Unifying Vision-Language Representation Space with Single-tower Transformer

Jiho Jang*[1], Chaerin Kong*[1], Donghyeon Jeon[2], Seonhoon Kim[3], Nojun Kwak[1]
Seoul National University[1], NAVER[2], Coupang[3]

Motivation

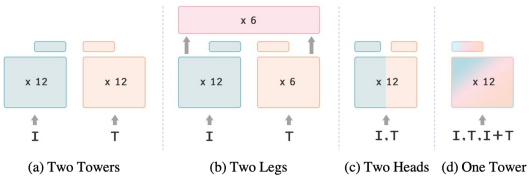
Modality-Agnostic Multimodal Representation Learning

- Mapping visual and linguistic information into a unified representation space.
- Mixing information within the input sequence in a modality-blind manner with generic token attention.

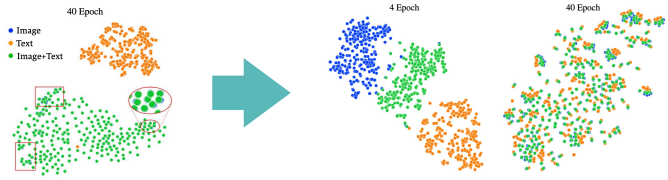
→ Better scalability, cross-modal / cross-task transferability, wider applications (ex: document understanding)

Modality Gap

- Previous frameworks train either modality-specific transformer blocks or linear projections.
- Training a single-tower representation learner is challenging due to the inherent modality gap of vision and language.



(a) Two Towers (b) Two Legs (c) Two Heads (d) One Tower



One Representation (OneR)

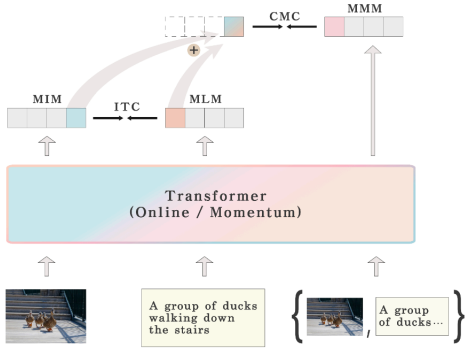
3-step modification to Image-Text Contrast

1/ Cross-Modal Mixup (XMC)

2/ Contextual Invariance (CIC)

3/ Contextual Mixup (CMC)

Method	Formulation	
ITC	$\mathcal{F}(I)$	$\mathcal{F}(T)$
XMC	$(\mathcal{F}(I)+\mathcal{F}(T))/2$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$
CIC	$(\mathcal{F}(I\|T)+\mathcal{F}(T\|I))/2$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$
CMC	$\mathcal{F}(I,T\|I,T)$	$(\mathcal{F}(I)+\mathcal{F}(T))/2$



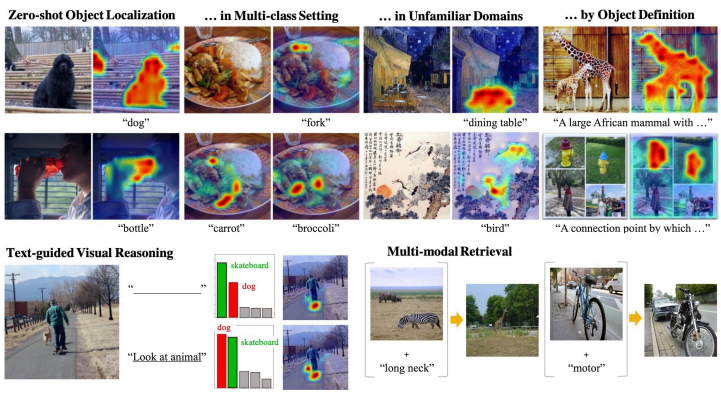
OneR Framework

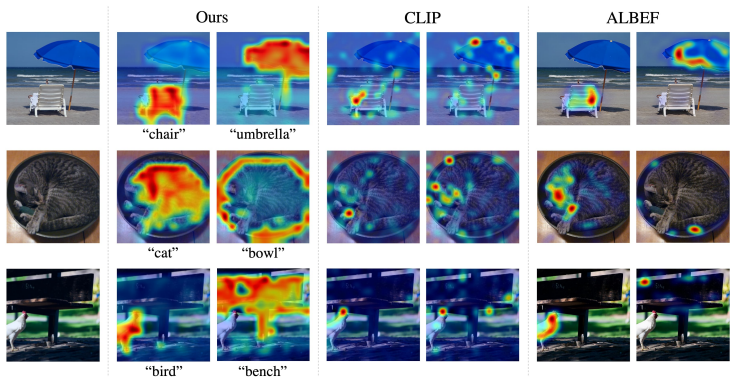
- Generic one-tower architecture with no modality specific modules or tokens.
- Understands Image, Text, (Image, Text) input in simplest way.
- Trained with MLM + MIM + ITC + CMC

Training Details

- Trained on 4M image-text pairs
- Trained with 32 A100 GPUs (26 hours)

Experiments





Method	Architecture	Pre.	#Images	Zero-shot MS-COCO (5K) Text Retrieval R@1	Zero-shot MS-COCO (5K) Text Retrieval R@5	Zero-shot MS-COCO (5K) Image Retrieval R@1	Zero-shot MS-COCO (5K) Image Retrieval R@5	Fine-tuned MS-COCO (5K) Text Retrieval R@1	Fine-tuned MS-COCO (5K) Text Retrieval R@5	Fine-tuned MS-COCO (5K) Image Retrieval R@1	Fine-tuned MS-COCO (5K) Image Retrieval R@5
ImageBert†	One Tower	O	6M	44.0	71.2	32.3	59.0	**66.4**	**89.8**	**50.5**	**78.7**
ViLT	One Tower	O	4M	56.5	82.6	40.4	70.0	61.5	86.3	42.7	72.9
Uni-Perceiver	One Tower	X	44.3M	57.7	85.6	46.3	**75.0**	64.7	87.8	48.3	75.9
OneR	One Tower	X	4M	**62.9**	**86.3**	**47.0**	74.7	66.1	87.8	48.3	76.0
CLIP	Two Towers	X	400M	58.4	81.5	37.8	62.4	-	-	-	-
FLAVA	Two Legs	O	70M	42.7	76.8	38.4	67.5	-	-	-	-
ALBEF	Two Legs	O	4M	68.7	89.5	50.1	76.4	73.1	91.4	56.8	81.5
TCL	Two Legs	O	4M	71.4	90.8	53.5	79.0	75.6	92.8	59.0	83.2

Method	INet Acc.	MS-COCO TR@1	MS-COCO TR@5	MS-COCO IR@1	MS-COCO IR@5
CLIP	17.1	15.0	34.8	10.9	26.7
SLIP	23.0	21.7	45.1	15.6	35.2
ITC (two heads)	17.5	10.4	26.8	10.7	26.4
ITC	1.6	0.8	2.5	0.7	2.2
+ XMC	22.1	25.2	48.1	15.2	33.6
+ XMC + CIC	22.9	25.4	48.1	16.3	35.5
+ CMC (OneR)	**23.7**	**25.5**	**48.2**	**16.9**	**36.9**

Conclusion

We present OneR, a simple framework that enables single-tower training on image-text pairs. The resulting unified representation naturally yields intriguing properties, such as word-patch level correspondence and multimodal retrieval.