Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment

Mayug Maniparambil*† Raiymbek Akshulakov*‡ Yasser Abdelaziz Dahou Djilali §†
Sanath Narayan § Ankit Singh § Noel E. O'Connor †

Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 2025

*joint first authors

†ML Labs, Dublin City University ‡University of California Berkeley §Technological Innovation Institute

1. 研究背景
2. 核心思路
3. 原理基础
4. 具体实现
5. 测试结果
6. 研究结论

研究背景

以CLIP为代表的使用对比学习驱动的视觉语言模型能力强，但训练成本过高，训练复杂度高，学术界和工业界迫切找到一种能够实现CLIP一样的强多模态能力，但训练成本低的模型构建方法

表1：训练时长对比

模型 (MODEL)	训练天数 (TRAINING DAYS)
RN50×64	18 天
ViT-L/14	12 天

表2：GPU资源对比

模型 (MODEL)	GPU 数量 (GPU COUNT)	GPU 型号 (GPU TYPE)
RN50×64	592	V100
ViT-L/14	256	V100

非本篇论文内容，仅用于展示
数据来源于CLIP原始论文中公布数据

研究背景

前序工作

本篇论文是该团队2024年发表在CVPR上的文章的后续工作

Do Vision and Language Encoders Represent the World Similarly?

Mayug Maniparambil[*1] Raiymbek Akshulakov[*2] Yasser Abdelaziz Dahou Djilali[3,1]
Mohamed El Amine Seddik[3] Sanath Narayan[3] Karttikeya Mangalam[2] Noel E. O'Connor[1]

Abstract

Aligned text-image encoders such as CLIP have become the de-facto model for vision-language tasks. Furthermore, modality-specific encoders achieve impressive performances in their respective domains. This raises a central question: does an alignment exist between uni-modal vision and language encoders since they fundamentally represent the same physical world? Analyzing the latent spaces structure of vision and language models on image-caption benchmarks using the Centered Kernel Alignment (CKA), we find that the representation spaces of unaligned and aligned encoders are semantically similar. In the absence of statistical similarity in aligned encoders like CLIP, we show that a possible matching of unaligned encoders exists without any training. We frame this as a seeded graph-matching problem exploiting the semantic similarity between graphs and propose two methods - a Fast Quadratic Assignment Problem optimization, and a novel localized CKA metric-based matching/retrieval. We demonstrate the effectiveness of this on several downstream tasks including cross-lingual, cross-domain caption matching and image classification. Code available at github.com/mayug/0-shot-llm-vision.

1. Introduction

The recent success of deep learning on vision-language tasks mainly relies on jointly trained language and image encoders following the success of CLIP and ALIGN [20, 40]. The standard procedure for training these models aims at aligning text and image representation using a contrastive loss that maximizes the similarity between image-text pairs while pushing negative captions away [10, 19, 36]. This achieves a statistical similarity across the two latent spaces, which is key to retrieving the closest cross-modal representations using cosine similarity. This property is not valid for unaligned encoders, hence, extra transformations are needed to bridge the gap. These transformations can be

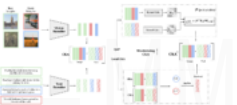


Figure 1. For matching, we calculate the kernels for image and text embeddings and employ QAP-based seeded matching to maximize CKA for obtaining the optimal permutation P. For retrieval, we append query embeddings to base embeddings and retrieve the best caption that maximizes the local CKA for a query image.

training a mapping network that captures the prior distribution over the text and image representations [31, 34, 35]. The work of [31] has shown that it is possible to train a linear mapping from the output embeddings of vision encoders to the input embeddings of language models and exhibit impressive performance on image captioning and VQA tasks. This indicates that the representations between the unaligned uni-modal vision and language encoders are sufficiently high level and differ only by a linear transformation. However, this linear layer is trained on CC-3M [9] consisting of three million image-caption pairs.

Is this training step necessary? In an ideal scenario, we anticipate an alignment between vision and language encoders as they inherently capture representations of the same physical world. To this end, we employ Centered Kernel Alignment (CKA) [12, 22, 42], which is known for measuring representation similarity both within and between networks. As shown in Figure 2, we measure the CKA between a variety of unaligned vision and language encoders [8, 16, 28, 37, 47], on the image-caption pairs of the COCO [27] dataset and observe that some have comparable scores to that of aligned encoders like CLIP [40], affirmative of semantic similarities.

We then ask the question: If the unaligned image and text encoders are semantically similar, is there a way to connect them in a *zero-shot manner?* Do they build a similar representation graph over the same information coming from

[*]joint first authors
[1]ML Labs, Dublin City University
[2]University of California Berkeley
[3]Technological Innovation Institute

研究背景

前序工作

- 表征相似性：学术界的研究显示训练好的视觉和语义编码器具有语义相似性，这为将其连接起来奠定了理论基础
- 自动数据策展：使用训练好的CLIP文本编码器在数据集中筛选出高质量的图文配对数据，为准备训练数据提供可能
- 多模态预训练：受CLIP启发，后续工作不断改进，学术界内有已经有尝试冻结单模态的编码器，通过连接方式实现多模态能力的工作，但其都受限于随机的编码器对的选择导致效果不佳

研究背景

核心支撑理论

- 以CLIP为代表的对比多模态视觉语言模型，展现了强大的开放世界语义理解能力，验证了用对比学习驱动多模态学习是有效的
- 最近的研究显示训练好的单模态编码器之间具有很高的语义相似性，为连接单模态编码器提供了理论基础

研究背景

现有研究的问题

- 现有的单视觉和语言编码器能力远强于CLIP中的编码器能力
- 现有的对齐策略需要对视觉和语言编码器完全重新训练，开销大
- 改进现有CLIP模型需要大量数据和计算资源，任一无法满足都无法获取媲美CLIP的能力

核心思路

主要任务

如何将训练好的单模态编码器连接在一起?

核心步骤

找出语义相似度高的视觉和语言编码器

↓

构建适应训练投影器的数据集

↓

训练投影器

原理基础

找出语义相似度高的视觉和语言编码器 → 如何找出语义相似度高编码器

↓ ↓

构建适应训练投影器的数据集 → 合适数据集的对于模型最终效果的影响

↓

训练投影器

基础原理介绍与消融实验

原理基础

1.如何找出语义相似度高编码器

引入衡量标准

中心核对齐CKA
- Algorithms for learning kernels based on centered alignment 2012年
- Similarity of Neural Network Representations Revisited 2019年

线性核 $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$

高斯核 $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j), \quad L_{ij} = l(\mathbf{y}_i, \mathbf{y}_j)$$

$$\mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(N-1)^2} \mathrm{tr}\,(\mathbf{KCLC}) \longrightarrow \mathrm{CKA}(K, L) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K) \cdot \mathrm{HSIC}(L, L)}}$$

HSIC代表希尔伯特 - 施密特独立性判别准则，
用于衡量向量集之间的依赖关系，其中 $\mathbf{C} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$

原理基础

1.如何找出语义相似度高编码器

验证标准有效性

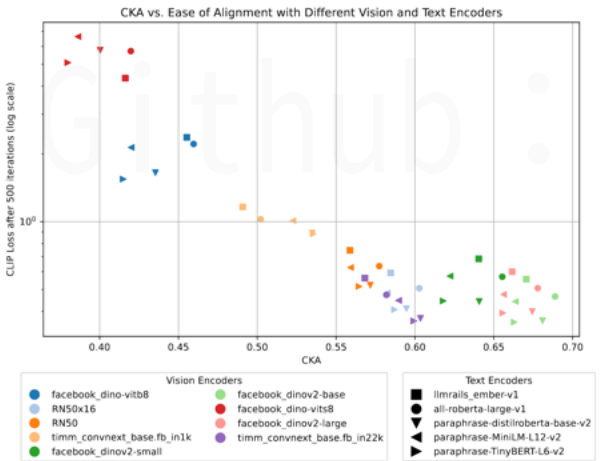
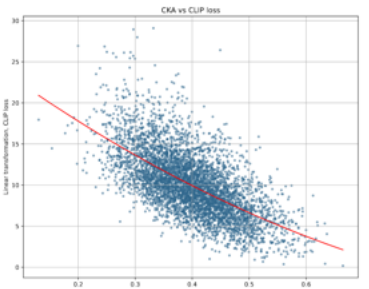
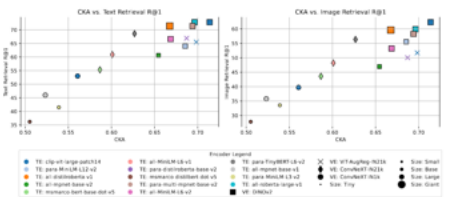


图1 不同编码器上的实验

实验验证：通过固定轮数的训练来比较CKA与CLIP损失之间的关系，可以见得CKA值与CLIP损失之间呈现负相关，可以得出CKA能够指示出编码器的结构相似性



图A.2 使用人造随机数据的实验

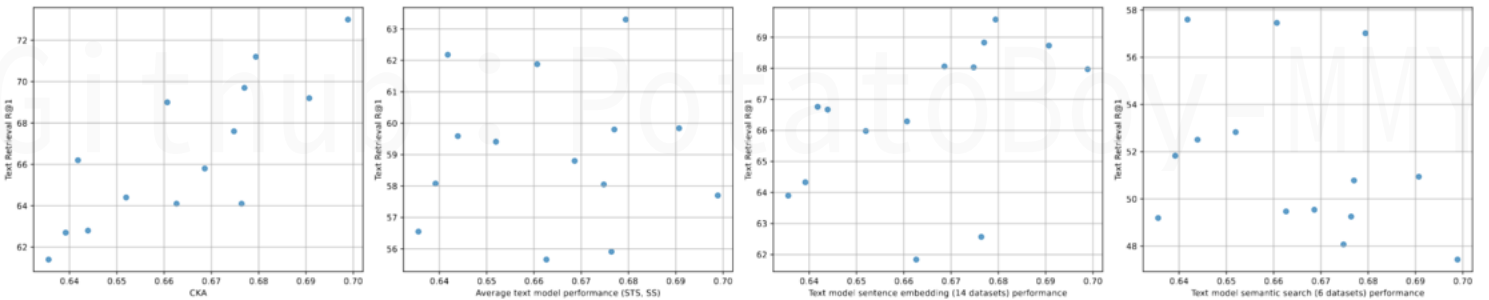


图A.5 CKA在不同编码对上的实验

原理基础

1.如何找出语义相似度高编码器

验证标准优势性



图A.6 相较于其他标准，CKA与检索性能的相关性最大，故选择CKA

原理基础

2.合适数据集的对于模型最终效果的影响

投影器的训练需要一个合适的数据集，需要满足三个条件

数据量较**小**：投影器的参数量较小，故不需要CLIP训练数据集那么大的数据集

概念覆盖**全**：数据集要尽可能覆盖足够多的概念，保证能够尽可能的覆盖单模态嵌入空间的各个区域

图文对齐**质量高**：以确保能够高质量的对齐两个编码器

论文主要通过消融实验来证明数据集构建对于训练的有效性，由于解释此部分与数据集的具体构建有关，此部分放于具体实现部分展开解释

具体实现

找出语义相似度高的视觉和语言编码器

↓

构建适应训练投影器的数据集

↓

训练投影器

具体实现

1.找出语义相似度高的视觉和语言编码器 ?

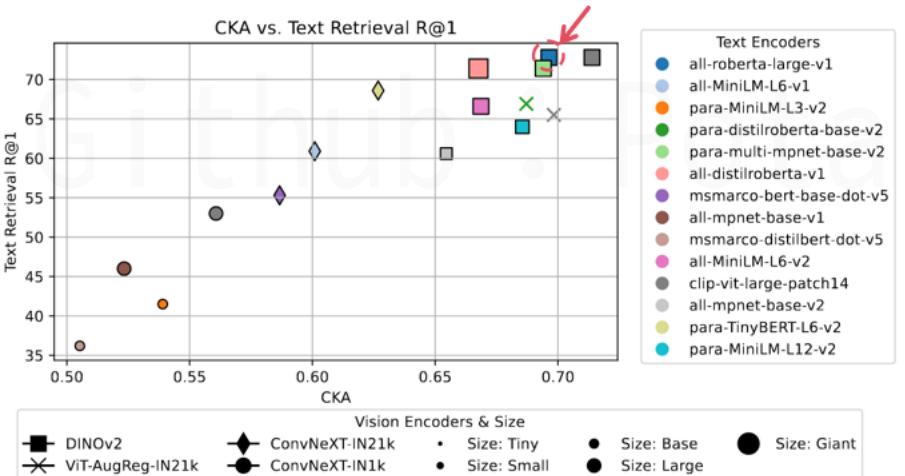


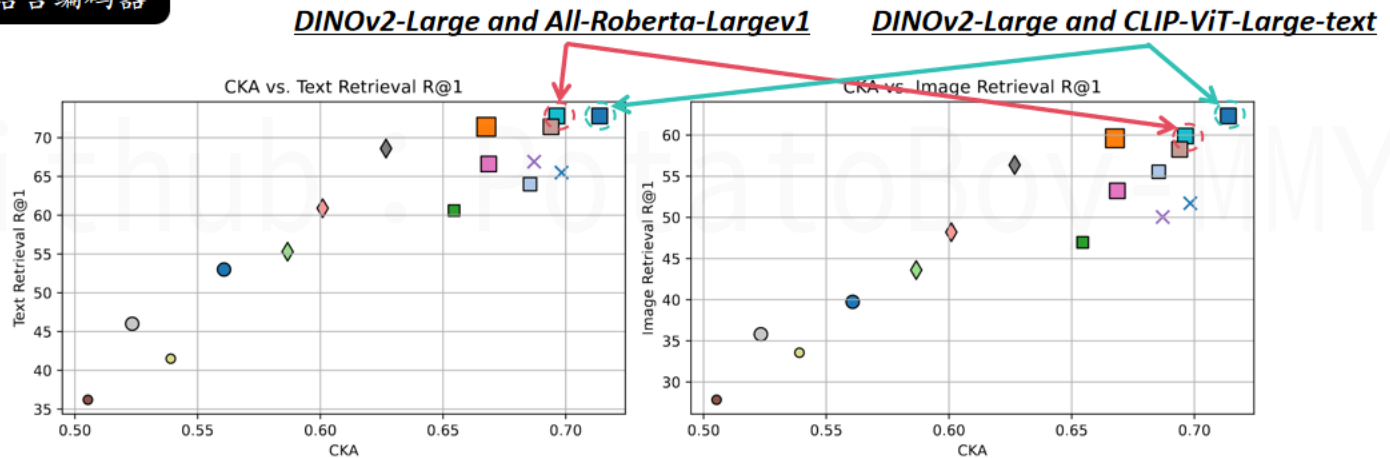
图4

前面的研究已经证明了CKA作为对齐编码器的标准的有效性和优势性，从图4中的检索精度和CKA的关系也可以看出CKA可以直接指示编码器的对齐，基于对左图数据的分析，最论文选择了DINOv2-Large and All-Roberta-Largev1

具体实现

1.找出语义相似度高的视觉和语言编码器

?

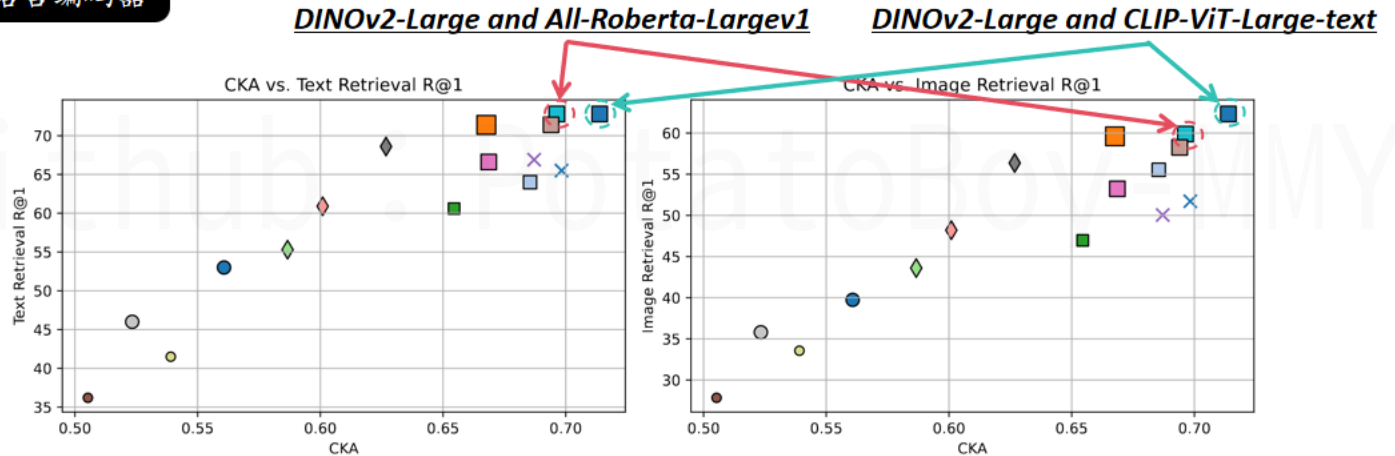


从图A.5.上看无论是这两个检索任务的精度还是CKA上看，均是后者更好，为何选择前者，论文并未做出解释

具体实现

1.找出语义相似度高的视觉和语言编码器

?



我的猜测：1.用一个本身不具备对齐特性的文本编码器来证明有效性
2.前者的纯文本能力强，在后续各项任务的实验中效果更好

具体实现

2.构建适应训练投影器的数据集

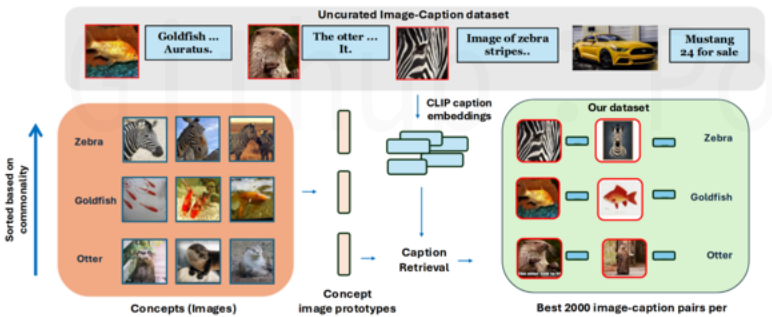


图2 数据策展的概览图

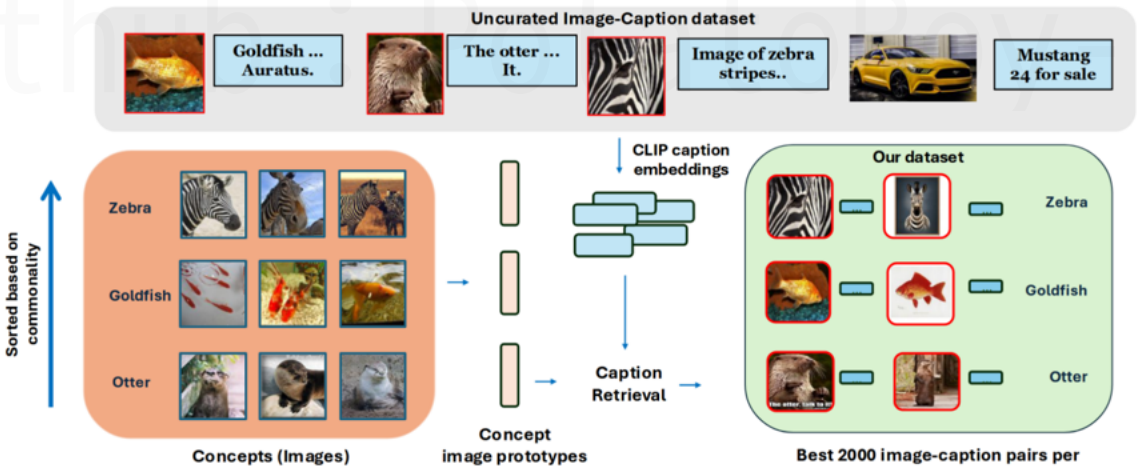
投影器中的参数仅有11M，故所需的数据量较小，但由于是要训练一个用于对齐两个编码器的投影器，所以这个数据集要满足两个要求：

- 数据集要尽可能覆盖足够多的概念，保证能够尽可能的覆盖单模态嵌入空间的各个区域
- 保证图片和文本的高度对齐，以便于能够有助于学习视觉和嵌入空间之间的映射

具体实现

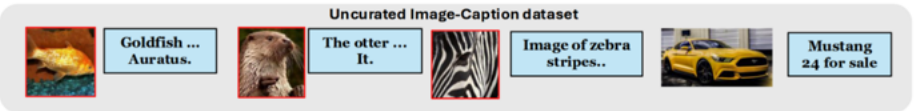
2.构建适应训练投影器的数据集

“谷歌检索式构建”

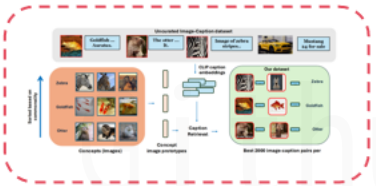


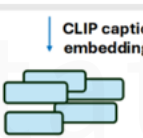
具体实现

2.构建适应训练投影器的数据集



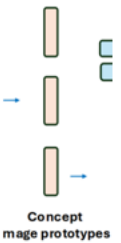
LAION400M数据集中的图片文本对





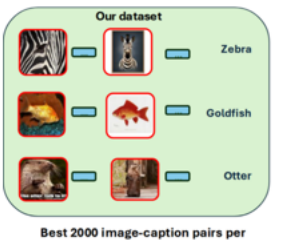
使用CLIP-VIT-Large的文本编码器抽出文本向量

精选出的概念，配少量的图片



使用CLIP-VIT-Large的图像编码器抽出图片概念

文本和概念匹配检索



构建出的数据集

具体实现

2.构建适应训练投影器的数据集

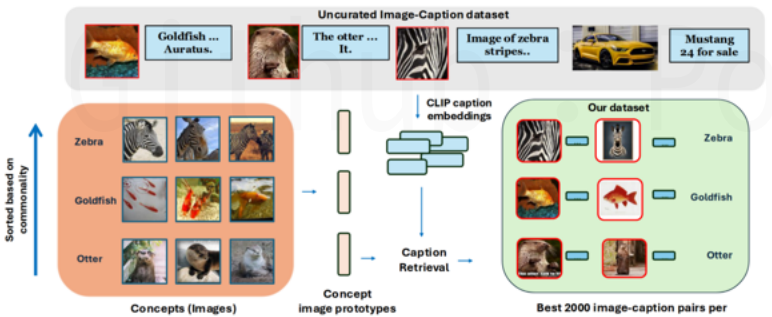


图2 数据策展的概览图

LAION-CLASS-Collected

此数据集的构建由左侧概念图流程显示的一样构建，

① 从多个数据集中精选出3000个左右的概念图片

② 然后使用CLIP中的视觉编码器构建出概念向量

③ 再对 LAION-400M 中所有的 caption 用 CLIP 文本编码器编码，与概念原型做余弦相似度检索

④ 然后保证冷门概念先挑选从而保证样本数量足够

根据以上流程，从而建立一个3000概念，每个概念2000样本，大小为6M的精选数据集

具体实现

2.构建适应训练投影器的数据集

***LAION-CLASS-Collected*的缺陷**

该数据集脱胎于LAION数据集，而LAION数据集本身的文本和图像对齐质量以及图像本身的质量不高，所以用这个训练投影器会出现一些问题

加入一些高质量图像和对齐质量高的数据

MIX-CLASS-Collected

该数据集将***LAION-CLASS-Collected***和***CC3M***, ***CC12M***和***SBU***这些具有比LAION更高质量的图像和更好的图像 - 标题对齐的数据集，从而提高模型的检索能力，且能够保证概念的基本覆盖

具体实现

构建适应训练投影器的数据集

为了满足基本要求的数据集

LAION-CLASS-Collected

仅使用 ***LAION-CLASS-Collected***

Ima-geNet零样本域迁移准确率较高为76.12%，但检索率较低，为52.70%和42.48%

消融实验

Data Source	N	ImageNet	I2T	T2I
LAION-CLASS-Collected	6M	76.12	52.70	42.48
CC3M, CC12M, SBU	14M	54.17	85.30	72.44
Both	20M	75.04	81.32	71.38
Both longer training	20M	**76.30**	**87.54**	**74.17**

MIX-CLASS-Collected

仅使用 ***CC3M***, ***CC12M***和 ***SBU***

真正用于训练的数据集

Ima-geNet零样本域迁移准确率较低为54.17%，但检索率较高，为85.30%和72.44%

具体实现

2.构建适应训练投影器的数据集

Data Source	N	ImageNet	I2T	T2I
LAION-CLASS-Collected	6M	76.12	52.70	42.48
CC3M, CC12M, SBU	14M	54.17	85.30	72.44
Both	20M	75.04	81.32	71.38
Both longer training	20M	**76.30**	**87.54**	**74.17**

结合***LAION-CLASS-Collected***和***CC3M***, ***CC12M***, ***SBU***这些数据集，可以使得ImageNet零样本域迁移准确率较高的同时使得检索任务准确率较高，也验证了前者为投影器训练提供较好的概念覆盖，后者为训练提供了检索能力的提高

具体实现

3.训练投影器

投影器通过变化
将语义空间对齐

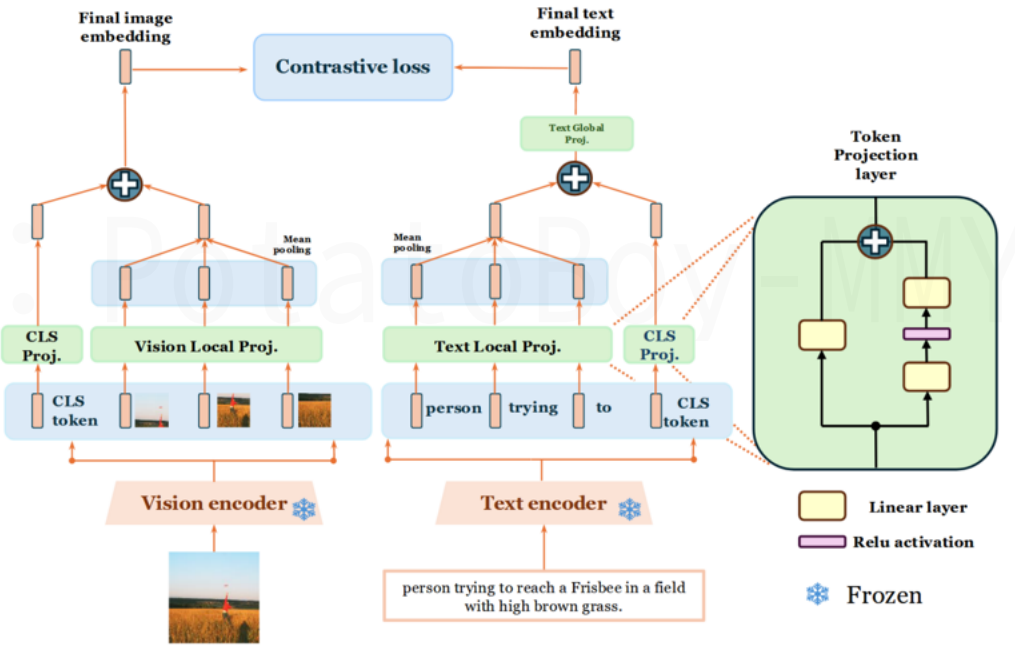


图3 投影器架构

具体实现

3.训练投影器

投影器效果

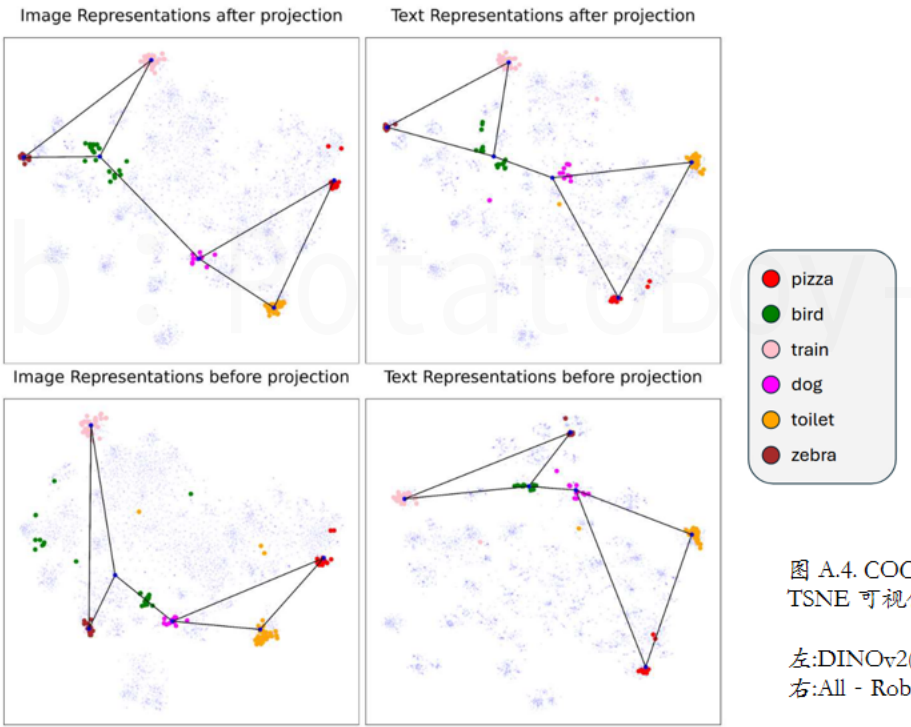
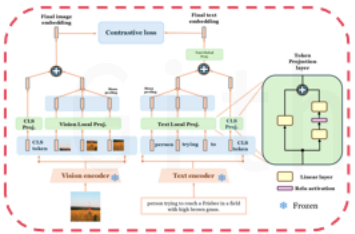


图 A.4. COCO 检测类别的编码器输出 TSNE 可视化。

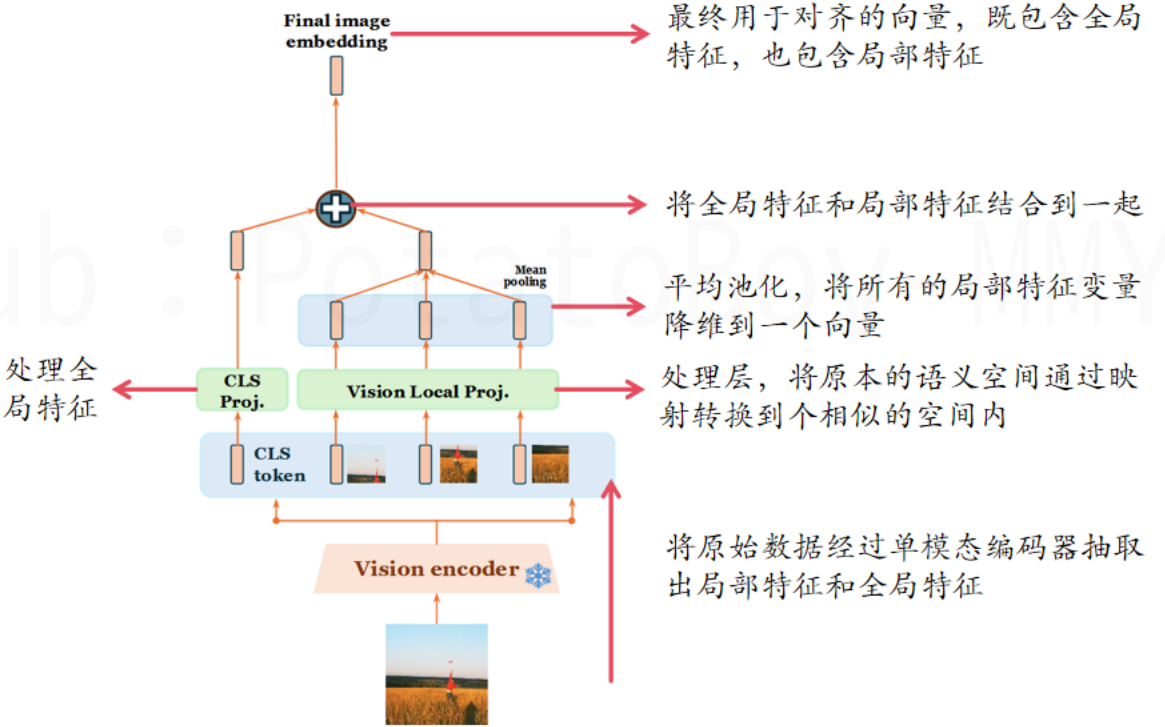
左:DINOv2(视觉)
右:All - Roberta - Large - v1(文本)。

具体实现

3.训练投影器

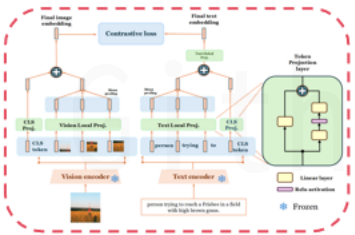


视觉一侧

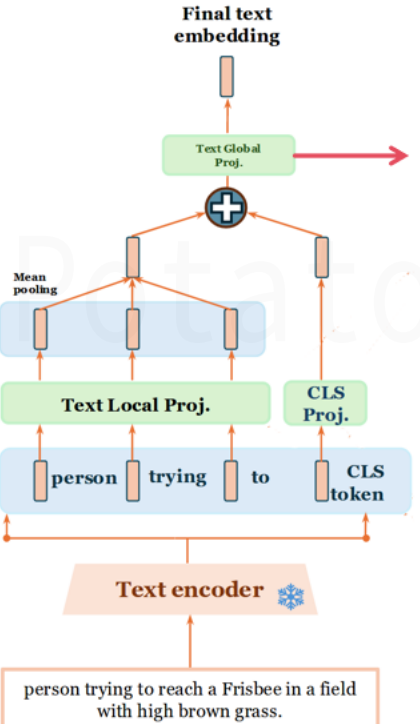


具体实现

3.训练投影器



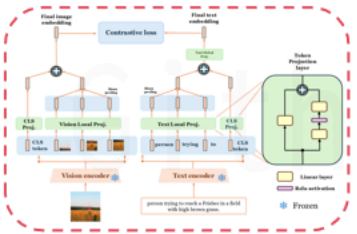
文本一侧

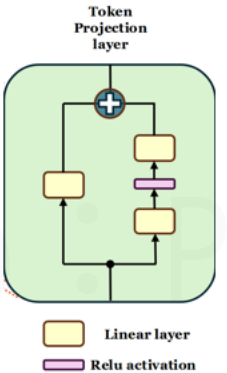


相较于图像一侧，多出了一个两层的MLP，将文本一侧进一步变化从而更好对齐，其有效性在消融实验中证明

具体实现

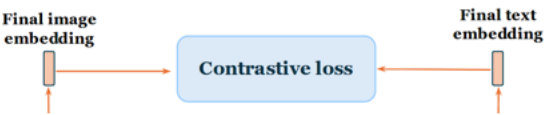
3.训练投影器





Token投影模块，将特征变换到相同空间，同时使用残差结构，既保证了原本信息的不丢失，又引入激活函数增加非线性

将最终处理完的嵌入空间整合到一起，计算对比损失



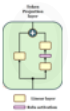
具体实现

3.训练投影器

V Proj. Local	V Proj. CLS	T Proj. Local	T Proj. Global	INet 0-shot
mlp	identity	identity	identity	68.81
token	identity	identity	identity	68.84
token	identity	identity	mlp	70.90
token	identity	patch	identity	71.85
token	identity	token	mlp	72.15
identity	token	token	mlp	75.53
token	token	token	mlp	**76.12**

表1 投影器结构的消融

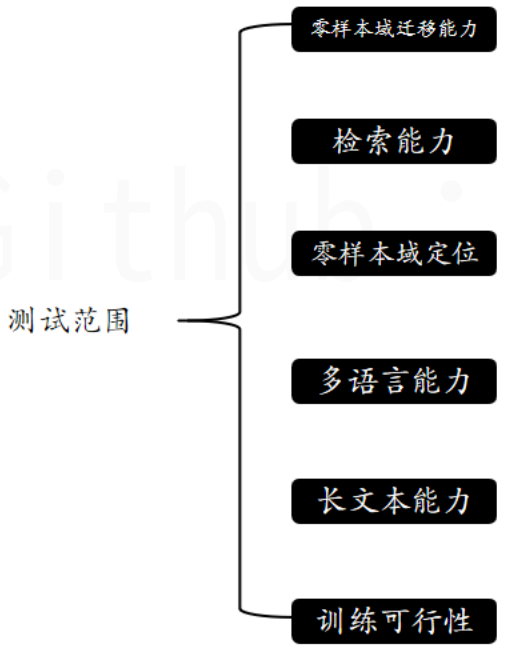
mlp 两层mlp

token 

identity 跳过

patch 文本直接共享图像那一侧参数

测试结果



主要显示了该模型的能力强，成本低，灵活性高

测试结果

零样本域迁移能力

Model	N	ImageNet	ImageNetv2	Caltech	Pets	Cars	Flowers	Food	Aircrafts	SUN	CUB	UCF101
LAION-CLIP VIT-L	400M	72.7	65.4	92.5	91.5	**89.6**	73.0	90.0	24.6	70.9	**71.4**	71.6
OpenAI-CLIP VIT-L	400M	75.3	**69.8**	92.6	**93.5**	77.3	**78.7**	**92.9**	**36.1**	67.7	61.4	**75.0**
LiT L16L	112M	75.7	66.6	89.1	83.3	24.3	76.3	81.1	15.2	62.5	58.7	60.0
DINOv2-MpNet (Ours)	20M	74.8	68.0	91.8	91.7	71.0	75.8	87.5	23.0	71.9	63.2	71.0
DINOv2-ARL(Ours)	20M	**76.3**	69.2	**92.8**	92.1	73.9	78.4	89.1	28.1	**72.6**	66.1	73.2

表3

- 在ImageNet上的结果显示了虽然论文提出的模型使用了更少的数据量训练，但是效果并不输于400M
- 在有些数据集上，论文中模型的效果部分弱于CLIP等模型，主要是由于数据集的概念主要选自ImageNet，有些概念没对齐

测试结果

检索能力

Model	Flickr		COCO	
	I2T	T2I	I2T	T2I
LAION-CLIP VIT-L	**87.6**	70.2	59.7	43.0
OpenAI-CLIP VIT-L	85.2	64.9	56.3	36.5
LiT L16L	73.0	53.4	48.5	31.2
DINOv2-MpNet (Ours)	84.6	71.2	58.0	42.6
DINOv2-ARL (Ours)	87.5	**74.1**	**60.1**	**45.1**

表4

在各项检索能力上，优于或持平于CLIP模型，论文将其归结于模型所使用的单模态编码器的能力强于CLIP的编码器

测试结果

零样本域定位

Model	Pascal VOC	Pascal Context
OpenAI-CLIP-VIT-L*	23.46	14.25
SPARC	27.36	21.65
DINOv2-ARL	**31.37**	**24.61**

表5 数值为平均IOU

*代表这里的CLIP使用了MaskCLIP方法使得CLIP分割定位能力更高

论文模型的能力强于针对分割能力增强的CLIP和针对定位能力优化过的SPARC，且论文中的投影器并没有针对定位能力单独优化过，仅仅是依赖于DINOv2的强大定位能力在此结构中被继承了下来，后续若在投影仪中针对定位能力单独优化，有更大的潜力

测试结果

多语言能力

model	classification						retrieval					
	EN	DE	FR	JP	RU	average	EN	DE	FR	JP	RU	average
nllb-clip-base@v1	25.4	23.3	23.9	21.7	23.0	23.5	47.2	43.3	45.0	37.9	40.6	42.8
M-CLIP/XLM-Roberta-Large-Vit-B-32	46.2	43.3	43.3	31.6	38.8	40.6	48.5	46.9	46.1	35.0	43.2	43.9
M-CLIP/XLM-Roberta-Large-Vit-L-14	54.7	51.9	51.6	37.2	47.4	48.6	56.3	52.2	51.8	41.5	48.4	50.0
xlm-roberta-base-ViT-B-32@laion5b	63.0	55.8	53.8	37.3	40.3	50.0	63.2	54.5	55.7	47.1	50.3	54.2
nllb-clip-large@v1	39.1	36.2	36.0	32.0	33.9	35.4	59.9	56.5	56.0	**49.3**	50.4	54.4
M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	48.0	46.1	45.4	32.9	40.3	42.5	63.2	**61.4**	59.3	48.3	**54.8**	57.4
ViT-L-14@laion400m	72.3	48.2	49.9	2.7	4.5	35.5	64.5	26.7	38.3	1.4	1.7	26.5
openai/clip-vit-large-patch14	**75.6**	46.7	49.6	6.6	3.5	36.4	59.4	19.9	28.5	4.1	1.3	22.6
DINOv2-MpNet (Ours)	73.4	**61.6**	**58.3**	**43.2**	**49.3**	**57.1**	**70.7**	60.6	**60.6**	45.6	52.7	**58.0**

使用了多语言数据的模型

仅使用英语数据

表6

在这个测试之中，文本编码器使用了具有多语言能力的 **paraphrase - multilingual - MpNetv2**(简称 MpNet)，且只使用英语数据来对齐，验证了使用投影器对齐后，编码器的多语言能力得到了保留，也证明了投影器架构的灵活性

测试结果

长文本能力

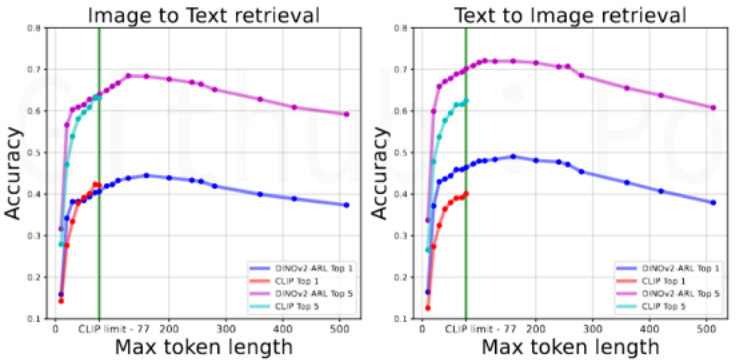


图5

得益于文本编码器自带的长文本能力，模型的长文本能力强于CLIP

测试结果

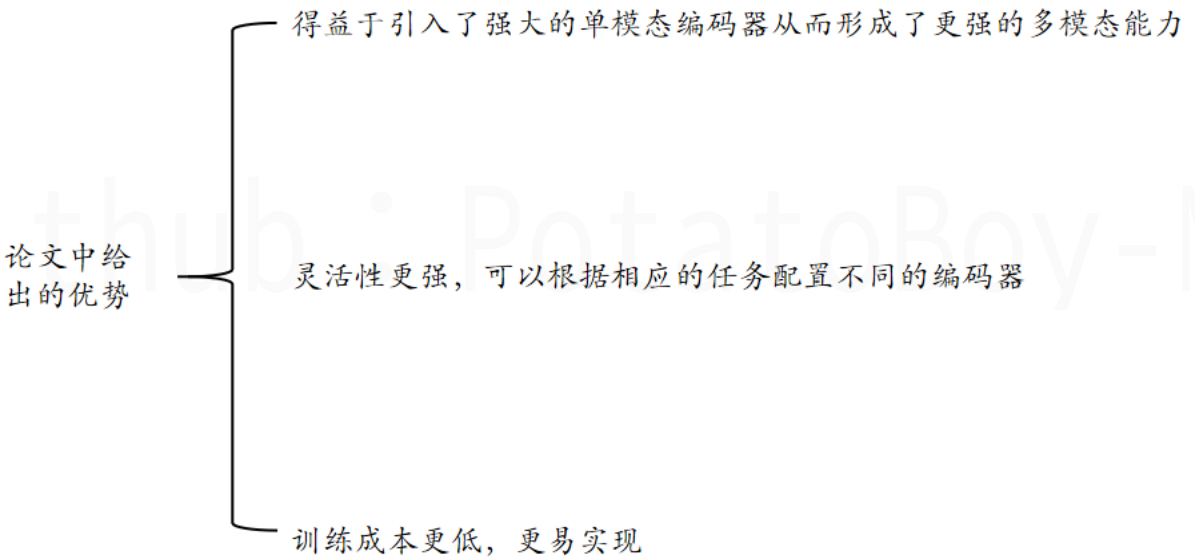
训练可行性

Model	Data	SS	Trainable / Total	Compute	IN 0-shot
OpenAI CLIP	400M	12.8B	427M / 427M	21,845	72.7%
LAION400M CLIP	400M	12.8B	427M / 427M	25,400	75.3%
DINOv2-ARL	20M	0.6B	11.5M / 670M	400	76.3%

数据需求更小　训练样本更少　参数量更小　计算量小　能力更强

训练成本更小

研究结论



研究结论

个人看法

- 不可否认，该结构的灵活性和低成本性确实有优势，但是低成本性有被作者刻意夸大的成分，因为强调的低成本一直没有把训练单编码器的成本计算进去；而灵活性会导致在切换各种编码器的时候一种任务能力强了，另一种任务能力差了，犹如跷跷板

- 该结构只能接受图像和文本输入，多模态能力尚且欠佳，但是若只是一味的加入其他模态的编码器又可能会使对齐工作变得十分困难

谢谢！

恳请批评指正