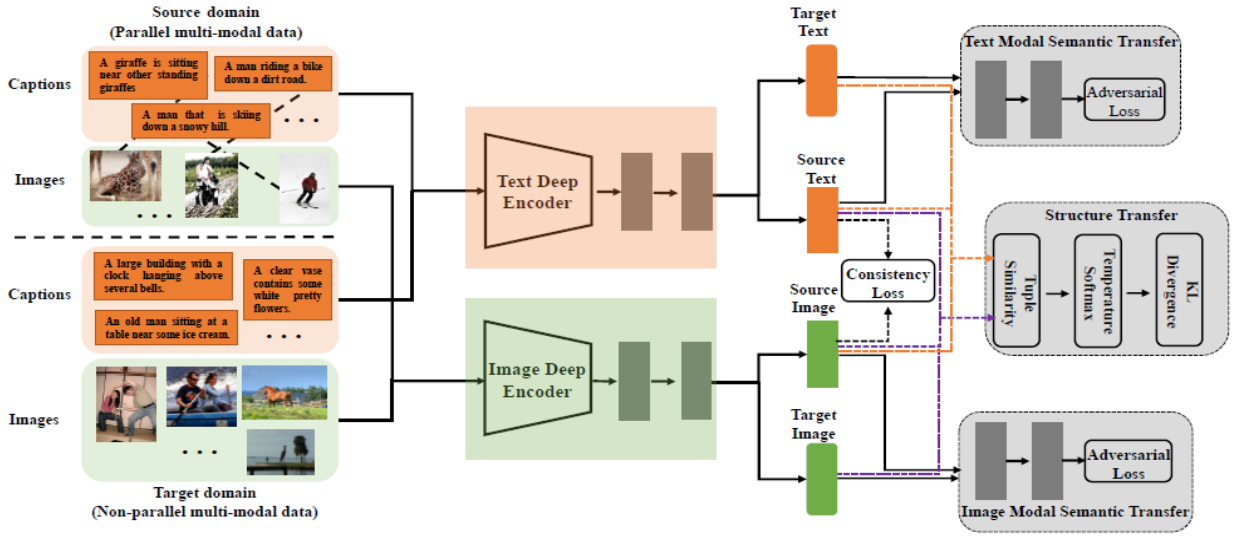
Alignment Efficient Image-Sentence Retrieval Considering Transferable Cross-Modal Representation Learning

Yang YANG, Jinyi GUO, Guangyu LI, Lanyu LI, Wenjie LI, Jian YANG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3186-6

Problems & Ideas

- Problems of cross modal retrieval:
  - In many real-world applications, a large amount of parallel data for new scenarios is difficult to obtain, leading the non-parallel multi-modal data and existing methods cannot be used directly.
  - Traditional transfer learning methods cannot handle cross modal retrieval transfer learning properly.
- Ideas: Develop semantic transfer for modal representation learning and structure transfer for modal consistency learning.



In detail, AEIR learns the consistent representations for target domain from three aspects: 1) Consistent representation learning of source domain, which utilizes the matching loss for supervised learning in source domain; 2) Semantic transfer, which aligns source and target representations for each modality with a domain adversarial loss; and 3) Structure transfer, which further generalizes the consistent representations to target domain using a cross-domain cross-modal metric based constraint.

Main Contributions

- Contributions:
  - 1) Use the domain discriminators to learn the common representations of source and target domains for each modality.
  - 2) Use the cross-domain consistency of relational distribution

    to constrain the learning of cross-modal consistent representations in target domain.

Methods	FLICKR30K-to-MSCOCO(1K)						FLICKR30K-to-MSCOCO(5K)						MSCOCO-to-FLICKR30K					
	Image2Text			Text2Image			Image2Text			Text2Image			Image2Text			Text2Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CCA	13.1	34.4	47.5	9.3	29.8	43.1	3.6	12.8	19.9	2.8	10.4	17.3	6.6	18.4	25.6	5.6	16.1	22.5
UVCL	37.9	**71.5**	**82.7**	27.0	58.9	71.4	16.7	39.3	52.0	11.3	29.2	40.4	18.6	41.0	51.9	11.9	29.2	38.0
DCCA	14.0	35.5	48.3	9.4	30.1	44.5	3.7	12.9	20.1	2.8	11.0	18.4	6.8	19.4	29.1	6.5	20.9	30.9
VSE0	15.5	35.4	47.1	10.6	30.1	42.2	5.2	15.2	22.5	3.5	11.5	18.2	20.6	39.8	54.1	14.9	36.1	47.1
UGACH	13.9	32.0	43.5	9.1	27.1	39.1	4.4	13.2	20.3	3.1	10.4	16.3	14.8	35.1	45.5	11.0	29.5	39.7
VSEPP	17.1	38.0	50.7	12.5	33.9	47.6	7.2	19.7	27.5	4.9	14.6	22.2	24.4	48.9	60.1	16.5	38.2	48.9
SCAN	30.5	58.6	70.8	23.7	52.0	66.2	16.5	35.0	45.7	10.6	26.4	36.2	48.4	75.2	83.9	35.4	62.2	72.3
IMRAM	35.3	62.0	74.4	26.2	53.9	66.9	18.5	40.4	51.9	12.8	29.9	40.0	53.1	80.2	87.9	41.1	66.3	75.2
SGRAF	38.6	66.8	76.7	29.6	56.6	68.2	20.2	42.5	54.4	15.1	33.5	44.3	57.7	82.8	89.4	41.9	68.2	77.4
RACG	39.3	67.2	76.6	29.8	57.0	68.1	20.8	42.7	54.1	15.5	33.5	44.2	58.1	82.5	89.2	41.1	68.5	77.0
A3VSE	27.2	49.5	58.1	15.1	41.5	49.3	12.3	31.2	40.2	3.1	19.3	26.4	34.1	55.4	67.1	28.4	43.4	56.9
DMTL	13.3	34.7	44.2	8.8	27.0	39.5	4.8	14.8	20.9	2.6	9.1	14.9	N/A	N/A	N/A	N/A	N/A	N/A
CAPQ	12.8	30.6	41.5	9.5	27.2	37.9	2.1	3.2	4.3	1.0	1.4	3.2	21.2	46.7	59.0	15.8	37.9	49.4
MME	40.2	66.7	77.7	31.0	59.3	71.0	21.1	44.2	55.6	17.1	36.5	47.1	59.7	84.2	90.7	44.4	70.8	79.4
DMTL-A	41.2	69.4	76.8	29.9	57.6	69.0	21.9	45.9	57.0	14.4	34.4	45.5	43.8	70.8	79.5	31.7	58.8	69.3
CDCMR	35.5	61.5	75.4	25.9	52.5	66.0	11.6	27.4	37.0	7.8	20.3	28.5	48.2	73.7	82.6	37.8	63.1	72.9
AEIR	**43.1**	69.7	79.6	**33.0**	**61.9**	**72.5**	**25.1**	**47.8**	**58.5**	**18.3**	**37.8**	**48.8**	**61.8**	**86.4**	**91.7**	**45.2**	**71.2**	**79.7**