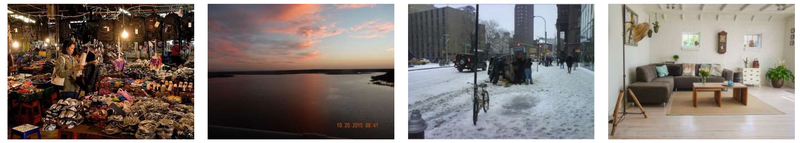
BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency

Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, Min Xu

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

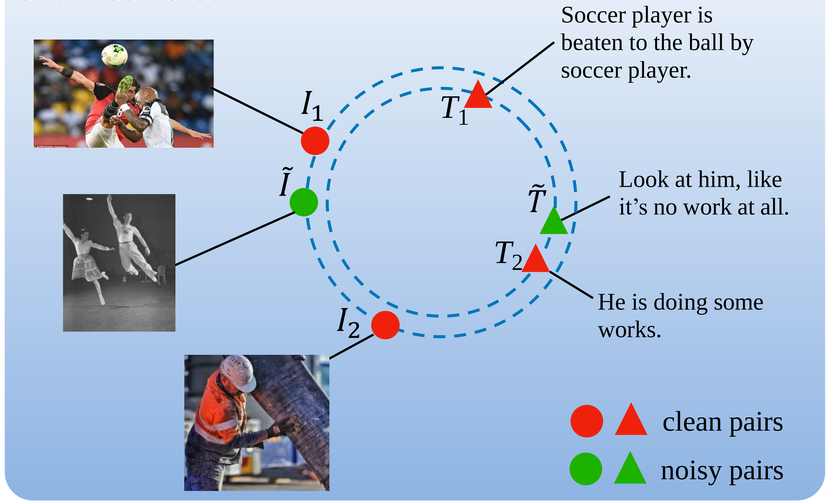
Ⅰ. Problems

- Multimodal datasets (e.g., image-text pairs) collected from the Internet contains many mismatched data pairs.
- The key challenge in this problem is how to estimate accurate soft correspondence labels for those noisy data pairs.

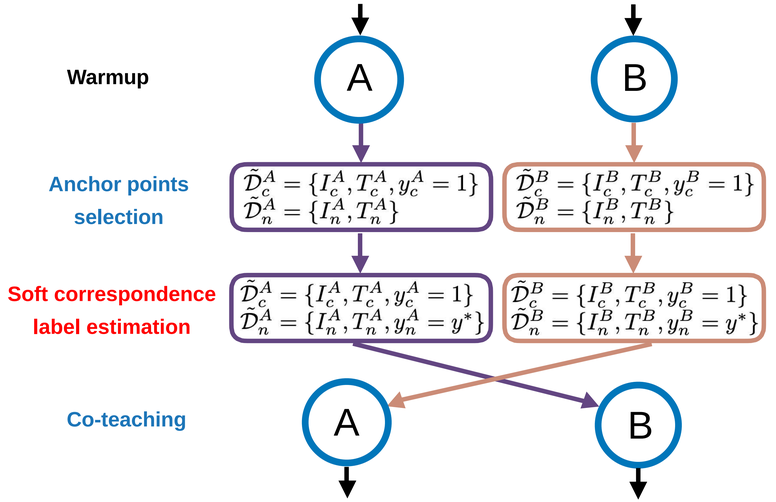


Ⅱ. Motivation

Similar images should have similar textual descriptions and vice versa



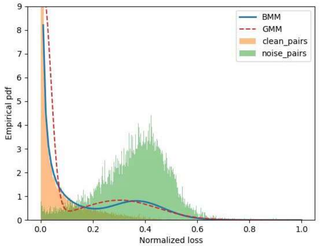
Ⅲ. Method



Soft correspondence label:

$$y_i^* = \left(\frac{D(I_n^i, I_c^{\triangle})}{D(T_n^i, T_c^{\triangle})} + \frac{D(T_n^i, T_c^{\Diamond})}{D(I_n^i, I_c^{\Diamond})}\right)/2$$

Anchor points :



Co-teaching :

$$\hat{\alpha}_i = \frac{m^{y_i^*} - 1}{m - 1}\alpha$$

$$L_{soft}(I_i, T_i) = \left[\hat{\alpha}_i - S(I_i, T_i) + S\left(I_i, \hat{T}_h\right)\right]_+ + \left[\hat{\alpha}_i - S(I_i, T_i) + S\left(\hat{I}_h, T_i\right)\right]_+$$

Ⅳ. Results

a helicopter hovers above structure. (0.85)

1. military fighter aircraft at the international exhibition
2. a helicopter hovers above structure .
3. aerial view taken from the window a descending airplane arriving
4. trucks were out in force in the area
5. airplanes moving in the same direction

Query: Cooling off the cars and riders

Table 1. Image-Text Retrieval on Flickr30K and MS-COCO 1K.

		Flickr30K							MS-COCO						
		Image⟶Text			Text⟶Image				Image⟶Text			Text⟶Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
20%	SCAN	58.5	81.0	90.8	35.5	65.0	75.2	406.0	62.2	90.0	96.1	46.2	80.8	89.2	464.5
	VSRN	33.4	59.5	71.3	25.0	47.6	58.6	295.4	61.8	87.3	92.9	50.0	80.3	88.3	460.6
	IMRAM	22.7	54.0	67.8	16.6	41.8	54.1	257.0	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	SAF	62.8	88.7	93.9	49.7	73.6	78.0	446.7	71.5	94.0	97.5	57.8	86.4	91.9	499.1
	SGR	55.9	81.5	88.9	40.2	66.8	75.3	408.6	25.7	58.8	75.1	23.5	58.9	75.1	317.1
	NCR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	DECL	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	BiCro	**78.3**	94.1	97.3	60.0	83.7	89.5	502.9	78.2	95.9	98.4	62.5	89.8	95.5	520.3
	BiCro*	78.1	**94.4**	**97.5**	**60.4**	**84.4**	**89.9**	**504.7**	**78.8**	**96.1**	**98.6**	**63.7**	**90.3**	**95.7**	**523.2**
40%	SCAN	26.0	57.4	71.8	17.8	40.5	51.4	264.9	42.9	74.6	85.1	24.2	52.6	63.8	343.2
	VSRN	2.6	10.3	14.8	3.0	9.3	15.0	55.0	29.8	62.1	76.6	17.1	46.1	60.3	292.0
	IMRAM	5.3	25.4	37.6	5.0	13.5	19.6	106.4	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	SAF	7.4	19.6	26.7	4.4	12.2	17.0	87.3	13.5	43.8	48.2	16.0	39.0	50.8	211.3
	SGR	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	NCR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	DECL	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	98.3	59.5	88.3	94.8	512.0
	BiCro	73.6	**93.0**	**96.4**	**56.0**	80.8	87.4	487.2	76.4	95.2	**98.6**	61.5	**89.4**	**95.5**	516.6
	BiCro*	**74.6**	92.7	96.2	55.5	**81.1**	**87.4**	**487.5**	**77.0**	**95.9**	98.3	**61.8**	89.2	94.9	**517.1**

Ⅴ. Conclusion

- We propose a general framework called bidirectional crossmodal similarity consistency (BiCro) for soft correspondence label estimation given only noisily-collected data
- The effectiveness of the proposed framework was verified on both synthetic noisy datasets and real noisy dataset.

Ⅵ. Contact

Email：20b903054@stu.hit.edu.cn