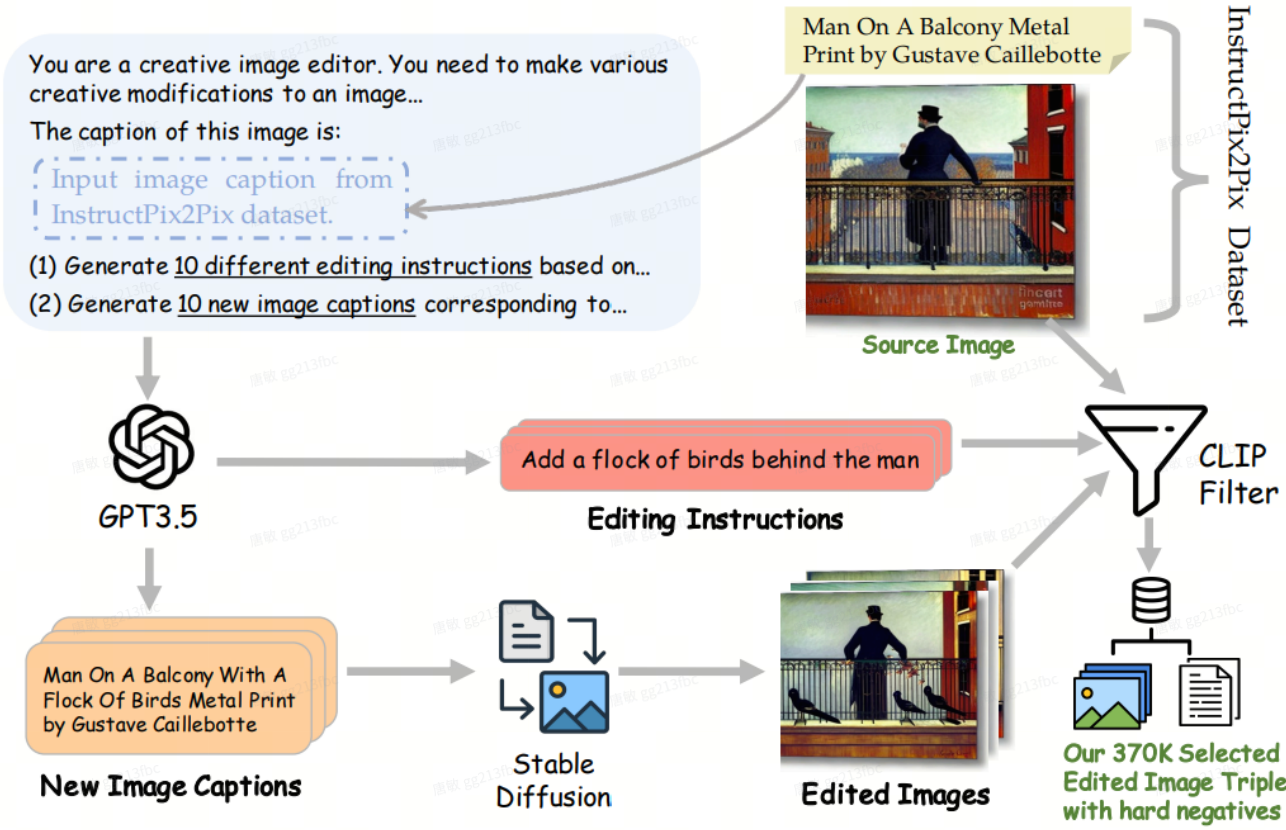
文献

1.VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval

期刊：acl 2024

论文链接：https://arxiv.org/abs/2406.04292



本章提出了一个新的跨模态预训练模型，名为FILIP，是一个**双流模型**，具有**基于Transformer的图像和文本编码器**。对于视觉模态，图像编码器采用**Vision Transformer**作为输入。对于文本模态，遵循Radford等人（2021年）的方法，使用小写的字节对编码（BPE）对文本进行分词，词汇量为49,408。在词嵌入层之后，将标记嵌入送入修改后的仅解码器Transformer模型模型。在图像和文本编码器之上，文本标记和视觉标记的表示被线性映射到多模态公共空间，并分别进行L2归一化。引入了一种新颖的细粒度对比学习目标，配备了跨模态后期交互，它考虑了图像块与文本标记之间的细粒度交互。

图像编码器：在 FILIPbase 中为 ViT-B/32，在 FILIPlarge 中为 ViT-L/14

model沿用了CLIP的model

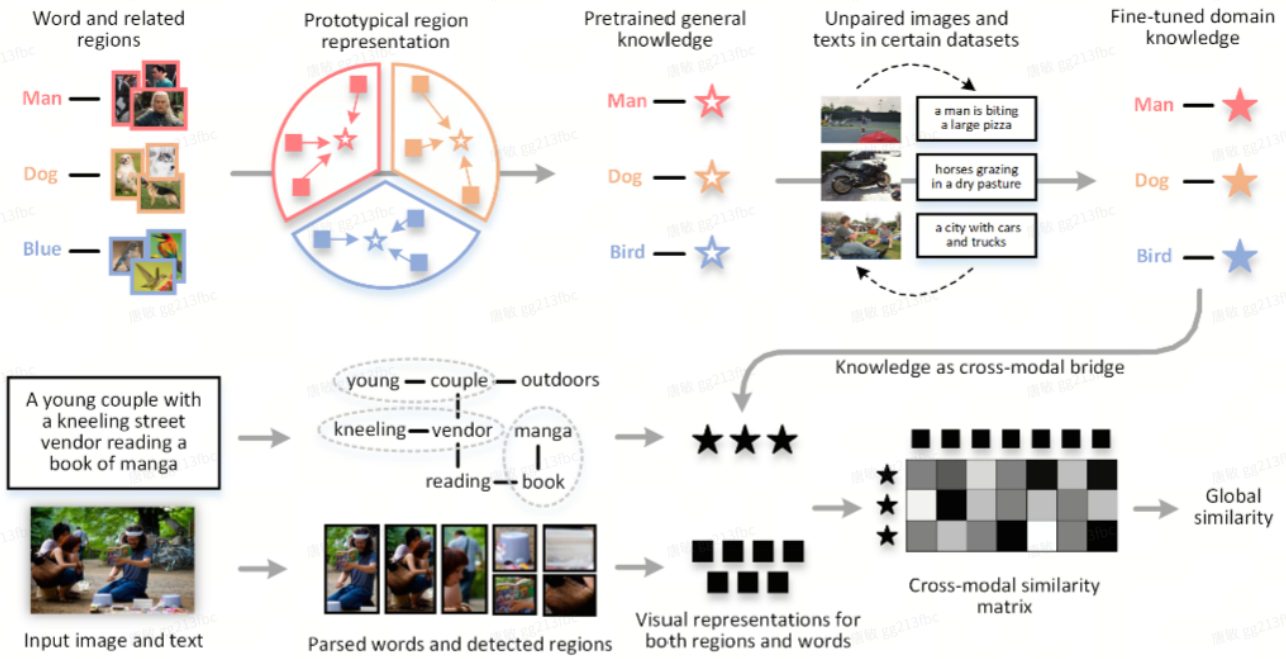
A Benchmark for CompositionalText-to-image Retrieval

期刊: icml

2. MACK: Multimodal Aligned Conceptual Knowledgefor Unpaired Image-text Matching

NeurIPS-2022

链接:

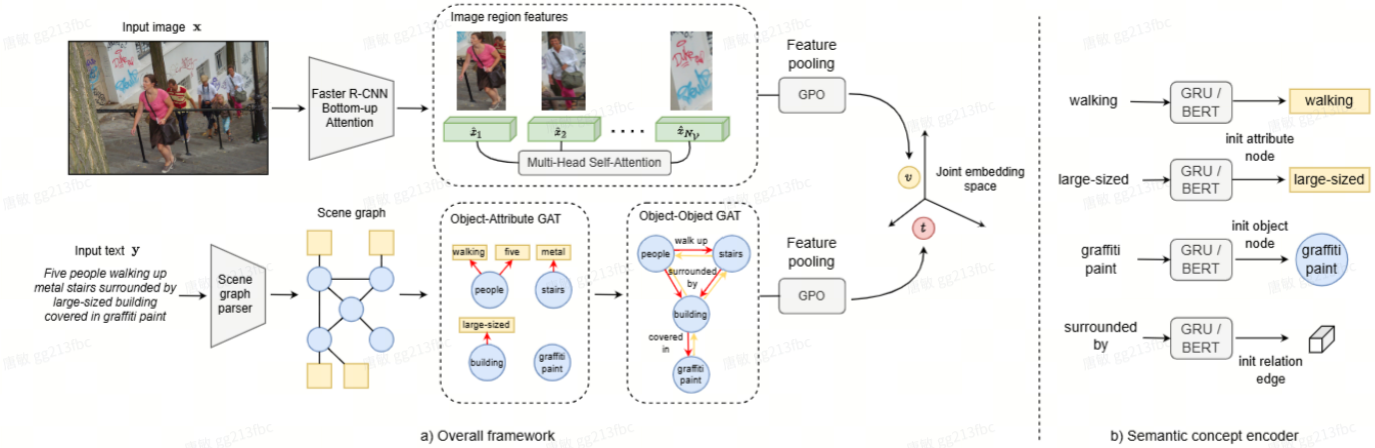


这项工作试图在新场景的背景下研究图像-文本匹配，它的灵感来源于：人类的大脑能够很好地将任意的图像与文本相关联，而无需从如此大规模的成对图像和文本中学习。相反，它存储了关于物体、动作、属性等的语义知识，这些知识是多模态对齐的，可以用来关联视觉和语言信息。受到以上启发，这项工作试图通过模拟类似人脑的知识来处理未配对的图像-文本匹配。论文提出了一种新方法，为了去除图像中与单词无关的内容，该方法专注于语义概念，1. 从公开可用的数据集中收集单词及其语义相关的图像区域对。然后，2. 它计算原型区域（通过平均所有相关区域的表示来获得）表示并将它们与单词对齐，以减轻外观变异的影响。基于对齐的概念知识（即单词-原型区域对），3. MACK能够在同一特征空间中桥接图像和文本，以测量它们的跨模态相似性。为了使预计算的一般知识更好地适应某些数据集，4. 论文进一步根据区域级循环一致性原则对其进行微调，这不需要配对的图像和文本。由于提出的MACK既简单又有效，它可以很好地与现有的图像-文本匹配模型结合，作为一种重排方法，进一步提高它们的性能。

3. Composing Object Relations and Attributes for Image-Text Matching

来源：CVPR 2024

链接：
https://openaccess.thecvf.com/content/CVPR2024/html/Pham_Composing_Object_Relations_and_Attributes_for_Image-Text_Matching_CVPR_2024_paper.html



在这项工作中，文章提出了一种不同于序列模型的方法，将标题表示为一个由对象和属性节点组成的场景图，这些节点通过关系边连接。场景图展示了对象-属性和对象-对象对这样的语义结构已经是组织好的。为此，提出了CORA，一种用于图像-文本匹配的双编码器模型。

在图像方面：重用了GPO ，它是图像-文本匹配的最新 pooling 操作符，将图像嵌入为一个向量。

在文本方面：使用图注意力网络，具有强大的关系归纳偏置，为标题生成整体场景图嵌入。

损失函数：对比损失，以指导CORA在整体图像-标题级别和局部图像-对象实体级别进行对齐。

特点：基于标题的场景图表示来开发双编码器模型中的文本编码器。论文模型显式地学习如何通过对象之间的关系来组合对象及其属性，以及场景中的所有对象，以生成一个富含语义信息的文本嵌入向量。

4.How to Make Cross Encoder a Good Teacherfor Efficient Image-Text Retrieval?

来源：CVPR 2024

链接：
https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_How_to_Make_Cross_Encoder_a_Good_Teacher_for_Efficient_CVPR_2024_paper.pdf

文章提出了一种新颖的对比部分排序蒸馏（Contrastive Partial Ranking Distillation, CPRD）方法，以实现有效的排序知识蒸馏。具体来说，通过对比学习学习难负样本的排序：给定一张图像，首先使用双编码器识别前K个难负文本？并获取这些文本的排序。然后，将图像和负文本输入到交叉编码器中计算匹配分数，并将负文本分为有效和无效负文本。有效负文本与图像的匹配分数更高，它们的相对顺序包含了丰富的跨模态匹配知识。因此，文章使用对比学习将排名较高的有效负文本拉向图像，同时将排名较低的文本推开，确保双编码器中有效负文本的排序与交叉编码器中的排序一致性。另一方面，无效负文本与图像的匹配分数较低，它们的相对顺序不包含有效信息。因此，只对所有的无效负文本使用对比学习推开，而不考虑它们之间的相对顺序。这种方法不要求双编码器和交叉编码器的相似度分布相似，克服了由于相似度分布之间巨大差异导致的蒸馏困难。此外，通过对比学习实现了难负样本排序学习的目标，与双编码器的训练过程无缝对接。

本工作的贡献可以概括如下：

· 对从交叉编码器到双编码器有效知识蒸馏进行了全面研究，并确定了三个关键方面。

· 提出了对比部分排序蒸馏（CPRD）方法，该方法通过对比学习实现了学习有效难负样本之间相对顺序的目标，使从交叉编码器到双编码器的知识传递有效。

5.Object-Aware Query Perturbation for Cross-Modal Image-Text Retrieval

来源：ECCV 2024

链接：https://arxiv.org/abs/2407.12346v2

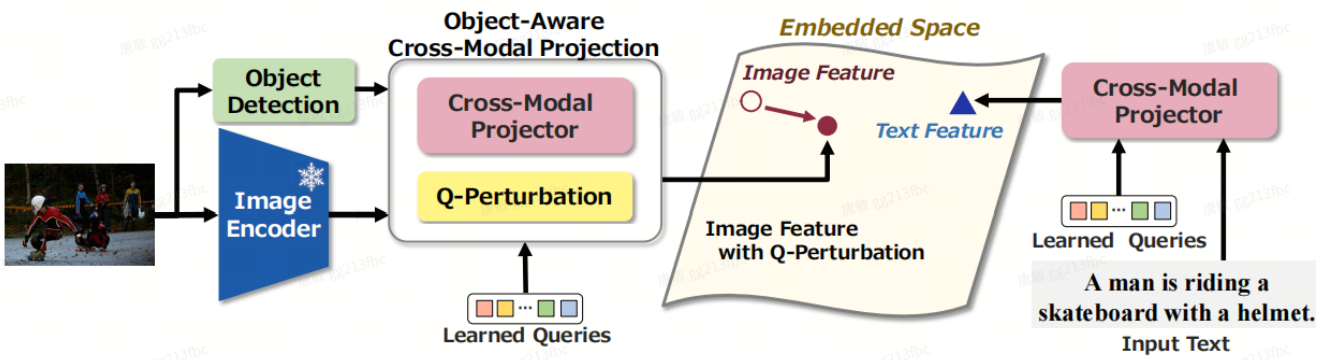


Fig. 3: Overview of the proposed framework. The proposed framework constructs an object-aware cross-modal projector by incorporating localization cues from object detection into the existing cross-modal projector.

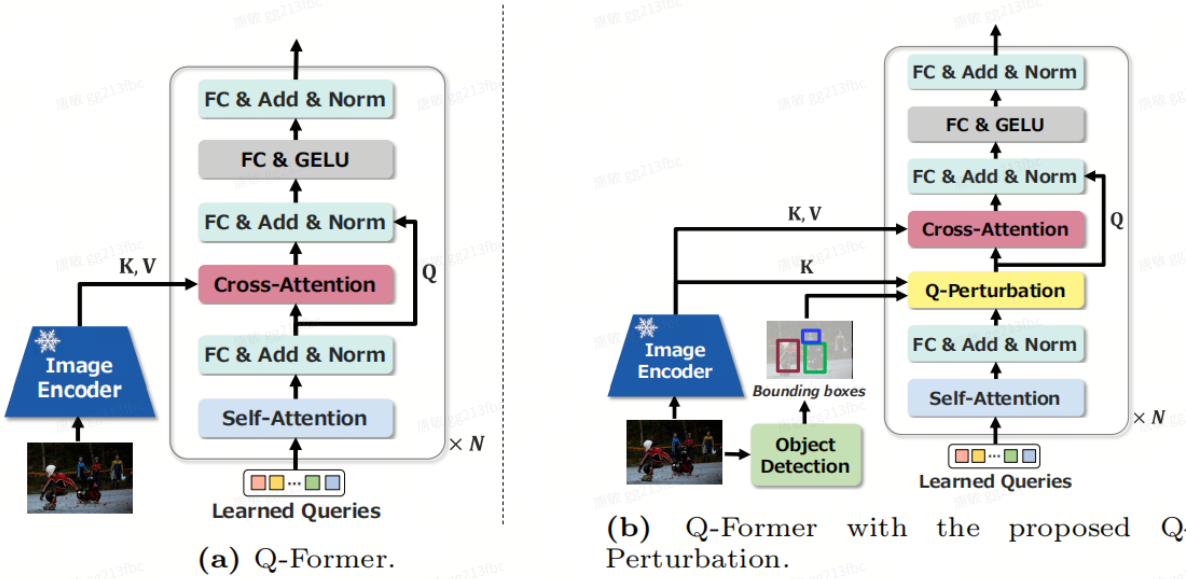


Fig. 4: Overview of our object-aware mechanism with Q-Perturbation module. Our framework can incorporate the object localization cues from the bounding boxes into a cross-modal projection module, e.g., Q-Former, with minimal modification.

这篇论文提出了一种面向对象的查询扰动方法。查询扰动（Query-Perturbation，简称Q-Perturbation）通过关注图像中感兴趣的对象信息，即使对象相对较小，也能增强视觉和语言模型的对象感知能力。能够对捕捉到小对象的图像进行准确检索。Q-Perturbation的核心机制是在视觉和语言模型中的交叉注意力模块，通过增强与对象区域相对应的关键字来提升查询效果。这个方法适用于各种V&L模型，且由于该方法无需训练，易于实施，可以避免因数据更新而增加的计算成本以及因重新训练而导致的灾难性遗忘。

旨在扩展现有的V&L模型，同时继承这些模型的高表现力，来提高包含小对象的图像的检索性能。

用Q-Perturbation来扰动已经获得的查询，以突出对象区域特征，即利用对象定位（边界框）来实现。通过在跨注意力模块之前插入所提出的Q-Perturbation模块，以最少的修改实现一个面向对象的跨模态投影模块。接下来，首先描述针对单个对象情况的Q-Perturbation模块，然后将其扩展到多个对象

6.High-Order Semantic Alignment for Unsupervised Fine-Grained Image-Text Retrieval

来源：ACL Anthology 2024

链接：https://aclanthology.org/2024.lrec-main.714.pdf

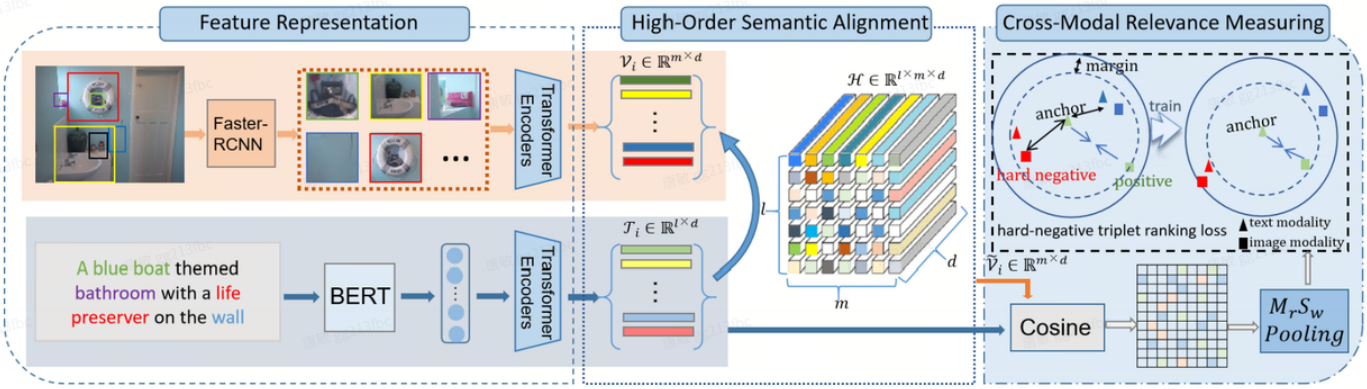


Figure 1: The framework of the **HOSA** model. An image with m regions and a text with l words are encoded in a $d-$dimensional common space via a stack of transformer layers, and $\mathcal{H} \in \mathbb{R}^{l \times m \times d}$ is the reconstruction coefficient tensor, characterizing the high-order semantic alignment across modalities.

新颖的高阶语义对齐（HOSA）模型，用于细粒度图像-文本检索。图1展示了所提出的框架，它包含三个主要模块：具有嵌入模态特定片段的特征表示，通过探索全局和局部对应关系以及局部-全局交互实现的高阶语义对齐，以及通过聚合局部相似度的跨模态相关性度量。

特征表示：通过自底向上和自顶向下的注意力机制，Faster-RCNN来选择和提取图像的显著区域的特征，用预训练的BERT模型来处理句子文本

语义对齐：高阶语义对齐（High-Order Semantic Alignment，HOSA）模块，给定一个视觉特征集以及一个文本特征集 ，引入了一个映射函数 来对齐视觉和文本表示。假设图像中的每个区域都可以由对应句子的单词线性表示。得益于t-积操作，所获得的映射函数H能够表征在多个实例中隐藏的片段之间的局部和全局结构。

7.MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training

来源：CVPR 2024

链接：

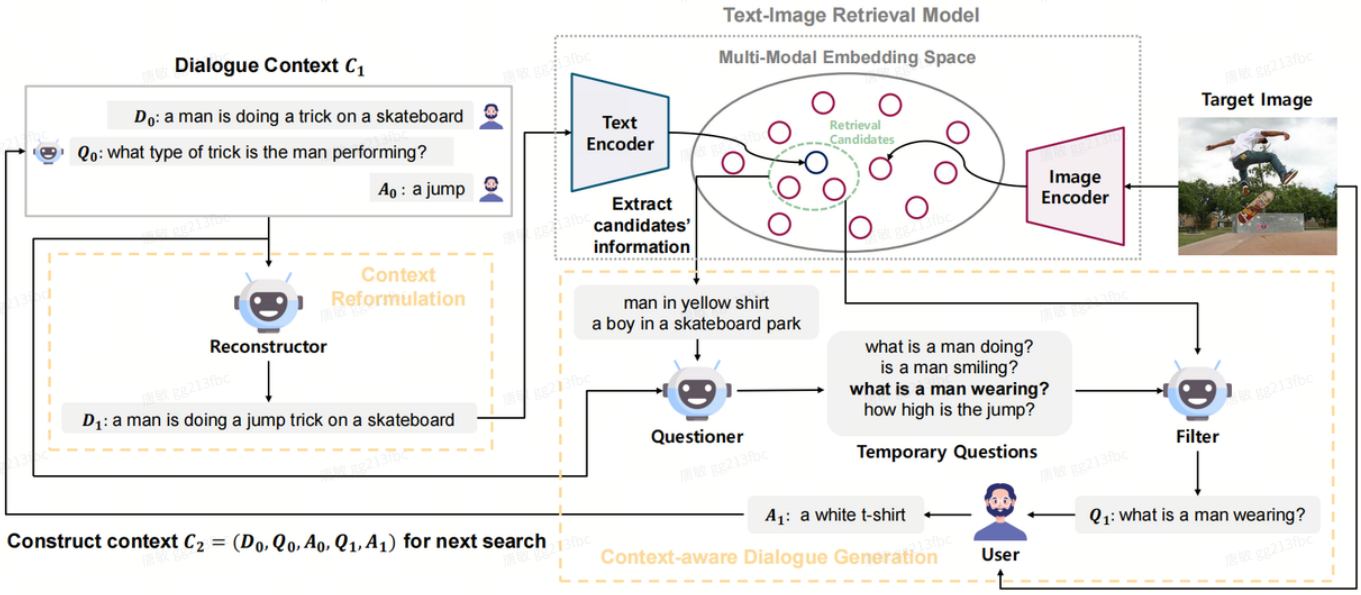


Figure 1: The main framework of the plug-and-play interactive text-to-image retrieval system.

一种针对运行时性能优化的高效图像文本模型新系列，以及一种新颖高效的训练方法，即多模态强化训练。所提出的训练方法利用了来自图像描述模型和强CLIP编码器集合的知识迁移，通过在强化数据集中存储额外知识，避免了训练时的计算开销。

一种基于数据集强化方法的新型训练方法：i)用额外的信息强化数据集一次，ii)在实验中使用强化过的数据集多次；多模态数据集强化变体：通过添加来自一组预训练的强CLIP模型的合成标题和嵌入，强化了图像-文本DataComp 数据集，获得了DataCompDR。

- 设计了一种新的移动设备友好的CLIP模型系列，MobileCLIP。MobileCLIP的变种使用了混合的CNN-Transformer架构，并在图像和文本编码器中应用了结构重参数化技术，以减小模型大小和延迟。
- 引入了多模态强化训练，这是一种新颖的训练策略，它结合了从预训练的图像描述模型和一组强大的CLIP模型中迁移知识，以提高学习效率。
- 引入了我们的强化数据集的两个变种：DataCompDR-12M和DataCompDR-1B。使用DataCompDR，展示了与DataComp相比，学习效率提高了10倍到1000倍。

8. Fine-Grained Image-Text Alignment in Medical ImagingEnables Explainable Cyclic Image-Report Generation

来源：ACL 2024

细粒度视觉语言模型（VLM）已被广泛用于预定义固定图块与文本单词之间的跨模态局部对齐。然而，在医学分析中，病变表现出不同的尺寸和位置，使用固定图块可能导致病变表示不完整。此外，这些方法通过使用热图来显示可能与文本相关的一般图像区域，而不是特定区域，使得它们的解释不

够明确和具体。为了解决这些问题，我们提出了一种新颖的自适应图块-单词匹配（AdaMatch）模型，用于将胸片（CXR）图像区域与医学报告中的单词相关联，并将其应用于CXR报告生成，为生成过程提供解释性。AdaMatch利用自适应图块与单词之间的细粒度关系，为特定图像区域提供相应单词的解释。为了捕捉不同尺寸和位置的异常区域，我们引入了一个自适应图块提取（AdaPatch）模块，以自适应地获取这些区域的自适应图块。为了为CXR报告生成任务提供明确的解释性，我们提出了一个基于AdaMatch的循环CXR报告生成双向语言模型（AdaMatch-Cyclic）。它使用AdaMatch获取CXR图像的关键词和医学报告中的“关键图块”作为提示，指导CXR报告的生成。

9.RS5M and GeoRSCLIP: A Large Scale Vision-Language Dataset and A Large Vision-Language Model for Remote Sensing