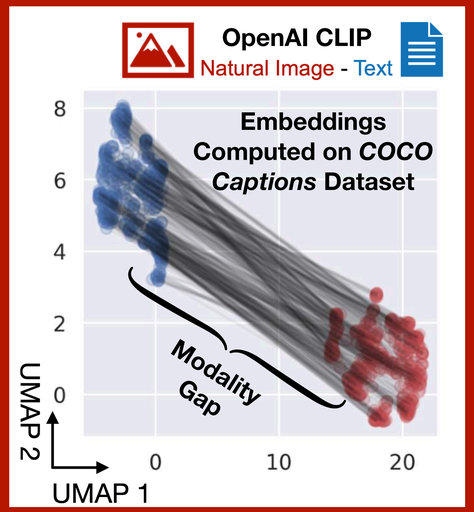
Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning

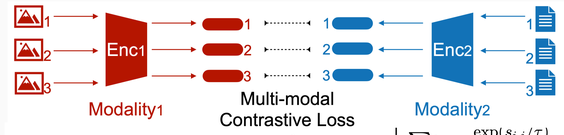
Yuhui Zhang*, Weixin Liang*, Yongchan Kwon*, Serena Yeung, James Zou

We present **modality gap**, an intriguing geometric phenomenon of multi-modal representation space, caused by a combination of **model initialization** and **contrastive learning optimization**.



Paper

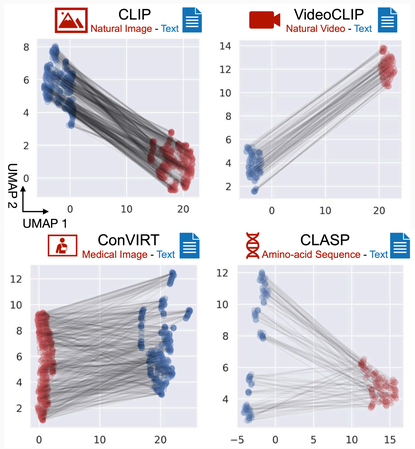
Modality Gap in Multi-Modal Contrastive Learning



$$(x_k \in \text{Modal}_1, y_k \in \text{Modal}_2) \sim D$$
$$\mathbf{x}_k = \text{Normalize}(\text{Enc}_1(x_k))$$
$$\mathbf{y}_k = \text{Normalize}(\text{Enc}_2(y_k))$$
$$s_{i,j} = \mathbf{x}_i \cdot \mathbf{y}_j$$

$$\mathcal{L}_{\mathcal{M}_1 \to \mathcal{M}_2} = -\frac{1}{N}\sum_i \log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{i,j}/\tau)}$$
$$\mathcal{L}_{\mathcal{M}_2 \to \mathcal{M}_1} = -\frac{1}{N}\sum_i \log \frac{\exp(s_{i,i}/\tau)}{\sum_j \exp(s_{j,i}/\tau)}$$
$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{\mathcal{M}_1 \to \mathcal{M}_2} + \mathcal{L}_{\mathcal{M}_2 \to \mathcal{M}_1})$$

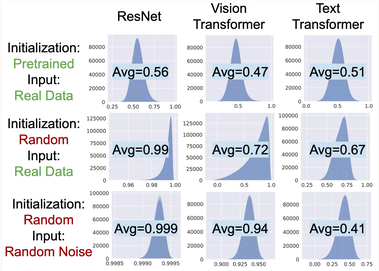
Multi-modal Contrastive Learning: Paired inputs from different data modalities are mapped into a shared representation space (e.g., OpenAI CLIP).



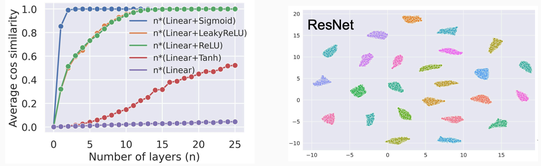
Pervasive Modality Gap: Clear gap is observed when visualizing embeddings from different modalities in 2D using UMAP (lines indicate pairs).

Modality gap is caused by ① + ②

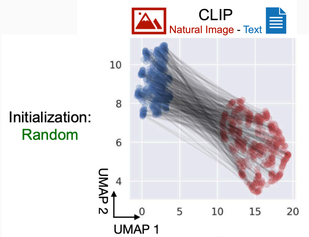
① Initialization: The Cone Effect Induces A Modality Gap



Cone Effect: The average cosine similarity between all pairs of embeddings is substantially larger than 0, indicating that the embedding space is a narrow cone.

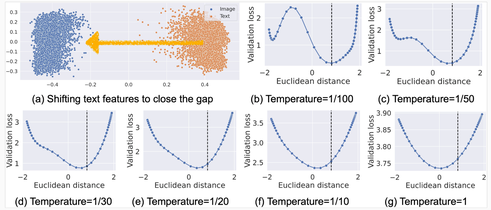


Analysis of Cone Effect: Deeper networks create narrower cones; different random initializations create distinctively different cones.

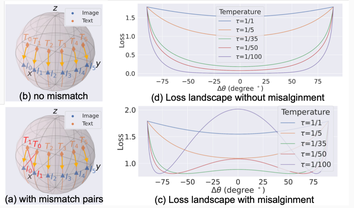


Cone Effect Induces Modality Gap: For multi-modal models with two encoders, representations of two modalities are embedded into two different cones at the initialization stage, thus creating the modality gap.

② Optimization: Contrastive Learning Preserves Modality Gap



Embedding Shift Analysis: The landscape of contrastive loss when manually shifting the embeddings from two modalities. The gap is optimal at default temperature.



Simulation Analysis: Six embedding pairs on a 3D sphere; text embeddings are shifted towards closing the modality gap. Mismatched pairs encourages the gap.

Modality Gap Implications

Dataset	Original gap	Modified gap	Direction
Coarse-grained Classification			
CIFAR10	0.9013	**0.9081**	↑
CIFAR100	0.6658	**0.6737**	↓
Fine-grained Classification			
EuroSAT	0.5410	**0.5645**	↓
Optical Character Recognition			
SVHN	0.5389	**0.5396**	↑
HatefulMemes	0.5800	**0.5811**	↑

Denigration Biases	Original gap			Modified gap		
	Crime related	Non human	Sum	Crime related	Non human	Sum
Black	1.0%	0.1%	**1.1%**	0.8%	0.1%	**1.0%**
White	15.5%	0.2%	**15.7%**	13.2%	0.4%	**13.7%**
Indian	1.2%	0.0%	**1.2%**	1.1%	0.0%	**1.1%**
Latino	2.8%	0.1%	**2.8%**	1.9%	0.1%	**2.0%**
Middle Eastern	6.3%	0.0%	**6.3%**	5.2%	0.0%	**5.2%**
Southeast Asian	0.5%	0.0%	**0.5%**	0.3%	0.0%	**0.3%**
East Asian	0.7%	0.0%	**0.7%**	0.6%	0.0%	**0.6%**

Modifying the modality gap can improve zero-shot performances for downstream tasks and reduces model biases.