R^3Net: Relation-embedded Representation Reconstruction Network for Change Captioning

Yunbin Tu[1], Liang Li[2], Chenggang Yan[3], Shengxiang Gao[1] and Zhengtao Yu[1]

[1]Kunimg University of Science and Technology, [2]Institute of Computing Technology, CAS, [3]Hangzhou Dianzi University

Goal and Application

- **Goal:** Describing the change between two similar images.
- **Practical Applications:**
  - Medical imaging: Comparing CT images, locating the lesion, and generating the report of the patient's physical abnormalities;
  - Facility monitoring: Generating the report about whether there is a change of the monitored facility;
  - Aerial photography: Monitoring and describing land dynamics.

Challenge

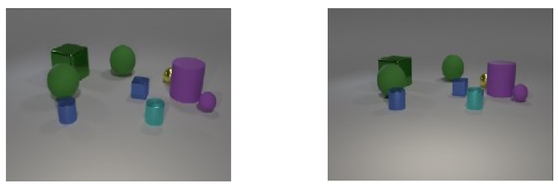
Fine-grained difference

<Before> <After>

- **Ground truth:** A person on sidewalk is now gone.
- **Baseline:** There is no difference.

Distraction of viewpoint change

<Before> <After>



- **Ground truth:** The large green matte sphere that is behind the purple cylinder is in a different location.
- **Baseline:** The scene is the same as before.

Motivation

Previous work（ICCV'19, ECCV'20）

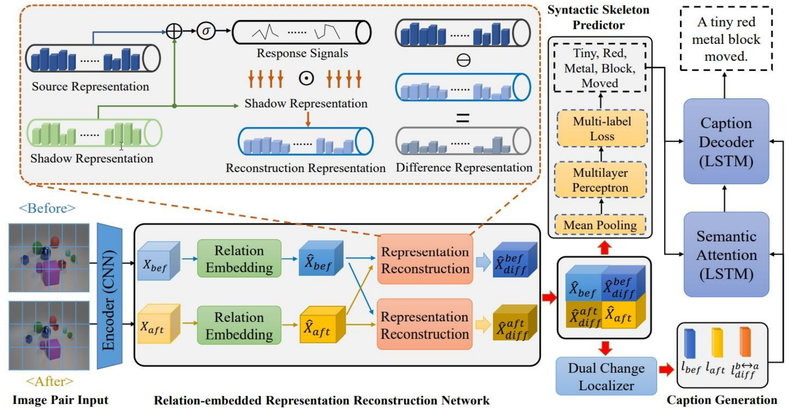
- Modeling the difference representation only at feature level, which is difficult to discriminate fine-grained change;
- Applying simple subtraction between two unaligned images, which computes the difference representation with much noise;
- Conducting change localization and caption generation separately.

Our idea

- **Embedding semantic relations among object features to help explore the fine-grained change;**
- **Modeling the difference representation based on the semantic similarities in the corresponding locations of two images;**
- **Leveraging syntactic skeletons to enhance the interaction between change localization and caption generation.**

Approach

Overall framework



1 Relation-embedded Module

1) Learning semantic relations among object features via self-attention;

2) Modeling the difference representation at both feature and relation levels.

2 Representation Reconstruction Module

1) A "shadow" representation ("after" or "before") is used to reconstruct a "source" representation ("before" or "after");

2) The "difference" representation is computed with the changed feature between "source" and "reconstruction" representation.

3 Syntactic skeletons Predictor

Enhancing the semantic interaction between change localization and caption generation.

Results

CLEVR-change dataset (Total performance on change and none-scene change)

Method	RL	Total				
		BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Capt-Dual (ICCV'19)	×	43.5	32.7	-	108.5	23.4
DUDA (ICCV'19)	×	47.3	33.9	-	112.3	24.5
M-VAM (ECCV'20)	×	50.3	37.0	69.7	114.9	30.5
M-VAM+RAF (ECCV'20)	√	51.3	37.8	70.4	115.8	30.7
R^3Net+SSP (Ours, EMNLP'21)	×	**54.7**	**39.8**	**73.1**	**123.0**	**32.6**

*RL is short for reinforcement learning

CLEVR-change dataset (The performance of scene change)

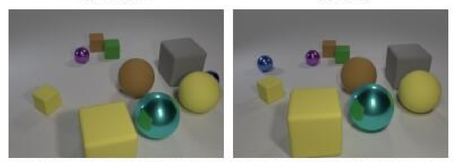
Method	RL	BLEU-4	METEOR	CIDEr	SPICE
Capt-Dual (ICCV'19)	×	38.4	28.5	89.8	18.2
DUDA (ICCV'19)	×	42.9	29.7	94.6	19.9
M-VAM+RAF (ECCV'20)	√	-	-	-	-
R^3Net+SSP (Ours, EMNLP'21)	×	**52.7**	**36.2**	**116.6**	**30.3**

CLEVR-change dataset (The performance of none-scene change)

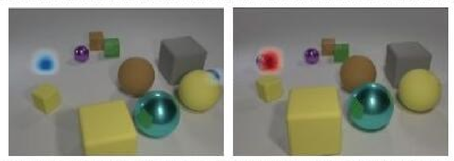
Method	RL	BLEU-4	METEOR	CIDEr	SPICE
Capt-Dual (ICCV'19)	×	56.3	44.0	108.9	28.7
DUDA (ICCV'19)	×	59.8	45.2	110.8	29.1
M-VAM+RAF (ECCV'20)	√	-	**66.4**	**122.6**	33.4
R^3Net+SSP (Ours, EMNLP'21)	×	**61.9**	50.5	116.4	**34.8**

Qualitative results

<Before> <After>



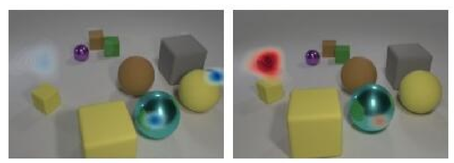
Ground Truth: The blue metal ball is in a different location.



DUDA: The small blue metal ball that is behind the tiny yellow rubber thing has been newly placed.

Predicted Syntactic Skeletons

small (0.782), sphere (0.767), ball (0.756), metal (0.692), object (0.663), rubber (0.648), changed (0.605), yellow (0.516), location (0.425), blue (0.383), different (0.342), moved (0.333), newly (0.100), placed (0.099)



R^3Net + SSP The small blue metal sphere that is behind the small yellow rubber object is in a different location.