Contrastive Alignment for Multimodal Representation Learning[1]

Hangke Sui
hangkes2@illinois.edu

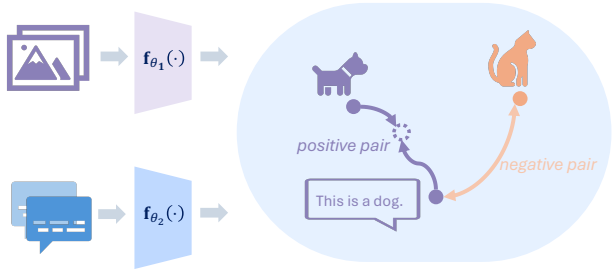
PhD Student
Department of Electrical & Computer Engineering
University of Illinois Urbana-Champaign

Supervisor: Prof. Minh Do

[1]Hangke Sui*, Yuqing Wang*, Minh Do. *UniCon: Unified Framework for Efficient Contrastive Alignment via Kernels*, ICLR2026

Outline

Introduction

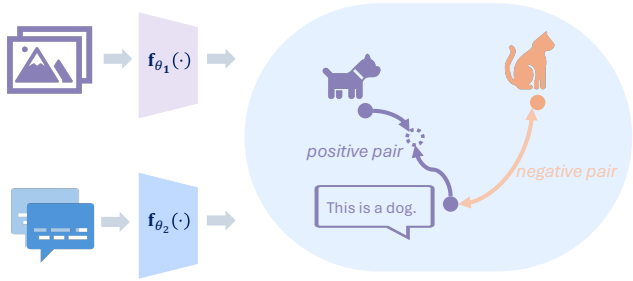


In multimodal representation learning, contrastive learning (CL) seeks a shared representation space with two behaviors:

- Alignment: matched pairs should be close
- Uniformity: mismatched pairs should be separated

Practical challenge: training often relies on large batches, long schedules, and heavy augmentation.

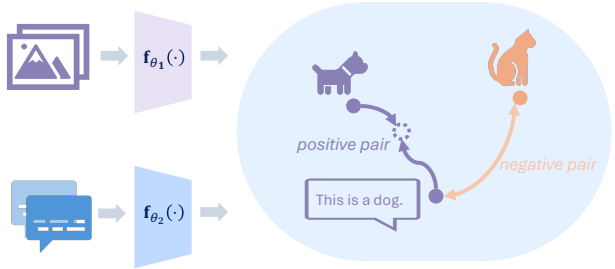
Motivation



Key questions

- Theoretical understanding - What is contrastive alignment really doing?
  Prior theory (mainly in the linear regime) suggests that contrastive objectives implicitly recover a low-rank spectral structure.
- Practical efficiency - Can a better theoretical understanding lead to more efficient multimodal alignment?

Problem Setup



- **Input:** N paired samples$\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^{d_1},\ \mathbf{y}_i \in \mathbb{R}^{d_2}$.
- **Encoders:** learn modality-specific mappings

$$\mathbf{f}_{\theta_1} : \mathbb{R}^{d_1} \to \mathbb{R}^r, \qquad \mathbf{f}_{\theta_2} : \mathbb{R}^{d_2} \to \mathbb{R}^r.$$

- **Shared r-dimensional representation space.**
- **Contrastive objective:** align two modalities in a shared representation space.

Outline

Generalized contrastive loss[2]

Definition (Generalized contrastive loss)

Given N paired samples $\{(x_i, y_i)\}_{i=1}^N$, let $[s_{ij}]$ denote the similarity matrix. With monotonically increasing functions $\phi, \psi: \mathbb{R} \to \mathbb{R}_+$, scaling factor $\nu \geq 1$, and weights $\epsilon_{ij} \in [0, 1]$, the bidirectional generalized contrastive loss is

$$L(\theta_1, \theta_2) = \frac{1}{2N} \sum_{i=1}^{N} \frac{1}{|\mathcal{P}_x(i)|} \sum_{k \in \mathcal{P}_x(i)} \phi\left(\sum_{j \notin \mathcal{P}_x(i)} \epsilon_{ij}\, \psi(s_{ij} - \nu s_{ik}) + \epsilon_{ik}\, \psi(s_{ik} - \nu s_{ik}) \right)$$

$$+ \frac{1}{2N} \sum_{i=1}^{N} \frac{1}{|\mathcal{P}_y(i)|} \sum_{k \in \mathcal{P}_y(i)} \phi\left(\sum_{j \notin \mathcal{P}_y(i)} \epsilon_{ij}\, \psi(s_{ji} - \nu s_{ki}) + \epsilon_{ik}\, \psi(s_{ki} - \nu s_{ki}) \right) + R(\theta_1, \theta_2).$$

- $\mathcal{P}_x(i)$, $\mathcal{P}_y(i)$: positive index sets for the two directions; $|\mathcal{P}_x(i)|$ and $|\mathcal{P}_y(i)|$ are their cardinalities
- $R(\theta_1, \theta_2)$: optional regularizer.

[2] Generalized from Tian, Yuandong. "Understanding deep contrastive learning via coordinate-wise optimization." Advances in Neural Information Processing Systems.

Generalized contrastive loss

$$L(\theta_1,\theta_2)=\frac{1}{2N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_x(i)|}\sum_{k\in\mathcal{P}_x(i)}\phi\left(\sum_{j\notin\mathcal{P}_x(i)}\epsilon_{ij}\,\psi(s_{ij}-\nu s_{ik})+\epsilon_{ik}\,\psi(s_{ik}-\nu s_{ik})\right)$$

$$+\frac{1}{2N}\sum_{i=1}^{N}\frac{1}{|\mathcal{P}_y(i)|}\sum_{k\in\mathcal{P}_y(i)}\phi\left(\sum_{j\notin\mathcal{P}_y(i)}\epsilon_{ij}\,\psi(s_{ji}-\nu s_{ki})+\epsilon_{ik}\,\psi(s_{ki}-\nu s_{ki})\right)+R(\theta_1,\theta_2).$$

Example.

$$\phi(x)=\tau\log x,\quad \psi(x)=e^{x/\tau},\quad |\mathcal{P}_x(i)|=|\mathcal{P}_y(i)|=1,\quad \nu=1,\quad \epsilon_{ij}=1-\delta_{ij},\quad R=0.$$

This gives the CLIP loss

$$\mathcal{L}_{CLIP}=\frac{\tau}{2N}\sum_{i=1}^{N}\log\left(\sum_{j\in[N]}\exp\left(\frac{s_{ij}-s_{ii}}{\tau}\right)\right)+\frac{\tau}{2N}\sum_{i=1}^{N}\log\left(\sum_{j\in[N]}\exp\left(\frac{s_{ji}-s_{ii}}{\tau}\right)\right)$$

$$=\frac{\tau}{2N}\sum_{i=1}^{N}\left[-\log\left(\frac{\exp\left(\frac{s_{ii}}{\tau}\right)}{\sum_{j\in[N]}\exp\left(\frac{s_{ij}}{\tau}\right)}\right)-\log\left(\frac{\exp\left(\frac{s_{ii}}{\tau}\right)}{\sum_{j\in[N]}\exp\left(\frac{s_{ji}}{\tau}\right)}\right)\right]$$

Generalized contrastive loss

$$L(\theta_1, \theta_2) = \frac{1}{2N} \sum_{i=1}^{N} \frac{1}{|\mathcal{P}_x(i)|} \sum_{k \in \mathcal{P}_x(i)} \phi\left(\sum_{j \notin \mathcal{P}_x(i)} \epsilon_{ij}\, \psi(s_{ij} - \nu s_{ik}) + \epsilon_{ik}\, \psi(s_{ik} - \nu s_{ik}) \right)$$

$$+ \frac{1}{2N} \sum_{i=1}^{N} \frac{1}{|\mathcal{P}_y(i)|} \sum_{k \in \mathcal{P}_y(i)} \phi\left(\sum_{j \notin \mathcal{P}_y(i)} \epsilon_{ij}\, \psi(s_{ji} - \nu s_{ki}) + \epsilon_{ik}\, \psi(s_{ki} - \nu s_{ki}) \right) + R(\theta_1, \theta_2).$$

Unified in two aspects:

1. **A unified family of contrastive losses.**
   - **CLIP / InfoNCE** [3, 2]: $\phi(x) = \tau \log x, \quad \psi(x) = e^{x/\tau}$
   - **Triplet loss** [4]: $\phi(x) = x, \quad \psi(x) = [\epsilon - x]_+$
2. **Handle both one-to-one and many-to-many alignment.** The positive index sets $\mathcal{P}_x(i)$ and $\mathcal{P}_y(i)$ define the matched samples in the two directions.

Next?

Gradient equivalance

Lemma (Gradient equivalence)

Consider minimizing the general contrastive loss, the gradient of the contrastive loss with respect to encoder parameters satisfies:

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = -\left.\frac{\partial \operatorname{tr}\left(\mathcal{F}_{\theta_1}(\mathbf{X}) S(\gamma) \mathcal{F}_{\theta_2}^{\top}(\mathbf{Y})\right)}{\partial \theta_k}\right|_{\gamma=\gamma(\theta_1,\theta_2)} + \frac{\partial R\left(\theta_1, \theta_2\right)}{\partial \theta_k}, \quad k \in \{1,2\} \tag{1}$$

where

$$\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_1 \times n}, \mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n] \in \mathbb{R}^{d_2 \times n}$$

$$\mathcal{F}_{\theta_1}(\mathbf{X}) = [\mathbf{f}_{\theta_1}(\mathbf{x}_1) \cdots \ \mathbf{f}_{\theta_1}(\mathbf{x}_n)] \in \mathbb{R}^{r \times n}, \mathcal{F}_{\theta_2}(\mathbf{Y}) = [\mathbf{f}_{\theta_2}(\mathbf{y}_1) \ \cdots \ \mathbf{f}_{\theta_2}(\mathbf{y}_n)] \in \mathbb{R}^{r \times n}.$$

$$\arg\min_{\theta_1,\theta_2} \mathcal{L} \quad \Rightarrow \arg\max_{\theta_1,\theta_2} \operatorname{tr}\left(\mathcal{F}_{\theta_1}(\mathbf{X}) S(\gamma) \mathcal{F}_{\theta_2}^{\top}(\mathbf{Y})\right) - R\left(\theta_1, \theta_2\right) \tag{2}$$

Contrastive similarity weight matrix

Denote $\{\mathbf{e}_i\}_{i=1}^n$ as the elementary basis vectors of $\mathbb{R}^n$. The contrastive similarity weight is then defined as:

$$S(\gamma) = -\frac{1}{n}\sum_{i,j}\frac{1}{2}\left(\frac{\gamma_{ij}}{|\mathcal{P}_x(i)|} + \frac{\bar{\gamma}_{ji}}{|\mathcal{P}_y(j)|}\right)\mathbf{e}_i\mathbf{e}_j^\top, \tag{3}$$

with *weight coefficients*

$$\gamma_{ij} = \begin{cases} \phi'_{ij}\cdot\left(\epsilon_{ij}(1-\nu)\psi'((1-\nu)s_{ij}) - \nu\sum_{m\notin P_x(i)}\epsilon_{im}\psi'(s_{im}-\nu s_{ij})\right), & \text{if } j\in P_x(i) \\ \sum_{k\in P_x(i)}\phi'_{ik}\cdot(\epsilon_{ij}\psi'(s_{ij}-\nu s_{ik})), & \text{if } j\notin P_x(i) \end{cases} \tag{4}$$

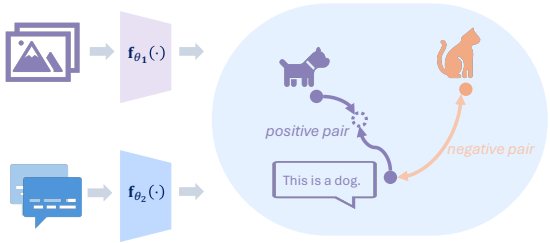
$$\bar{\gamma}_{ij} = \begin{cases} \bar{\phi}'_{ij}\cdot\left(\epsilon_{ji}(1-\nu)\psi'((1-\nu)s_{ji}) - \nu\sum_{m\notin P_y(i)}\epsilon_{mi}\psi'(s_{mi}-\nu s_{ji})\right), & \text{if } j\in P_y(i) \\ \sum_{k\in P_y(i)}\bar{\phi}'_{ik}\cdot(\epsilon_{ji}\psi'(s_{ji}-\nu s_{ki})), & \text{if } j\notin P_y(i) \end{cases} \tag{5}$$

where

$$\begin{aligned} \phi'_{ij} =& \phi'(\epsilon_{ij}\psi((1-\nu)s_{ij}) + \sum_{m\notin P_x(i)}\epsilon_{im}\psi(s_{im}-\nu s_{ij})), \\ \bar{\phi}'_{ij} =& \phi'(\epsilon_{ji}\psi(1-\nu)s_{ji} + \sum_{m\notin P_y(i)}\epsilon_{mi}\psi(s_{mi}-\nu s_{ji})) \end{aligned} \tag{6}$$

Outline

Linear representation setting



- Consider the linear encoder setting

$$\mathbf{f}_{\theta_1}(\mathbf{x}) = F_1\mathbf{x}, \qquad \mathbf{f}_{\theta_2}(\mathbf{y}) = F_2\mathbf{y},$$

where

$$F_1 \in \mathbb{R}^{r\times d_1}, \qquad F_2 \in \mathbb{R}^{r\times d_2}$$

are learnable projection matrices.

- Core idea: the generalized contrastive objective admits a spectral interpretation.

Linear representation setting: spectral solution

$$\arg\min_{F_1,F_2} \mathcal{L} \quad \Longrightarrow \quad \arg\max_{F_1,F_2} \operatorname{tr}\left(F_1\, \mathbf{X}S(\gamma)\mathbf{Y}^\top\, F_2^\top\right) - R(F_1,F_2).$$

Weighted contrastive covariance

$$C(\gamma) = \mathbf{X}S(\gamma)\mathbf{Y}^\top = -\frac{1}{n}\sum_{i,j}\frac{1}{2}\left(\frac{\gamma_{ij}}{|\mathcal{P}_x(i)|} + \frac{\bar{\gamma}_{ji}}{|\mathcal{P}_y(j)|}\right)\mathbf{x}_i\mathbf{y}_j^\top \tag{7}$$

$$C(\gamma) = U\Sigma V^\top, \quad R(F_1,F_2) = \frac{\rho}{2}||F_1^T F_2||_F^2 \tag{8}$$

$$\Rightarrow F_1^\top F_2 = \frac{1}{\rho}\sum_{j=1}^{r}\sigma_j U_j V_j^\top \tag{9}$$

$$\Rightarrow F_1 = \frac{1}{\sqrt{\rho}}\Sigma_r^{\frac{1}{2}}U_r^T, \quad F_2 = \frac{1}{\sqrt{\rho}}\Sigma_r^{\frac{1}{2}}V_r^T \tag{10}$$

Linear representation setting: spectral update

$$F_1 = \frac{1}{\sqrt{\rho}} \Sigma_r^{\frac{1}{2}} U_r^T, \quad F_2 = \frac{1}{\sqrt{\rho}} \Sigma_r^{\frac{1}{2}} V_r^T \tag{11}$$

Takeaway: under the linear setting, the convergence of the contrastive loss can be replaced by a closed-form spectral update.

Why leave the linear world?

- Cross–modal relations (e.g., vision $\leftrightarrow$ language) are typically nonlinear in general.
- With frozen or partially frozen pretrained encoders, the residual alignment is rarely captured by mere linear heads.

Outline

Nonlinear representation setting: kernel generalization

Core Idea: Lift the nonlinear representation to a high-dimensional feature space where the problem becomes operator-linear.

Let $(\mathcal{H}_X, k_X)$ and $(\mathcal{H}_Y, k_Y)$ be RKHSs with canonical feature maps

$$\phi_X(\mathbf{x}) = k_X(\cdot, \mathbf{x}) \in \mathcal{H}_X, \qquad \phi_Y(\mathbf{y}) = k_Y(\cdot, \mathbf{y}) \in \mathcal{H}_Y, \tag{12}$$

satisfying the reproducing property $f(\mathbf{x}) = \langle f, \phi_X(\mathbf{x}) \rangle_{\mathcal{H}_X}$ for all $f \in \mathcal{H}_X$ (and analogously for $\mathcal{H}_Y$).

RKHS is a Hilbert space of functions in which evaluating a function at a point is itself a continuous linear operation.

For r-dimensional outputs, the a-th coordinate $(a = 1, \ldots, r)$ admits the representer form

$$f_{\theta_1}^{(a)}(\cdot) = \sum_{i=1}^{n} A_{ia}\, k_X(\mathbf{x}_i, \cdot), \qquad f_{\theta_2}^{(a)}(\cdot) = \sum_{j=1}^{n} B_{ja}\, k_Y(\mathbf{y}_j, \cdot), \tag{13}$$

with $A, B \in \mathbb{R}^{n \times r}$.

Nonlinear representation setting: kernel generalization

Let $K_X = [k_X(\mathbf{x}_i, \mathbf{x}_j)]$ and $K_Y = [k_Y(\mathbf{y}_i, \mathbf{y}_j)]$.

The batch embeddings are

$$\mathcal{F}_{\theta_1}(\mathbf{X}) = A^\top K_X \in \mathbb{R}^{r \times n}, \qquad \mathcal{F}_{\theta_2}(\mathbf{Y}) = B^\top K_Y \in \mathbb{R}^{r \times n}. \tag{14}$$

The contrastive trace term becomes

$$\operatorname{tr}\big(\mathcal{F}_{\theta_1}(\mathbf{X})\, S(\gamma)\, \mathcal{F}_{\theta_2}(\mathbf{Y})^\top\big) = \operatorname{tr}\big(A^\top K_X\, S(\gamma)\, K_Y B\big). \tag{15}$$

Kernelized spectral characterization

With the RKHS parameterization, the trace objective becomes

$$\operatorname{tr}\big(\mathcal{F}_{\theta_1}(\mathbf{X})\, S(\gamma)\, \mathcal{F}_{\theta_2}(\mathbf{Y})^\top\big) = \operatorname{tr}\big(A^\top K_X\, S(\gamma)\, K_Y B\big).$$

Theorem (Kernelized spectral characterization)

Let

$$R(A,B) = \frac{\rho}{2}\big\| (K_X^{1/2}A)^\top (K_Y^{1/2}B) \big\|_F^2, \qquad \rho > 0.$$

Then minimizing the contrastive loss is equivalent to

$$\max_{A,B\in\mathbb{R}^{N\times r}} \operatorname{tr}\big(A^\top K_X\, S(\gamma)\, K_Y B\big) - \frac{\rho}{2}\big\| (K_X^{1/2}A)^\top (K_Y^{1/2}B) \big\|_F^2.$$

Let $M := K_X^{1/2}\, S(\gamma)\, K_Y^{1/2}$ *and let* M_r *be its best rank-*r *approximation. Then the maximizers satisfy*

$$AB^\top = \frac{1}{\rho}\, K_X^{-1/2} M_r K_Y^{-1/2}.$$

Core idea

Contrastive loss minimization $\iff$ **best rank-r approximation with RKHS**

Standard view

Iterative optimization

- optimize encoder parameters by SGD

$\Longrightarrow$

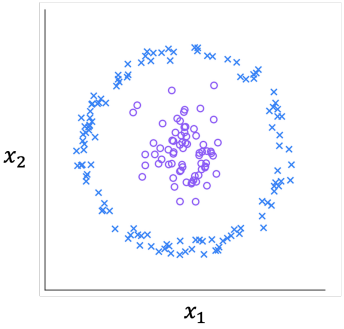
UniCon (Ours)

Spectral characterization

- Directly identifying the dominant singular subspace where the converged representation lies in.

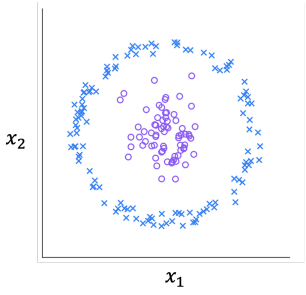
Takeaway: nonlinear contrastive alignment can be understood as r-rank structure discovery in a kernel feature space.

Intuition

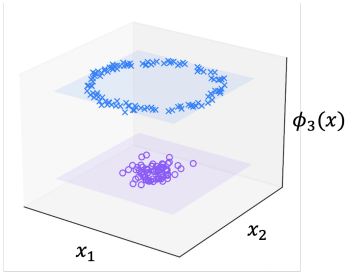


(a) $\mathbb{R}^2$

Intuition

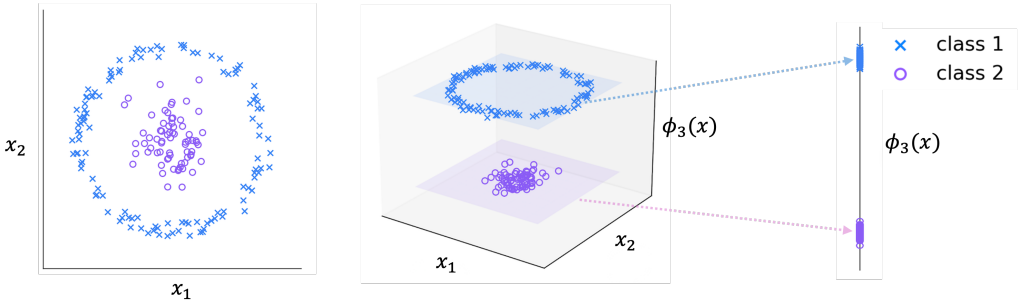


(a) $\mathbb{R}^2$



(b) $\mathbb{R}^3$

Intuition



(a) $\mathbb{R}^2$ (b) $\mathbb{R}^3$ (c) $\mathbb{R}^1$

Outline

Outline

Outline

Image–Text Retrieval on Flickr30k and MSCOCO

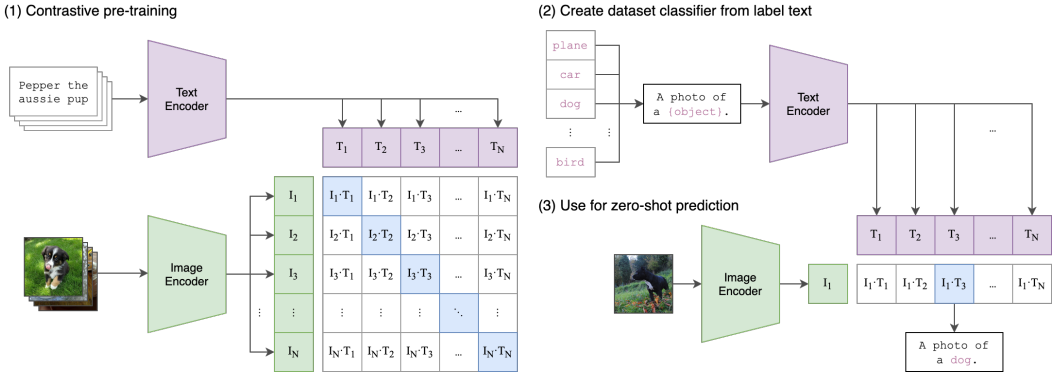


Figure 1: CLIP model.[3]

[3] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

Image–Text Retrieval on Flickr30k and MSCOCO

Table 1: Recall@1 and Recall@10 for both image→text and text→image directions, together with average retrieval performance.

Dataset	Backbone	Method	Train time	Image→Text		Text→Image		Average	
				R@1	R@10	R@1	R@10	R@1	R@10
Flickr30k	RN-18 + SBERT	SGD–CLIP	45.6 s	.043	.221	.041	.217	.042	.219
		UniCon	**1.7 s**	.020	.145	.087	.361	.054	.253
	RN-50 + SBERT	SGD–CLIP	45.0 s	.043	.221	.041	.217	.042	.219
		UniCon	**0.81 s**	.134	.464	.188	.567	.161	.515
	CLIP ViT-B/32	SGD–CLIP	45.3 s	.231	.595	.241	.600	.236	.597
		UniCon	**0.76 s**	.284	.636	.421	.777	.353	.701
MSCOCO	RN-50 + SBERT	SGD–CLIP	5121 s	.053	.253	.060	.286	.057	.270
		UniCon	**11.1 s**	.105	.388	.129	.439	.117	.414
	CLIP ViT-B/32	SGD–CLIP	1066 s	.128	.415	.123	.427	.126	.421
		UniCon	**11.2 s**	.329	.685	.292	.644	.311	.665

Takeaway: UniCon achieves comparable or better retrieval with **orders-of-magnitude faster training**.

Zero-shot transfer to Flickr30k

- **Training set:** all models are trained on MSCOCO.
- **Evaluation:** zero-shot transfer to FLICKR30K, with no fine-tuning.
- **Goal:** test whether the learned alignment transfers across datasets.

Backbone	Method	Flickr30k R@5	R@10
RN50 + SBERT	SGD–CLIP	.060	.286
	UniCon	.249	.353
CLIP ViT-B/32	SGD–CLIP	.123	.427
	UniCon	.766	.848

Takeaway: UniCon produces transferable representations, even without fine-tuning.

Conclusion

Unified framework: The analysis is general for the family of contrastive loss, encoders (linear & nonlinear) and multiple-positive settings.

Theoretical insights: Contrastive learning optimization is equivalent to r-rank structure discovery with RKHS.

Empirical efficiency: Computation efficiency (up to 461x speedup in multimodal alignment training) and data efficiency.

Takeaway: Contrastive learning can be viewed as **discovering r-rank structure in high-dimensional feature spaces.**

Discussion and Future Directions

- **Kernel Selection**
  - Developing **learnable kernels** to adaptively capture complex data manifolds.
- **Scalability for Massive Data**
  - Exploring mini-batch aggregation strategies to balance analytical precision with global scalability.
- **Cross-modal + intra-modal learning**
  - Current focus: cross-modal alignment
  - Next step: integrate intra-modal self-supervision into a unified framework
- **Toward joint representation and generation**
  - Contrastive learning: strong for discriminative tasks
  - Diffusion / flow matching: strong for distribution modeling
  - Goal: a unified model with both representation and generation ability

References

[1] R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang.
Understanding multimodal contrastive learning and incorporating unpaired data.
In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380. PMLR, 2023.

[2] A. v. d. Oord, Y. Li, and O. Vinyals.
Representation learning with contrastive predictive coding.
arXiv preprint arXiv:1807.03748, 2018.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.
Learning transferable visual models from natural language supervision.
In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[4] F. Schroff, D. Kalenichenko, and J. Philbin.
Facenet: A unified embedding for face recognition and clustering.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.