Do Vision and Language Encoders Represent the World Similarly?

Mayug Maniparambil, Raiymbek Akshulakov, Yasser Dahou, Sanath Narayan, Mohamed Seddik, Karttikeya Mangalam, Noel E. O'Connor

ML Labs
Berkeley UNIVERSITY OF CALIFORNIA
TII Technology Innovation Institute

Introduction

- Given vision and language represent the same physical reality, how similar are their encoders' representations?
- Semantic similarity as measured by CKA is high for well-trained vision and language encoders.
- We propose 2 novels methods to connect unimodal vision and language representation spaces in a 0-shot manner.
- We showcase the 0-shot connection on cross-domain, and cross-lingual retrieval tasks.

CKA Metric

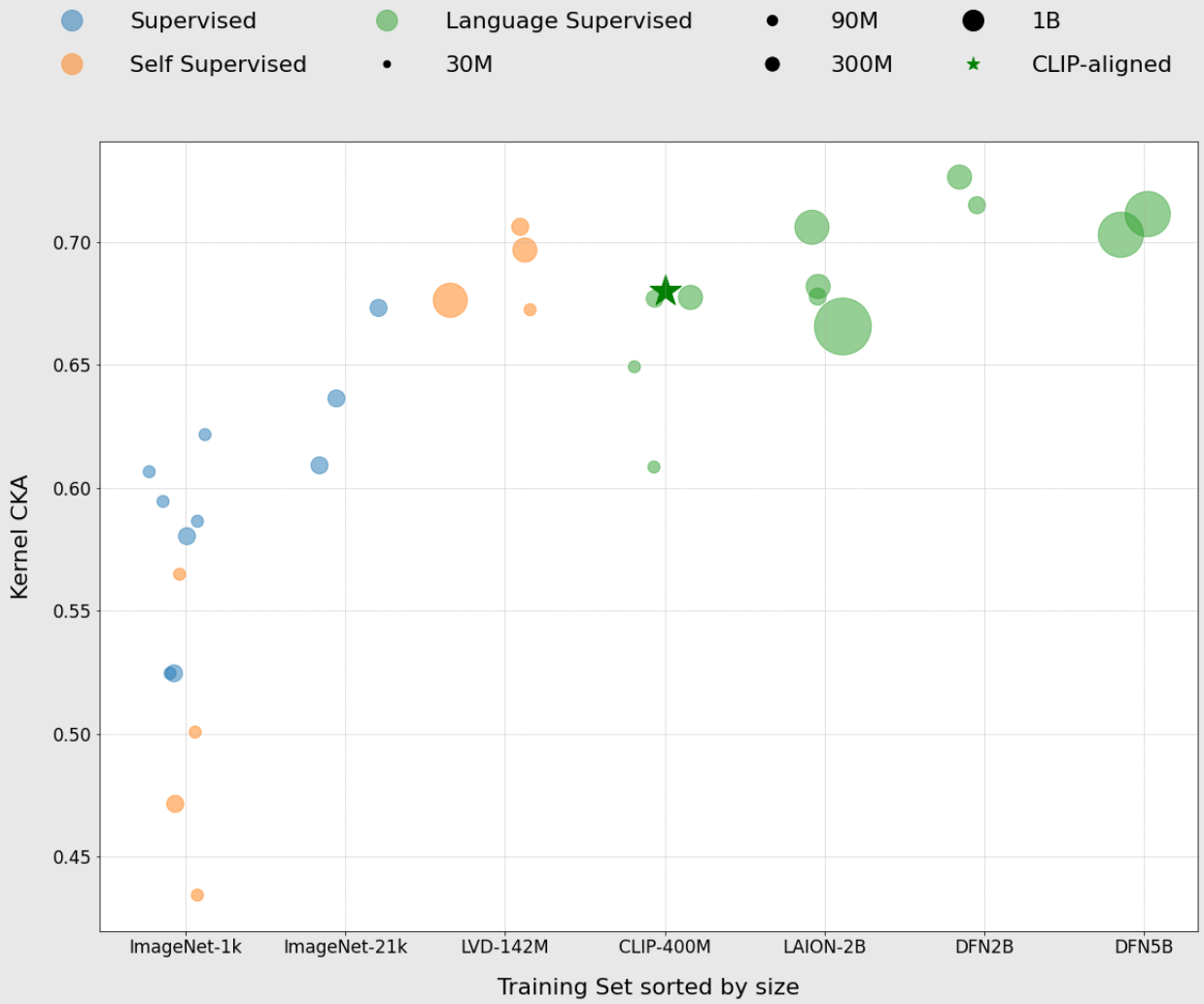
$$\mathrm{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\mathrm{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathrm{HSIC}(\mathbf{K}, \mathbf{K})\,\mathrm{HSIC}(\mathbf{L}, \mathbf{L})}}$$

CKA (Centered Kernel Alignment) measures the semantic similarity between the representation spaces of vision and language encoders by comparing the centered kernel matrices of their embeddings. K,L are the kernels of N vision (dimension p) and text embeddings (dimension q) as below.

$$\mathbf{K} = k(\mathbf{X}^\top, \mathbf{X}) \qquad \mathbf{L} = \ell(\mathbf{Y}^\top, \mathbf{Y})$$

$$\mathbf{X} \in \mathbb{R}^{p \times N} \qquad \mathbf{Y} \in \mathbb{R}^{q \times N}$$

Semantic Alignment with text encoder



- CKA between several vision encoders and a text encoder- All-Roberta-Large
- Well trained Vision encoders exhibit high semantic similarity with text encoder
- Beyond a certain scale, vision encoders exhibit comparable or better CKA to unaligned language encoder than that of CLIP's image, text embeddings(green star).

CKA reduces with shuffling

- DinoV2 and All-Roberta-Large have high semantic similarity (CKA Score) between vision and text embeddings on COCO.
- How can we exploit this semantic similarity to connect unaligned vision/language encoders?

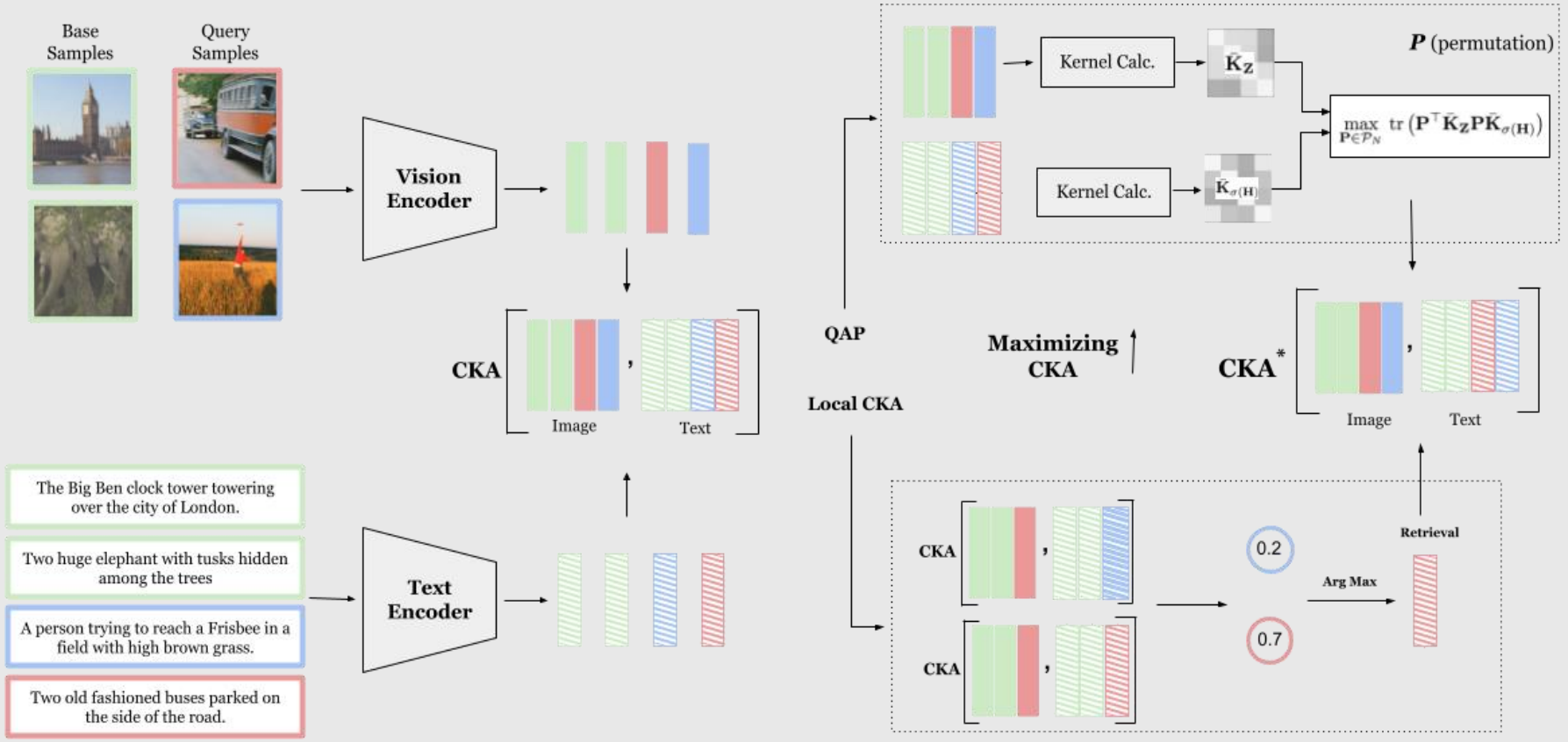
Shuffling (%)	0	20	40	60	80	100
CKA Score	0.72	0.46	0.27	0.13	0.04	0.01

CKA reduces with shuffling

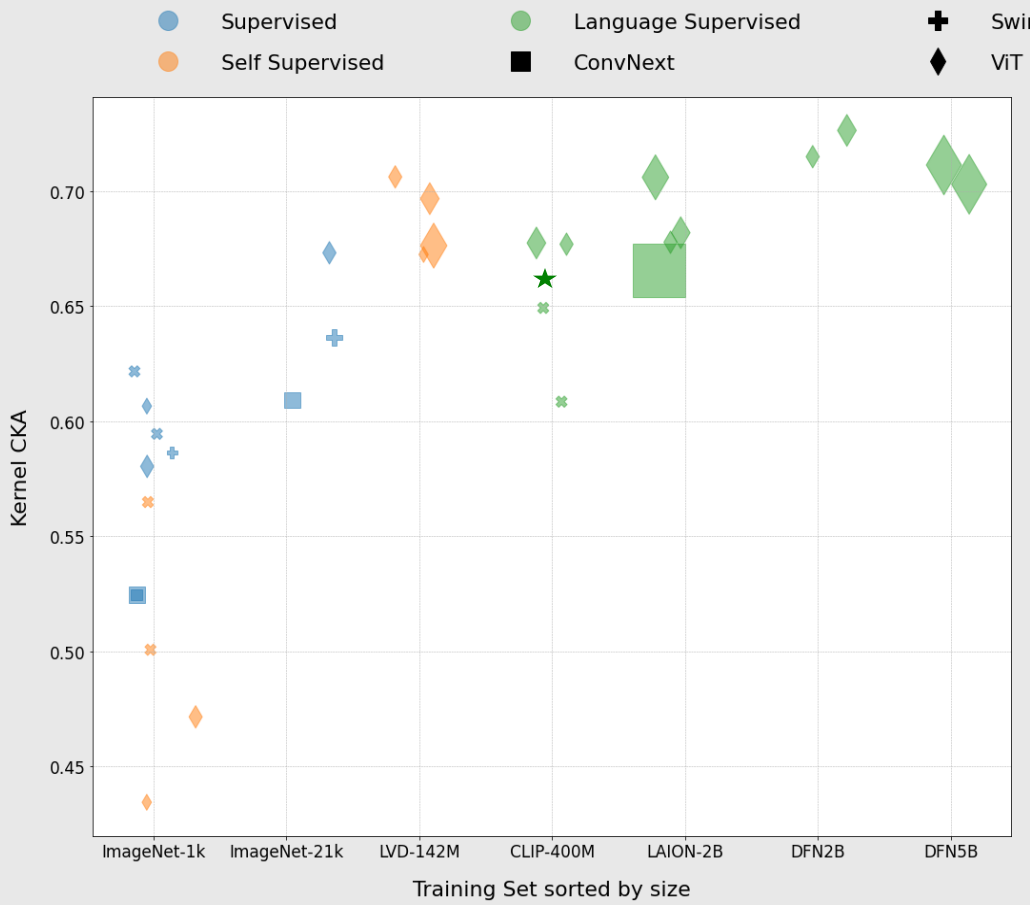
CKA scores between image-caption representations pairs. The exact ordering yields the best score, whereas shuffling the representations reduces the CKA score.

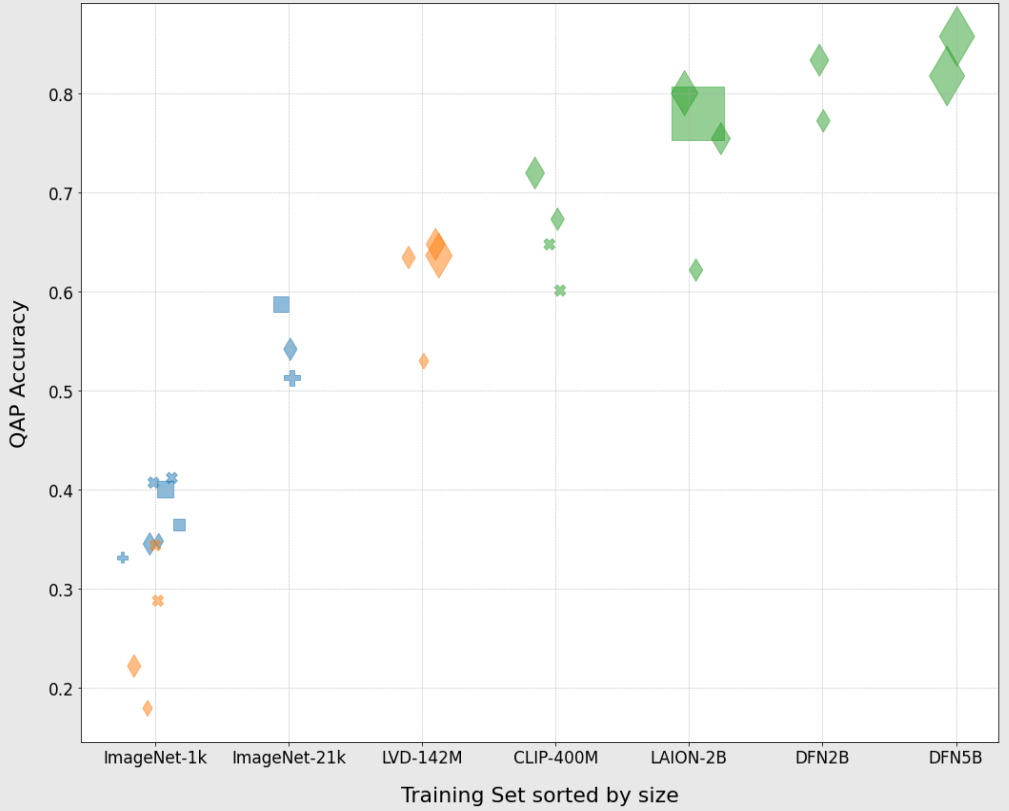
- We can maximize CKA to find the alignment between representation spaces of unaligned vision/language encoders.

Methodology



Different vision encoders comparison





COCO / NoCaps Caption Retrieval and Matching

Method	Vision Model	NoCaps [2]		COCO [27]	
		Matching accuracy	Top-5 retrieval	Matching accuracy	Top-5 retrieval
Cosine Similarity*	CLIP [40]	99.5	99.6	97.1	96.1
Linear regression	CLIP-V [40]	29.3	44.7	42.7	59.1
	ConvNeXt [47]	19.0	28.5	31.3	46.1
	DINOv2 [37]	38.1	50.3	45.1	65.4
Relative representations [34]	CLIP-V [40]	61.3	37.6	61.6	41.3
	ConvNeXt [47]	25.5	17.8	38.6	34.1
	DINOv2 [37]	46.0	46.4	47.7	52.3
Ours: QAP	CLIP-V [40]	**67.3**	-	**72.3**	-
	ConvNeXt [47]	46.7	-	66.1	-
	DINOv2 [37]	57.7	-	66.0	-
Ours: Local CKA	CLIP-V [40]	65.1	60.5	71.9	69.9
	ConvNeXt [47]	43.7	44.4	64.8	65.5
	DINOv2 [37]	58.7	**61.8**	64.3	**70.5**

The DINOv2 model, trained solely through self-supervision, demonstrates the formation of semantic concepts independently of language supervision

Cross-Lingual caption matching and retrieval

Language		Kernel CKA		Matching Accuracy				Retrieval @ 5	
		CLIP	Ours	CLIP	Relative[34]	Linear	Ours (QAP)	CLIP	Ours (Local)
Latin	de	0.472	0.627	41.8	35.0	34.0	39.6	65.1	56.7
	en	0.567	0.646	81.5	52.5	40.9	51.6	92.5	69.0
	es	0.471	0.634	50.2	37.8	31.7	41.4	68.5	61.6
	fr	0.477	0.624	49.4	37.5	30.7	40.2	68.7	57.6
	it	0.472	0.638	41.0	37.2	34.9	38.5	61.3	59.7
Non-Latin	jp	0.337	0.598	13.2	28.3	23.5	30.5	30.0	49.4
	ko	0.154	0.620	0.50	30.4	23.5	30.9	3.30	53.4
	pl	0.261	0.642	5.40	36.6	30.2	40.2	18.8	59.5
	ru	0.077	0.632	0.80	31.9	30.7	35.1	4.10	53.2
	tr	0.301	0.624	4.30	35.8	29.6	38.9	15.2	59.3
	zh	0.133	0.641	2.70	36.5	31.1	40.3	8.90	57.8
	Avg.	–	–	26.4	36.3	30.9	**38.8**	39.6	**57.9**

QAP and Local-CKA-based methods perform comparably in Latin languages while outperforming non-Latin languages, highlighting the efficacy of our training-free transfer approach

Conclusion

- Well Trained Vision Encoders on sufficiently large datasets exhibit high semantic similarity with language encoders comparable to aligned encoders.
- Semantic similarity to language encoders (as measured by CKA) increases with dataset size and dataset quality irrespective of the training paradigm.
- We develop 2 novel methods to perform matching / retrieval between unaligned vision language encoders by maximizing the CKA
- Demonstrate superior performance than SOTA on 0-shot latent space communication on cross-domain, cross-lingual caption matching/retrieval tasks.

github.com/mayug/0-shot-llm-vision