为什么要做跨模态对齐?

跨模态对齐的核心问题在于不同模态的数据涵盖不等量的信息。跨模态对齐的目标是使得机器能够理解和关联不同模态之间的语义，从而实现更高级的智能计算任务。这包括两个主要类型的对齐方式：

1. **显式对齐**：在特征向量域中，将一个模态的信息映射到另一个模态的特征表示。例如，可以通过对文本描述的特征向量进行处理，使其能够与对应的图像特征向量匹配。
2. **隐式对齐**：利用一个模态的信息，对同一模态之间的表示进行重新加权。一个直观的例子是通过文本模态的信息，对视觉特征向量进行重加权操作，以提升对齐效果。

显式对齐技术

相关性分析

相关性分析旨在描述不同模态（如图像和文本）信息的对齐问题，目的是建立两种模态信息之间的相关性。相关性分析的核心在于通过统计和数学方法，找到不同模态数据之间的最佳匹配，从而实现跨模态的语义关联。

1. **2004 - Neural Computation: Canonical correlation analysis: An overview with application to learning methods**：提出了典型相关性分析（CCA）的概念，用于挖掘两组向量之间的线性相关关系。首次系统化地提供了一种挖掘两组向量之间线性相关关系的方法，为跨模态对齐奠定了理论基础。
2. **2010 - ACM MM: A New Approach to Multimedia Retrieval**：扩展了CCA，提出语义相关匹配模型（SCM），通过联合学习语义类别推断与CCA空间来改进模型性能。在CCA的基础上加入了语义信息的考虑，通过同时学习语义类别和模态对齐关系，提高了跨模态检索的效果。
3. **2006 - TNNLS: Facial expression recognition using kernel canonical correlation analysis**：提出了核典型相关性分析（KCCA），利用核函数克服传统CCA处理非线性关系的限制，显著提高了对复杂跨模态数据的处理能力。
4. **2013 - ICML: Deep canonical correlation analysis**：提出了深度典型相关性分析（DCCA），利用深度神经网络嵌入的多层次映射解决非线性问题，增强了图像和文本的对齐效果。

度量学习

相关性分析方法主要关注配对样本的整体相关性，但在实际应用中，了解哪些特征导致不匹配同样重要。度量学习通过直接度量样本之间的距离关系，解决了这一问题。它不仅关注匹配样本之间的相似性，还通过优化样本距离，使得相似样本距离更近，不相似样本距离更远，从而提升模型的区分能力。以下是度量学习领域的几个关键研究：

1. **2006 - CVPR: Dimensionality Reduction by Learning an Invariant Mapping**：提出了对比损失（Contrastive loss），通过最大化正样本对距离最小化负样本对距离，来增强语义对齐。提供了一种直观且有效的方法来度量样本之间的相似性，通过明确的距离度量，提升了模型在跨模态任务中的表现。
2. **2018 - BMVC: Vse++: Improving visual-semantic embeddings with hard negatives**：改进了对比损失，引入困难负样本挖掘（Hard Negatives），通过选择最难区分的负样本来优化模型学习，显著提升了模型的区分能力。
3. **2018 - CVPR: Triplet-Center Loss for Multi-View 3D Object Retrieval**：提出了中心三元组损失（Center Triplet Loss），结合锚点与类别中心的距离来进一步优化模型，提高了模型的聚类效果和识别能力。

4. **2016 - NIPS: Improved deep metric learning with multi-class n-pair loss objective**：提出N对多分类损失（N-Pair Loss），通过结合多个负样本对正样本进行学习，提高了模型的收敛速度、训练效率和稳定性，在大规模数据集上的表现尤为显著。。
5. **2016 - CVPR: Learning Deep Structure-Preserving Image-Text Embeddings**：提出对称四元组损失（Symmetric Quadruplet Loss），通过增加模态内部结构约束关系，增强了模型对不同模态间语义对齐的性能。不仅关注跨模态的对齐，还保证了模态内部的结构一致性，使得模型在处理复杂跨模态任务时更具鲁棒性。

隐式对齐技术

在跨模态学习中，不同模态的数据处理过程需要自适应地感知并利用来自其他模态的信息，以实现更有效的特征对齐。为此，有两种主要的解决思路：注意力机制和无参交互。

注意力机制

注意力机制通过在不同模态之间建立隐式关联，捕捉潜在的关联关系，实现更精细的跨模态对齐。例如，基于一个模态的信息，可以配置一个注意力模块，这个注意力模块以另一个模态的特征表示作为输入输出，从某种意义上说，这一个信息处理模块的权重系数是被另一个模态的信息动态决定的。这类似于编程中的runtime，在这种机制下，一个模态的信息处理过程能够根据另一个模态的信息进行自适应调整。这是系统架构层面的解决方案，通过添加基于注意力机制的交互模块、改变系统数据通路，让不同模态的数据之间实现各种模式的交互。当通过架构设计实现了数据交互后，研究者才能进一步考虑如何用具体的数据集与训练算法使得这个交互能够work。

1. **2018 - ECCV: Stacked Cross Attention for Image-Text Matching**：提出了跨模态交叉注意力机制（Cross Attention Mechanism），通过计算图像区域和文本单词之间的注意力分布，实现了图像和文本的隐式关联。模型能够在不同的语义层次上进行图像和文本的对齐，从而捕捉到更细粒度的跨模态关联信息。
2. **2020 - CVPR: IMRAM: Iterative Matching with Recurrent Attention Memory for Cross-Modal Image-Text Retrieval**：改进了单轮交叉注意力机制，提出多轮交叉注意力机制，并设计了记忆单元以更好地传播对齐信息。模型能够更好地捕捉长距离的跨模态依赖关系，提高了图像和文本检索的准确性。
3. **2021 - SIGIR: Dynamic Modality Interaction Modeling for Image-Text Retrieval**：进一步改进，设计了动态模态交互模型，通过不同单元的路径规划为不同样本建立不同的学习路径，提高了模型的灵活性和对齐效果。使模型能够根据不同样本的特性，自适应地调整交互策略，从而更有效地进行跨模态对齐。
4. **2016 - CVPR: Stacked Attention Networks for Image Question Answering**：将注意力机制应用于视觉问答任务，关联图像区域信息和文本编码信息，以辅助生成问题答案。模型能够逐步聚焦到图像中与问题相关的区域，从而提高了视觉问答的准确性和鲁棒性。
5. **2015 - ICML: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention**：提出了基于视觉注意力的图像字幕生成方法，通过注意力机制为每个词关注图像的显著性区域，提升图像字幕生成的质量。

无参交互

无参交互类似于编程中的compile time，通过训练使得不同模态的数据在处理过程中能够相互影响，旨在不通过引入额外的模块，基于对比学习方式进行局部关联信息的挖掘，是在算法层面的解决方案。可以通过稠密对比学习等方法，实现不同模态之间的隐式对齐，避免大量交互参数带来的计算负担。

1. **2021 - CVPR: Dense Contrastive Learning for Self-Supervised Visual Pre-Training**：提出了稠密对比学习方法，通过挖掘潜在的正样本对，进行局部关联信息的挖掘，实现了细粒度的图像信息对齐。

2. **2022 - ECCV: LocVTP: Video-Text Pre-training for Temporal Localization**：将稠密对比学习应用于视频文本特征学习，挖掘视频片段和文本片段之间的细粒度对齐信息。模型能够更准确地进行视频片段的时间定位，提高了跨模态时间对齐的效果。
3. **2023 - TMM: Towards Fast and Accurate Image-Text Retrieval with Self-Supervised Fine-Grained Alignment**：将稠密对比学习应用于跨模态检索，改进对比学习方法，通过聚类方式进行特征对齐，提高了检索效率和准确性。