Conditional Representation Learning for Customized Tasks

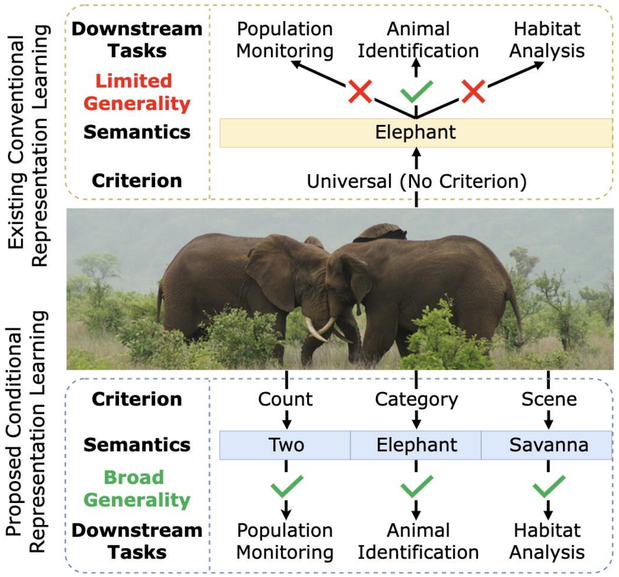
Honglin Liu[1], Chao Sun[2], Peng Hu[1], Yunfan Li[1,*], Xi Peng[1,*]

[1]Sichuan University, [2]Aerospace Information Research Institute, Chinese Academy of Sciences

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Existing representation learning learns a universal representation that prioritizes the dominant semantics while overlooking other meaningful features.



In contrast, our proposed conditional representation learning (CRL) extracts representations conditioned on specific criteria, significantly enhancing its generality.

ContactInfo

TristanLiuHL@gmail.com

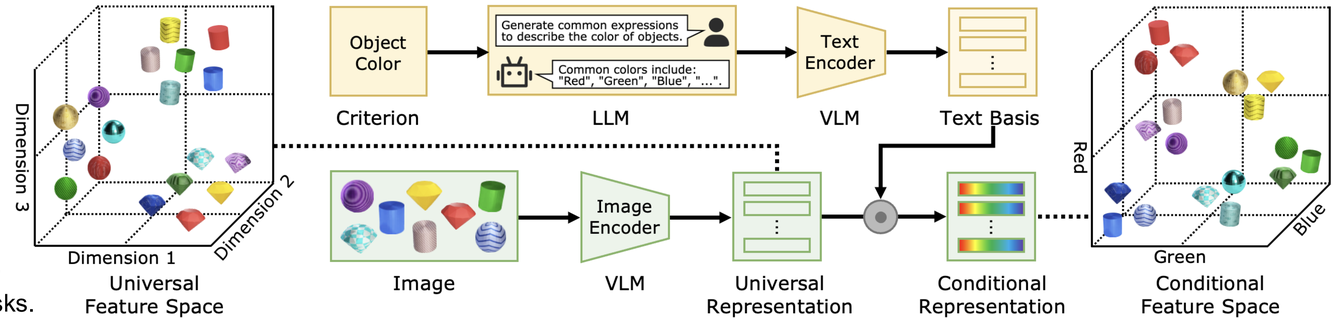
WeChat | Paper | Code

Method

1. Given a criterion, let LLM output related descriptive texts.

2. Feed images and texts into the VLM to get embeddings I and T.

3. Obtain conditional representations $\mathrm{R} = IT^\top$ and apply to downstream tasks.



Experiment

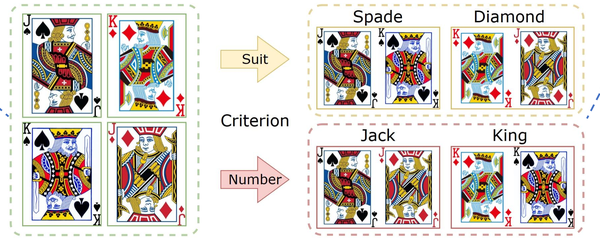
❖ Few-shot Classification

Method	Clevr4-10k									Mean
	Texture			Shape			Color			
	1	5	10	1	5	10	1	5	10	
CLIP [34]	17.46	29.39	36.26	58.16	83.17	89.47	26.85	57.33	70.00	52.01
ALIGN [19]	18.80	34.35	45.22	**73.40**	91.82	95.02	20.08	41.89	56.45	53.00
MetaCLIP [48]	17.68	30.96	39.03	70.13	91.69	95.47	22.37	46.71	61.74	52.86
BLIP2 [23]	15.93	25.23	32.58	72.91	**95.18**	97.88	28.96	60.53	73.25	55.83
CLIP+CRL	18.76	35.54	45.54	58.67	86.61	92.29	**65.28**	**88.89**	**93.08**	64.96
ALIGN+CRL	**20.91**	**41.77**	**54.92**	63.05	92.74	96.25	60.26	87.38	92.56	**67.76**
MetaCLIP+CRL	18.14	34.89	44.69	66.36	92.01	95.50	62.41	88.45	92.50	66.11
BLIP2+CRL	16.35	34.67	47.28	73.22	95.12	**97.90**	63.75	86.16	92.13	67.40

Method	Clevr4-10k			Cards						Mean
	Count			Number			Suits			
	1	5	10	1	5	10	1	5	10	
CLIP [34]	17.50	23.43	25.45	20.63	33.73	41.84	37.65	56.36	65.98	35.84
ALIGN [19]	14.64	21.63	25.16	16.97	24.70	29.15	34.67	52.75	61.78	31.27
MetaCLIP [48]	16.61	22.64	24.92	37.47	55.03	65.16	20.71	35.16	42.97	35.63
BLIP2 [23]	16.92	25.63	29.38	27.21	45.54	55.94	44.61	70.14	78.16	43.73
CLIP+CRL	**23.38**	29.59	32.40	17.66	44.52	51.09	37.10	67.16	72.64	41.73
ALIGN+CRL	18.16	32.62	36.80	17.39	30.61	35.93	42.13	76.36	80.11	41.12
MetaCLIP+CRL	17.36	26.29	29.93	**42.32**	**71.88**	**77.32**	25.30	50.53	56.90	44.20
BLIP2+CRL	23.06	**34.86**	**39.07**	23.47	61.19	70.05	**49.57**	**80.44**	**84.06**	**51.75**

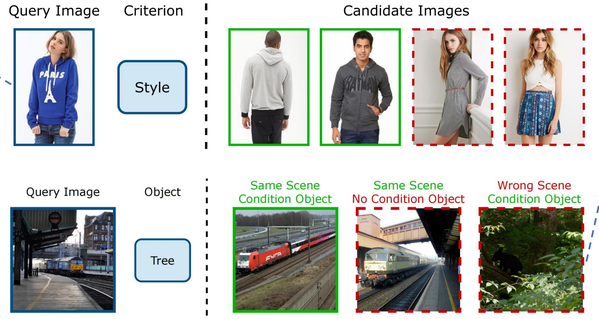
❖ Fashion Retrieval

Method	Texture	Fabric	Shape	Part	Style	Mean
Random	6.69	2.69	3.23	2.55	1.97	3.38
Triplet [43]	13.26	6.28	9.49	4.43	3.33	7.36
CSN [43]	14.09	6.39	11.07	5.13	3.49	8.01
ASEN [29]	15.13	7.11	12.39	5.51	3.56	8.74
ASEN++ [7]	15.60	7.67	14.31	6.60	4.07	9.64
RPF [8]	15.62	8.30	15.02	7.38	4.77	10.22
CLIP [34]	9.14	4.68	7.86	4.26	4.48	6.08
CLIP+CRL†	11.03	6.76	11.80	5.56	4.42	7.93
CLIP+CRL	**16.88**	**9.31**	**16.98**	**7.54**	**5.95**	**11.33**



Now we could apply CRL to...

And more ?



❖ Unsupervised Classification

Method	Clevr4-10k									Mean
	Texture			Shape			Color			
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
CC [25]	0.16	11.34	0.00	**94.66**	**96.89**	**93.90**	16.54	11.42	0.07	36.11
SCAN [40]	0.41	11.97	0.86	90.99	89.10	84.03	0.20	11.51	0.01	32.12
Multi-Map [54]	3.77	17.25	1.81	67.48	66.01	57.40	56.83	56.46	45.73	41.42
CLIP [34]	1.11	13.09	0.41	74.22	73.19	64.15	0.83	12.23	0.27	26.61
ALIGN [19]	1.36	13.30	0.41	89.33	86.77	83.37	0.47	11.79	0.10	31.88
MetaCLIP [48]	1.44	12.75	0.42	80.54	77.17	71.58	0.32	11.85	0.06	28.46
BLIP2 [23]	0.79	12.32	0.28	86.98	85.68	81.17	0.99	11.92	0.24	31.15
CLIP+CRL	10.74	25.11	6.35	78.69	83.05	72.42	**88.67**	**88.05**	**82.30**	59.49
ALIGN+CRL	**15.08**	**26.08**	**9.18**	88.27	87.63	81.83	85.07	76.15	72.69	60.22
MetaCLIP+CRL	12.74	25.89	7.28	87.32	88.15	82.98	88.35	86.27	81.08	**62.23**
BLIP2+CRL	6.46	18.77	3.37	90.11	88.91	84.52	84.67	81.97	74.85	59.29

Method	Clevr4-10k			Cards						Mean
	Count			Number			Suits			
	NMI	ACC	ARI	NMI	ACC	ARI	NMI	ACC	ARI	
CC [25]	2.08	14.67	1.09	24.91	26.34	12.30	24.94	39.21	16.87	18.05
SCAN [40]	3.42	14.29	1.23	11.11	18.21	17.60	15.01	32.02	9.48	13.60
Multi-Map [54]	11.38	20.13	7.67	16.32	20.61	7.95	14.02	46.65	11.08	17.31
CLIP [34]	9.50	19.02	5.70	16.84	18.91	8.44	16.52	43.74	12.93	16.84
ALIGN [19]	0.63	12.50	0.19	14.86	17.51	6.47	3.49	31.72	2.31	9.96
MetaCLIP [48]	7.62	17.27	3.97	17.39	19.78	9.04	15.48	38.72	13.11	15.82
BLIP2 [23]	6.11	16.36	3.13	24.34	25.25	13.08	31.26	47.04	22.25	20.98
CLIP+CRL	25.57	26.24	12.54	24.79	28.19	12.14	39.71	67.15	37.59	30.44
ALIGN+CRL	22.78	26.59	12.18	20.12	25.32	10.24	42.94	50.79	34.47	27.27
MetaCLIP+CRL	12.22	20.80	6.18	39.07	41.63	24.37	45.19	58.71	36.97	31.68
BLIP2+CRL	**28.55**	**30.92**	**16.28**	**46.55**	**48.35**	**32.31**	**60.86**	**76.07**	**55.94**	**43.98**

❖ Conditional Similarity Retrieval

Method	Focus			Change			Mean
	R@1	R@2	R@3	R@1	R@2	R@3	
$\text{CLIP}_{\text{image}}$	9.4	17.0	25.4	7.6	17.1	25.5	17.0
$\text{CLIP}_{\text{text}}$	7.4	14.0	23.0	8.1	16.4	24.7	15.6
$\text{CLIP}_{\text{image+text}}$	11.5	20.1	29.2	9.8	20.0	28.9	19.9
Pic2Word [35]	9.9	19.3	27.4	8.6	18.2	26.1	18.3
SEARLE [1]	10.8	18.2	27.9	8.3	15.6	25.8	17.8
LinCIR [15]	10.1	19.1	28.1	7.9	16.3	25.7	17.9
CIG [46]	10.6	19.2	27.4	7.9	16.9	25.4	17.9
CLIP+CRL	15.4	26.7	35.8	17.0	27.8	37.8	26.8
Combiner* [41]	16.6	27.7	37.2	18.0	32.2	41.6	28.9
CLIP+CRL*	**19.7**	**32.7**	**41.3**	**21.0**	**35.9**	**44.8**	**32.6**