Topological Alignment of Shared Vision-Language Embedding Space

paper

Junwon You [1], Dasol Kang [2], Jae-Hun Jung [1]

E-mail: {jwyou627, jung153} @ postech.ac.kr, english4118@gmail.com

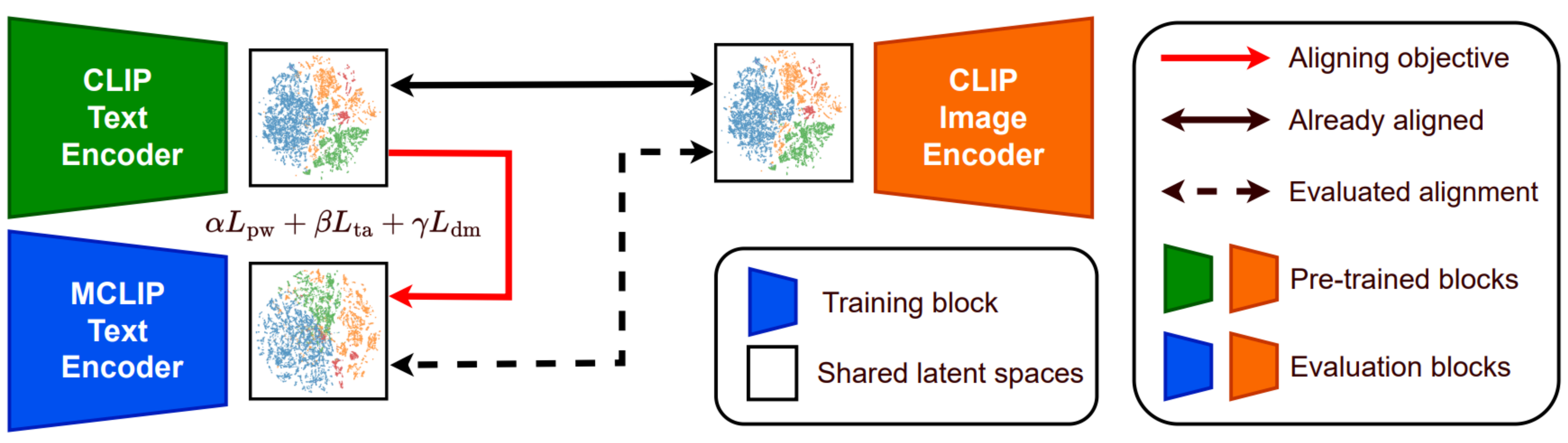
[1] Department of Mathematics, POSTECH , [2] Dololo Research Engineer

We propose **ToMCLIP** (**To**pological Alignment for **M**ultilingual **CLIP**), which **aligns embedding spaces** through **topology-preserving constraints**. ToMCLIP applies **persistent homology** to define topological alignment loss and **approximates persistence diagrams (PDs)** with theoretical error bounds using graph sparsification strategy.

Motivation

The **multilingual extension of CLIP** extends visual-language understanding beyond English, enabling consistent multimodal intelligence across diverse languages and cultures. However, a performance gap compared to English still remains. We visualized the **vision-language embedding spaces** for each language and observed that non-English languages exhibit **structural differences compared to English**. We hypothesize that this performance gap arises from topological discrepancies and propose a **topological alignment loss** to align these structures.

Methodology



To align Multilingual CLIP (MCLIP) text encoder with CLIP Text Encoder, we propose **topological alignment loss** composed of L_{ta} and L_{dm}, which aligns the **topological structural features** of the embedding space. The L_{pw} represents MCLIP loss, which performs **pointwise alignment** between embeddings.

- $L_{pw} = MSE(E_T(X), E_S(X^*))$
- $L_{ta} = SW_p(D_T, D_S)$
- $L_{dm} = MSE(M_T, M_S)$

MSE : mean squared error
SW_p : p-sliced Wasserstein distance

E_T (E_S): CLIP (MCLIP) text encoder.
X^*: machine translation of X.
D_T, D_S: PDs of $E_T(X)$ and $E_S(X^*)$.
M_T, M_S : pairwise distance matrices of $E_T(X)$ and $E_S(X^*)$.

Approximation of PDs. Let $X = \{x_1, \cdots, x_N\}$ be a point cloud. We construct a complete graph $G = (V, E, w)$ and sparse graph $G_\epsilon = (V, E, w_\epsilon)$ where $V = \{x_i\}_{i=1}^N$, $E = \{(x_i, x_j) \mid x_i, x_j \in X, i \neq j\}$,

$$w\left((x_i, x_j)\right) = \frac{d(x_i, x_j)}{M} \quad \text{and} \quad w_\epsilon(e) = \begin{cases} w(e), & if\ w(e) \le \epsilon \\ 1, & if\ w(e) > \epsilon \end{cases}.$$

Here, $M = \max_{(x_i,x_j)\in E} d(x_i, x_j)$. By construction, $0 \le w(e) \le 1$.

Theorem (Approximation Error Bound)

Let $0 \le \epsilon \le 1$ and $m(\epsilon) := \#\left\{(0, d) \in D_0^{Rips}(G) \mid \epsilon < d < \infty\right\}$, i.e. the number of finite 0-dimensional persistence points of G whose death times exceed ϵ. Then,

$$W_p(D_0^{Rips}(G), D_0^{Rips}(G_\epsilon)) \le m(\epsilon)^{\frac{1}{p}}(1 - \epsilon)$$

and $0 \le m(\epsilon) \le N - 1$ where W_p denotes the p-Wasserstein distance. In addition, $m(\epsilon) = c(\epsilon) - 1$ where $c(\epsilon)$ is the number of connected components in $VR_\epsilon(G)$.

$c(\epsilon)$ and sparsity of G_ϵ constructed from random point clouds.

N	Connected components $c(\epsilon)$										Sparsity									
	Uniform (λ)					Gaussian (λ)					Uniform (λ)					Gaussian (λ)				
	1.0	0.5	0.0	−0.5	−1.0	1.0	0.5	0.0	−0.5	−1.0	1.0	0.5	0.0	−0.5	−1.0	1.0	0.5	0.0	−0.5	−1.0
64	1.6	1.1	1.0	1.0	1.0	4.1	1.4	1.1	1.0	1.0	0.158	0.306	0.496	0.690	0.840	0.157	0.309	0.504	0.693	0.840
128	1.7	1.0	1.0	1.0	1.0	3.1	1.2	1.0	1.0	1.0	0.160	0.310	0.499	0.692	0.841	0.160	0.311	0.502	0.694	0.841
256	1.1	1.0	1.0	1.0	1.0	3.2	1.2	1.1	1.0	1.0	0.159	0.308	0.499	0.692	0.841	0.159	0.310	0.503	0.693	0.842
512	1.0	1.0	1.0	1.0	1.0	2.2	1.0	1.0	1.0	1.0	0.158	0.308	0.499	0.690	0.841	0.159	0.310	0.502	0.692	0.841

- $\lambda \in \{1.0, 0.5, 0.0, -0.5, -1.0\}$
- $\epsilon = \mu - \lambda\sigma$ where μ and σ are the mean and standard deviation of w.

When λ is 0.5, the approximation error is close to 0 and G_ϵ has 0.3 sparsity. We take this value for experiments.

Results

The model trained with Topological alignment loss, **ToMCLIP**, improves the performance on zero-shot classification and multilingual retrieval. In zero-shot classification, L_{ta} shows improvement without L_{dm}, but not L_{dm}.

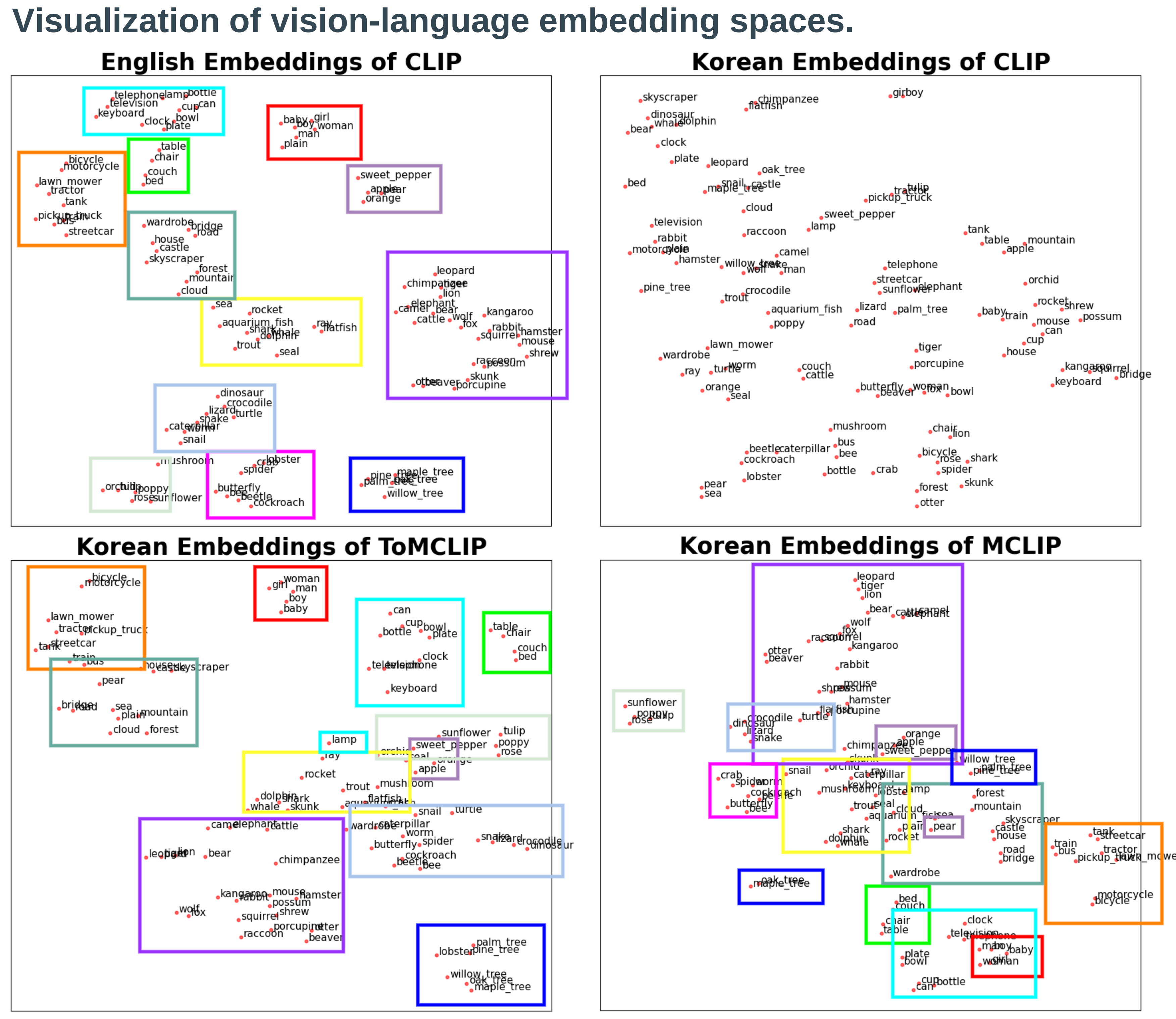
Zero-shot classification on CIFAR-100.

Setting	Model	Languages (13)													*Avg*
		En	Fr	Es	De	It	Ru	Pl	Tr	Da	Ja	Zh	Ko	Vi	
Full	CLIP	91.06	66.18	63.69	64.05	49.33	11.95	22.03	24.73	32.42	32.80	21.56	12.38	15.32	39.04
	MCLIP	91.97	85.66	87.10	85.74	88.23	87.98	**85.38**	87.65	87.83	53.60	89.50	87.20	86.26	84.93
	ToMCLIP(L_{dm})	**91.99**	84.77	84.63	**89.63**	86.17	87.78	84.86	87.35	86.88	56.27	88.11	87.94	86.98	84.87
	ToMCLIP(L_{ta})	91.48	85.41	84.23	87.85	88.49	**89.43**	84.35	**88.76**	**87.98**	**58.57**	**89.75**	**88.76**	**89.41**	85.73
	ToMCLIP	91.40	**87.59**	**87.37**	89.30	**89.11**	87.66	83.59	88.59	87.79	57.95	88.68	88.36	88.17	**85.81**
Low	CLIP	**91.06**	66.18	63.69	64.05	49.33	11.95	22.03	24.73	32.42	32.80	21.56	12.38	15.32	39.04
	MCLIP	79.72	67.60	62.20	71.41	59.68	69.80	64.55	58.71	73.31	60.68	78.27	65.43	71.38	67.90
	ToMCLIP(L_{dm})	79.46	67.99	62.51	70.81	60.75	69.30	64.02	57.21	72.64	59.20	77.43	67.42	70.07	67.60
	ToMCLIP(L_{ta})	80.00	67.37	62.66	70.09	60.88	70.31	65.22	59.50	72.68	60.94	77.36	67.01	**73.37**	68.26
	ToMCLIP	80.75	**68.56**	**63.85**	**71.49**	**62.91**	**71.23**	**65.50**	**60.80**	**73.75**	**62.39**	**78.82**	**67.96**	72.44	**69.26**

Multilingual retrieval on xFlicker&CO.

Direction	Model	Low			Full		
		R@1	R@5	R@10	R@1	R@5	R@10
IR	CLIP	12.08	22.12	27.19	12.08	22.12	27.19
	MCLIP	33.51	62.04	73.70	50.13	77.51	85.86
	ToMCLIP(L_{dm})	34.49 (▲0.98)	62.93 (▲0.89)	**74.50** (▲0.80)	**50.85** (▲0.72)	**78.25** (▲0.74)	**86.56** (▲0.70)
	ToMCLIP(L_{ta})	**34.50** (▲0.99)	**62.96** (▲0.93)	74.45 (▲0.74)	50.79 (▲0.66)	78.01 (▲0.50)	86.19 (▲0.33)
	ToMCLIP	34.03 (▲0.52)	62.59 (▲0.56)	74.00 (▲0.30)	50.76 (▲0.63)	77.99 (▲0.48)	86.48 (▲0.62)
TR	CLIP	16.01	28.75	35.40	16.01	28.75	35.40
	MCLIP	39.39	68.02	78.65	53.38	79.48	87.34
	ToMCLIP(L_{dm})	39.71 (▲0.32)	68.63 (▲0.61)	79.38 (▲0.74)	54.01 (▲0.63)	**80.38** (▲0.90)	**88.08** (▲0.74)
	ToMCLIP(L_{ta})	**40.29** (▲0.90)	**69.18** (▲1.16)	**79.61** (▲0.97)	53.83 (▲0.45)	79.91 (▲0.43)	87.80 (▲0.46)
	ToMCLIP	39.51 (▲0.12)	68.42 (▲0.40)	78.96 (▲0.32)	**54.07** (▲0.69)	79.98 (▲0.50)	87.67 (▲0.33)

Visualization of vision-language embedding spaces.



Sorted pairwise distance curves of English and Korean embeddings

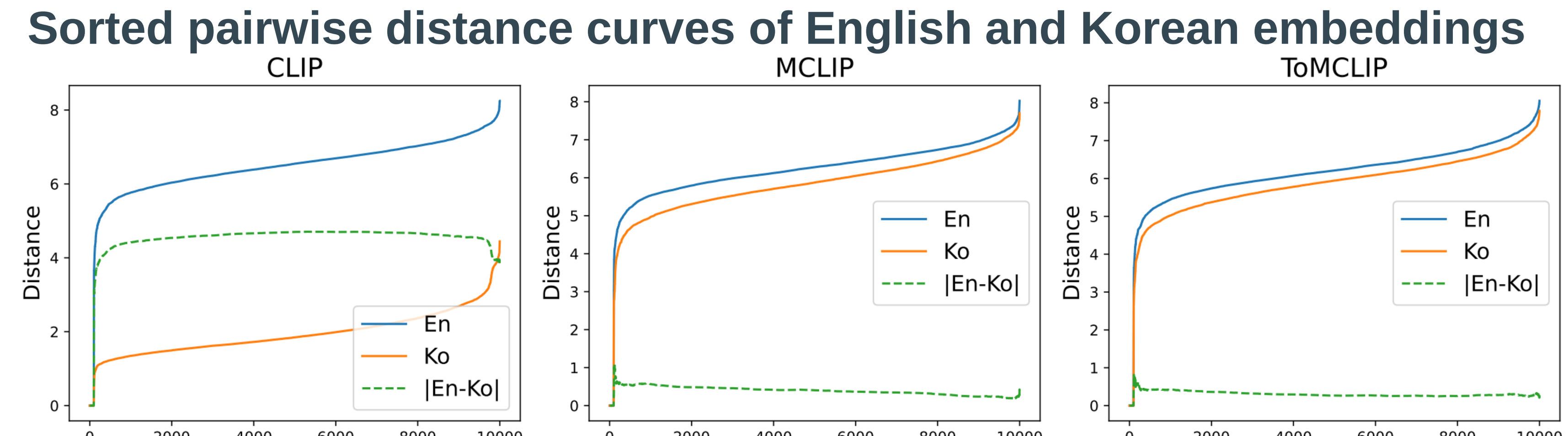


Table 4. Comparison of topological distances between English and Korean embeddings.

Comparison	W_2^ϵ	W_2		$SW_2^{(50)}$	
		0-dim	1-dim	0-dim	1-dim
CLIP (En) vs. CLIP (Ko)	7.7870	34.5016	1.0468	2.8261	4.1593
MCLIP (En) vs. MCLIP (Ko)	2.5988	5.1995	0.9250	0.3670	0.5964
ToMCLIP (En) vs. ToMCLIP (Ko)	**2.4929**	**4.2072**	**0.7444**	**0.3056**	**0.3393**

Visualizations of embedding spaces and distance curves verify that the topological alignment loss helps the model reduce the topological discrepancy between languages. The reduction of topological discrepancy is quantitatively confirmed by the lower Wasserstein distance (Table 4).

Conclusion

We proposed a topological alignment loss integrated into MCLIP to reduce topological discrepancies across languages. ToMCLIP reduces the topological discrepancies between embedding spaces and improves classification and retrieval performance. However, a gap still remains. We plan to extend this approach for multimodal alignment.

Reference

Radford, A., et al. (2021, July). Learning transferable visual models from natural language supervision. ICML2021
Carlsson, et al. (2022, June). Cross-lingual and multilingual clip. LREC2022
Kim, J., et al. (2024, July). Do topological characteristics help in knowledge distillation?. ICML2024

Acknowledgements.

This work was supported by National Research Foundation of Korea under grant number 2021R1A2C3009648, Basic Science Research Institute Fund, whose NRF grant number is RS -2021-NR060139, and partially by NRF grant by the Korea government (MSIT) (RS-2023-00219980).