Build pre-training model (in medical)

2023.12.3

目录

CONTENTS

第一章 Language Model

第二章 Vision Model

第三章 Vision Language Model

第四章 Future Work

Foundation Model

Foundation Model：the base models trained on large-scale data in a self-supervised or semi-supervised manner that can be adapted for several other downstream task

--------- Bommasani et al. at Stanford Institute for Human-Centered AI. Foundational models

Part 1

Language Model

语言模型的技术发展

Artificial General Intelligence

- GPT 3.5 / 4：人机接口
- GPT 3.0：生成模型也能实现强大的理解能力
- BERT：预训练 + Fine-tuning (理解类任务)
- GPT：自回归模型 + Zero / Few Shot Prompt (生成类任务)

Transformer

模型参数量

训练数据

BERT & GPT

	Unidirectional language model	Bidirectional language model	Sequence-to-sequence model
Architecture	Transformer decoder	Transformer encoder	Transformer
Pre-training	Language modeling (2)	Mask language modeling (3)	Sequence-to-sequence learning
Tasks	Language generation	Language understanding	Sequence-to-sequence
Models	GPTs[3,25,26]	BERT,[8] RoBERTa,[17] ALBERT,[14] XLNet,[36] Electra[7]	BART,[15] T5[24]

GPT：自回归模型。单向架构。

主要特点：使用上文的信息来预测下一个单词的概率。不微调，使用Zero / Few Shot Prompt的方法进行生成。

BERT：自编码模型。双向架构。

主要特点：使用上下文的信息来填补缺失的内容。需要微调，以实现在下游任务中的理解能力。

GPT

自回归语言模型：

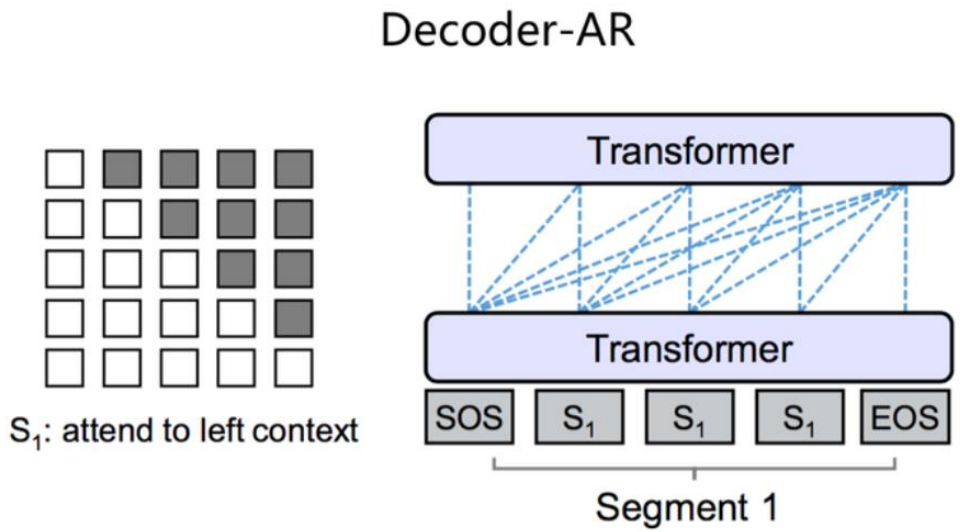
文本序列联合概率的密度估计，即为传统的语言模型，天然适合处理自然生成任务；

输入是一系列单词 w1，w2，... ，wN，首先，通过输入层，创建一个输入表征序列，表示为矩阵 H (0)。通过 L 层Transformer blocks解码层后，生成一系列中间表征，表示为矩阵 H (L)。

$$H^{(L)} = \text{transformer\_decoder}(H^{(0)})$$

最后，根据每个位置的最终中间表示，计算每个位置的单词概率分布。计算并最小化交叉熵或负对数似然来估计参数:

$$\max_{\theta} \quad \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)},$$



与下游任务交互的方式：

- Zero Shot Prompting : Instruct
  - 构建的专业Instruct
  - 人类真实需求的Instruct
- Few Shot Prompting : In Context Learning
  - 提供若干基本示例来指示任务
  - 有趣的是，输入的x和y不需要是一个映射，只需要满足符合候选分布即可

BERT

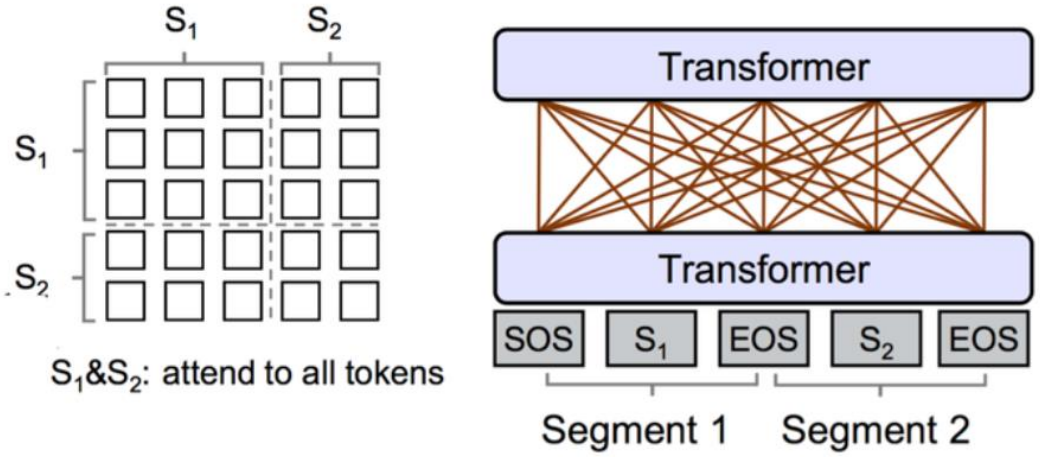
自编码语言模型：

降噪自编码特征表示，通过引入噪声[MASK]构建MLM，获取上下文相关的双向特征表示；

输入是一个单词序列，可以是单个文档中的连续句子。输入句中的未被Mask的任意单词两两可见，但是被Mask掉的单词之间都相互独立，互不可见。在预测某个被Mask掉的单词的时候，所有其它被Mask的单词都不起作用，但是句内未被Mask掉的所有单词，都可以参与当前单词的预测。

$$\max_{\theta} \quad \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)},$$

Encoder-AE



与下游任务交互的方式：

- 微调

由于BERT并不是生成式的，所以输出和输入的长度被限制成相同的。这就导致了不能完成机器翻译、机器人对话等高阶的自然语言生成任务。

如果是以fine-tuning方式解决下游任务，Bert模式的效果优于GPT模式；若是以zero shot/few shot prompting这种模式解决下游任务，则GPT模式效果要优于Bert模式。

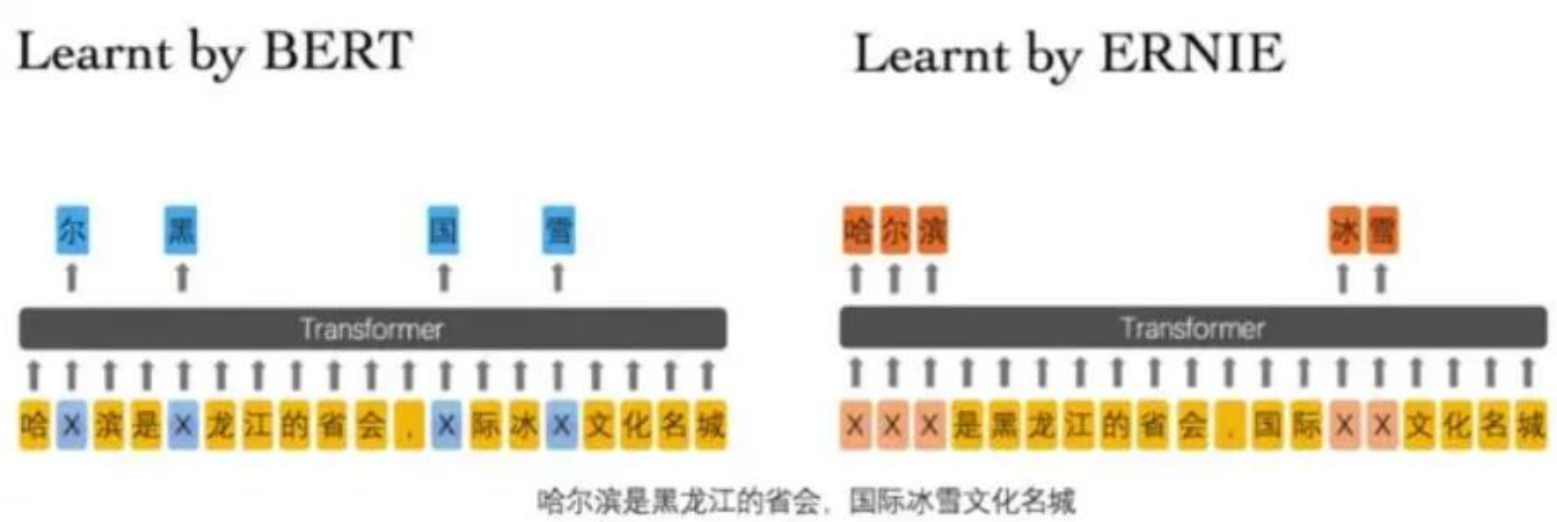
BERT

如何在BERT中融入显示的知识：

ERNIE - 百度

5. Entity-level Masking 预训练

实体信息包括人名，地名，组织名称，产品名称等， 而实体又是一种抽象的概念，且通常包含着一些重要的信息，且实体之间的关系也十分重要。 ERNIE 先用命名实体识别找出句子中的实体，然后与 Phrase-level 一样， mask 其中的一些实体并预测这些mask掉的 word (字)。



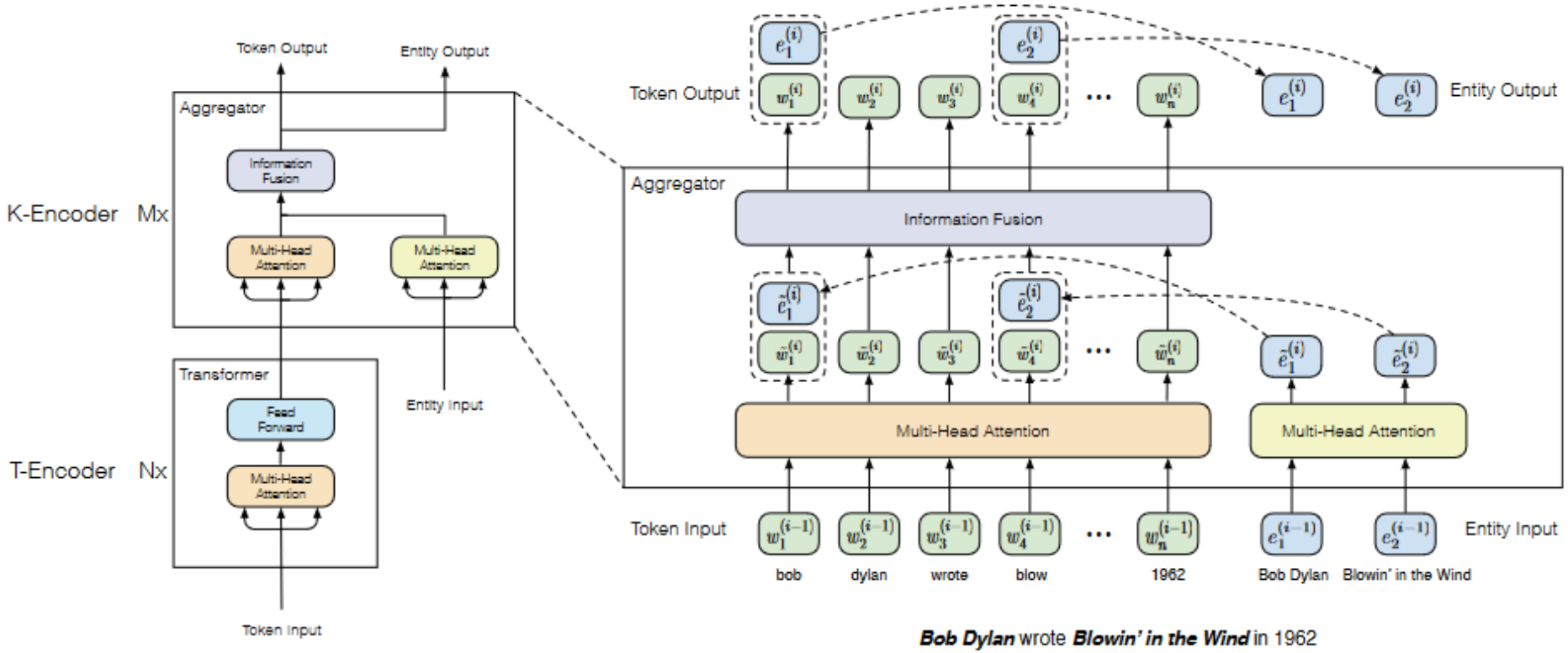
BERT

如何在BERT中融入显示的知识：

ERNIE – 清华：
如何融入知识图谱的信息

知识图谱本质是 实体 + 实体间关系， 其中实体为点，实体间关系为边。
而将知识图谱引入到预训练语言模型，有两个主要的挑战：

- Structed Knowledge Encoding： 对于给定的文本，如何有效的提取其中的知识图谱信息并对其进行 encode。
- Information Fusion： 即如何将 encode 后的知识图谱信息融入预训练模型。



(a) Model Achitecture

(b) Aggregator

BERT in Medical

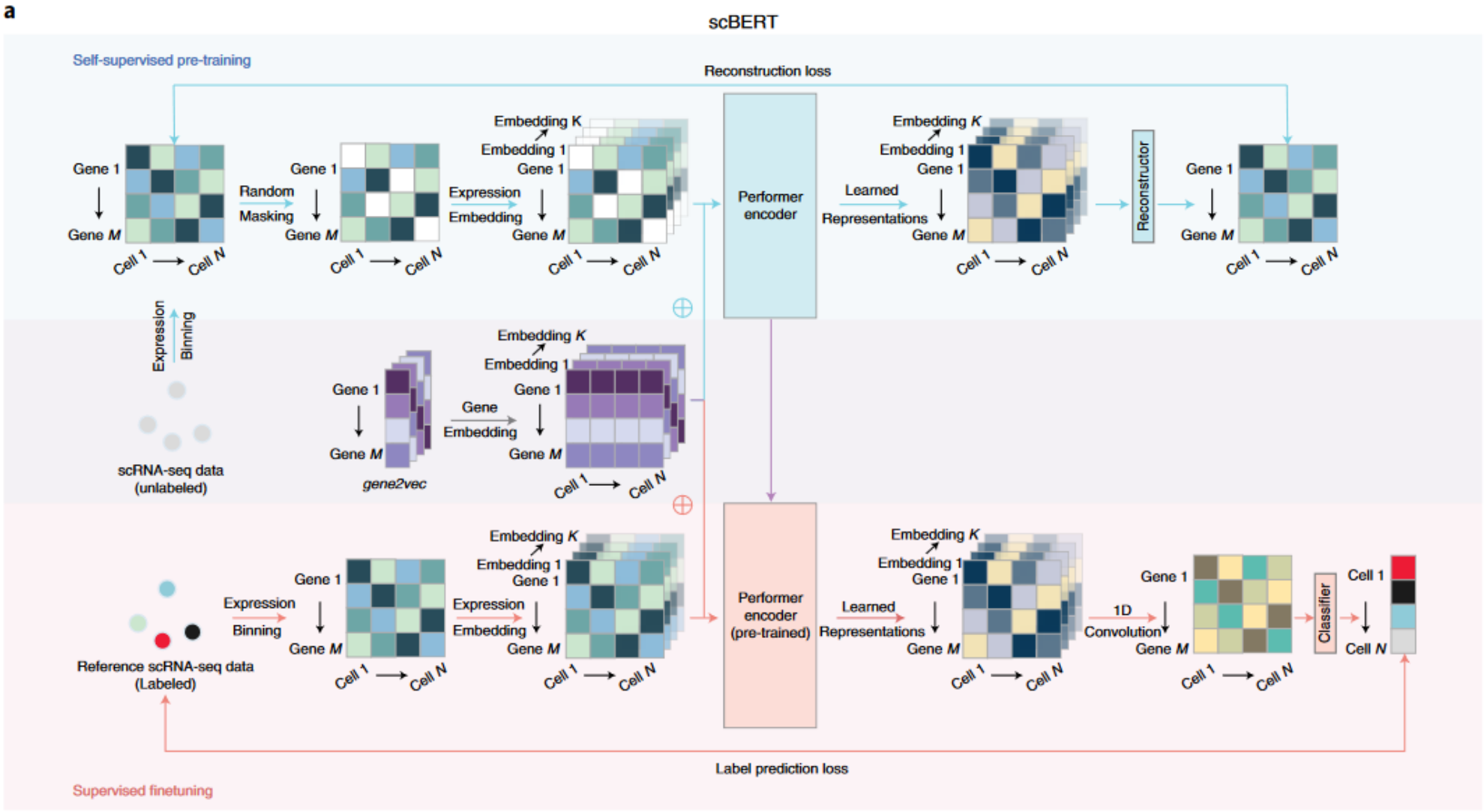
scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. (NMI 2022)

出发点：
为单细胞 RNA-seq 数据注释细胞类型是研究疾病进展和肿瘤微环境的先决条件。

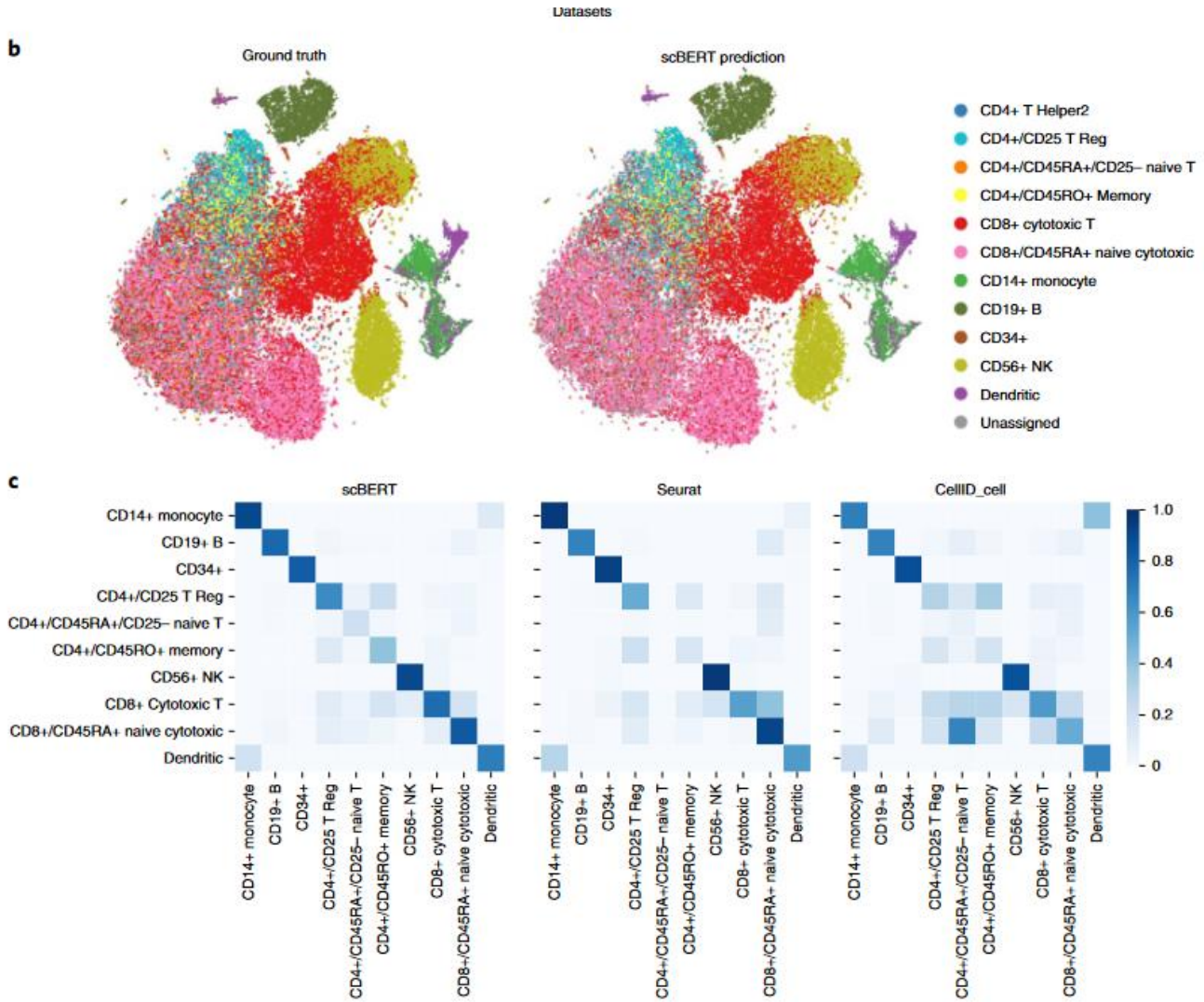
问题：
现有的注释方法通常都存在缺乏经过策划的标记基因列表、批次效应处理不当以及难以利用
潜在的基因-基因相互作用信息等问题，从而影响了其通用性和鲁棒性

贡献：
scBERT 通过在大量未标记的 scRNA-seq 数据上进行预训练，获得了对基因-基因相互作用的一般理解；然后将其转移到未见过的用户特定 scRNA-seq 数据的细胞类型注释任务中，进行监督微调。

BERT in Medical



BERT in Medical



Part 2

Vision Model

视觉模型的技术发展

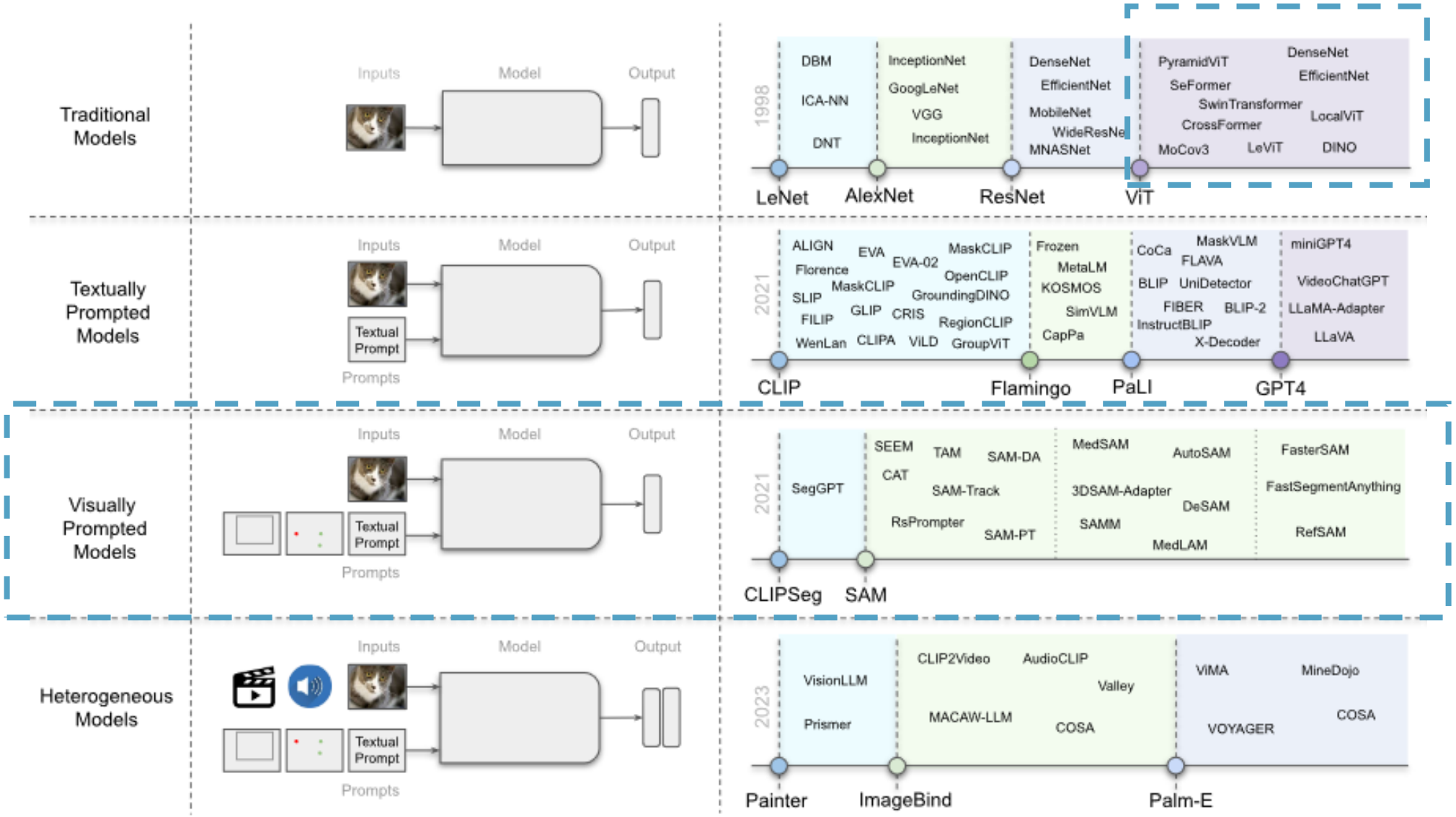
- 传统无监督训练：MAE，DINO等等
- 视觉/文本输入提示：SAM， CLIP，SegGPT等 (prompt engineering)

Transformer

模型参数量

训练数据

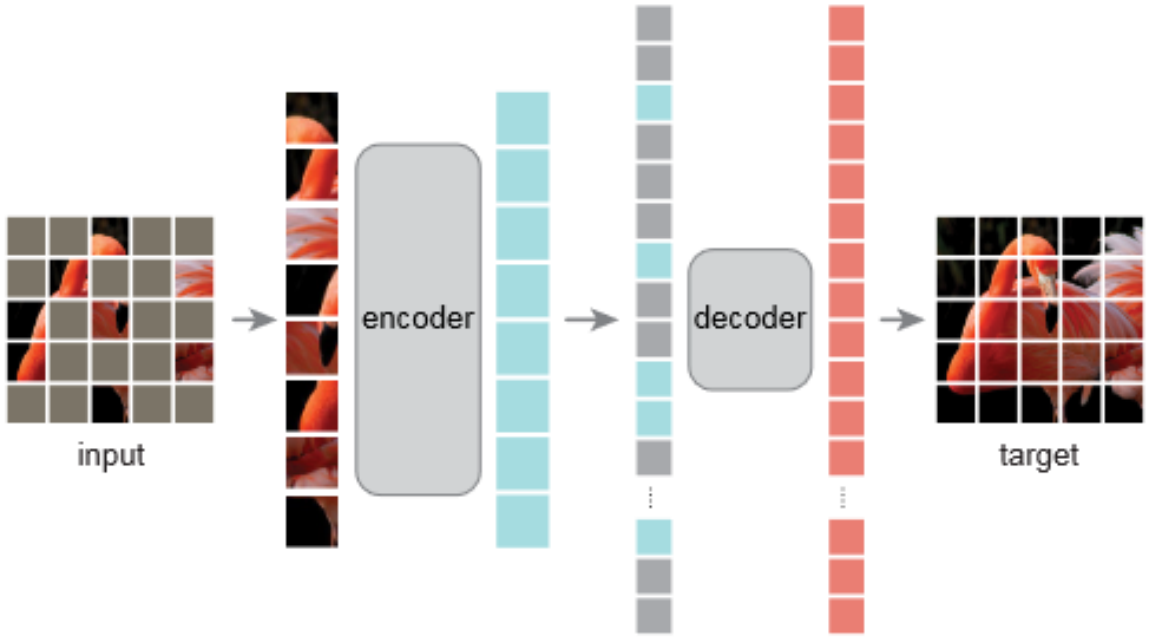
视觉基础模型

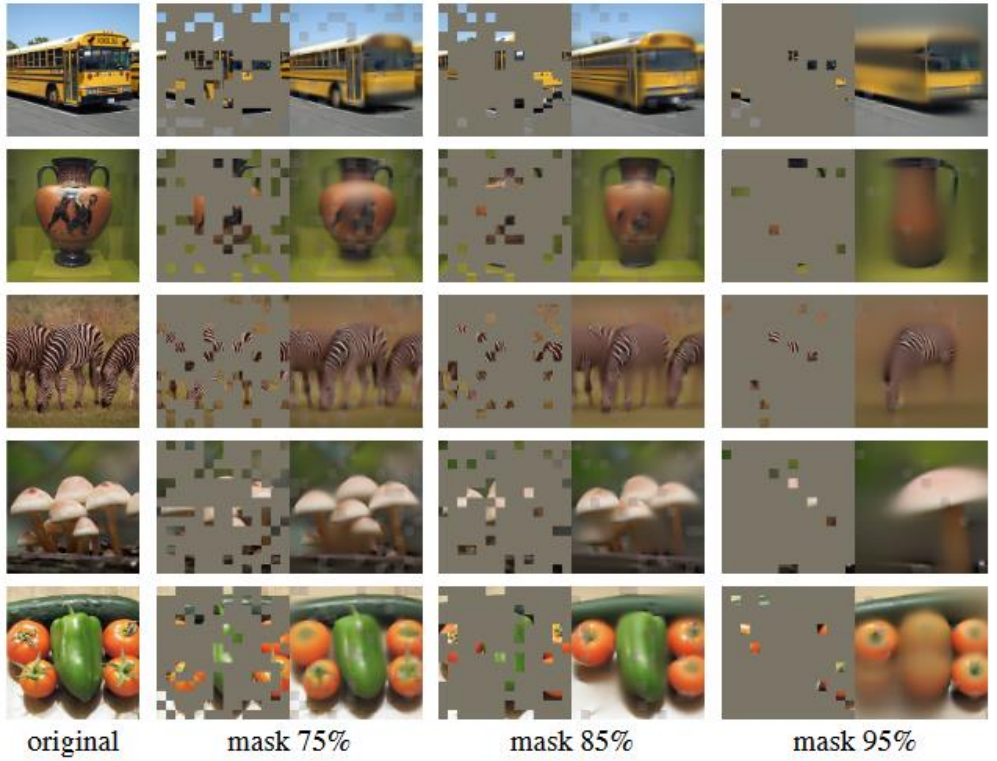


Foundational Models Defining a New Era in Vision: A Survey and Outlook

传统预训练模型

MAE：





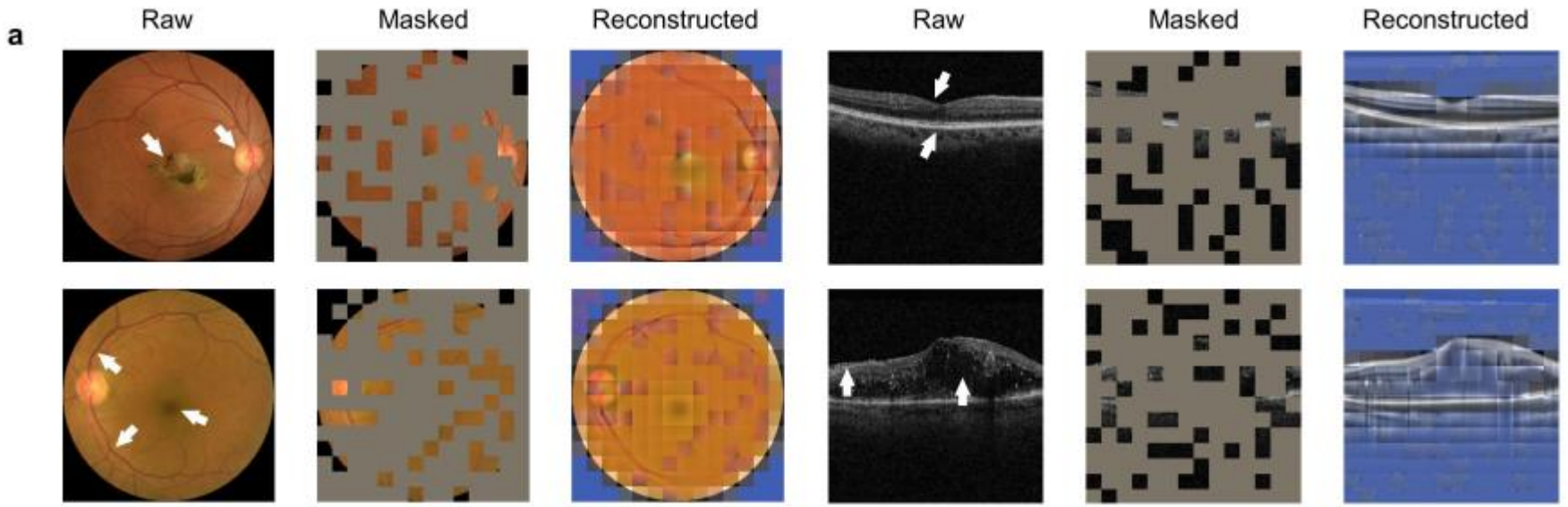
Masked Autoencoder使用了掩码机制，利用编码器将像素信息映射为语义空间中的特征向量，而使用解码器重构原始空间中的像素。

MAE使用的是非对称的Encoder-Decoder架构，即编码器只能看到未被遮蔽的部分像素块信息，以节省计算开销，而解码器解码的是所有像素块的特征信息。

传统预训练模型

MAE in Medical：

RETFound

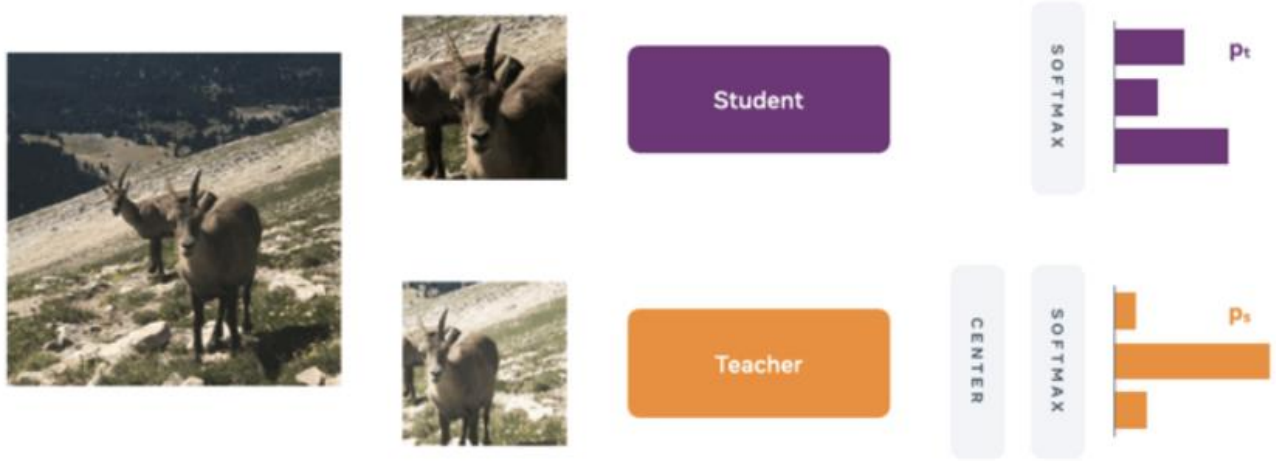


数据量：
90w, 70w

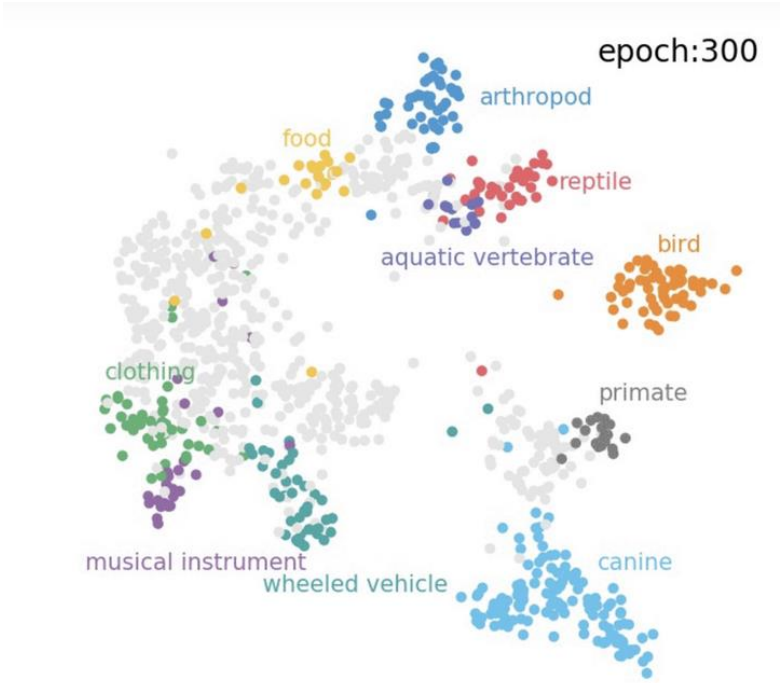
从个人感觉来说，MAE的效果同时受到预训练图像数量和预训练时间的影响，并且受到预训练时间的影响更大。少量的数据只需要训练充分，就能够实现与大量数据相同的结果。LLM里面的一些经验是：同时增加模型参数和训练数据量，此时模型效果最佳，训练时间一般采用Early Stop的方式。但是在视觉领域，模型训练得充分是一个相对来说更加重要的问题。可能是因为MAE这种视觉表示学习的方法还不够成熟。

传统预训练模型

DINO：



无监督下的聚类效果：



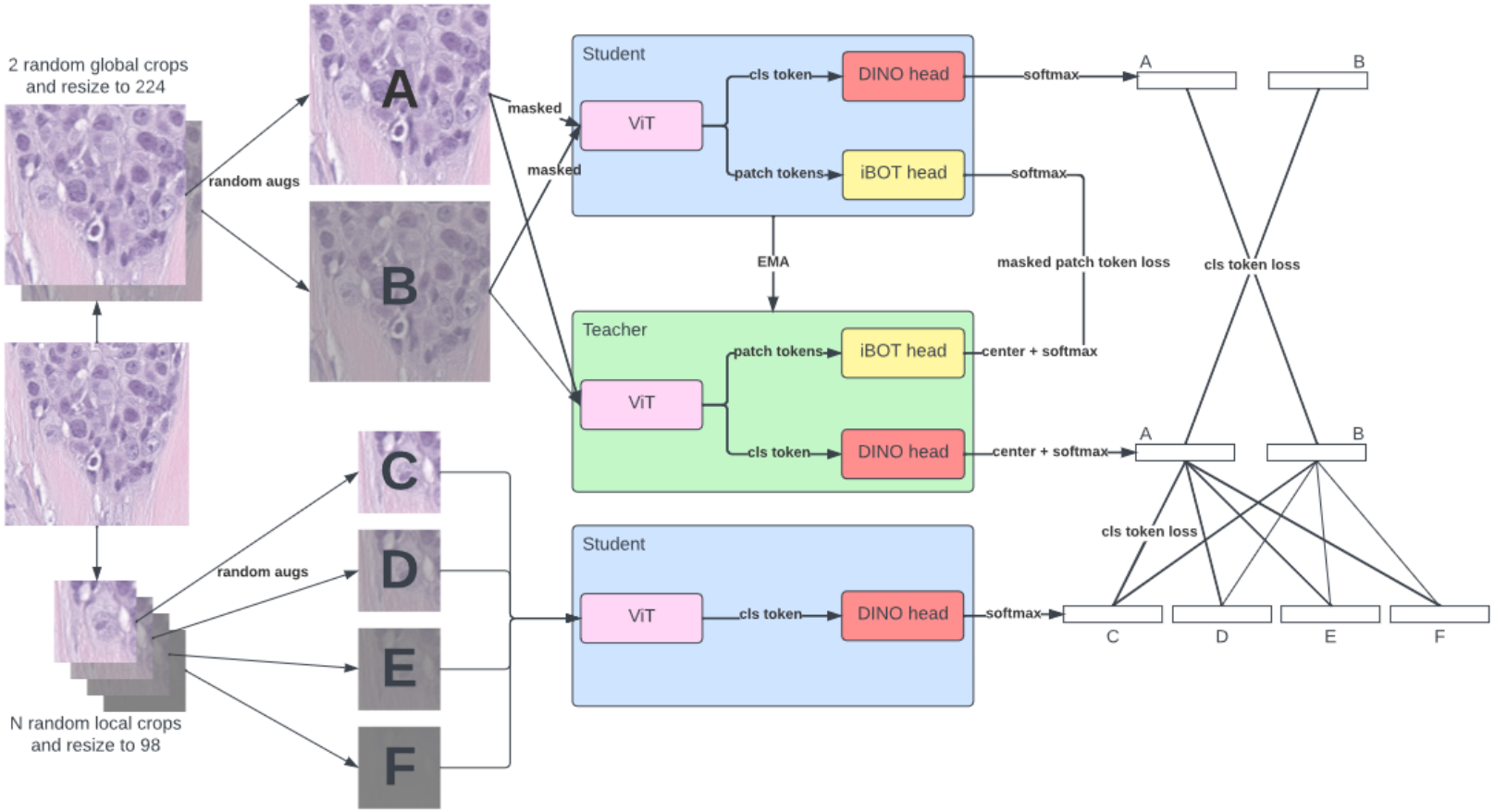
DINO 采用一种称为自蒸馏的方法。自蒸馏创造了一个教师和一个学生网络。 这两个网络都具有完全相同的模型架构。

训练过程：一张图片被裁剪成两种尺寸，然后输入学生和教师网络。 对教师的输出应用居中操作，并且两个输出都通过 softmax 层归一化整理。为了交叉熵作为损失函数为模型反向传播提供更新参数的策略。反向传播是通过学生网络执行的，这时教师的权重尚未更新的原因。 为了更新教师模型，DINO 对学生权重使用指数移动平均 (EMA)，将学生网络的模型参数传输到教师网络。

传统预训练模型

DINO in Medical：

VIRCHOW: A MILLION-SLIDE DIGITAL PATHOLOGY FOUNDATION MODEL

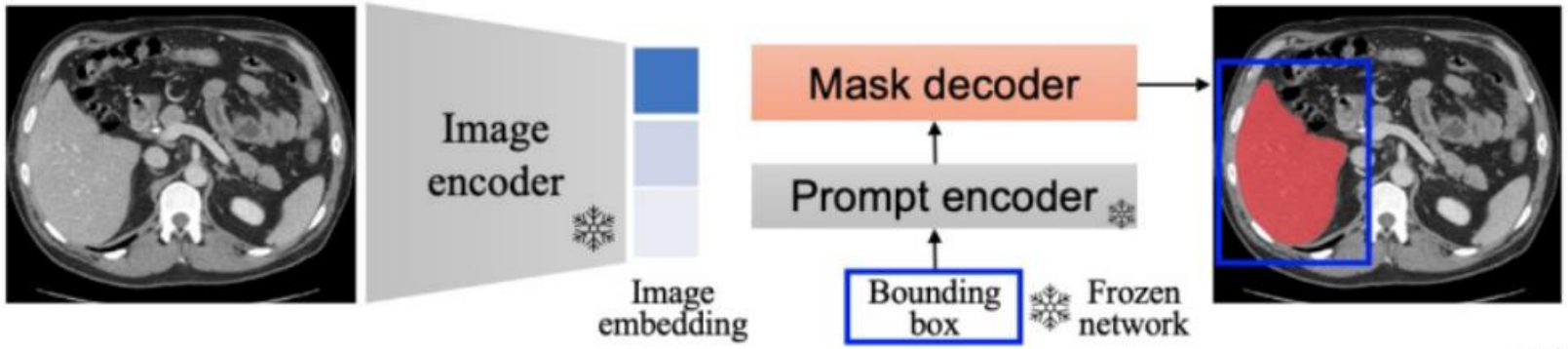


1,488,550 WSIs derived from 119,629 patients.

Visually Prompted Models

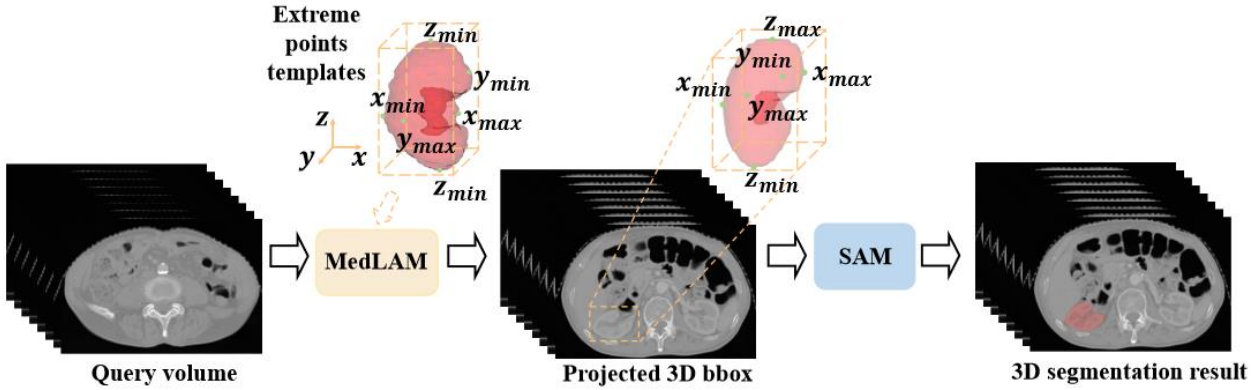
SAM in Medical：

Segment Anything in Medical Images



1,090,486 medical image-mask pairs, covering 15 imaging modalities, over 30 cancer types

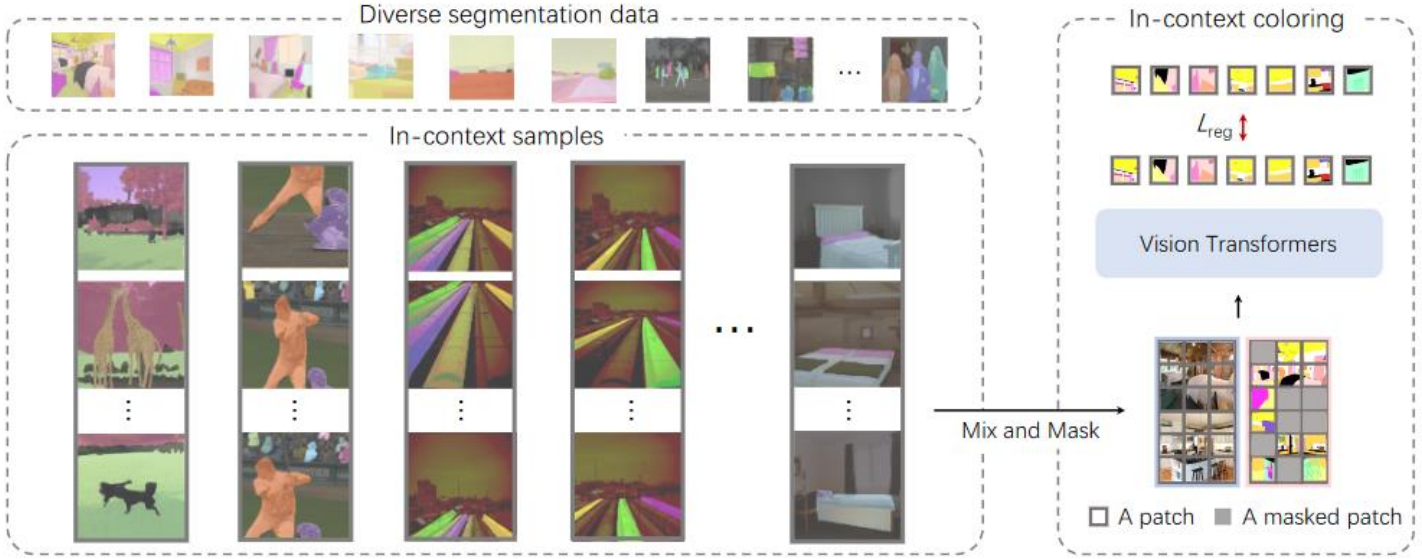
MedLSAM: Localize and Segment Anything Model for 3D CT Images



避免对3D数据的逐帧处理，先定位，然后根据定位结果利用MedSAM和SAM来生成分割结果

Visually Prompted Models

SegGPT：



即使不训练，也能达到非常好的效果

method	venue	mIoU one-shot	mIoU few-shot
trained on FSS-1000			
DAN [43]	ECCV'20	85.2	88.1
HSNet [35]	ICCV'21	86.5	88.5
SSP [15]	ECCV'22	87.3	88.6
VAT [19]	ECCV'22	90.3	90.8
DACM [50]	ECCV'22	90.8	91.7
not trained on FSS-1000			
Painter	CVPR'23	61.7	62.3
SegGPT	this work	85.6	89.3

SegGPT 将各种分割任务统一成一个通用的上下文学习框架，可用于分割上下文中的所有事物。SegGPT 的训练被制定为一个上下文着色问题，为每个数据样本随机分配颜色映射。目标是根据上下文完成不同的分割任务，而不是依赖于特定的颜色。

与SAM的不同在于，SAM类似于Zero Shot的分割，Prompt是交互式提供的。SegGPT是Few Shot的分割，需要提供示例说明任务。

Visually Prompted Models

SegGPT in medical：

CHASE_DB1 [16], **DRIVE** [40], **HRF** [4] and **STARE** [20] provide annotations for retinal vessel segmentation. We augment the high-resolution raw images with random cropping.

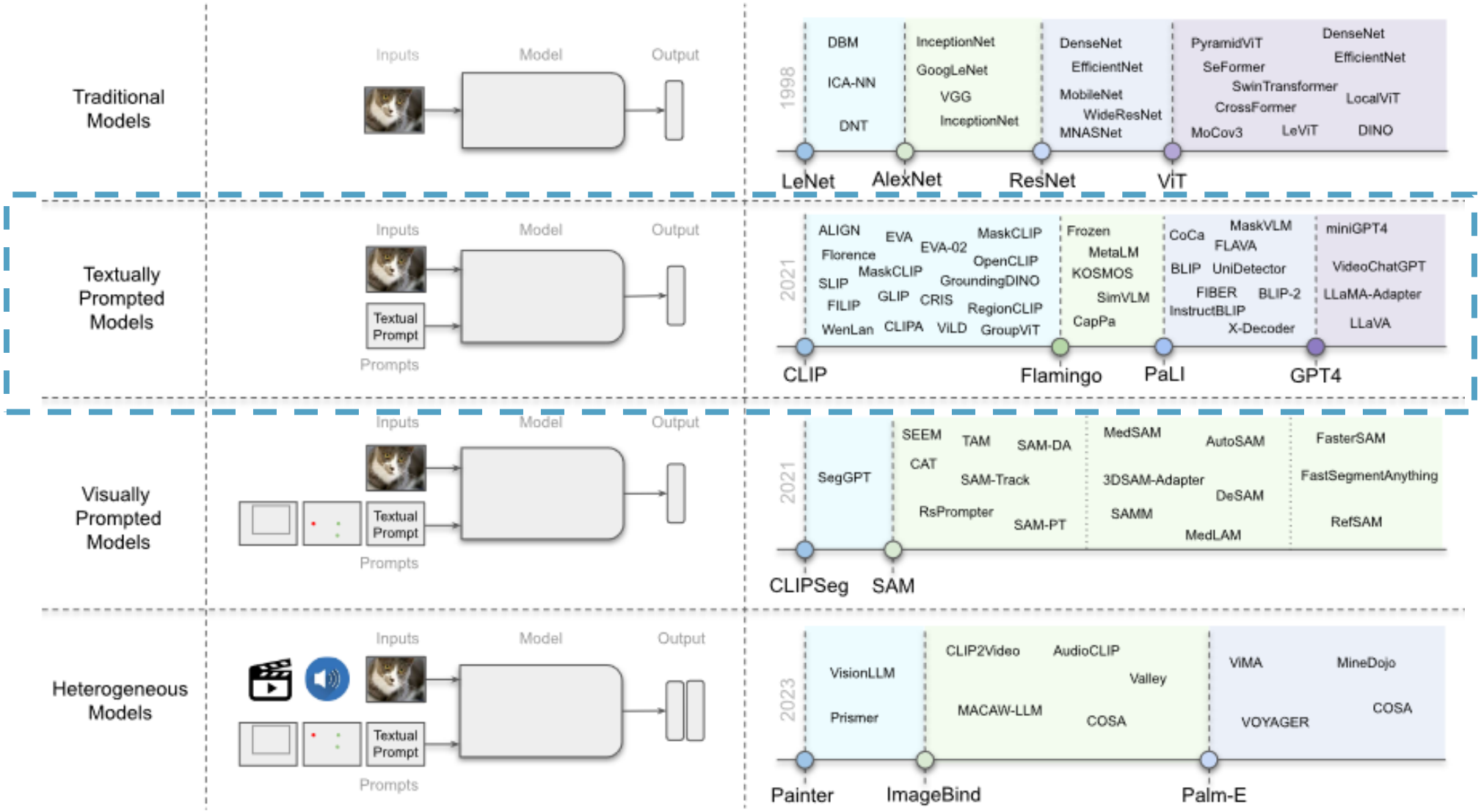
SegGPT本身使用了一些眼底图像进行训练。

但是目前还没有专门针对医疗图像上的SegGPT

Part 3

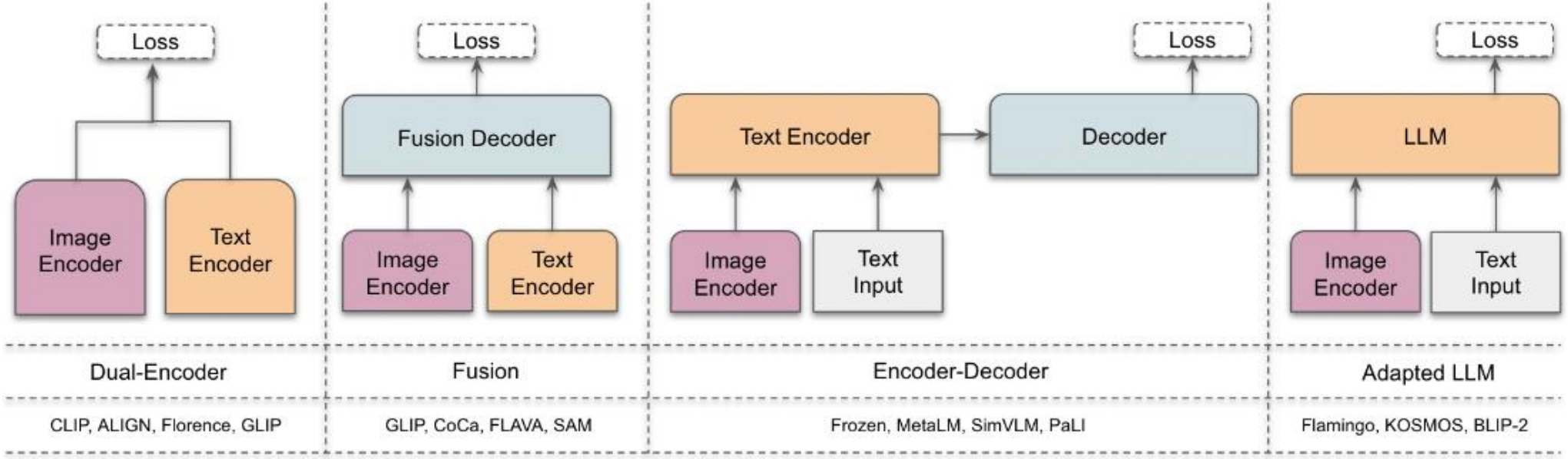
Vision Language Model

Textually Prompted Models



Textually Prompted Models

关于图像-文本模态对齐的集中结构：



训练损失：

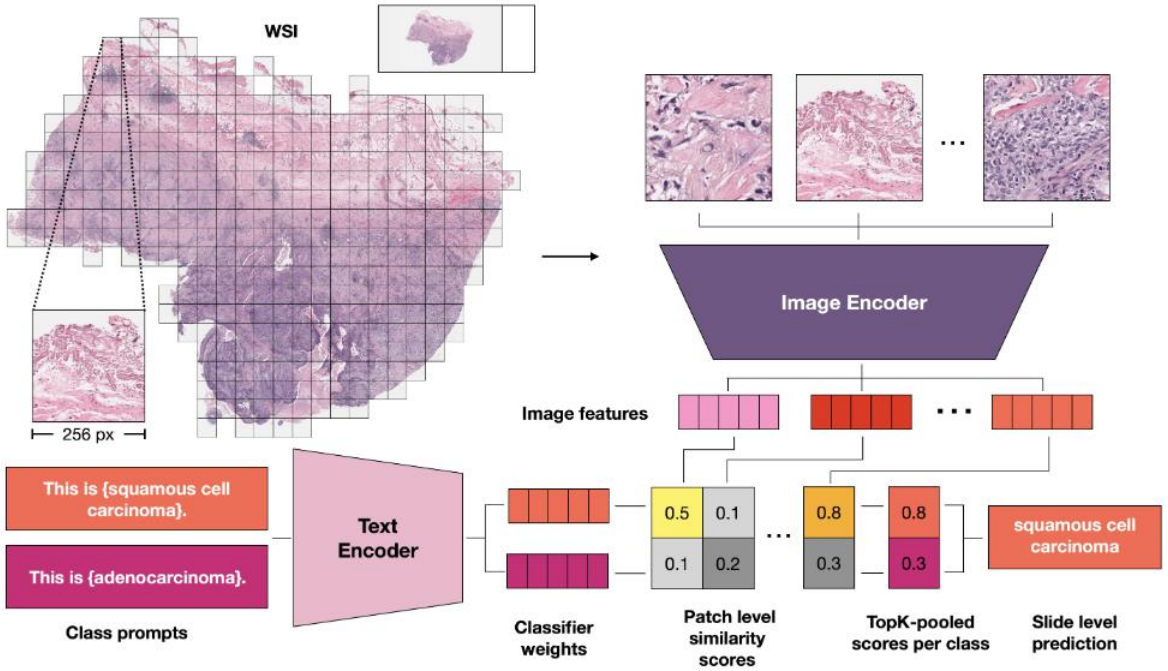
1、对比学习

2、生成式方法（包括Masked Language Model）

3、混合式

Textually Prompted Models

CLIP in medical：



其他工作：

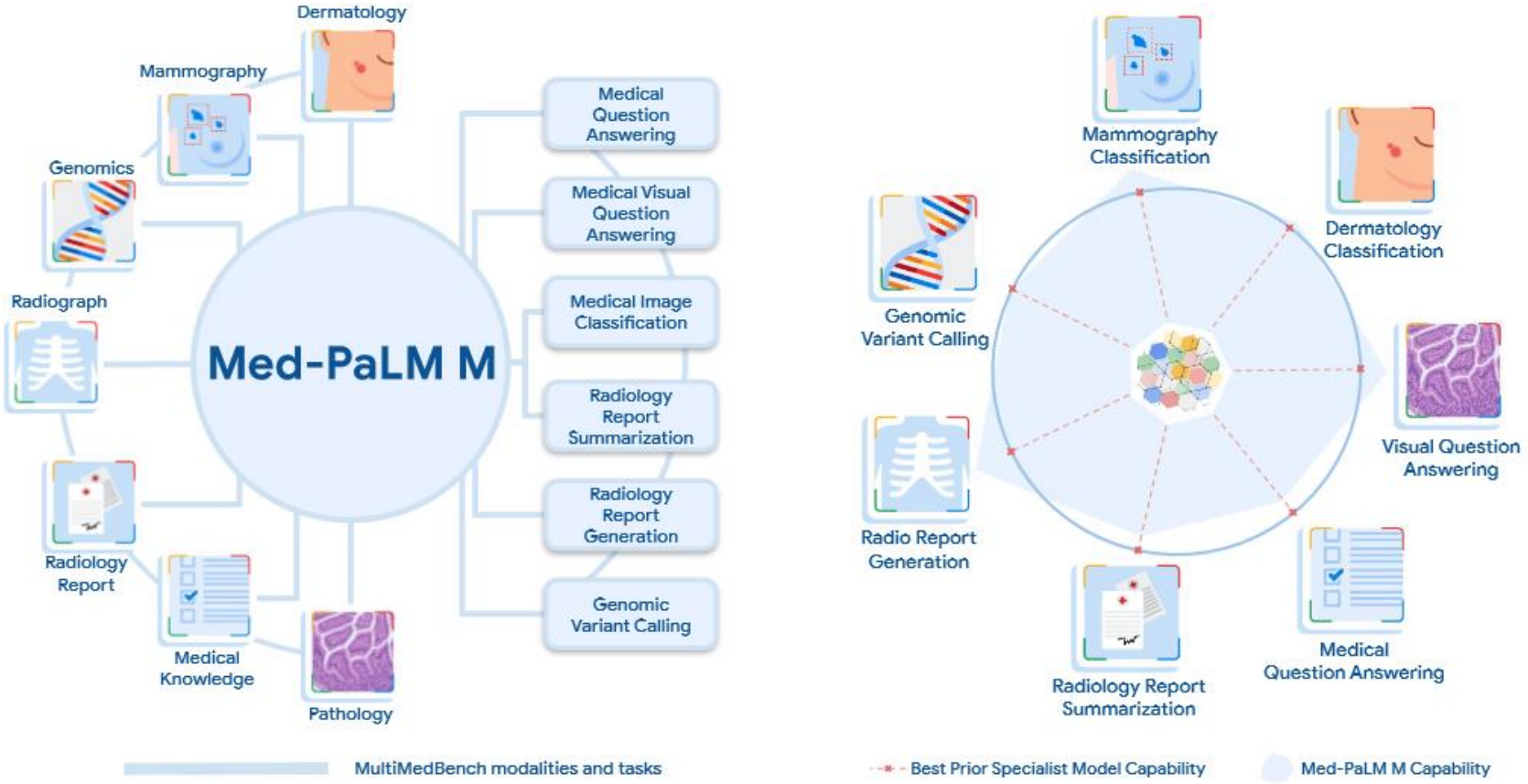
- **MedCLIP: Contrastive Learning from Unpaired Medical Images and Text**
- **LARGE-SCALE DOMAIN-SPECIFIC PRETRAINING FOR BIOMEDICAL VISION-LANGUAGE PROCESSING**

针对潜空间中的每个文本提示，独立计算每个实例在补丁级的余弦相似度得分。在测试时，合并不同实例的分数，生成最终分数。在三个不同的真实世界癌症分型任务中进行测试时，MI-Zero 的表现要么与基线相当，要么优于基线，达到了 70.2% 的平均中位零点准确率。

Textually Prompted Models

Generalist Model in Medical：

Towards Generalist Biomedical AI（DeepMind） **第一个医学领域的通用基础模型**



Textually Prompted Models

Generalist Model in Medical：

BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks

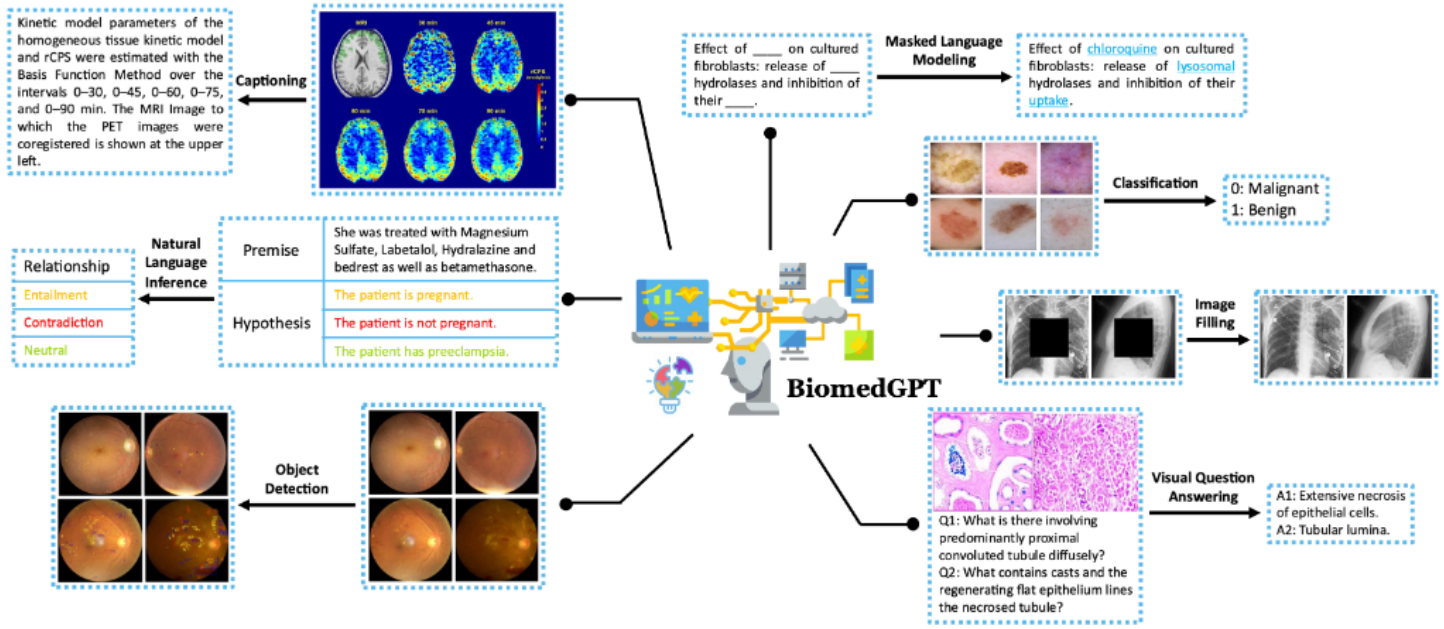


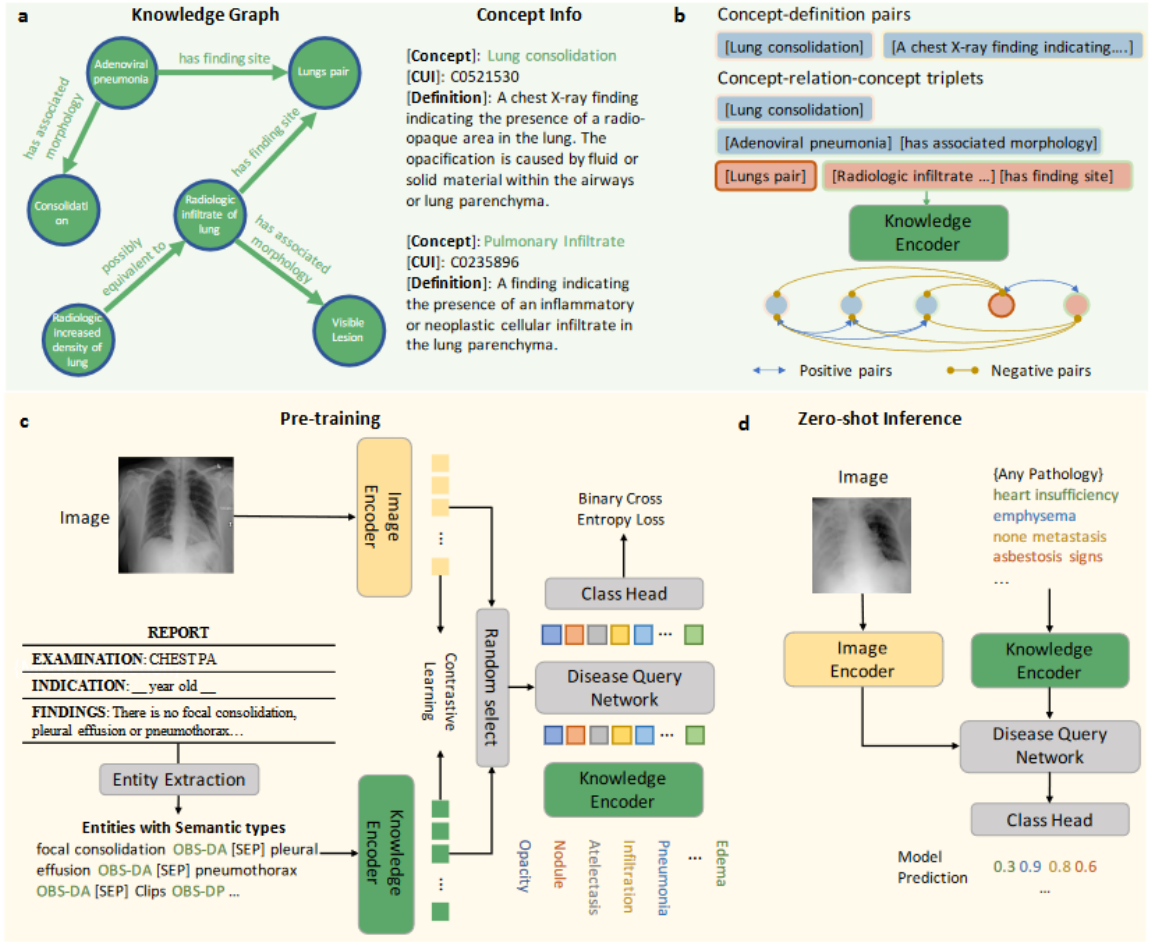
Figure 1: Illustration of the diverse range of tasks supported by BiomedGPT during pretraining and subsequent fine-tuning. During the pretraining phase, we employ prevalent unimodal strategies, including masked language modeling and masked image infilling, and multimodal techniques, such as visual question answering and captioning. Object detection is also incorporated into the pretraining to infuse locational data. Following pretraining, the enhanced model is leveraged for a suite of five downstream tasks, encompassing image classification and natural language inference, demonstrating its efficient utilization of data.

BiomedGPT是一个统一的框架，可通过多种模式（包括射线照片、数字图像和文本）进行训练，以执行生物医学领域的各种任务。

Textually Prompted Models

Knowledge-Based VL Model in Medical：

Knowledge-enhanced Visual-Language Pre-training on Chest Radiology Images, NC2023



本文的目标是通过对成对的图像和报告进行训练来构建胸部x光片的基础模型，称为知识增强自动诊断(KAD)。如图1所示，本文明确地利用成熟的医学知识图来训练知识编码器。具体来说，所提出的 KAD 遵循两阶段框架：

- 首先，学习知识图谱的神经表示，实体表示为节点，它们之间的关系为边，为实体的结构化、多步推理提供支架；
- 其次，通过启发式地定义的规则或使用 ChatGPT 从放射学报告中提取临床实体和关系，
- 然后，利用预训练的知识编码器来指导使用图像和放射学报告的视觉表示学习，有效地将领域知识注入视觉编码器

Part 4

Future Work

挑战

Multimodal Model

基础模型在医学成像中的未来方向具有很大的前景，主要是由于它们无缝集成了不同的数据模式。这种集成创造了在多个尺度上探索医学概念的机会，并利用各种知识源的见解，包括成像、文本和音频数据。这种多模态集成使得仅使用单模态数据难以实现的医学发现成为可能，同时也促进了跨域的知识转移。

衍生出挑战是：

1、跨膜态配对的数据集
2、图像模型和语言模型之间的对齐

挑战

Extensive Data and Computational Demands

基础模型虽然强大，但具有开发、培训和部署的大量计算成本。在特定情况下，较小的模型可以以更低的成本实现相似或更好的结果。

训练大规模模型是数据和计算密集型的，获取广泛的标记数据可能既昂贵又耗时，尤其是对于资源较少专业领域。更大的数据集通常会导致模型性能和泛化能力的提高，但由于隐私问题和医疗数据注释的劳动密集型性质，它们在医学领域很难获得。未来的研究应该探索高效数据集收集、增强和利用的技术，从而能够开发能够在保持患者隐私的同时在有限的数据下表现良好的模型。

另一方面，由于参数众多，使用这些模型进行推理也很昂贵。这些计算需求阻碍了它们在现实应用中的实用性，特别是那些需要实时推理或运行在资源受限的边缘和移动设备上的人。例如，基于视觉提示的模型，如 Segment Anything[10]，虽然具有健壮的图像编码器，但目前缺乏实时处理速度，这是实际使用的关键要求。

另一方面，FastSAM 实现了与 SAM 方法相当的性能，但通过将 Transformer 架构替换为 YOLOv8-seg，运行速度提高了 50 倍，显着扩展了此类模型在现实场景中的效用。

因此，有可能开发更有效的后继来解决这个问题，特别是在边缘设备上运行模型提供实质性优势的医疗应用中，特别是在服务不足的领域。

挑战

Interpretablity

了解模型的能力、推理和机制提供了对其输出的深刻见解。在医疗保健中，可解释性对于患者症状、临床试验和知情同意的决定至关重要。透明的 AI 推理有助于解决 AI 系统和人类专家之间的分歧，解释创建决策背后的原因。然而，目前大多数基础模型缺乏内置的可解释性，需要未来的研究。通过将 AI 输出与医学知识联系起来，模型变得更加易于理解，使用户能够不仅掌握模型预测什么，还可以掌握为什么。**这种跨学科的方法，将 AI 与领域专业知识相结合**，促进疾病理解，提高患者护理，并促进医疗保健中负责任的 AI 使用。

Bias and Variance in Foundational Model

偏见：基础模型面临的主要挑战之一是数据和预测中都存在偏差。与视觉和语言模型一样，医学成像的基础模型可以继承和放大训练数据中存在的偏差。这些偏见可能与种族、民族、性别或社会经济因素有关，它们可以体现在模型的预测和行为中。因此，解决和减轻基础模型偏差对于确保医学领域的公平性、包容性和伦理部署至关重要。

方差：方差与模型对训练数据波动的敏感性有关。在医学成像的背景下，方差可以表现为模型无法在不同的患者群体或不同的医疗保健环境中有效地进行泛化。高方差的模型在一个数据集上的表现可能非常好，但在另一个数据集上表现不佳，阻碍了它们在现实世界临床应用中的可靠性。因此，提高基础模型鲁棒性和泛化能力的策略对于广泛采用和实用性至关重要。

挑战

Prompt Engineering

提示工程是医学成像中基础模型的一个关键方面，其重要性在于它有可能弥合这些模型和医生之间的差距，最终增强患者护理。

在医学图像解释的背景下，医生和AI模型之间的有效通信可以带来几个值得注意的好处。

- 首先，提示工程允许医生与 AI 模型具有自然和交互式的对话。这种能力特别有价值，因为它使医生能够寻求澄清、提供额外的上下文并提出后续问题，反映现实世界的临床场景。例如，在回顾复杂的医学图像时，医生可能需要要求 AI 模型进一步解释其发现、请求替代视图或探索差异诊断。提示工程促进了这种对话流程，使 AI 模型更易于放射科医生访问和协作工具。

- 通过良好构建的提示与 AI 模型交谈的能力使医生具有更具交互性和直观的工作流程。医生不仅可以仅依靠固定的查询或预定义的提示，还可以根据每种情况的具体细微差别来调整他们的交互。这种适应性允许更动态和个性化的用户体验，最终提高诊断精度和效率。

- 提示工程有助于 AI 模型的可解释性和透明度。医生可以通过制作引发详细解释的提示来深入了解模型如何得出其结论。这种透明度在临床环境中至关重要，放射科医生需要了解 AI 模型推荐背后的推理并信任其诊断见解。

挑战

Enhancing Feature Representation from Frequency Perspective

鉴于大多数基础模型使用 ViT 模型作为主干，从频率的角度评估这些模型以确保它们能够捕获和学习对象识别所需的不同频率信息变得至关重要。最近的研究阐明了ViT中传统的自我注意机制虽然能有效地减轻局部特征差异，但往往忽略了重要的高频细节，如纹理和边缘特征。这种疏忽在肿瘤检测、通过放射组学分析和治疗反应评估等任务中尤其成问题，因为这些任务通常依赖于识别细微的纹理异常。鉴于这些考虑，新基础模型的设计应考虑这些限制并探索潜在的增强。这可能涉及合并 CNN 层或采用更有效的 ViT 架构在计算效率和保留高频信息之间取得平衡。

THANKS !