Distractors-Immune Representation Learning with Cross-modal Contrastive Regularization for Change Captioning Supplementary Material

Yunbin Tu[1], Liang Li[2*], Li Su[1*], Chenggang Yan[3], and Qingming Huang[1]

[1] University of Chinese Academy of Sciences, Beijing, China
tuyunbin22@mails.ucas.ac.cn, {suli,qmhuang}@ucas.ac.cn
[2] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China liang.li@ict.ac.cn
[3] Hangzhou Dianzi University & Lishui Institute, China

1 Experiments

1.1 Implementation Details

For a fair comparison, we follow the state-of-the-art methods [5, 6] to utilize a pre-trained ResNet-101 [1] model to obtain the features of a pair of images, with the dimension of $1024 \times 14 \times 14$. We project them into a lower dimension of 512. The hidden size of the model and word embedding size are set to 512 and 300, separately. Temperature τ in Eq. (14) of main paper is set to 0.5.

In the training phase, the batch sizes and learning rates of our method on the four datasets are shown in Table 1. We use Adam optimizer [2] to minimize the negative log-likelihood loss of Eq. (16) of main paper. In the inference phase, we use the greedy decoding strategy to generate captions. Both training and inference are implemented with PyTorch on an RTX 3090 GPU. The used training resources on the all datasets are shown in Table 2. We find that our method does not require much training time and GPU memory, so it can be easily reproduced by the researchers. To facilitate future research, the code is publicly available at https://github.com/tuyunbin/DIRL.

Table 1: The training parameters of our method on the four datasets.

	batch size	learning rate
CLEVR-Change	128	2×10^{-4}
CLEVR-DC	128	2×10^{-4}
Spot-the-Diff	64	1×10^{-4}
Image Editing Request	16	1×10^{-4}

* Corresponding authors

Table 2: The used training time and GPU memory on the four datasets.

	Training Time	GPU Memory
CLEVR-Change	150 minutes	14 GB
CLEVR-DC	90 minutes	8.4 GB
Spot-the-Diff	25 minutes	6.5 GB
Image Editing Request	10 minutes	3.2 GB

Table 3: Effect of trade-off parameter α in DIRL on the CLEVR-DC dataset.

α	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
0.001	**50.6**	32.3	65.7	**81.8**	16.1
0.002	50.4	**32.5**	**65.9**	81.2	16.1
0.003	50.5	**32.5**	65.8	**81.8**	16.2
0.004	50.4	32.3	65.7	80.5	**16.3**
0.005	50.1	31.9	65.4	79.7	16.2
0.006	50.5	31.5	64.9	79.8	15.8

1.2 Study on the Trade-off Parameter α

In this section, we discuss the effect of trade-off parameter α in Eq. (4) of main paper. As mentioned in Sec. 3.2 Distractors-Immune Representation Learning, α is a trade-off parameter to balance the importance between the two terms. The first term equates the diagonal of $\mathcal{C}$ to one, *i.e.,* the corresponding feature channels of two image representations will be correlated and thus have similar semantics under distractors. The second term equates the off diagonal of $\mathcal{C}$ to zero, *i.e.,* the different channels will be decorrelated. This enhances the discrimination of each image representation. Both terms are key to handle the influence of distractors. Since the CLEVR-DC dataset is with extreme distractors and more challenging, we conduct experiment on this dataset. In Table 3, we find that the captioning results are close when setting the value of α from 0.001 to 0.006, and the model's overall performance is better under the value of 0.003. This shows that the proposed DIRL is robust. Empirically, we set the value of α as 0.003 on the four datasets.

Table 4: Ablation of DIRL with/without the MLP on the CLEVR-DC dataset.

Ablation	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
DIRL w/ MLP	**51.4**	**32.3**	**66.3**	**84.1**	**16.8**
DIRL w/o MLP	48.0	32.0	65.2	78.5	15.7

1.3 Study on the MLP Function in DIRL

In Eq.(1) of the main paper, a convolution function transforms the features of two images into a low dimension, and then the trainable position encodings are integrated into the transformed features along their height and width. Further, the MLP in Eq.(2) projects the two position-embedded features into a shared embedding space. To study the effect of the MLP, we carry out the ablation study of DIRL with/without MLP on the CLEVR-DC dataset, which is show in Table 4. It is noted that DIRL with MLP is much better than it without MLP, which shows that adding the MLP helps DIRL learn a pair of distractors-immune representations in terms of semantics and position.

1.4 Comparison between DIRL and Static Methods

The proposed DIRL aims to learn a pair of distractors-immune representations by correlating the corresponding feature channels and decorrelating different ones. In fact, there are some simpler static methods such as PCA or ZCA whitening to remove the degree of correlation. Here, we conduct the experiment to show performance comparison of cross-channel decorrelation among the transformer-based model with PCA/ZCA whitening and our DIRL.

Table 5: Performance comparison among Transformer-based model with PCA/ZCA whitening and our DIRL.

Ablative Variants	BLEU-4	ROUGE-L	CIDEr	SPICE
Transformer	48.9	65.6	79.6	15.7
Transformer w/ PCA whitening	40.6	57.9	22.0	10.0
Transformer w/ ZCA whitening	38.6	57.3	33.3	13.1
Transformer w/ DIRL	**50.5**	**65.8**	**81.8**	**16.2**

The two static methods are used as data preprocessing strategies before model training, while the proposed DIRL is jointly trained with the model. The comparison results are shown in Table 5. We can find that the model with DIRL outperforms the others by a large margin, indicating that joint training strategy (DIRL) does help the model learn two stable image features under distractors.

1.5 Qualitative Analysis

In this supplementary material, we will show more qualitative examples on the CLEVR-Change, CLEVR-DC, Spot-the-Diff, and Image Editing Request datasets, which are shown in Fig. 1-5. In Fig. 1-2, on the four datasets, we visualize shared objects matching between two images, which are yielded by the classic match-based method MCCFormers-D [4] and the proposed DIRL+CCR. We can find that MCCFormers-D is unable to sufficiently or correctly align the shared

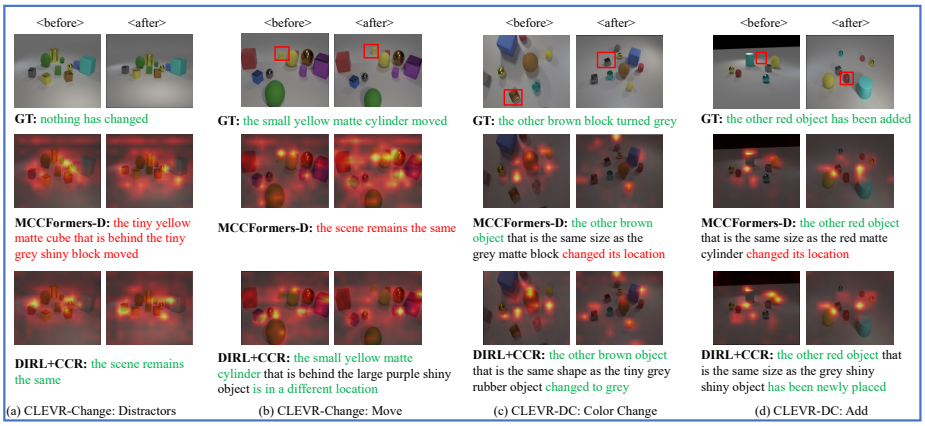

Fig. 1: Visualization of shared objects matching between two images on the CLEVR-Change and CLEVR-DC with distractors to varying degrees. For each example, we visualize matching results by the classic match-based method MCCFormers-D [4] and the proposed DIRL+CCR. The changed objects are shown in the red boxes. Besides, we visualize the ground-truth caption (GT), and the generated captions by MCCFormers-D and our DIRL+CCR. The correct words are in green color, while incorrect words are in red color.

objects between two images. By contrast, the proposed DIRL+CCR can better match the shared objects. These examples validate that the proposed DIRL can make the representations of image pair non-perturbational and discriminative under the distractors to varying degrees. Based on the two distractors-immune image representations, the model can better interact and mine their shared features.

To intuitively validate the change localization and caption capabilities of our method, we visualize the generated captions by DIRL+CCR under different change types on the four datasets. Meanwhile, we visualize the change localization results that are obtained from the cross-attention maps between the difference features and generated words in the decoding process. These are shown in Fig. 3-5. When the attention score is higher, the region is brighter. We observe that the proposed DIRL+CCR can accurately localize and describe the actually changed objects under different scenarios. This superiority mainly benefits from the facts that 1) DIRL is able to learn two distractors-immune image representations for matching the shared objects, so as to learn the reliable difference features for caption generation; 2) CCR is capable of regularizing the cross-modal alignment by maximizing the contrastive alignment between the generated words and attended difference features, so as to improve the quality of yielded captions.

Fig. 2: Visualization of shared objects matching between two images on the Image Editing Request (IER) and Spot-the-Diff datasets. For each example, we visualize matching results by the classic match-based method MCCFormers-D [4] and the proposed DIRL+CCR. The changed objects are shown in the red boxes. Besides, we visualize the ground-truth caption (GT), and the generated captions by MCCFormers-D and our DIRL+CCR. The correct words are in green color, while incorrect words are in red color.

Limitation

Fig. 6 shows a failure case that derives from our DIRL+CCR. This image pair is from the surveillance cameras and with underlying distractors (illumination change), where there are three people newly appearing in the "after" image. DIRL+CCR can accurately localize the region containing the added people and describe the change type, which benefit from the proposed DIRL and CCR. However, we note that the generated sentence wrongly describes the number of added people. Our conjecture is that under surveillance cameras, the changed objects are commonly small. Further, the distance of the left two people is very close, and the color of one man's coat is similar to the other's pants. In this situation, the model risks recognizing two people as one person, so as to generate the incorrect result. In our opinion, a possible solution is to leverage finer-level visual modality for the representation of such small objects, *e.g.,* image segmentation features [3, 7].

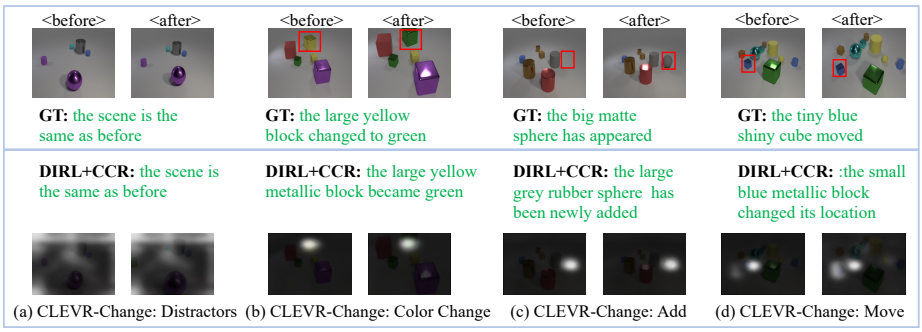

Fig. 3: Four cases from the CLEVR-Change dataset. We visualize the ground-truth caption (GT), and the captions yielded by our DIRL+CCR. We also visualize its change localization results. The ground-truth changed objects are shown in red boxes.

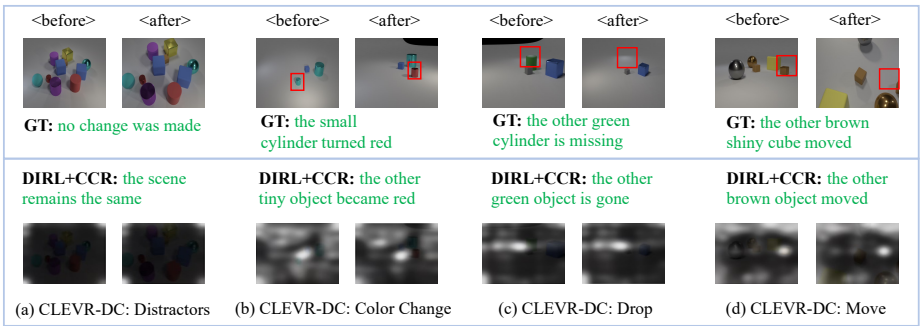

Fig. 4: Four cases from the CLEVR-DC dataset. We visualize the ground-truth caption (GT), and the captions yielded by our DIRL+CCR. We also visualize its change localization results. The ground-truth changed objects are shown in red boxes.

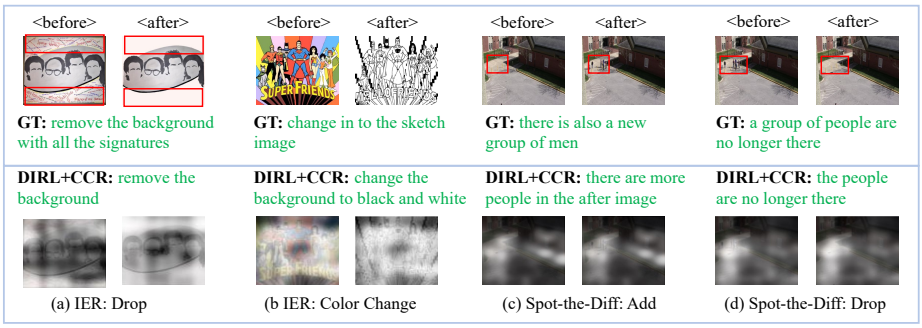

Fig. 5: Four cases from the Image Editing Request (IER) and Spot-the-Diff datasets. We visualize the ground-truth caption (GT), and the captions yielded by our DIRL+CCR. We also visualize its change localization results. The ground-truth changed objects are shown in red boxes.

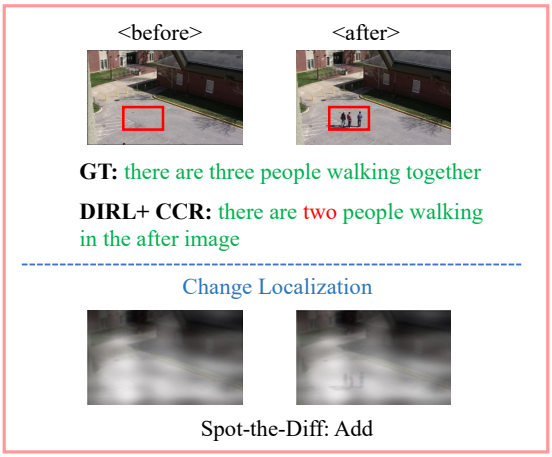

Fig. 6: The failure case that derives from the proposed DIRL+CCR.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 1
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 1
3. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023) 5
4. Qiu, Y., Yamamoto, S., Nakashima, K., Suzuki, R., Iwata, K., Kataoka, H., Satoh, Y.: Describing and localizing multiple changes with transformers. In: ICCV. pp. 1971–1980 (2021) 3, 4, 5
5. Tu, Y., Li, L., Su, L., Zha, Z.J., Huang, Q.: Smart: Syntax-calibrated multi-aspect relation transformer for change captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence p. Early Access (2024) 1
6. Tu, Y., Li, L., Su, L., Zha, Z.J., Yan, C., Huang, Q.: Self-supervised cross-view representation reconstruction for change captioning. In: ICCV. pp. 2805–2815 (2023) 1
7. Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. NeurIPS **33**, 655–666 (2020) 5