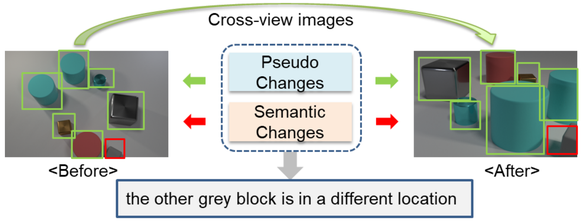
Self-supervised Cross-view Representation Reconstruction for Change Captioning

Yunbin Tu[1], Liang Li[2], Li Su[1], Zheng-Jun Zha[3], Chenggang Yan[4] and Qingming Huang[1]

[1]University of Chinese Academy of Sciences, [2]Institute of Computing Technology, CAS,
[3]University of Science and Technology of China, [4]Hangzhou Dianzi University

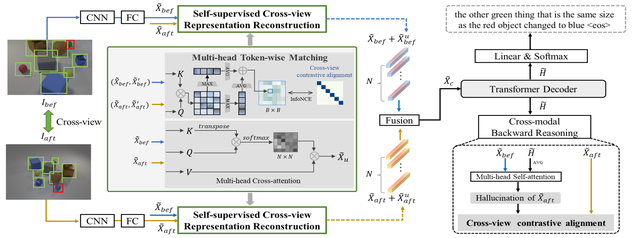
ICCV23 PARIS

1. Motivation



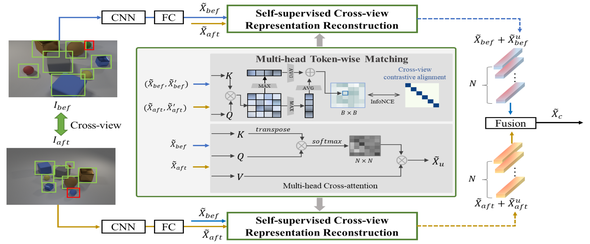
- Reconstructing unchanged representations to capture a stable difference representation for cross-view images
  - interact cross-view features: multi-head token-wise matching
  - learn view-invariant representations: cross-view contrastive alignment
- Rebuilding the "after'' image with the full representations of caption and "before" image to improve captioning quality
  - model "hallucination" via the caption and "before"
  - match "hallucination" with "after": cross-view contrastive alignment

2. Approach Overview



- Stage 1: cross-view feature extraction
- Stage 2: self-supervised cross-view representation reconstruction (SCORER)
- Stage 3: difference representation modeling and caption generation
- Stage 4: cross-modal backward reasoning (CBR)

3. SCORER



- For a pair of images, multi-head token-wise matching (MTM) is

$$\mathrm{MTM}(Q,K) = \underset{i'=1\ldots h}{\mathrm{Concat}}\left(\mathrm{head}_{i'}\right), \qquad \mathrm{head}_{i'} = \mathrm{TM}\left(QW_{i'}^{Q}, KW_{i'}^{K}\right).$$

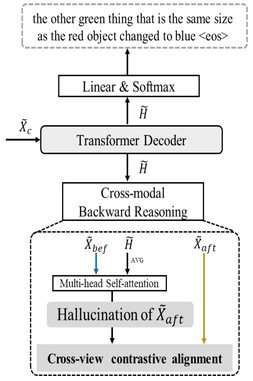
where

$$\mathrm{TM}(Q,K) = \left[\frac{1}{N}\sum_{i=1}^{N}\max_{j=1}^{N}(e_{i,j}) + \frac{1}{N}\sum_{j=1}^{N}\max_{i=1}^{N}(e_{i,j})\right]/2, \qquad e_{i,j} = (q_i)^{\top} k_j.$$

- Maximizing cross-view contrastive alignment

$$\mathcal{L}_{b2a} = -\frac{1}{B}\sum_{k}^{B}\log\frac{\exp\left(\mathrm{MTM}\left(\tilde{X}_k^b, \tilde{X}_k^a\right)/\tau\right)}{\sum_{r}^{B}\exp\left(\mathrm{MTM}\left(\tilde{X}_k^b, \tilde{X}_r^a\right)/\tau\right)} \qquad \mathcal{L}_{a2b} = -\frac{1}{B}\sum_{k}^{B}\log\frac{\exp\left(\mathrm{MTM}\left(\tilde{X}_k^a, \tilde{X}_k^b\right)/\tau\right)}{\sum_{r}^{B}\exp\left(\mathrm{MTM}\left(\tilde{X}_k^a, \tilde{X}_r^b\right)/\tau\right)}, \qquad \mathcal{L}_{\mathrm{cv}} = \frac{1}{2}(\mathcal{L}_{b2a} + \mathcal{L}_{a2b}),$$

4. CBR



- Rebuilding a "hallucination" representation with the sentence and "before" image

$$\hat{X}_{hal} = \mathrm{conv}_2([\tilde{X}_{bef}; \tilde{T}]), \hat{X}_{hal} \in \mathbb{R}^{D\times H\times W}.$$

$$\tilde{X}_{hal} = \mathrm{conv}_2[\mathrm{MHSA}\ (\hat{X}_{hal}, \hat{X}_{hal}, \hat{X}_{hal})],$$

- Using MTM to match the representations of "hallucination" and "after" image
- Maximizing cross-view contrastive alignment

$$\mathcal{L}_{\mathrm{cm}} = \frac{1}{2}(\mathcal{L}_{h2a} + \mathcal{L}_{a2h}).$$

5. Experimental Results

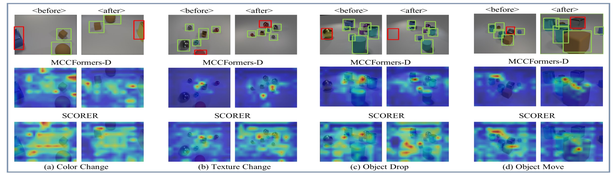
- Comparison with existing methods

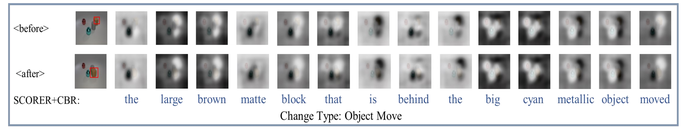
Method	Total B	M	R	C	S	Semantic Change B	M	R	C	S
PCL w/ Pre-training (AAAI 2022) [38]	51.2	36.2	71.7	**128.9**	-	-	-	-	-	-
M-VAM+RAF (ECCV 2020) [25]	51.3	37.8	70.4	115.8	30.7	-	-	-	-	-
DUDA (ICCV 2019) [22]	47.3	33.9	-	112.3	24.5	42.9	29.7	-	94.6	19.9
DUDA+ (CVPR 2021) [7]	51.2	37.7	70.5	115.4	31.1	49.9	34.3	65.4	101.3	27.9
R^{3}Net+SSP (EMNLP 2021) [31]	54.7	39.8	73.1	123.0	32.6	52.7	36.2	69.8	116.6	30.3
VACC (ICCV 2021) [11]	52.4	37.5	-	114.2	31.0	-	-	-	-	-
SGCC (ACM MM 2021) [15]	51.1	40.6	73.9	121.8	32.2	-	-	-	-	-
SRDRL+AVS (ACL 2021) [32]	54.9	40.2	73.3	122.2	32.9	52.7	36.4	69.7	114.2	30.8
MCCFormers-D (ICCV 2021) [24]	52.4	38.3	-	121.6	26.8	-	-	-	-	-
IFDC (TMM 2022) [9]	49.2	32.5	69.1	118.7	-	47.2	29.3	63.7	105.4	-
NCT (TMM 2023) [30]	55.1	40.2	73.8	124.1	32.9	53.1	36.5	70.7	118.4	30.9
VARD-Trans (TIP 2023) [28]	55.4	40.1	73.8	126.4	32.6	-	-	-	-	-
SCORER (Ours)	55.8	40.8	74.0	126.0	33.0	54.1	37.4	71.5	122.0	31.2
SCORER+CBR (Ours)	**56.3**	**41.2**	**74.5**	**126.8**	**33.3**	**54.4**	**37.6**	**71.7**	**122.4**	**31.6**

- Ablation study

Method	Semantic Change B	M	R	C	S	Only Pseudo Change B	M	R	C	S
Subtraction	50.2	34.1	67.1	108.0	28	57.3	48.4	74.7	113.8	34.0
RR	53.3	37.1	70.8	119.1	30.4	61.1	50.7	76.4	114.9	34.6
SCORER	54.3	37.5	71.5	122.0	31.2	61.4	50.6	76.5	116.4	34.7
RR+CBR	54.1	37.4	71.5	**122.4**	31.2	60.7	51.2	76.9	114.9	34.6
SCORER+CBR	**54.4**	**37.6**	**71.7**	**122.4**	**31.6**	**62.0**	**51.7**	**77.4**	**117.9**	**35.0**

- Visualization for shared object matching and caption generation





- More experimental results shown in our paper
- Code available at: https://github.com/tuyunbin/SCORER