为什么要做跨模态预训练?

从BERT开始，预训练模型与微调的范式开启了自然语言处理领域的新篇章，这一方法被称为预训练-微调范式。该范式通过在大规模无标注数据上进行预训练，再在特定任务上进行微调，大大提高了模型的泛化能力和性能。不同于这一范式出现之前的工作，当前的预训练大模型更多地在present一个产品，而不是回答某个科学问题。这些模型的开发过程更像是技术报告和系统论文（如OSDI等），注重系统的实现和工程实践。

在预训练过程中，研究人员面临一系列共性问题，这些问题在文本预训练、视觉预训练以及跨模态预训练中都有体现，包括：

- **模型架构设计与预训练算法**：需要设计能够有效处理和融合不同模态数据（如图像和文本）的模型架构，并开发相应的预训练算法，以确保模型在不同任务中的表现。
- **高质量预训练语料的构建与配比**：为了训练出性能优越的模型，必须构建和选择高质量的预训练数据集。这包括从海量的无标注数据中提取有用的信息，并合理配比不同模态的数据，以充分利用它们的互补优势。

每个预训练模型的研究成果在一定程度上都回答了这些共性问题，推动了跨模态预训练技术的发展。通过不断优化模型架构和预训练算法，改进预训练数据的质量和选择方法，研究人员逐步提升了预训练模型在实际应用中的表现。

我们观察到，生成式语言模型在参数量scaling up的情况下涌现出了推理能力。如何运用这份涌现的推理能力？事实上，我们在考虑：如何设计、实现一个从视觉生成文本内容的模型？

模型架构

单流模型

单流模型不区分图像和文本输入，图像和文本直接输入到一个统一的模型中进行学习。这种方法的优点是能够在同一模型中捕捉到图像和文本之间的全局关联。

- **VL-BERT (2020-ICLR)**：VL-BERT提出了将视觉与文本先拼接，然后输入一个基于Transformer的模型。利用注意机制进行多模态交互，使得模型能够同时处理图像和文本信息。Transformer架构的自注意力机制有效地捕捉了图像和文本之间的关联，提升了模型的跨模态理解能力。通过预训练大规模无标注数据的image-text pair，VL-BERT能够学习到丰富的跨模态特征。在固定参数得到VL-BERT模型后，可以在下游任务中进行微调，提高模型在具体任务上的表现。
- **Unicoder-VL (2020-AAAI)**：Unicoder-VL在VL-BERT的基础上进行改进，提出了输入图像块表征和文本单词的拼接。引入了多任务学习，包括文本掩码预测（MLM）、图像遮掩区域预测（MOC）和图文匹配任务（VLM）。这种设计使模型能够更好地捕捉多模态信息，提高泛化能力。同样利用大规模的image-text pair数据，Unicoder-VL通过多任务学习方式提升模型的鲁棒性和适应性。
- **OSCAR (2020-ECCV)**：OSCAR进一步发展了单流模型的架构，通过将相同语义的物体作为图像和语言对齐的锚点，简化了图像和文本之间的语义对齐难度。利用Faster-RCNN提取的物体区域和标签作为图像模态特征进行跨模态预训练。OSCAR使用Faster-RCNN提取的物体区域和标签，确保模型在预训练过程中能够学习到高质量的视觉和文本特征对齐，从而提升下游任务的性能。

双流模型

双流模型分别对图像和文本进行编码，在某个阶段再将图像和文本的特征融合起来。这种方法的优点是可以独立优化图像和文本的特征提取器，并在融合阶段进行更细粒度的跨模态对齐。

- **ViLBERT (2019-NeurIPS)**：ViLBERT提出了双流跨模态模型，图像和文本先分别进行编码，之后利用跨模态自注意力机制进行交互。这种架构设计允许模型在融合前独立学习每种模态的特征，确保各自的特征提取器能够充分优化。利用多任务学习，包括文本掩码预测（MLM）、图像遮掩区域预测（MOC）和图文匹配任务（VLM），进一步提升模型的跨模态理解能力。通过大规模的多模态数据进行训练，确保模型在多个下游任务中表现优异。
- **CLIP (2021-ICML)**：CLIP在ViLBERT的基础上，通过对比学习训练，利用Vision Transformer提取图像模态表示，并基于自然语言监督学习可迁移的视觉模型。对比学习使得图像和文本在共享的表示空间中对齐，提高了模型的泛化能力和迁移性能。利用大量的自然语言监督数据，确保模型在多个下游任务中表现优异。
- **ALBEF (2021-NeurIPS)**：ALBEF进一步改进了CLIP，通过在融合前学习更好的单模态表示，同时引入了遮掩单词预测（MLM）和图文匹配任务（VLM）。利用MoCo动量模型的思路增强数据，使模型在训练中更具鲁棒性。使用大量带有噪声的图像文本对，通过引入对比学习和动量模型的方法，提高数据质量。

预训练语料构建

预训练语料的质量直接影响模型的性能。网上爬取的语料往往含有大量噪声，需要设计有效的算法来提升数据集的质量。

- **BLIP (2022-ICML)**：BLIP提出了在文本输入形式是prompt + [caption]的情况下，利用two-stream模型分离Encoder-Decoder模型中的不同输入。通过对比学习拉近视觉空间和语言空间，并以视觉特征为输入进行图像字幕序列生成。结合对比学习和编码-解码结构，提升模型在生成任务和理解任务上的表现。通过bootstrapping算法和高质量的文本输入形式，提升了数据集的质量和模型的训练效果。这种方法能够有效过滤噪声数据，确保模型在预训练阶段学习到更准确和有用的特征。

基于大语言模型的跨模态预训练语言模型

知识蒸馏在跨模态预训练中起到了至关重要的作用，通过提取和转移大规模预训练语言模型中的知识，可以显著改善预训练效果。

- **Flamingo (2022-NeurIPS)**：Flamingo提出了基于大语言模型的跨模态预训练框架，采用Perceiver Resampler将图片和视频进行统一的表征。这个过程确保了不同尺寸和形式的视觉数据能够以一致的格式输入模型。然后，利用Gated XATTN-DENSE机制，将视觉信息嵌入到语言模型中，保持了大语言模型的推理能力，同时增强了其处理多模态任务的能力。Flamingo通过在大规模、多样化的图文数据集上进行训练，确保了模型在多模态任务中的泛化能力。通过这种方式，Flamingo在few-shot学习任务中表现出色。
- **BLIP-2 (2023-ICML)**：BLIP-2在前一版本的基础上进行了改进，通过阶段式地选择和对齐视觉表示，逐步拉近视觉和文本空间的距离。在预训练阶段，利用视觉语言损失函数，使得视觉和文本特征能够更好地对齐。BLIP-2通过引入更精细的视觉特征和文本特征对齐方法，提高了数据的质量和模型的训练效果。这种方法确保了模型在处理复杂多模态任务时的表现。
- **VisualChatGPT (2023-arXiv)**：VisualChatGPT引入了一个控制器模块，将多模态信息转换为文本输入，并通过思维链（Chain-of-Thought）阶段式地解决复杂Query。控制器模块负责协调不同的视觉基础模型，确保在处理复杂任务时能够有序地整合视觉和文本信息。VisualChatGPT利用多模态数据的协同作用，通过逐步迭代的方式，提高了预训练数据的质量，增强了模型在复杂任务中的鲁棒性和适应性。
- **LLaVA (2023-NeurIPS)**：LLaVA通过构造视觉指令用于微调开源的大语言模型（如LLaMA），使其适应视觉语言输入。这种指令微调方法有效地将视觉信息嵌入到语言模型中，利用ChatGPT的推理能力，解决了视觉语言任务。LLaVA通过构建大量的视觉指令数据集，提高了模型在视觉语言任务中的表现。指令数据集的构建方式确保了数据的多样性和质量，使模型能够处理多种复杂的视觉语言任务。

- **MiniGPT-4 (2023-arXiv)**：MiniGPT-4采用了一种简化的线性层对齐方法，通过大量的image-text pairs进行训练。这种方法确保了模型能够快速适应多模态输入。通过ChatGPT重述生成的图片描述，MiniGPT-4实现了更连贯且信息量丰富的文本生成。MiniGPT-4利用ChatGPT生成的高质量描述文本，改善了预训练数据的质量。通过这种方法，模型能够更好地理解和生成多模态内容，提高了在实际应用中的表现。
- **MiniGPT-5 (2023-arXiv)**：MiniGPT-5引入了Voken的概念，通过阶段式选择和对齐视觉表示，进一步优化了多模态对齐过程。利用Diffusion模型实现了从文本到图像的生成任务，增强了模型的生成能力。MiniGPT-5通过Diffusion模型的生成能力，创建了高质量的文本和图像对齐数据。这种方法确保了模型能够在多模态生成任务中表现出色，进一步提升了模型的应用价值。