Multi-modal Alignment using Representation Codebook

Jiali Duan*, liqun Chen*, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, Trishul Chilimbi

Amazon

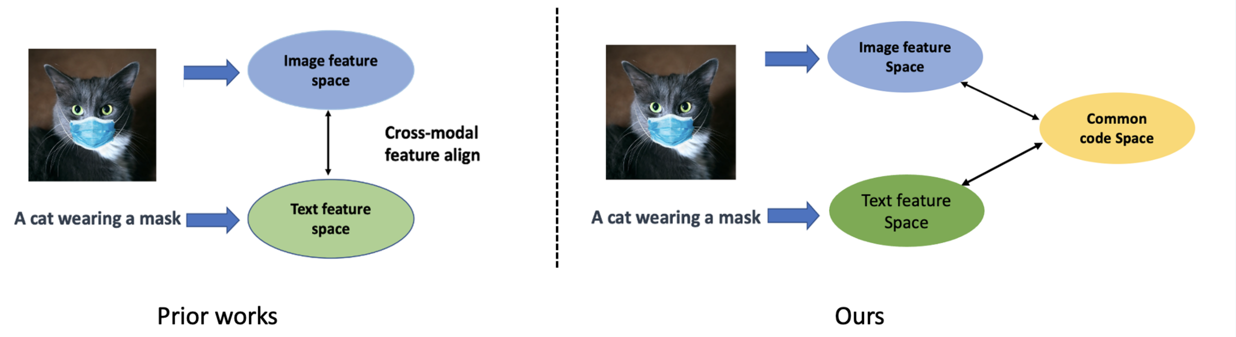
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Problem Definition

Goal:

- Improve vision language pretraining by learning better image/text alignment.

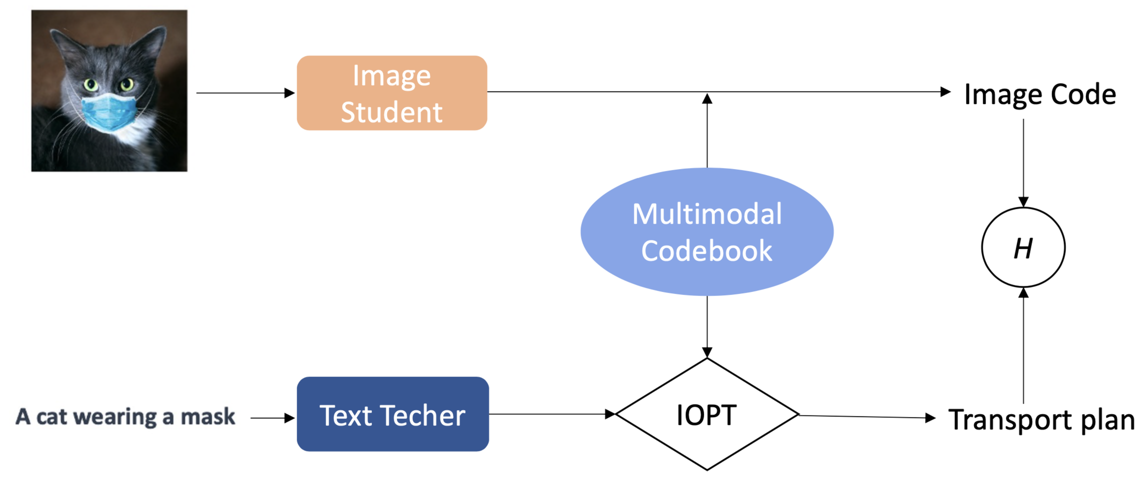
Motivation: i) alignment with multimodal codebook as semantic bridge; ii) image and text as two views of the same entity.



Contributions:

- Propose multi-modal codebook to align image and text at the cluster level.
- Connect SSL with VLP by generalizing teacher-student distillation to multimodal setting.

Multimodal Codebook Learning



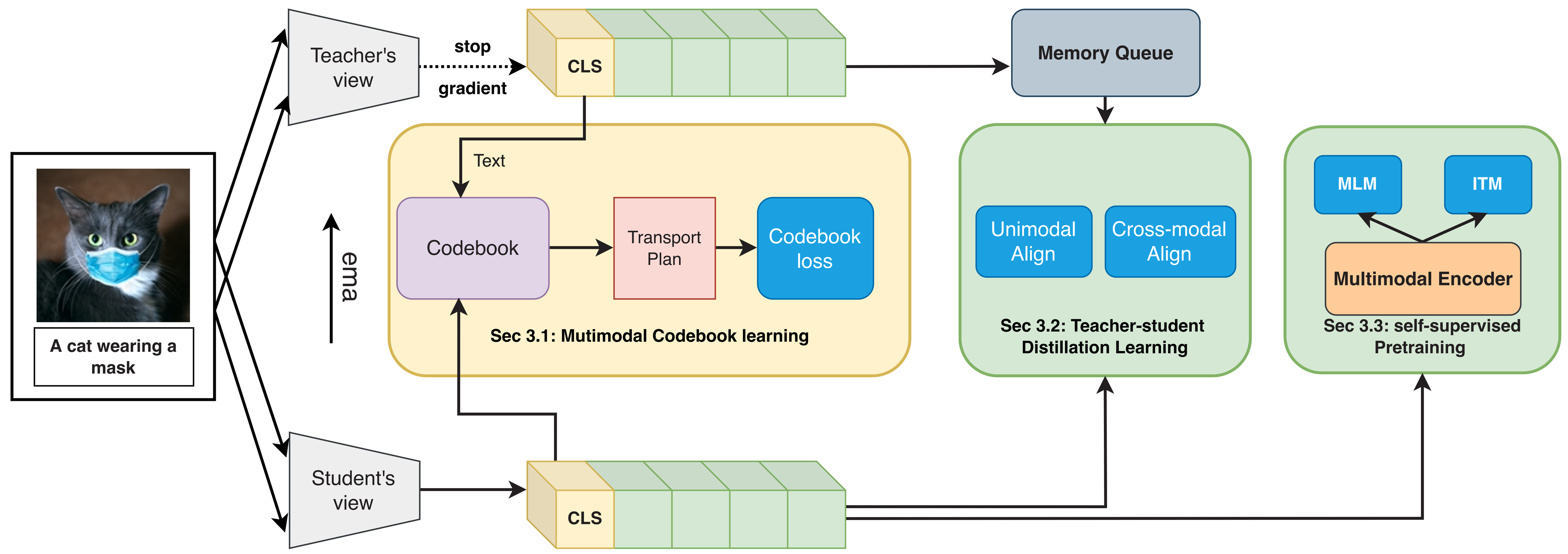
Image/text Alignment:

- Image instances should distribute to clusters proportionally to optimal text transport plan
- Text instances should distribute to clusters proportionally to optimal image transport plan

Optimal Transport Plan:

- Optimal cost to transport image instances to clusters
- Optimal cost to transport text instances to clusters

Method



Multimodal Codebook Learning:

$$L_{\text{code}} = L_{\text{t2p}}(\mathbf{Z_t}, \mathbf{C}, \mathbf{T_{i2p}}) + \mathbf{L}_{\text{i2p}}(\mathbf{Z_v}, \mathbf{C}, \mathbf{T_{t2p}})$$

Optimal Transport Plan:

$$L_{\text{ot}} = L_{\text{ot}}(\mathbf{Z_v^m}, \mathbf{C}) + \mathbf{L}_{\text{ot}}(\mathbf{Z_t^m}, \mathbf{C})$$

Self-supervised Pretraining:

$$L_{\text{itm}} = \mathbb{E}_{I,T\sim p_{\text{data}}} H(\boldsymbol{p}_{\text{itm}}, \boldsymbol{y}_{\text{itm}})$$
$$L_{\text{mlm}} = \mathbb{E}_{I,T\sim p_{\text{data}}} H(\boldsymbol{p}_{\text{mlm}}, \boldsymbol{y}_{\text{mlm}})$$

Teacher Student Contrastive Learning:

$$L_{\text{align}} = H(\boldsymbol{p}_{\text{t2i}}, \boldsymbol{y}_{\text{t2i}}) + H(\boldsymbol{p}_{\text{i2t}}, \boldsymbol{y}_{\text{i2t}}) + H(\boldsymbol{p}_{\text{i2i}}, \boldsymbol{y}_{\text{i2i}}) + H(\boldsymbol{p}_{\text{t2t}}, \boldsymbol{y}_{\text{t2t}})$$

- L_{t2i}: text student to image teacher
- L_{i2t}: image student to text teacher
- L_{t2t}: text student to text teacher
- L_{i2i}: image student to image teacher

Experiments & Results

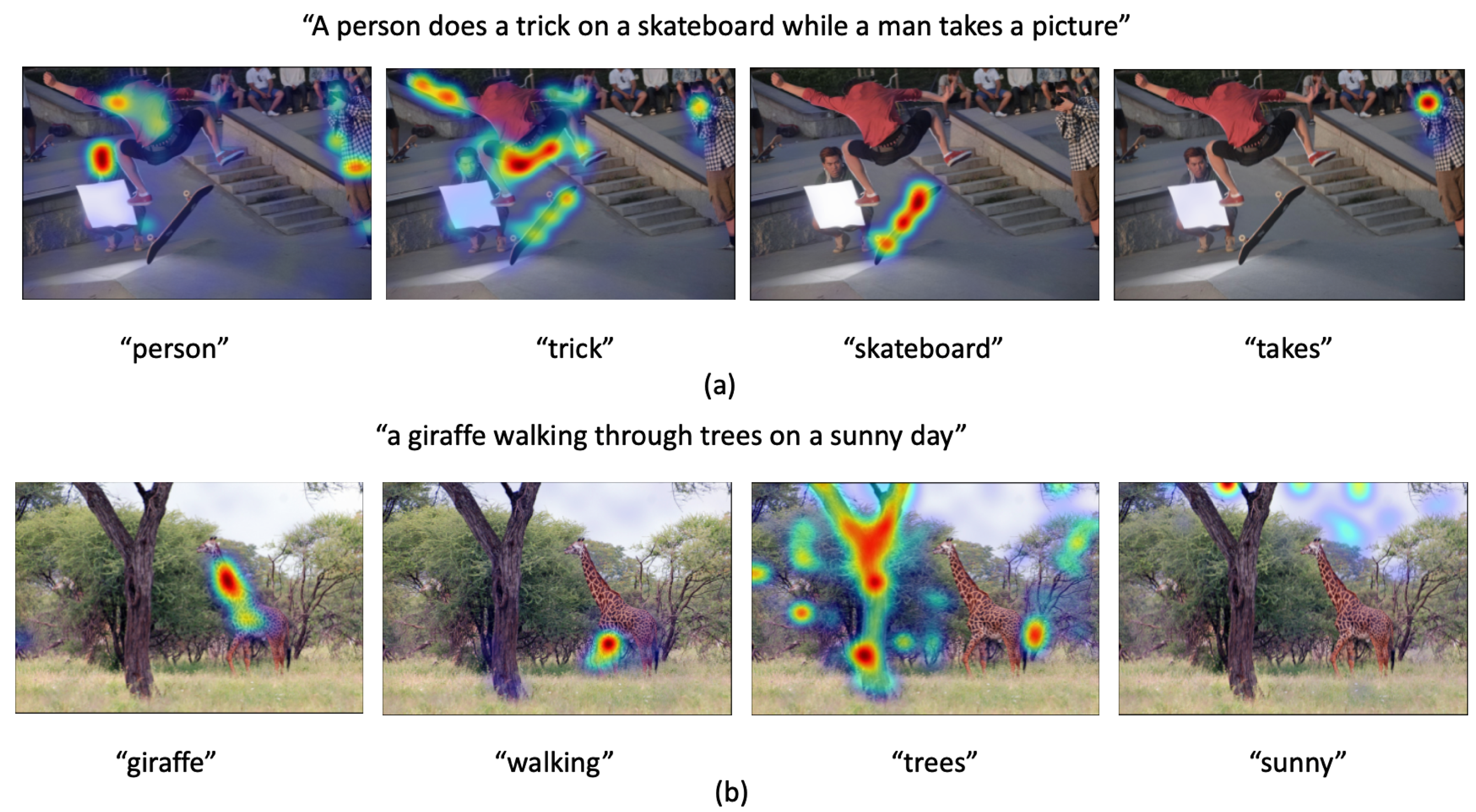
Performance comparison of zero-shot image-text retrieval on MSCOCO and Flickr30K datasets.

Method	MSCOCO (5K)						Flickr30K (1K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	44.0	71.2	80.4	32.3	59.0	70.2	70.7	90.2	94.0	54.3	79.6	87.5
Unicoder-VL	-	-	-	-	-	-	64.3	85.8	92.3	48.4	76.0	85.2
UNITER	-	-	-	-	-	-	80.7	95.7	98.0	66.2	88.4	92.9
ViLT	56.5	82.6	89.6	40.4	70.0	81.1	73.2	93.6	96.5	55.0	82.5	89.8
CLIP	58.4	81.5	88.1	37.8	62.4	72.2	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN	58.6	83.0	89.7	45.6	69.8	78.6	88.6	98.7	**99.7**	75.7	93.8	**96.8**
ALBEF 4M	68.6	89.5	94.7	50.1	76.4	84.5	90.5	98.8	**99.7**	76.8	93.7	96.7
Ours	**71.5**	**91.1**	**95.5**	**53.9**	**79.5**	**87.1**	**91.7**	**99.3**	99.8	**79.7**	**94.8**	97.3

Performance comparison of finetuned image-text retrieval on MSCOCO and Flickr30K datasets.

Method	MSCOCO (5K)						Flickr30K (1K)					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ImageBERT	66.4	89.8	94.4	50.5	78.7	87.1	87.0	97.6	99.2	73.1	92.6	96.0
UNITER	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
OSCAR	70.0	91.1	95.5	54.0	80.8	88.5	-	-	-	-	-	-
ViLT	61.5	86.3	92.7	42.7	72.9	83.1	83.5	96.7	98.6	64.4	88.7	93.8
UNIMO	-	-	-	-	-	-	89.7	98.4	99.1	74.6	93.4	96.0
SOHO	66.4	88.2	93.8	50.6	78.0	86.7	86.5	98.1	99.3	72.5	92.7	96.1
ALBEF 4M	73.1	91.4	96.0	56.8	81.5	89.2	94.3	**99.4**	99.8	82.8	96.7	**98.4**
Ours	**75.3**	**92.6**	**96.6**	**58.7**	**82.8**	**89.7**	**95.1**	**99.4**	**99.9**	**83.3**	96.1	97.8

Qualitative Results:



Paper ID: 11488

Paper Link: https://arxiv.org/abs/2203.00048