w/o DI)	METEOR	38.3	36.0	38.1	37.1	36.2	-
DUDA [22]	SPICE	21.2	18.3	22.4	22.2	15.4	28.7
M-VAM + RAF (w DI)	SPICE	28.0	26.7	30.8	32.3	22.5	33.4
M-VAM + RAF (w/o DI)	SPICE	29.7	29.9	32.2	33.3	30.7	-

scene, which is the most challenging case as reported in [22]. The model can also match the objects well when objects are missing from the ‘before’ image or added to the ‘after’ image. Referring to the METEOR score, we get the highest improvement for the “Distractor” cases. This means that our method can better discriminate the scene only with the viewpoint change from the other cases.

Comparing the model trained with and without the distractors, we can see that the performance is further boosted in “Color”, “Texture” and “Move” cases. The SPICE scores also show an increase in the “Move” case (from 22.5 to 30.7). It means that such cases are easier to be influenced by the distractor data. This is because the movement of objects can be confused with the viewpoint movement. As for “Color” and “Texture”, it may be caused by the surrounding description mismatching according to the observation. When the model is trained without the distractors, our RAF process, which is proposed to capture a better semantic surrounding information, is more efficient and thus improves the performance in “Color” and “Texture”.

Qualitative Results We show the visualization of some examples from the CLEVR-Change dataset in Fig. 5. The images at the top of the figure are the given image pairs. The images in the middle two rows are the changed probability maps of M-VAM and M-VAM+RAF, respectively. Comparing the probability maps, we can see that the maps after the RAF process are brighter and more concentrated at the changed regions than the ones from M-VAM, which indicates RAF can force the model to focus on the actual changed regions. Note that a rigorous overlap comparison is not meaningful here since the model does not perform an explicit change detection to find the objects in the pixel level (the size of the attention map is only 14×14).

In the first example, we can see that the attention map from M-VAM covers the changed object in the image only partially, while also attending to some irrelevant portions above the changed object. Thus, with the shifted attention region in the “after” image, the model mistakenly infers that the cube is behind (as opposed to in the front) the cyan cube. The RAF fine-tuning helps the model focus more appropriately and model the change correctly. In the second

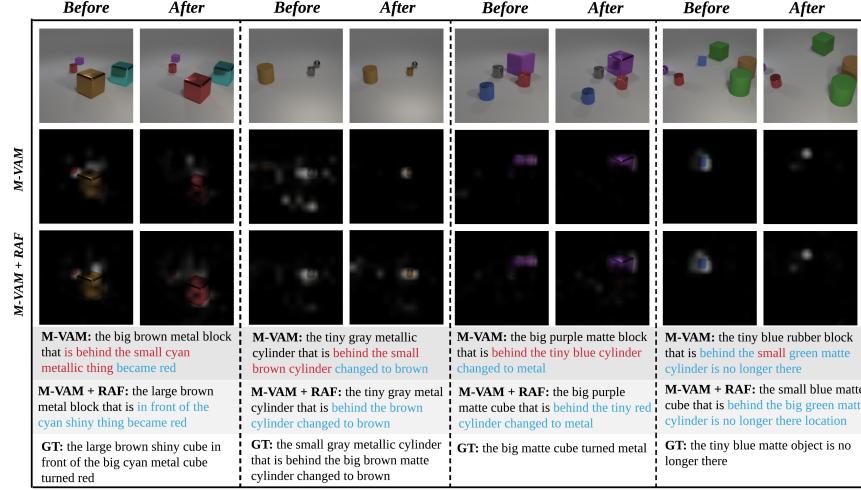


Fig. 5. Visualization of examples from the CLEVR-Change dataset. Red and blue color respectively denote the inconsistent and consistent words. The middle two rows show the changed probability maps of the two models.

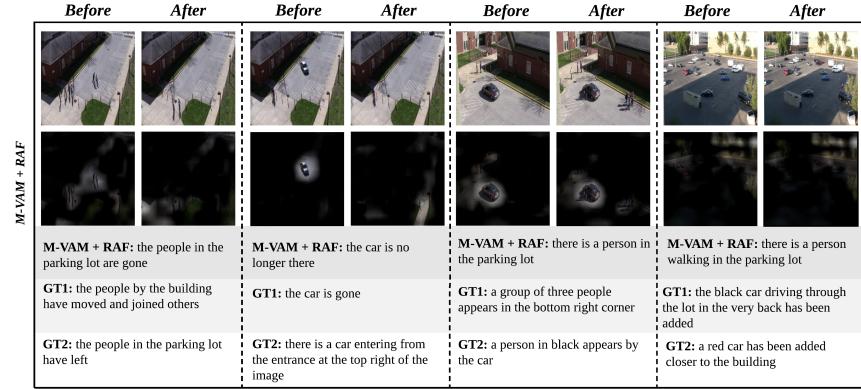


Fig. 6. Visualizations of the generated captions on Spot-the-Diff.

column, we observe that RAF can disregard small irrelevant attentions and keep concentration on the semantic change that matters (e.g., the brown cylinder in the “after” image in the second column). In the third column, the example shows that our model can detect changes in *Texture* successfully, which has a low reported score in Park *et al.* [22]. In the last column, we can see that our model shows robustness to large viewpoint change; it can predict a clear and accurate changed probability map, filtering the distraction caused by the viewpoint shift.

The small bright region in the “after” image represents a location where the blue cube is missing.

When summarizing an image with a short description like caption, humans tend to describe the surroundings and relationships with respect to the central entities. In the third column, the M-VAM picks “tiny blue cylinder” to describe the change of “big purple matte cube”, whereas the M-VAM+RAF picks “tiny red cylinder” which is closer to the central entity.

Fig. 6 shows examples from the Spot-the-Diff dataset. From the first to the third column, we can see that our model is also good at finding the difference in real-world scenes. In the first example, our model could successfully detect the people and their non-existence. In the second example, our model can detect the car and the fact that it is not in its position anymore. Although there is a mismatch at the changed probability map of the “after” image, the information from the ‘before’ image is strong enough to filter such noise and generate an accurate caption. Similarly, our model can successfully detect the changes between the images in the third example and focus on the most salient information (i.e., the appearance of a person near the car) to generate the caption somewhat correctly.

A negative sample is shown in the fourth column. We suspect this is because of the resize operation, due to which some objects become too small to be recognized. As a result, although the matching map is able to find all the changes, the semantic change information is overshadowed by the information in the surrounding, leading to a misunderstanding.

5 Conclusion

In this paper, we have introduced the novel Mirrored Viewpoint-Adapted Matching encoder for the change captioning task to distinguish semantic changes from viewpoint changes. Compared with other methods, our M-VAM encoder can accurately filter out the viewpoint influence and figure out the semantic changes from the images. To further utilize the language information to guide the difference judgment, we have further proposed a Reinforcement Attention Fine-tuning process to supervise the matching map with the language evaluation rewards. Our proposed RAF process can highlight the semantic difference in the changed map. The test results have shown that our proposed method surpasses all the current methods in both Spot-the-Diff and CLEVR-Change datasets.

Acknowledgement This research is partially supported by the MOE Tier-1 research grants: RG28/18 (S) and the Monash University FIT Start-up Grant.

References

1. Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R., Gherardi, R.: Street-view change detection with deconvolutional networks. *Autonomous Robots* **42**(7), 1301–1322 (2018)

2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Proceedings of the European Conference on Computer Vision. pp. 382–398. Springer (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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. pp. 6077–6086 (2018)
4. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1171–1179 (2015)
5. Bruzzone, L., Prieto, D.F.: Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote Sensing* **38**(3), 1171–1182 (2000)
6. Chen, C., Mu, S., Xiao, W., Ye, Z., Wu, L., Ju, Q.: Improving image captioning with conditional generative adversarial nets. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence. vol. 33, pp. 8142–8150 (2019)
7. Daudt, R.C., Le Saux, B., Boulch, A.: Fully convolutional siamese networks for change detection. In: Proceedings of the 2018 25th IEEE International Conference on Image Processing. pp. 4063–4067. IEEE (2018)
8. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on Statistical Machine Translation. pp. 376–380 (2014)
9. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: Proceedings of the European Conference on Computer Vision. pp. 15–29. Springer (2010)
10. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
11. Gu, J., Joty, S., Cai, J., Wang, G.: Unpaired image captioning by language pivoting. In: Proceedings of the European Conference on Computer Vision. pp. 503–519 (2018)
12. Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
14. Jhamtani, H., Berg-Kirkpatrick, T.: Learning to describe differences between pairs of similar images. In: arXiv preprint arXiv:1808.10584 (2018)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems. pp. 1097–1105 (2012)
17. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
18. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3153–3160. IEEE (2011)

19. Oluwasanmi, A., Aftab, M.U., Alabdulkreem, E., Kumeda, B., Baagye, E.Y., Qin, Z.: Captionnet: Automatic end-to-end siamese difference captioning model with attention. In: IEEE Access. vol. 7, pp. 106773–106783 (2019)
20. Oluwasanmi, A., Frimpong, E., Aftab, M.U., Baagye, E.Y., Qin, Z., Ullah, K.: Fully convolutional captionnet: Siamese difference captioning attention model. In: IEEE Access. vol. 7, pp. 175929–175939 (2019)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. Proceedings of Association for Computational Linguistics (2002)
22. Park, D.H., Darrell, T., Rohrbach, A.: Robust change captioning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4624–4633 (2019)
23. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015)
24. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
25. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
26. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3156–3164 (2015)
27. Wang, Q., Yuan, Z., Du, Q., Li, X.: Getnet: A general end-to-end 2-d cnn framework for hyperspectral image change detection. IEEE Transactions on Geoscience and Remote Sensing **57**(1), 3–13 (2018)
28. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning. pp. 2048–2057 (2015)
29. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2019)
30. Yang, X., Zhang, H., Cai, J.: Deconfounded image captioning: A causal retrospect. arXiv preprint arXiv:2003.03923 (2020)
31. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: Sequence generative adversarial nets with policy gradient. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (2017)