【DyFo：一种无需训练的动态聚焦视觉搜索方法，用于增强大语言模型的细粒度视觉理解能力】

——【DyFo: A Training-Free Dynamic Focus Visual Search for Enhancing LMMs in Fine-Grained Visual Understanding】

1 相关资源

pdf: https://arxiv.org/pdf/2504.14920v1
ppt:
短视频:
数据集:
源码: https://github.com/PKU- ICST- MIPL/DyFo_CVPR2025
网站:
【除了网站，其他资源尽量下载】

2 论文属性

论文来源: CVPR 2025
【给出具体会议名称和年限，不要仅仅写 ACM，IEEE】
论文类别: Large Language Model
【论文的类别，比如移动计算、轨迹处理、深度学习等】
论文关键字: LLMs, Hallucination Mitigation
推荐程度：3 （其他说明可标注）
(5 非常棒，建议认真研读、小组讨论和复现；4 好，建议细读，考虑复现；3 可以，部分内容值得注意；2 一般，简单浏览即可；1 没有意义，不建议阅读)

3 工作团队

作者: Geng Li, Jinglin Xu, Yunzhen Zhao, Yuxin Peng
单位:

1. Wangxuan Institute of Computer Technology, Peking University
2. School of Intelligence Science and Technology, University of Science and Technology Beijing
3. Tencent Beijing Research, Beijing, 100193, China

团队情况描述:

4 论文介绍

(1) 研究目的

【研究背景是什么？本文工作有什么用?】

大型视觉-语言模型（LVLM）已能在图像描述、视觉问答等任务中生成流畅且上下文相关的文本，但在医疗、自动驾驶、机器人等高可靠性场景中，模型经常“脑补”出图像中并不存在的物体（即对象幻觉），导致错误决策。

本文旨在通过模拟人类的视觉搜索过程，提出一种训练无关的动态聚焦视觉搜索方法（DyFo），以增强 LMMs 在细粒度视觉理解任务中的表现，减少幻觉现象。

(2) 研究现状

【当前的最好研究做到什么程度了？存在的问题是什么？这里采信论文的说法，可以给出自己的评点】

当前，减少 LMMs 幻觉的研究已经取得了一定的进展。现有的方法主要分为三类：后处理当前，多模态模型在视觉和语言任务中取得了显著进展，但仍然面临一些挑战。例如，固定分辨率的编码器（如 ViT）在处理高分辨率图像时会丢失细节，而动态分辨率模型虽然能够处理高分辨率图像，但在复杂场景中容易引入大量无关信息，导致幻觉问题加剧。一些研究通过引入区域边界框或掩码来增强局部区域理解，但这需要额外的训练和用户提供的空间提示，限制了其在实际应用中的可行性。此外，现有的视觉搜索方法（如 SEAL）虽然通过结合定位模块和视觉记忆实现了视觉搜索功能，但需要对 LMM 进行额外的词汇扩展和模块依赖，增加了训练成本和模型复杂性。总体来看，尽管现有研究在多模态视觉理解方面取得了进展，但在细粒度视觉任务中，如何在不增加额外训练负担的情况下有效聚焦关键视觉信息仍然是一个亟待解决的问题。

(3) 本文解决的问题

【一句话概括本文解决的核心问题】

本文解决的核心问题是减少大型多模态模型在生成文本时的幻觉现象，同时增强其在细粒度视觉理解任务中的表现。

(4) 创新与优势

【本文的创新之处是什么？新场景？新发现？新视角？新方法？请明确指出】

【本文工作的贡献或优点是什么?】

1. 提出了 DyFo，这是一种无需训练的视觉搜索方法，它显著改进了细粒度视觉下的
2. 提出了一种基于蒙特卡洛树搜索（MCTS）的协作范式，无需进行昂贵的大语言模型（LMM）训练或架构更改，就能使用高性能的视觉专家模型。
3. 在一般幻觉和细粒度视觉理解任务中验证了方法的优越性能，展示了其增强视觉能力的潜力。

(5) 解决思路

【本文是怎样解决问题的？包括方法、技术、模型等，以自己理解的方式表述清楚】

视觉搜索

一个具有参数　的大型多模态模型（LMM）接收图像 $I \in \mathbb{R}^{H\times W\times 3}$ 和文本 $X = [x_1, ..., x_n]$ 作为输入，其中文本包含 n 个标记，并输出具有 m 个标记的文本 $Y = [y_1, ..., y_m]$。在生成输出文本 Y 时，模型根据输入 I 和 X 的概率分布进行自回归采样。在时间步 t，标记 y_t 的概率为

$$y_t \sim p_\theta(y_t|I, X, y_{<t}) \propto explogit_\theta(y_t|I, X, y_{<t}) \tag{1}$$

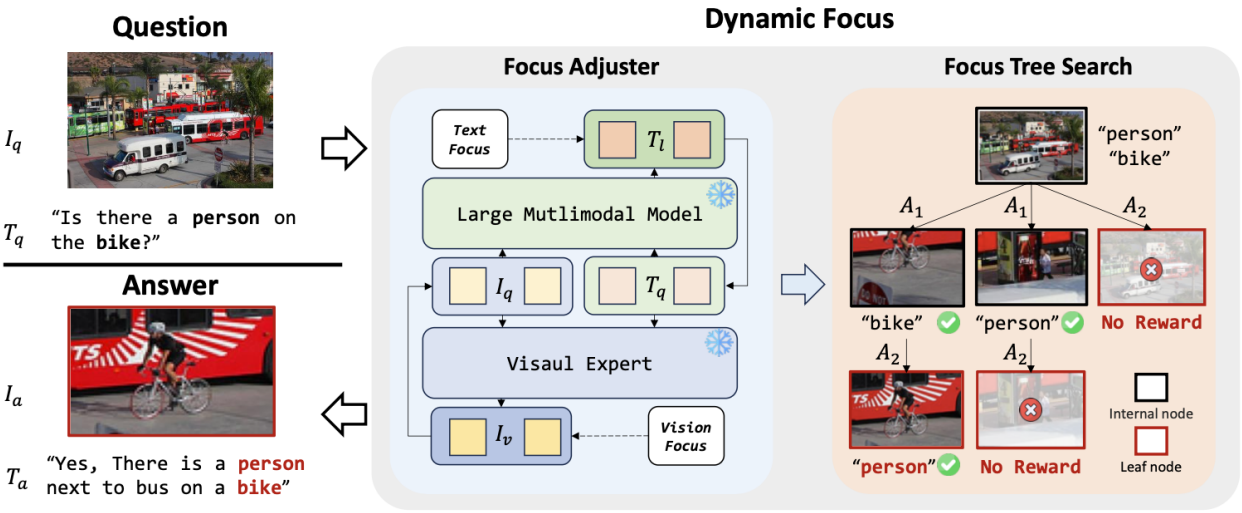

图 1: An illustration of DyFo framework.

其中 $logit_\theta$ 是标记 y_t 的非归一化对数概率，即 softmax 之前的值。

视觉搜索的目标是找到对应的图像区域 $I \in \mathbb{R}^{H' \times W' \times 3}$，其中 $H' < H$ 和 $W' < W$，从而使 $logit_\theta(\hat{y}_t|I, X, y_{<t})$ 尽可能高。$\hat{y}_t$ 是正确答案的标记。

最近的研究引入了定位模块，该模块输出热力图和目标框，以突出显示特定区域，从而增强模型的视觉定位能力。然而，这种视觉搜索方法依赖于带有额外定位模块的大语言模的特定架构，并且在定位数据收集和大语言模型微调方面引入了大量成本，阻碍了其向 Qwen2-VL 等其他模型的迁移性。

动态聚焦视觉搜索

作者提出了一种无需训练的视觉搜索方法，名为动态聚焦 DyFo，旨在通过与视觉专家协作，帮助大型多模态模型聚焦于图像的特定区域。具体而言，采用蒙特卡洛树搜索 MCTS 算法，将聚焦区域和文本视为节点，以模拟类人的、基于聚焦的搜索过程。在此过程中，LMM 的文本输出引导视觉专家检索相关图像区域，而视觉专家的图像输出反过来指导 LMM 调整其文本聚焦，如图所示。

聚焦调节器

为了克服这些局限性，作者提出了一种聚焦调节器，这是一种由动作指令引导的协作机制，它结合了大语言模型和视觉专家的优势。通过实现文本和视觉层面的相互增强，聚焦调节器将大语言模型的多模态理解能力与视觉专家的精确性相融合，显著提升了细粒度视觉搜索能力。如图所示，这种整合在视觉搜索过程中得以实现。

将焦点定义为 $f = (I, T)$，其中 $I \in I_o$ 是从原始图像 I_o 中选取的相关图像区域，T 代表与 I 相关联的主要语义线索。焦点 f 的迭代更新过程如下：

$$\begin{cases} T^{i+1} = L\left(f^i, A^i\right) \\ I^{i+1} = E\left(T^{i+1}, I^i, A^i\right), \\ f^{i+1} = \left(I^{i+1}, T^{i+1}\right), \end{cases} \tag{2}$$

其中动作 A^i 是指导焦点调整的外部指令。大语言模型 L 根据当前焦点 f^i 和动作 A^i 优化文本焦点，而视觉专家 E 则根据优化后的文本线索 T^{i+1} 更新图像区域。这一迭代过程使 A 能够动态指导视觉和文本的更新，从而实现动态焦点调整。

为了模拟人类的注意力转移行为，我们设计了一个动作空间，用于引导大语言模型和视觉专家进行

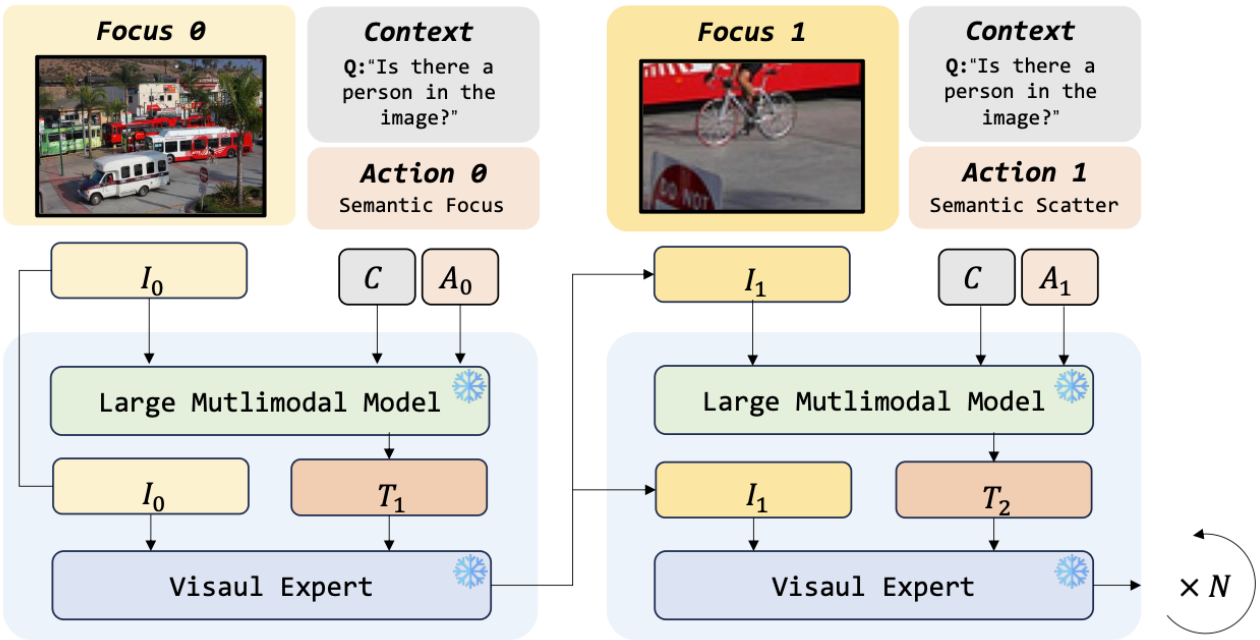

图 2: An illustration of Focus Adjuster of DyFo.

协同交互:

A1: 语义焦点。此动作基于语义信息（例如，与查询相关的对象）识别上下文相关的视觉目标，模拟联合搜索和内源性定向 [2]，以根据语义线索匹配目标。

A2: 语义分散。该动作会扩大当前的聚焦区域，避免因过度精确聚焦而导致关键信息丢失，模拟人眼焦点的发散动作。

焦点树搜索

作者提出了聚焦树搜 FTS，这是一种基于蒙特卡洛树搜索 MCTS 的视觉搜索方法，该方法能够导航通过具有焦点节点的视觉空间进行导航。FTS 有效地结合了焦点调整动作，平衡探索与利用，构建了一棵保留关键视觉信息同时消除无关细节的焦点树。

在此方法中，FTS 逐步构建一棵焦点树，其中每个节点代表一个焦点状态 f，每条边对应一个焦点快速切换动作，该动作会转换到新的焦点区域。为了有效地引导大型多模态模型和视觉专家探索最有前景的节点，维护了一个状态-动作价值函数 $Q : f \times A \mapsto \mathbb{R}$，该函数估计在焦点状态 f 中执行动作 a 的预期未来奖励。

选择阶段: 该算法从根节点（初始焦点状态 f^0）开始，通过平衡探索与利用，在每个树层级上迭代选择下一个节点。此阶段持续进行，直至到达叶节点。为了在探索（访问较少被访问的节点）和利用（访问高价值节点）之间取得平衡，采用成熟的树的上置信界来选择每个子节点。具体而言，在节点 f 处，通过同时考虑 Q 值（用于利用）和不确定性（用于探索）来选择一个动作:

$$a^* = argmax_{a \in A(f)} \left[Q(f,a) + w\sqrt{\frac{lnN(f)}{N(c(f,a))}} \right] \tag{3}$$

其中，$N(f)$ 表示在先前迭代中对节点 f 的访问次数，而 $c(f,a)$ 是在状态 f 中执行动作 a 后产生的子节点。第二项随着不确定性（对子节点的访问次数较少）而增大。权重 u 控制着探索与利用之间的权衡。

扩展阶段: 在此阶段，通过随机采样未探索的动作，将新的子节点添加到选定的叶节点。如果叶节点是终端节点（即搜索次数已达到最大值），则跳过扩展阶段，立即开始反向传播。

反向传播阶段: 一旦在上述阶段达到终端状态，就会获得从根节点到终端节点的搜索路径。此时，奖励会沿着该路径反向传播，以更新路径上每个状态-动作对的 Q 值。具体来说，$Q(f,a)$ 是向上的。通过聚合节点 f 所有后续步骤的奖励来确定日期，如下所示：

$$Q(f,a) = \sum_{f^i \in N_f} R_{f^i} \tag{4}$$

其中，N_f 是节点 f 的子节点集合，R_f 是每个子节点对应的奖励值。为了鼓励大语言模型和视觉专家过滤掉无关内容，同时避免引入偏差（例如，视觉专家定位错误目标或大语言模型产生视觉幻觉），将大语言模型和视觉专家之间的一致性作为评判标准，并将节点的有效区域作为搜索奖励：

$$R_{f^i} = \mathbb{I}_{\{I=T\}} \cdot \frac{s_{f^i}}{s_o}, \tag{5}$$

其中，R_{fi} 表示奖励函数的值。$\mathbb{I}_{I=T}$ 是一个指示函数，如果与节点 f^i 相关联的图像 I 在语义上一致，且文本为 T 时，则该函数取值为 1，否则为 0。$\frac{s_{fi}}{s_0}$ 表示节点 f^i 的有效面积比，其中 s_{fi} 是有效面积，s_0 是原始输入图像的面积。

多粒度投票: 完成上述阶段并构建聚焦树后，目标是充分利用该树中关键的视觉信息，以实现更精确的细粒度视觉理解。受自一致性方法的启发，作者采用多节点投票方式来获取最终答案。具体而言，最终结果是通过不同节点间的加权投票方案得出的，其中每个节点 f 会提供预测 $L(f)$，而每个节点的权重则是其获得的奖励 R_f。形式上，最终答案 A 的推导如下：

$$\mathcal{A} = argmax_{L(f)} \sum_{f \in \mathcal{T}_f} R_f \cdot L(f) \tag{6}$$

其中 T_f 是所构建的聚焦树。这种投票机制确保模型避免过度强调可能导致关键全球线索丢失，同时又能保留相关细节重要性的线索。

(6) 可改进的地方

【本文工作的局限性是什么？你觉得可以从哪些方面改进工作?】

虽然 MCTS 在搜索效率上表现出色，但在面对更复杂的视觉场景时，其搜索空间可能会迅速膨胀，导致计算负担加重。

(7) 可借鉴的地方

【你觉得本文哪些方面可以借鉴？比如思路、方法、技术等】

首先，其提出的动态聚焦视觉搜索思路为解决 LMMs 在复杂视觉任务中的幻觉问题提供了一种新的视角，这种通过模拟人类视觉搜索机制来优化模型注意力分配的方法可以为其他多模态任务提供参考。其次，基于 MCTS 的焦点树搜索方法为高效探索视觉空间提供了一种有效的技术手段，这种平衡探索与利用的搜索策略可以应用于其他需要在大规模数据中寻找最优解的场景。。

(8) 其他收获

【你有什么其他收获吗？比如了解了哪些团队和大牛在某领域做得很好，某类问题通常用什么技术解决，某些技术之间存在什么样的关联，某些会议和期刊在某领域很知名……】

我了解到蒙特卡洛树搜索在视觉搜索中的应用潜力，这种搜索算法通过模拟人类决策过程中的探索与利用平衡，为解决复杂的视觉问题提供了新的思路。

5 评阅人

姓名:
时间: