澳洲幸运5官方开奖历史记录

EMCm7DuGMf9lBRLV

澳洲幸运5官方开奖历史记录首个动态视觉-文本稀疏化框架来了，计算开销直降50%-75%

本文由华东师范大学和小红书联合完成，共同第一作者是华东师范大学在读硕士、小红书 NLP 团队实习生黄文轩和翟子杰，通讯作者是小红书 NLP 团队负责人曹绍升，以及华东师范大学林绍辉研究员。

多模态大模型（MLLMs）在视觉理解与推理等领域取得了显著成就。然而，随着解码（decoding）阶段不断生成新的 token，推理过程的计算复杂度和 GPU 显存占用逐渐增加，这导致了多模态大模型推理效率的降低。现有的方法通过减少预填充（prefill）阶段的视觉 token 冗余来实现推理加速。遗憾的是，这种在预填充阶段实现的视觉 token 稀疏化所带来的加速优势，在解码阶段会逐渐减弱。当解码输出的文本 token 数量增多时，这些方法仍然会遇到性能瓶颈。

为了解决上述问题，团队创新性地提出了一个全新的动态视觉 - 文本上下文稀疏化推理加速框架 —— Dynamic-LLaVA。该框架针对多模态大模型在不同推理模式下（包括预填充阶段以及有无 KV Cache 的解码阶段），设计了定制化的稀疏化推理方案，以实现多模态大模型的高效推理。实验结果表明，Dynamic-LLaVA 在几乎不损失视觉理解和生成能力的前提下，能够将预填充阶段的计算开销减少约 75%；在无 KV Cache 的解码阶段，计算开销减少约 50%；在有 KV Cache 的解码阶段，GPU 显存占用减少约 50%。Dynamic-LLaVA 为多模态大模型推理加速领域树立了新的标杆。

1 引言

1.1 前置信息：预填充与解码

本文主要围绕以 LLaVA 为范式的多模态大模型展开研究。一个多模态大模型的推理过程可以分为预填充和解码两个阶段：

在预填充阶段，不同模态的特征被映射到与大语言模型（LLM）输入 embedding 相同的特征分布空间中。这些多模态特征与文本 token 会一起被大语言模型处理，以生成初始输出文本 token。以图片理解场景为例，该阶段主要处理输入的图片和文本格式的问题。

在随后的解码阶段，预填充阶段生成的所有 token 以及后续生成的所有输出文本 token，将被用于自回归生成，从而产生完整的输出。同样以图片理解场景为例，该阶段生成针对整个问题的完整回答。

1.2 多模态大模型推理加速困境

图 1：多模态大模型生成过程（有 / 无 KV Cache）中 FLOPs（计算复杂度）和 GPU 显存开销的增长趋势

现有的多模态大模型大多以基于解码器架构的大语言模型（LLM）为核心，这些模型通常拥有庞大的参数规模。在生成输出文本 token 的过程中，模型计算负担会逐渐加重，导致对计算资源的巨大消耗。为了提升推理速度，现有模型通常会在解码过程中运用 KV Cache 技术，通过存储并复用之前计算的 KV 激活值来减少重复计算。然而，如图 1 (B) 所示，即使使用了 KV Cache，LLaVA 在输出 token 不断增加时，仍会迅速面临 GPU 显存耗尽的问题。

与文本不同，视觉信息往往包含大量冗余。因此，许多方法尝试通过减少视觉上下文来加速多模态大模型的推理，即对预填充阶段的视觉 token 进行剪枝处理。但这种方法存在局限性：其主要提升了多模态大语言模型在预填充阶段的推理效率，而在解码阶段，其效率提升会逐渐减弱。

如图 1 (B) 和 (C) 所示，FastV 这种针对视觉 token 剪枝的方法，虽然相较于原始的 LLaVA 能够节省一定的 GPU 显存和计算开销（FLOPs），但当输出 token 数接近 5K 时，它仍然会遭遇计算资源瓶颈。此外，FastV 和原始 LLaVA 的曲线斜率基本一致，这表明在长输出的解码阶段，这类方法并没有显著的推理效率优势。因此， 仅通过减少预填充阶段的视觉 token，在输出文本 token 数量远超视觉 token 时，难以实现整个推理效率的显著提升。

1.3 迈向全阶段推理加速：Dynamic-LLaVA

针对上述问题，我们认为：为了实现真正的全阶段推理加速，不仅需要对预填充阶段的视觉 token 进行剪枝，还必须对解码阶段输出的文本 token 进行稀疏化处理，限制参与自回归运算的 token 数量。为此，我们提出了 Dynamic-LLaVA ，针对多模态大模型的视觉 - 语言上下文稀疏化推理加速框架。该框架能够集成到多模态大模型推理的不同阶段中，实现以下目标：

通过这些创新，Dynamic-LLaVA 为多模态大模型的高效推理提供了一种全新的解决方案。

2 方法

图 2：Dynamic-LLaVA 整体框架

如图 2 所示，Dynamic-LLaVA 可以集成到多模态大模型推理流程中的不同阶段。具体而言，在预填充阶段，该框架对视觉 token 执行精准剪枝操作，剔除冗余信息；在不使用 KV Cache 的解码阶段，限制参与自回归运算的视觉与输出文本 token 数量，避免不必要的计算负担；而在使用 KV Cache 的解码阶段，Dynamic-LLaVA 则动态调控 KV Cache，自适应判断是否将当前输出文本 token 的 KV 激活值纳入 KV Cache，优化资源利用效率。为了使模型适应这种全新的稀疏化推理模式，Dynamic-LLaVA 在预训练的 LLaVA-1.5 基础上进行了 1 个 epoch 的监督微调（SFT），确保模型能够高效地运行在稀疏化的推理路径上。

2.1 预填充阶段

在预填充阶段，我们对输入的视觉 token 进行稀疏化操作。如图 2 左侧部分所示，我们引入一个可训练的轻量化的图像预测器（Image Predictor），来判断应当丢弃哪些视觉 token。该图像预测器的结构如下图：

图 3：图像预测器的结构示意图

图像预测器会对每个视觉 token 产生 “决策分数”，以决定对哪些视觉 token 进行保留。在端到端训练中，视觉 token 的剪枝通过 0-1 二值化的掩码操作实现（具体过程见 2.4 节）。在实际推理阶段中，通过保留 “决策分数” 前 k 大的视觉 token（即图 2 左侧部分的 “Yes” 分支），实现视觉 token 数量减少，以实现推理加速。

2.2 解码阶段

不使用 KV Cache 的解码过程：

对于视觉 token，采用和上一小节相同的做法，进行稀疏化处理。

对于输出的文本 token，分两类进行处理：

图 4：输出预测器的结构示意图

使用 KV Cache 的解码过程：

KV Cache 是节省冗余计算的一个关键推理加速技术，其思想是 “用 GPU 显存的空间换计算时间”。显而易见的是，KV Cache 也并非无限大，在长输出情况下，必须丢弃一些 KV Cache 以适应有限的 GPU 显存。目前在 LLM 领域已有大量的 KV Cache 压缩方案，以 方法为代表，这一类方法一般基于当前 token 和历史 KV Cache 进行重要性分数计算，以压缩历史 KV Cache。

与上述方法不同的是，我们对有 KV Cache 的解码阶段的设计，核心在于 “仅判断当前新 token 的 KV 激活是否需要加入 KV Cache 中”。如图 2 右侧所示，对于当前正在处理的新 token（Last output text token），使用和上一部分结构相同的输出预测器，以决定是否加入 KV Cache 集合中。这种 “Online KV Cache 压缩” 方法，判断是否保留 KV Cache 的过程计算复杂度更低，也更加适应多模态场景。在论文附录中，我们详细讨论了我们的方法和现有的 LLM KV Cache 压缩方法的区别。

需要特别说明的是，和不使用 KV Cache 的解码阶段相同，无论当前处理的 token 是否加入 KV Cache，其都会输入 LLM decoder 层进行计算，以保证输出的连贯性。

2.3 端到端训练

图 5：Dynamic-LLaVA 在端到端训练过程中的示意图

Dynamic-LLaVA 是一个需要训练的多模态大模型推理加速框架。我们基于 LLaVA 进行了一个 epoch 的指令微调，以实现对 token 动态选择的稳定性，保证最终的性能。为了保证端到端训练，在训练阶段的稀疏化操作通过 0-1 二值化掩码实现（在推理中的实现是直接从历史 token 序列中丢弃 token）。如图 5 所示，上半部分表示训练中进行 mask 的过程，在得到整个 token 序列的重要性分数后，我们选取前 k 重要的 token 进行保留，相对应的生成掩码向量，其中 0 对应丢弃的冗余 token（不参与注意力过程的计算），1 对应保留的重要 token，进一步基于掩码向量生成注意力过程的掩码矩阵。掩码矩阵用来对多头注意力机制进行掩码操作，以确保丢弃的 token 不参与注意力过程的计算。由于二值化操作会导致不可微问题，所以我们借助了 GumbalSoftmax 和梯度直通估计器（Straight Through Estimator, STE）来保证梯度流的正确传播，以进行端到端的训练，如图 5 下半部分所示。

3 实验

Dynamic-LLaVA 基于 LLaVA-1.5-7B 和 13B 的两个版本进行了 1 个 epoch 的指令微调，训练使用的数据和 LLaVA-1.5 相同。

3.1 视觉理解能力

我们首先评估了 Dynamic-LLaVA 在主要的视觉理解基准的性能，选取了目前主流的多模态大模型推理加速方法进行比较。

表 1：视觉理解基准效果对比。其中，Free 表示方法是否是 Training-Free 的。Dynamic-LLaVA 的下标 "I" 和 "I | T" 分别表示仅对视觉 token 做稀疏化和同时对视觉和文本 token 都做稀疏化（该标识适用于下文所有的表格）

如表 1 所示，Dynamic-LLaVA 在大部分视觉理解任务上取得了优越的性能。和其他对视觉内容稀疏化的方法相比，Dynamic-LLaVA 在能大幅减小计算复杂度的同时，能够实现相比原始的 LLaVA-1.5 性能几乎

不下降。此外，在 SciQA、POPE、MME 和 MMBench 上，Dynamic-LLaVA 相比 LLaVA-1.5 甚至有一定的性能提升。例如，在 SciQA 任务上，Dynamic-LLaVA 的 7B 和 13B 版本，相较于 LLaVA-1.5 实现了 2.3% 和 0.8% 的性能提升。

表 2：与其他高效视觉 projector 的 SOTA 方法对比

值得一提的是，Dynamic-LLaVA 并没有对 LLaVA-1.5 的视觉 projector 进行修改，就可以实现大幅降低预填充阶段计算复杂度，同时维持模型性能。在表 2 中，和其他针对视觉 projector 做高效设计（以提高推理效率）的 SOTA 方法进行了对比。相较于其他使用了高效的视觉 projector 的方法，Dynamic-LLaVA 使用和 LLaVA-1.5 相同的 MLP 结构作为视觉 projector，实现了更好的性能，同时也大幅降低了预填充阶段的计算复杂度。此外，Dynamic-LLaVA 也可以和其他使用高效视觉 projector 的方法集成。例如，表 2 中 Dynamic-LLaVA 使用 TokenPacker 这一高效视觉 projector 的版本，在原始的 TokenPacker 方法基础上，进一步减少了视觉 token。相较于其他基于 TokenPacker 的推理加速方法，性能损失最少。

3.2 生成能力

现有的视觉理解任务中，一般只要求模型给出简短的回复，这和现实世界中多模态大模型的应用场景仍然存在不小的区别。在现实使用中，多模态大模型多数情况下会被要求生成更长、更细致的描述。为了和现实世界的场景对齐，评估 Dynamic-LLaVA 在更长的输出情况下的生成能力和推理效率。我们额外构建了两个评估模型生成能力的基准：

我们以 PPL (Perplexity Metric) 指标评估模型生成内容的流畅度、以 METEOR (Metric for Evaluation of Translation with Explicit ORdering) 指标评估模型生成内容的质量。

表 3：生成能力基准比较。其中，解码阶段的 TFLOPs 和 Mem.（GPU 显存占用）分别在无 / 有 KV Cache 的情况下测量得出。PPL 越低越好，METEOR 越高越好

如表 3 所示，相比 LLaVA-1.5，只进行视觉内容稀疏化的 Dynamic-LLaVA 的生成流畅度（PPL）和生成质量（METEOR）几乎没有变化；同时对视觉和文本进行稀疏化的 Dynamic-LLaVA，PPL 仅变高了 0.3，METEOR 甚至略有提升，而在推理效率上，在无 KV Cache 的解码阶段降低了～50% 的 TFLOPs，在有 KV Cache 的解码阶段降低了～50% 的 GPU 显存占用。实验结果充分表明，Dynamic-LLaVA 针对视觉和文本同时进行稀疏化，几乎不影响实际生成能力，却可以实现大幅的推理效率提升。

3.3 实际推理效率

表 4：Dynamic-LLaVA-13B 推理效率实测。其中，2K/4K 表示输出的文本 token 数，所有结果均在一张 A100 (80G) 上测试得出，batch size 固定为 8。“” 表示 GPU 显存耗尽

在表 4 中，我们测试了多模态大模型实际推理的时间和 GPU 显存占用。Dynamic-LLaVA 实现了更快的推理速度和更低的显存占用。FastV 这种对预填充阶段的视觉 token 进行剪枝的方法，随着输出长度的增长，推理效率也逐渐降低。而我们提出的 Dynamic-LLaVA，随着输出变长，相比于 FastV 的推理效率优势也逐渐显现出来。

3.4 实例展示

图 6：Dynamic-LLaVA-13B 在 LVIS-VQA (single-round) 上的推理结果展示。图片的白色部分表示该位置的图像块被丢弃，文字中的灰色部分表示其在稀疏化过程中被丢弃，这表示它们不参与后续的自回归解码过程，但在模型的输出中都被完整保留

图 6 中展示了 Dynamic-LLaVA-13B 在 LVIS-VQA (single-round) 上的推理结果，以及对视觉和文本 token 的稀疏化情况。可视化结果表明，视觉 token 部分的主要信息得以保留；文本 token 中，一些不影响整体语义理解的连词、介词等被丢弃。这表明 Dynamic-LLaVA 能够实现关键的视觉、语义信息的保留，从而保证了模型整体的性能。

4 总结

针对当前多模态大模型推理效率受限的问题，团队通过分析多模态大模型推理过程中的不同阶段，针对性的设计了推理加速方案。提出了 Dynamic-LLaVA—— 第一个同时稀疏化视觉和语言上下文的多模态大模型推理加速框架，将不同推理模式的推理效率优化集成到统一框架中。

随着多模态大模型技术的发展，尤其是其在复杂推理、长思维链领域的不断进步。我们有理由相信，Dynamic-LLaVA 的应用场景正变得更加广泛，其对输出文本 token 进行稀疏化的模式，会在当前的更长输出、更复杂推理的场景下，体现出更明显的推理加速优势。

作者简介

黄文轩：小红书 NLP 团队算法实习生，现硕士就读于华东师范大学计算机科学与技术学院 2023 级。他在 ICLR、CVPR 等国际顶级会议上以第一作者身份发表了多篇学术论文，主要研究方向包括多模态大模型、大模型的高效训练与推理等。

翟子杰：小红书 NLP 团队算法实习生，现硕士就读于华东师范大学计算机科学与技术学院 2023 级。他在 ICML、ICLR、EMNLP 等国际顶级会议上发表过多篇学术论文，研究方向主要集中在多模态大模型、生成式搜索与推荐大模型等领域。

曹绍升：小红书 NLP 团队负责人，发表论文 30 余篇，授权专利 100 余项，引用近 4000 次，获得 ICDE 2023 年最佳工业论文奖、CIKM 2015-2020 年最高引用论文、AAAI 2016 最具影响力论文。此外，还荣获了中国发明协会创新成果一等奖（排名 1）、中国人工智能学会吴文俊科技进步二等奖（排名 1），连续 4 年入选世界人工智能学者榜单 AI-2000 新星榜前 100 名、Elsevier 中国区高被引学者，CCTV-13《新闻直播间》采访报道。

叶哲宇：硕士毕业于帝国理工学院计算机专业，小红书 NLP 团队算法工程师，专注于大模型算法与应用方向，开源社区 DMLC 成员。他在 ICLR、NAACL、EMNLP 等国际顶级会议上发表过多篇论文，研究领域涵盖大模型应用、多模态大模型、Agent 模拟等。

林绍辉：华东师范大学计算机学院研究员，紫江青年学者，2021 年扬帆计划获得者，曾获中国人工智能学会优秀博士论文提名奖、《中国科学：技术科学》最佳审稿人。在国际顶级期刊和会议发表超过 50 篇论文，包括 TPAMI、TNNLS、TMI、CVPR、ECCV、AAAI、IJCAI 等。担任 CVPR 2024 领域主席、IJCAI 2020 SPC 以及国际顶级期刊和会议审稿人。目前主要研究方向有计算机视觉、机器学习、图像视频理解、低层视觉等。

澳洲190打分表

澳洲幸运10要有多假就有多假

澳洲10冠军五码技巧

澳洲10人工必赢计划网

奇趣腾讯分分彩app下载

168飞艇官网开奖结果记录数据

名爵app澳洲幸运10

澳彩大数据分析软件

大小单双的死规律

幸运飞行艇app下载

幸运5稳赢的死方法

职业赌徒每天稳赢计划

大数据分析澳洲幸运10

168免费精准预测