Semantic Relation-aware Difference Representation Learning for Change Captioning

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

[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

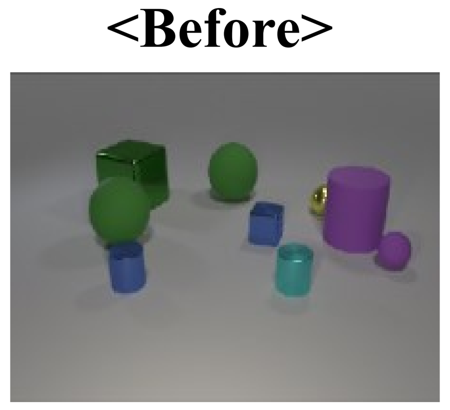
Fine-grained difference

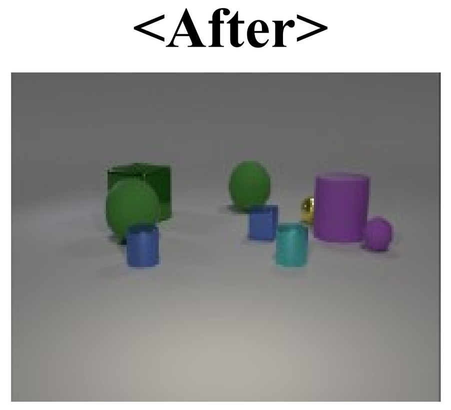
<Before> <After>

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

Distraction of viewpoint/illumination 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）

➢Capturing the semantic change only at feature level;

➢Misidentifying the distractor change as the real change;

➢Using visual information to generate each word;

Our idea

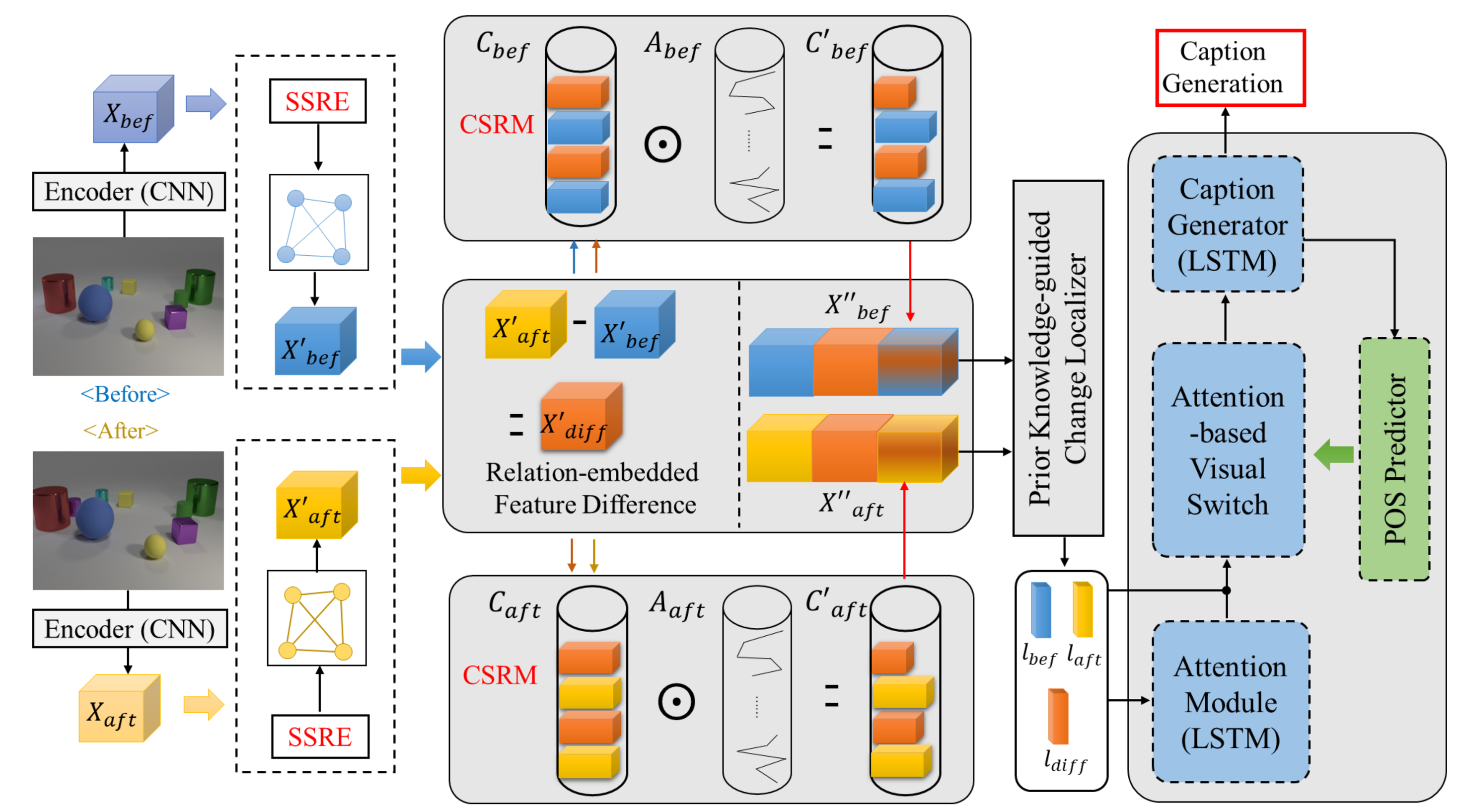
➢Capturing the semantic change at feature and relation levels.

➢Measuring semantic relation of candidate difference with respect to each image in the image pair.

➢Using visual information dynamically based on Part-of-Speech (POS) of words.

Approach

Overall framework



1 Self Semantic Relation Embedding block (SSRE)

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

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

2 Cross Semantic Relation Measuring block (CSRM)

1) Measuring relevance between each image and candidate difference;

2) Distinguishing the real change from irrelevant distractors.

3 Attention-based Visual Switch (AVS)

Exploiting visual information dynamically based on the POS of each word.

Results

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

		Total				
Method	RL	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
SRDRL+AVS (Ours, ACL'21)	×	**54.9**	**40.2**	**73.3**	**122.2**	**32.9**

*RL is short for reinforcement learning

CLEVR-change dataset (The performance of Semantic change)

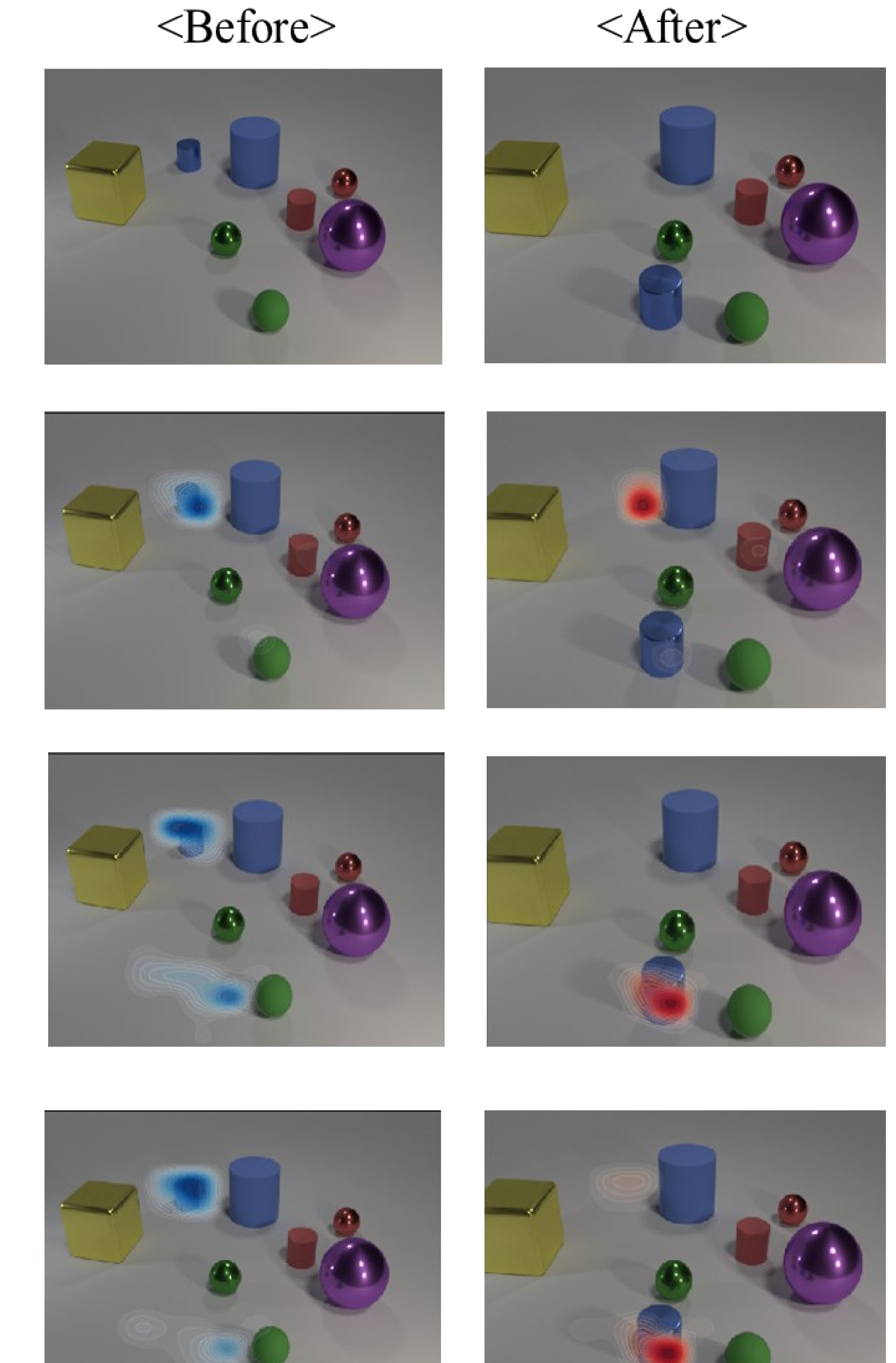
Method	RL	BLEU-4	METEOR	CIDEr	SPICE
Capt-Dual (ICCV'19)	×	38.4	28.5	89.8	18.2
DUDA (ICCV'19)	×	42.9	29.7	94.6	19.9
M-VAM+RAF (ECCV'20)	√	-	-	-	-
SRDRL+AVS (Ours, ACL'21)	×	**52.7**	**36.4**	**114.2**	**30.8**

CLEVR-change dataset (The performance of None-semantic 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
SRDRL+AVS (Ours, ACL'21)	×	**62.2**	51.3	117.0	**34.9**

Qualitative results

<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.