Related Work

全局对齐方法专注于直接学习整个图像和句子之间的跨模态相似性，通过将它们投影到一个共同的潜在空间（Zhang et al., 2018; Long et al., 2016）或利用视觉-语义嵌入（Faghri et al., 2018; Chen et al., 2021; Zheng et al., 2020; Radford et al., 2021）。因此，它们往往无法深入挖掘视觉对象和文本术语之间复杂的关系。因此，当面对涉及多个对象和更复杂描述的自然场景时，它们的性能可能不符合预期。

投影到一个共同的潜在空间:

1. bottom-up and top-downattention for image captioning and vqa　cvpr Zhang et al., 2018

视觉-语义嵌入

2. Learning the Best Pooling Strategy for Visual Semantic Embedding cvpr Chen et al., 2021
3. Vse++: Improving visual-semantic embeddings with hard negatives. Cvpr Faghri et al., 2018

局部对齐方法旨在探索图像区域与句子单词之间的局部相关性，以实现更精确的跨模态对齐。Karpathy和Fei-Fei（2015）开创了将多模态RNN检测到的局部图像区域与句子中的单词进行对齐的方法。随后，Lee等人（2018）利用堆叠交叉注意力来对齐显著区域和关键词，强调了区域-单词对齐的有效性，并激发了后续的研究。FPAN（Wang等人，2019）被提出，以强调每个区域内不同位置的重要性。CAMP（Wang等人，2020）创新地引入了跨模态信息传递中自适应调节信息流的机制。尽管这些方法提高了跨模态检索性能，但它们在细粒度片段的上下文中未能彻底挖掘模态内的相关性。与这些方法不同，我们自适应地建模多级对应关系，并全面探索细粒度的视觉-语义相似性，以实现更完整的对齐。

Karpathy和Fei-Fei（2015）开创了将多模态RNN检测到的局部图像区域与句子中的单词进行对齐的方法。

4. Deep visual-semantic alignments for generating image de-scriptions. Karpathy和Fei-Fei（2015） cvpr

Lee等人（2018）利用堆叠交叉注意力来对齐显著区域和关键词，强调了区域-单词对齐的有效性，并激发了后续的研究。

5. SCAN Lee （2018）

FPAN（Wang等人，2019）被提出，以强调每个区域内不同位置的重要性

6. Position Focused Attention Network for Image-Text Matching

CAMP（Wang等人，2020）创新地引入了跨模态信息传递中自适应调节信息流的机制。

7. Camp: Cross-modal adaptive messagepassing for text-image retrieval ICCV 2019

多阶对齐方法旨在利用全局和局部对应关系，以实现更精确的跨模态匹配。Ji等人（2020a）使用注意力机制定位局部对齐中有语义意义的部分，并使用记忆网络来捕捉全局对齐中的长期上下文知识。Wei和Zhou（2020）结合了对抗网络进行局部对齐，并利用注意力机制进行全局对齐。Qu等人（2020）设计了一个门控自注意力机制用于上下文建模，以及一个多视角摘要模块用于不对称匹配，以获得局部和全局对应关系。Messina等人（2021b，a）通过使用Transformer编码器推理网络，在同一模态内对区域和单词进行多阶推理。Wang等人（2023）利用不常见的文本内容减轻图像-文本匹配中局部对齐的长尾效应，然后利用注意力机制实现全局对齐。他们通过局部关联视觉语义来对齐图像区域和文本单词，并机械地汇总匹配区域-单词对之间的语义相似度，以测量整体的图像-文本相关性。

Ji等人（2020a）使用注意力机制定位局部对齐中有语义意义的部分，并使用记忆网络来捕捉全局对齐中的长期上下文知识

8. Multi-modal memory enhancement- attention network for image-text matching

Wei和Zhou（2020）结合了对抗网络进行局部对齐，并利用注意力机制进行全局对齐。

9. Adversarialattentive multi-modal embedding learning forimage-text matching.

Qu等人（2020）设计了一个门控自注意力机制用于上下文建模，以及一个多视角摘要模块用于不对称匹配，以获得局部和全局对应关系

- Context-aware multi-view sum-marization network for image-text matching

Messina等人（2021b，a）通过使用Transformer编码器推理网络，在同一模态内对区域和单词进行多阶推理

- Transformer reasoning network for image-text matching and retrieval.

Wang等人（2023）利用不常见的文本内容减轻图像-文本匹配中局部对齐的长尾效应，然后利用注意力机制实现全局对齐

- Rare-aware attention network for image-textmatching.

图像文本检索的预训练。图像文本检索的预训练可以分为双编码器方法和交叉编码器方法。双编码器方法[9, 21, 28, 31, 33]采用两个独立的编码器分别提取图像和文本特征，并使用对比学习来对齐共享嵌入空间中的全局表示。交叉编码器方法[1, 3, 14, 15, 17, 18, 22, 34, 35]为图像文本特征的联合编码采用单个编码器。为了促进交叉编码器中的跨模态交互，提出了许多代理任务，例如，掩码语言建模（MLM）、掩码区域建模（MRM）和掩码图像建模（MIM）等。为了提高双编码器的检索准确性，一些方法[21, 31]从交叉编码器中获得启示，通过将MLM、MRM和MIM任务适配到双编码器来增强跨模态交互。然而，另一种方法，即通过蒸馏从交叉编码器向双编码器迁移知识，尚未得到充分探索。

双编码器方法:

- Scaling up visual and vision-language representation learning with noisy text supervision. 2021

- Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieva cvpr 2022
- Learningtransferable visual models from natural language supervision. In international conference on machine learning 2021
- Pre-training visual-semantic embeddings for real-time image-text retrieval 2021
- Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. 2021

交叉编码器方法:

- Vlmo: Unified vision-languagepre-training with mixture-of-modality-experts 2022
- Learning universal image-text representation 2019
- Align before fuse: Vision and language representation learning with momentum distillation 2021
- Blip: Bootstrapping language-image pre-training for uni-fied vision-language understanding and generation. 2022
- A simple and perfor-mant baseline for vision and language 2019
- Oscar: Object-semantics aligned pre-training for vision-language tasks. Eccv 2020
- Pretraining task-agnostic visiolinguistic representations provision-and-language tasks 2019
- Multi-grained vi-sion language pre-training: Aligning texts with visual concepts 2021
- Vinvl: Revisiting visual representations in vision-language models. Cvpr 2021

知识蒸馏：从交叉编码器到双编码器。之前将知识从交叉编码器蒸馏到双编码器的工作可以分为注意力蒸馏方法[32]和 logits 蒸馏方法[24]。注意力蒸馏方法旨在对齐两个模型的跨模态注意力，这需要两个先决条件：（1）两个模型都必须采用基于注意力的主干网络，例如 ViT[4]、BERT[10]，以生成注意力图。（2）双编码器和交叉编码器的输入必须完全相同，以保证它们的注意力图具有相同的形状和语义。这使得注意力蒸馏方法的应用范围受限。

受到图像分类中蒸馏工作的启发，Miech 等人[24]和 Lei 等人[13]将 logits 蒸馏引入图像-文本检索。核心思想是通过基于 KL 散度的损失来约束双编码器和交叉编码器的图像-文本相似度得分分布的一致性。但是，双编码器和交叉编码器之间显著的相似度分布差异使得有效传递知识变得困难。

注意力蒸馏方法[32]和 logits 蒸馏方法[24]

- 32 Distilled dual-encoder model for vision-language understanding 2022
- 24 Thinking fast and slow: Effi-cient text-to-visual retrieval with transformers. 2021

神经排序中的排名蒸馏。在文本神经排序领域，已有一些关于排名蒸馏的研究[7, 23, 29]。Sashank等人[29]提出了带有交叉熵或均方误差（MSE）损失的排名蒸馏损失，以约束正样本分数的一致性，但由于图像-文本检索中相似度分数分布的显著差异，这种方法效果不佳。Sebastian等人[7]提出了Margin-MSE，它要求学生模型和教师模型在正样本和负样本分数之间保持相同的边际。Aditya等人[23]进一步

提出了M3SE，要求学生模型和教师模型在正样本和最难负样本之间保持相同的边际。然而，它只考虑了最难负样本，限制了可以转移的知识。此外，MSE损失与双编码器训练的对齐学习不协调，导致学习过程中的干扰。相比之下，我们提出的CPRD方法通过对比学习考虑多个难负样本之间的相对顺序，与双编码器的原始训练损失保持一致。

排名蒸馏的研究：

- 7 Improving efficient neural ranking models with cross-architecture knowledge distil-lation 2020
- 23 In defense of dual-encoders for neural ranking. 2022
- 29 Rankdistil: Knowledge distillation forranking. 2021

双编码器。这种方法在早期图像文本匹配研究[10,11,21,24,50]中占主导地位。图像和文本标题被独立地嵌入到一个联合度量空间中，匹配的图像-标题对彼此靠近。在这种范式下的现有工作通常通过引入新的损失函数[6,10]、为每个模态编码器提出新架构[24,52,54]，或学习更好的池化方法[4,26]来改进联合嵌入空间。例如，VSE++ [10]提出了一种带有难负样本挖掘的三元组损失，这一方法被后续所有图像文本匹配工作所采用。VSRN [24]、DSRAN [52]、SAEM [54]实现了图卷积和自注意力来改进编码器架构。GPO [4]通过设计一种可以从数据中学习的新池化操作符，取得了具有竞争力的结果。最近，MV-VSE [26]和SDE [20]提出对每个样本数据使用多个嵌入，而HREM [12]展示了一种双编码器模型，该模型可以训练以使用跨模态匹配损失来增强嵌入质量。

通过引入新的损失函数[6,10]

- Prob-abilistic embeddings for cross-modal retrieval cvpr2021
- Vse++: Improving visual-semantic embeddings with hard negatives. Cvpr Faghri et al., 2018 提出了一种带有难负样本挖掘的三元组损失

为每个模态编码器提出新架构[24,52,54],实现了图卷积和自注意力来改进编码器架构

- Visual semantic reasoning for image-text matching iccv 2019 VSRN [24]
- Learning dual semantic relations with graph attention for image-textmatching 2020 DSRAN [52]
- Learning fragment self-attention embeddings for image-textmatching. 2019 SAEM [54]

学习更好的池化方法[4,26]来改进联合嵌入空间:

- Learning the best pooling strategy for visual semantic embedding cvpr 2021 GPO [4]通过设计一种可以从数据中学习的新池化操作符，取得了具有竞争力的结果
- Multi-view visual semantic embedding 2022 MV-VSE [26] 对每个样本数据使用多个嵌入
- Learning semantic relationship among instances for image-text matching 2023 cvpr 训练以使用跨模态匹配损失来增强嵌入质量

交叉注意。与独立地嵌入图像和文本不同，这种方法在计算相似度之前，考虑了图像特征和文本标记之间的细粒度局部对应关系。SCAN[23]是第一个引入这种在两种模态之间使用交叉注意来寻找它们对齐的想法的代表性工作。CAAN[58]后来通过在跨模态交互之后增加一步额外的模态内交互来改进了这个想法。SGARF[9]提出从全局和局部对齐中共同学习，以突出重要的图像区域。最近，NAAF[57]鼓励不匹配的图像区域和单词对之间的不相似度，以增强相似度匹配，而CHAN[35]提出了一种新的跨模态对齐方法，可以忽略冗余的错误对齐。

CAAN[58]后来通过在跨模态交互之后增加一步额外的模态内交互来改进了这个想法

- Context-aware attention network for image-text retrieval cvpr 2020

SGARF[9]提出从全局和局部对齐中共同学习，以突出重要的图像区域

- Similarity reasoning and filtration for image-text matching 2021 AAAl

NAAF[57]鼓励不匹配的图像区域和单词对之间的不相似度，以增强相似度匹配

- Negative-aware attention framework for image-textmatching cvpr 2022

CHAN[35]提出了一种新的跨模态对齐方法，可以忽略冗余的错误对齐

- Fine-grainedimage-text matching by cross-modal hard aligning network cvpr 2023

基于图的图像-文本匹配。在双编码器（dual-encoder）和交叉注意力（cross-attention）方法中，有一些方法将场景图（scene graphs）作为其流程的一部分，以实现更准确的图像-文本对齐[25,28,30,51]。基于这种方法的框架利用图卷积网络（Graph Convolutional Networks, GCN）的能力来捕捉视觉区域和文本标记之间的空间和语义关系。例如，SGM[51]、GCN+DIST[25]、GraDual[30]使用了现成的视觉场景图生成器[56]从图像中提取场景图，然后执行视觉图和文本图之间的跨模态对齐。另一方面，GSMN[28]为视觉区域使用全连接图，但还额外使用区域的极坐标来编码它们的空间关系。

将场景图（scene graphs）作为其流程的一部分利用图卷积网络（Graph Convolutional Networks, GCN）的能力来捕捉视觉区域和文本标记之间的空间和语义关系，以实现更准确的图像-文本对齐[25,28,30,51]

使用现成的视觉场景图生成器:

- Visual-semantic matching by exploring high-order attention and distraction cvpr 2020 GCN+DIST[25]
- Graph structured network for image-text matching cvpr 2020 GSMN[28]为视觉区域使用全连接图
- Gradual: Graph-based dual-modal representation forimage-text matching 2022 GraDual[30]
- Cross-modal scene graph matching forrelationship-aware image-text retrieval 2020 SGM[51]

跨模态图像-文本检索。跨模态图像-文本检索的研究是视觉领域的基本任务，许多现有的方法已被提出[4,8,10,17,19,21,24,25,33,41,52,53,55,59,60,64,67,68]。一种标准方法是从预先准备好的图像和文本数据集中获取文本语言的公共空间作为训练数据[11,16,17,29,35,48,50]。为了获取准确的图像-文本公共空间，已经引入了多种方法来改进损失函数和距离空间，例如度量学习[17,68,69]和概率分布表示[11,31,56]。为了进行细粒度检索，已经提出了各种扩展[3,33,41,58,66]，通过引入对象检测[23,33]、对象之间的图基关系[13,42,65]、重新加权策略[3,57,58]和注意力机制[8,67]。这些现有的细粒度检索方法表明，对象意识是对局部细节进行跨模态检索的一个关键线索。本文重点关注预训练的V&L模型[14,20,26,36–38,51]中的对象意识。我们提出了一个简单而有效的新框架，能够有效提高包含在语义上重要的微小对象的图像-文本检索性能。

现有的方法:

- 2
- IMRAM: Iterative Matching With Recurrent Attention Memoryfor Cross-Modal Image-Text Retrieval. Cvpr 2020
- ViLEM: Visual-Language Error Modeling for Image-Text Retrieval.In: CVPR 2023
- VSE++
- Look, imagine and match: Improvingtextual-visual cross-modal retrieval with generative model cvpr 2018
- tance-aware image and sentence matching with selective multimodal lstm cvpr 2017
- Saliency-guided attention network for image-sentence matching cvpr 2019
- Step-Wise Hierarchical AlignmentNetwork for Image-Text Matching : IJCAI.2021
- Scan
- Focus your attention: A bidirectional focal attention network for image-text matching 2019
- Consensus-aware visual-semanticembedding for image-text matching iccv 2020
- Learning two-branch neural networks for image-text matching task TPAMI 2018
- Camp
- Multi-modality cross attention network for image and sentence matching cvpr 2020
- rete-continuous action space policy gradient-based attention for image-text matching cvpr 2021
- Learning Hierarchical Semantic Correspondences for Cross-ModalImage-Text Retrieval 2022 cvpr
- Context-aware attention network for image-text retrieval 2020 cvpr
- Deep cross-modal projection learning for image-text matching. Eccv 2018

一种标准方法是从预先准备好的图像和文本数据集中获取文本语言的公共空间作为训练数据[11,16,17,29,35,48,50]

- Probabilistic embeddings for cross-modal retrieval cvpr 2021
- Finding beans in burgers: Deepsemantic-visual embedding with localization cvpr 2018
- VSE++
- Deep fragment embeddings for bidirectionalimage sentence mapping 2014
- Visual semantic reasoning for image-text matching iccv 2019
- Polysemous visual-semantic embedding for cross-modal retrieval cvpr 2019
- Preserving semantic neighborhoods for robust cross-modal retrieval eccv 2020

运用度量学习来改进损失函数和距离空间

- VSE++
- Deep cross-modal projection learning for image-text matching. Eccv 2018
- Towards optimal finegrained retrieval via decorrelated centralized loss with normalized-scale layer AAAI 2019

概率分布表示改进[11,31,56]:

- Probabilistic embeddings for cross-modal retrieval cvpr 2021
- Improving Cross-Modal Retrieval With Set of Diverse Embeddings cvpr2023
- Multilateral Semantic Relations Modeling for Image Text Retrieval cvpr2023

为了进行细粒度检索，已经提出了各种扩展[3,33,41,58,66]

- Interclass-relativity-adaptivemetric learning for cross-modal matching and beyond 2020
- SCAN
- Focus your attention: A bidirectional focal attention network for image-text matching 2019
- Universal weighting metriclearning for cross-modal matching cvpr 2020
- Negative-aware attention framework forimage-text matching cvpr 2022

引入对象检测[23,33]:

- Step-wise hierarchical alignment network for image-text matching 2021
- SCAN

对象之间的图基关系[13,42,65]:

- Similarity reasoning and filtration for image-text matching AAAI2021
- Graph structured network for image-text matching cvpr 2020
- Cross-modalconfidence-aware network for image-text matching AAAI2022

重新加权策略[3,57,58]:

- Interclass-relativity-adaptivemetric learning for cross-modal matching and beyond 2020

- Meta self-paced learning for cross-modal matching. 2021
- Universal weighting metriclearning for cross-modal matching cvpr 2020

注意力机制[8,67]:

- IMRAM: Iterative Matching With Recurrent Attention Memoryfor Cross-Modal Image-Text Retrieval.
- Context-aware attention network for image-text retrieval.

预训练视觉与语言模型。近年来，使用V&L模型的跨模态图像-文本检索被提出作为一种新的范式[14,20, 26, 36–38, 51]。视觉语言预训练，如CLIP[45]，通过自监督任务从大量的图像-文本对中训练视觉语言对齐。在此范式之前，现有的图像-文本检索方法主要关注使用中等规模数据集（如Flicker 30K和COCO）来训练算法。相比之下，近期使用预训练的V&L模型的跨模态图像-文本检索性能优于那些传统的跨模态图像-文本检索方法，实现在多样化数据集上的高零样本性能，并实现了开放词汇检索。特别是最近提出的BLIP2[36]在跨模态图像-文本检索中表现出压倒性的性能。然而，最近指出，像CLIP这样的V&L模型在定位方面存在弱点，并提出了一些简单的改进[47, 70]。正如后面所述，这种弱点在跨模态图像-文本检索中也有所体现。所提出的方法是一个新颖的框架，它可以通过跨模态图像-文本检索来克服这一弱点，同时利用现有V&L模型的潜在能力。

使用V&L模型的跨模态图像-文本检索[14,20, 26, 36–38, 51]:

- Giving Text More Imagination Space for Image-textMatching. 2023
- PiTL: Cross-modal Retrieval with Weakly-supervisedVision-language Pre-training via Prompting 2023
- Scaling Up Visual and Vision-Language Representation Learning With NoisyText Supervision icml2021
- BLIP-2: Boot-strapping Language-Image Pre-training with Frozen Image Encoders and LargeLanguage Models icml2023
- BLIP: Bootstrap-ping Language-Image Pre-training for Unified Vision-Language Understanding andGeneration icml2022
- Align before Fuse: Vision andLanguage Representation Learning with Momentum Distillation : NeurIPS2021
- ProbVLM: Probabilistic Adapter for Frozen Vision-Language Models. iccv2023

对CLIP这样的V&L模型在定位方面存在弱点，提出了一些简单的改进[47, 70]:

- What does clip know about a red circle?visual prompt engineering for vlm 2023
- Regionclip: Region-based language-image pretraining cvpr 2022

最近，一些细粒度的视觉语言模型（VLM）通过利用视觉对象和文本单词之间的细粒度关系实现了局部对齐。一些方法（Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021）使用预训练的对象检测器从图像中获取对象特征，并与文本特征对齐，而其他方法（Kim et al., 2021; Yao et al., 2021; Wang et al., 2022a; Ji et al., 2021; Xue et al., 2023; Jiang et al., 2023）则试图在局部上将固定补丁与文本单词对齐。前者依赖于精确的对象检测器，后者关注的是预定义网格内固定补丁之间的单词关系。然而，不同的肺病变特征可能会导致这些方法将它们划分为单独的补丁，从而导致语义信息不完整。因此，我们设计了一种自适应补丁-单词匹配（AdaMatch）方法。

细粒度的视觉语言模型（VLM）通过利用视觉对象和文本单词之间的细粒度关系实现了局部对齐:

- Chen et al., 2020a Uniter: Universal image-text representation learning.
- Oscar: Object-semantics aligned pre-training for vision-language tasks. Li et al., 2020b,a
- Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. CVPR2021

试图在局部上将固定补丁与文本单词对齐:

- Vilt:Vision-and-language transformer without convolution or region supervision. Icml 2021
- Filip:Fine-grained interactive language-image pre-training 2021
- Multi-granularitycross-modal alignment for generalized medical visual representation learning. NeurIPS 2022a
- Improving joint learning ofchest x-ray and radiology report by word region alignment 2021
- Knowledge boosting: Rethinking medical contrastive vision-languagepre-training 2023
- Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment 2023

医疗领域的预训练语言模型（VLMs）在下游任务中广泛应用于胸部X光片，包括从胸部X光片到报告的生成（Chen et al., 2020b, 2021a; Yang et al., 2021; Wang et al., 2022b; Voutharoja et al., 2023; Yang et al., 2023; Shi et al., 2023; Huang et al., 2023）以及从报告到胸部X光片的生成（Rom-bach et al., 2022; Chambon et al., 2022b,a; Lee et al., 2023a,b; Han et al., 2024; Shentu and Al Moubayed, 2024; Hou et al., 2023; Hashmi et al., 2024; Chen et al., 2024）任务。在从胸部X光片到报告生成的任务中，Chen et al., 2021a 提出了一个具有共享内存的跨模态记忆网络，以将图像与文本对齐，从而提高报告生成的性能。在从报告到胸部X光片生成的任务中，先前的技术通过从医疗报告中创建注释过的胸部X光片图像来增强训练数据并解决隐私问题，这些方法被分为基于扩散的方法和基于变换器的方法。

从胸部X光片到报告的生成

从报告到胸部X光片的生成