LLaMA 进化史

大规模语言模型(Large Language Model, LLM)的快速发展正在以前所未有的速度推动人工智能(AI)技术的进步。

作为这一领域的先行者, Meta在其LLaMA(Large Language Model Meta AI)系列模型上取得了一系列重大突破。

近日, Meta官方正式宣布推出LLaMA-3, 作为继LLaMA-1、LLaMA-2和Code-LLaMA之后的第三代旗舰模型。

LLaMA-3在多个权威基准测试中均取得了全面领先的成绩, 超越了业界同类最先进模型, 再次刷新了LLM的性能上限。

LLaMA-1 系列

LLaMA-1是Meta公司在2023年2月发布的首款大规模语言模型，它提供了7B、13B、30B和65B四种不同的参数规模版本。

这些版本均在超过1T(万亿)个tokens的广泛语料库上进行了预训练，有效利用了海量数据的优势。

在这些版本中，参数量最大的65B版本在2048张A100 80G GPU上训练了21天，表现出色，在多数基准测试中超越了具有175B参数的GPT-3。

除了卓越的性能，LLaMA-1还采取了开源策略，极大地方便了研究者和开发者的获取与使用，迅速成为开源社区中极受欢迎的大规模语言模型之一。

围绕LLaMA-1，形成了一个活跃的生态系统，涌现了许多基于该模型的下游应用、微调模型和衍生变体。

LLaMA-2 系列

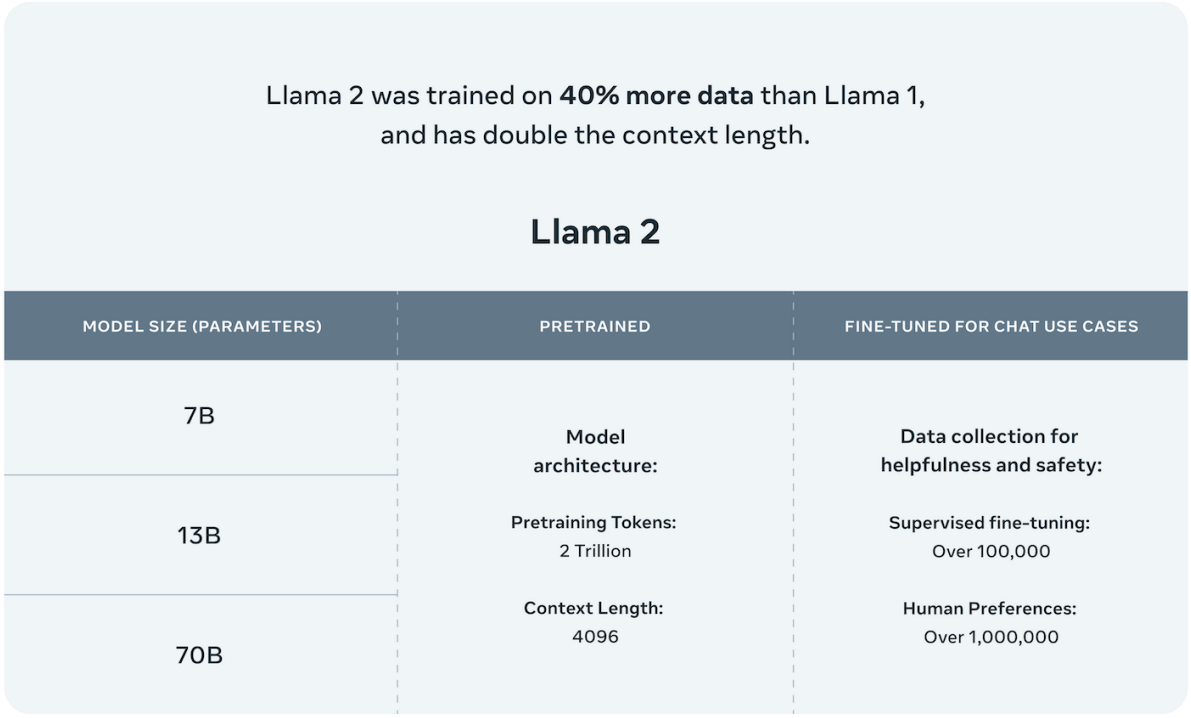
LLaMA-2 系列，Meta在2023年7月推出的第二代大规模语言模型，是对LLaMA-1系列功能的扩展和改进。

此系列包括7B、13B和70B三个不同规模的版本，并特别推出了经过指令微调的对话模型LLaMA-2-Chat，旨在增强模型在对话和任务执行方面的能力。

主要特点包括:

- **增强的模型性能：** 相较于LLaMA-1，LLaMA-2系列在更广泛的语料上进行了预训练，这大幅提升了其文本生成和理解能力，确保模型在各类NLP任务中的领先性能。
- **指令微调技术：** LLaMA-2-Chat通过对大量的高质量指令-回答对进行微调，优化了模型对复杂指令的响应能力，使其在对话和具体任务执行中表现更为精确和自然。
- **开源政策：** 继承LLaMA-1的开放精神，LLaMA-2以开源形式发布，不仅丰富了LLaMA的生态系统，也激励了基于这一平台的创新性研究和应用开发。

通过这些改进，LLaMA-2系列不仅在技术层面上提供了显著的升级，也在开放源代码和社区参与方面持续发挥着积极的推动作用。此举确保了LLaMA系列在全球语言模型领域的持续领先和影响力扩展。



LLaMA-3 系列

2024年4月, Meta 发布了最新一代的大规模语言模型系列 LLaMA-3。

作为 LLaMA 系列的第三代产品, LLaMA-3 在 LLaMA-1、LLaMA-2 和 Code-LLaMA 的基础上实现了全面升级和优化。

在多个权威基准测试中, LLaMA-3 的表现全面领先于业界同类最先进模型,再次刷新了大模型的性能上限。

LLaMA-3 在长序列建模能力上取得了重大突破。

相比 LLaMA-2 只能处理最长 2048 个 token 的文本, LLaMA-3 将这一上限提高到了 8192 个 token, 使其可以轻松应对超长文档、多轮对话等复杂场景。

同时, LLaMA-3 还采用了全新的 tokenizer, 将分词器更换为tiktoken，与GPT4保持一致。

其词表大小达到了 128K, 远超 LLaMA-2 的 32K。更大的词表意味着更精细的文本表示和更强的语言泛化能力。

注：Llama-1模型采用了BPE算法进行分词，使用sentencepiece实现。其词表大小为32k。

Llama-2维持了与Llama-1相同的架构和分词器，但将上下文长度扩展到了4k。

数据规模是大模型取得突破性进展的另一个关键因素。

LLaMA-3 的预训练语料规模超过了 15 T (150 万亿) token, 是 LLaMA-2 的 7 倍多。

海量高质量的预训练数据, 为 LLaMA-3 提供了更全面、更深入的世界知识和语言理解能力。

同时, Meta 还对预训练数据进行了更细粒度的筛选和清洗, 进一步提高了数据质量。

凭借优秀的模型架构和海量的预训练数据, LLaMA-3 在下游任务上实现了全面领先。

在自然语言推理、机器阅读理解、常识问答等多个基准测试中, LLaMA-3 的表现都超越了业界最先进的同规模模型。

此外, LLaMA-3 在推理能力、代码生成、指令跟随、多语言支持等方面也有长足进步, 使其成为更加全能和可控的通用人工智能助手。

随着 LLaMA-3 的发布, Meta 再次引领了大模型技术的发展潮流。

作为当前最先进的开源大模型, LLaMA-3 必将掀起新一轮的研究和应用热潮, 为人工智能的进步注入强大动力。

Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B Published	Mistral 7B Measured	Gemma 7B Published	Gemma 7B Measured
MMLU 5-shot	66.6	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	45.9	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	61.1	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	79.1
DROP 3-shot, F1	58.4	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0 Published	Mixtral 8x22B Measured
MMLU 5-shot	79.5	71.8	77.7
AGIEval English 3-5-shot	63.0	--	61.2
BIG-Bench Hard 3-shot, CoT	81.3	75.0	79.2
ARC-Challenge 25-shot	93.0	--	90.7
DROP 3-shot, F1	79.7	74.1 variable-shot	77.6

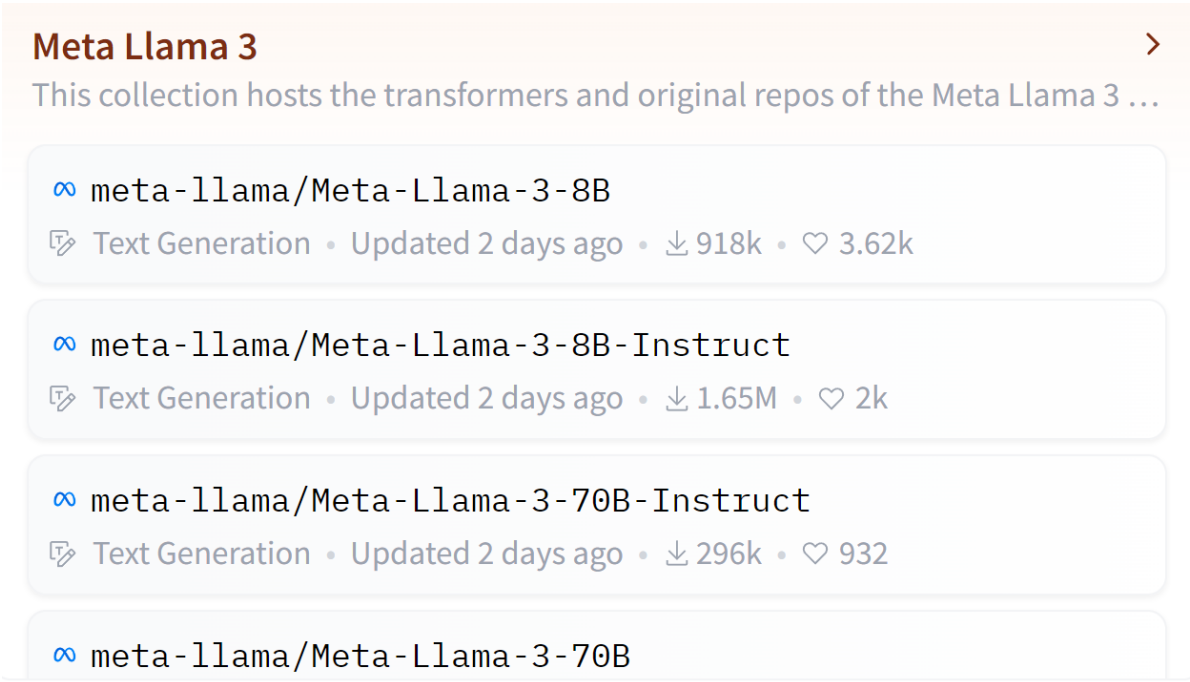
Meta Llama 3 400B+ (still training)

Checkpoint as of Apr 15, 2024

	PRE-TRAINED Meta Llama 3 400B+
MMLU 5-shot	84.8
AGIEval English 3-5-shot	69.9
BIG-Bench Hard 3-shot, CoT	85.3
ARC-Challenge 25-shot	96.0
DROP 3-shot, F1	83.5

	INSTRUCT Meta Llama 3 400B+
MMLU 5-shot	86.1
GPQA 0-shot	48.0
HumanEval 0-shot	84.1
GSM-8K 8-shot, CoT	94.1
MATH 4-shot, CoT	57.8

LLaMA-3 的基础版模型和指令微调模型



Meta发布了LLaMA 3模型的几个变体:

1. meta-llama/Meta-LLama-3-8B: 这是基础版的LLaMA 3模型, 有80亿个参数。
2. meta-llama/Meta-LLama-3-8B-Instruct: 这是一个经过指令微调的8B参数LLaMA 3模型变体。
3. meta-llama/Meta-LLama-3-70B-Instruct: 这是一个更大的经过指令微调的LLaMA 3模型, 有700亿个参数。
4. meta-llama/Meta-LLama-3-70B: 这是没有经过指令微调的基础版70B参数LLaMA 3模型。

总的来说, Meta发布了8B和70B两种参数规模的LLaMA 3模型, 每种规模都有提供经过指令微调的变体。

指令微调(Instruction Tuning)的变体和基础版模型之间有以下几个主要区别:

1. **训练目标不同**: 基础版模型通常使用语言建模(Language Modeling)作为训练目标, 即根据上文预测下一个单词。

   而指令微调的变体则引入了额外的指令数据, 通过监督学习让模型学会理解和执行自然语言指令。

2. **训练数据不同**: 基础版模型主要在大规模无标签文本语料上进行预训练。而指令微调的变体需要构建专门的指令数据集, 其中包含大量的自然语言指令及其对应的执行结果。

3. **应用场景不同**: 基础版模型可以用于各种NLP任务, 但在执行具体指令时可能表现欠佳。指令微调的变体则专门针对指令理解和执行进行了优化, 在问答、对话、任务完成等场景中表现更加出色。

4. **交互方式不同**: 使用基础版模型时, 用户需要根据具体任务设计prompts。而指令微调的变体允许用户使用自然语言直接下达指令, 交互更加直观和方便。

5. **可控性不同**: 基础版模型生成的内容可能不够可控, 容易出现幻觉或不合适的言论。指令微调引入了人类反馈, 可以更好地引导模型生成安全、可靠的内容。

6. **推理效率不同**: 指令微调通常会引入一些控制机制, 如提示工程、示例学习等, 可能降低推理速度。但一些高效的微调方法如LoRA、Prefix Tuning等可以在保证性能的同时加快推理。

总的来说, 指令微调赋予了语言模型更强的指令理解和执行能力, 使其在实际应用中更加智能、高效和可控。

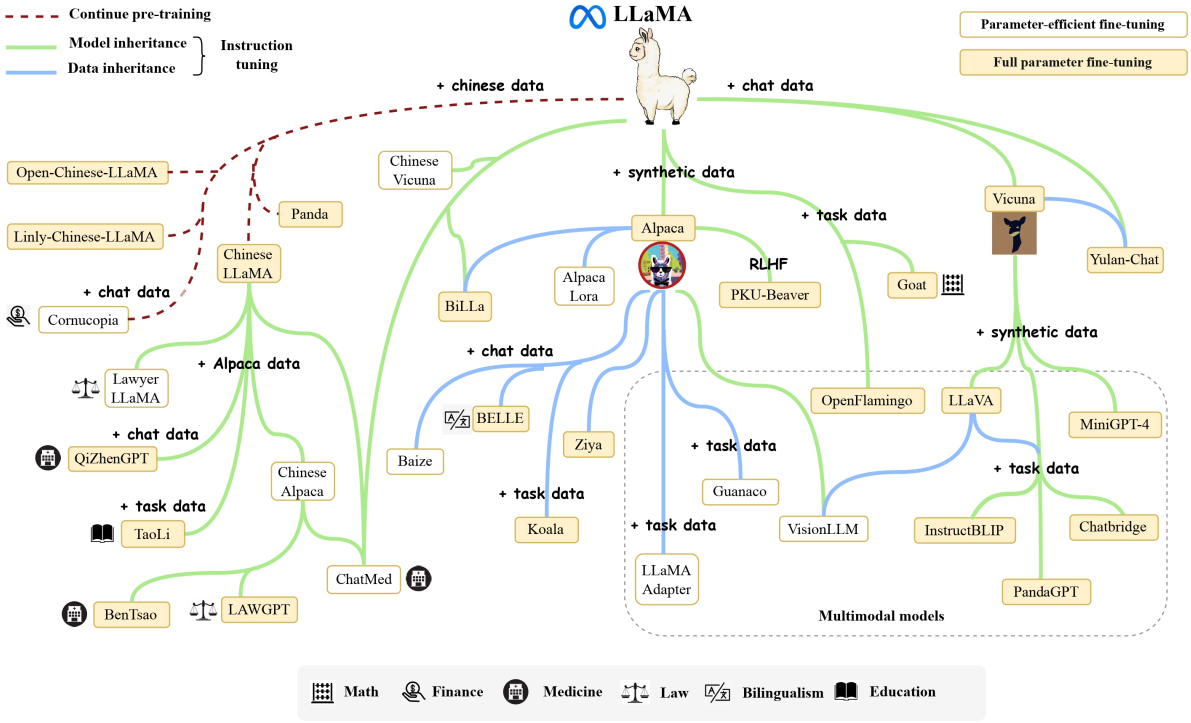
LLaMA衍生模型和生态

LLaMA系列模型的开源性和可访问性, 使其成为了LLM领域的重要研究和应用平台。

围绕LLaMA, 涌现出了一个繁荣的开源生态, 催生了如Alpaca、Vicuna、LLaVA等一系列优秀的衍生模型。

这些模型在LLaMA的基础上,结合领域数据和下游任务进行针对性优化, 在对话、多模态、开放域问答等方面取得了瞩目成绩。

LLaMA的开源, 极大地降低了LLM研究的门槛, 推动了AI技术的普惠进程。



图中展示的是LLaMA模型及其衍生模型的概览，包括了不同的训练方法、数据类型、以及一些特定的应用场景。

以下是对图中内容的描述和解释：

1. **继续预训练（Continue pre-training）**：指的是在现有模型基础上，使用更多数据继续训练以提升模型性能的过程。
2. **LLaMA**：是Meta公司开发的一个大型语言模型，图中显示了基于LLaMA模型的多种扩展和优化路径。
3. **参数高效微调（Parameter-efficient fine-tuning）**：这是一种微调技术，通过调整模型中较少的参数来适应新的任务，以提高效率。
4. **模型继承（Model inheritance）**：指新模型继承或基于旧模型的架构和参数进行开发。
5. **指令微调（Instruction tuning）**：通过指令来指导模型微调，使其更好地遵循给定的任务指令。
6. **全参数微调（Full parameter fine-tuning）**：指在微调过程中调整模型的所有参数。
7. **数据继承（Data inheritance）**：新模型使用旧模型训练过的数据集作为训练数据的一部分。
8. **中文数据（Chinese data）**：指使用了中文语言数据进行训练的模型。
9. **Open-Chinese-LLaMA**：可能是一个针对中文优化的LLaMA模型。
10. **合成数据（synthetic data）**：指通过技术手段生成的非真实世界数据，用于训练。
11. **Vicuna, Panda, Alpaca, Goat**：基于LLaMA模型的不同变种或特定用途的模型名称。
12. **RLHF**：指的是通过人类反馈进行强化学习的微调方法（Reinforcement Learning from Human Feedback）。
13. **Yulan-Chat**：专门为中文对话优化的模型。
14. **PKU-Beaver, BiLLa**：指北京大学开发的模型或者是双语语言模型。
15. **Cornucopia**：指包含多种数据类型的综合数据集。

16. **Lawyer, LLaVA, [BELLE]**：针对特定领域（如法律）或具有特定功能（如视觉语言模型）的模型。
17. **MiniGPT-4, Ziya, QiZhenGPT, Baize**：不同大小或针对特定任务优化的模型。
18. **Guanaco, Chatbridge, Koala**：特定的多模态模型或其他应用领域的模型。
19. **VisionLLM, TaoLi, InstructBLIP**：结合了视觉和语言任务的模型。
20. **ChatMed, Adapter, PandaGPT, LAWGPT, BenTsao**：针对医疗、适配器技术、法律等特定领域的模型。
21. **多模态模型（Multimodal models）**：能够处理并整合来自多种感官模式（如视觉、听觉、文本）的模型。
22. **数学（Math）、金融（Finance）、医学（Medicine）、法律（Law）、双语（Bilingualism）、教育（Education）**：模型应用的领域。

这张图展示了以 LLaMA 系列模型为核心的大语言模型生态系统。

1. **继续预训练(Continue pre-training)**: 一些模型如 Chinese LLaMA、Chinese Alpaca 等在 LLaMA 的基础上加入了中文数据继续预训练,以提高中文任务的表现。
2. **指令微调(Instruction tuning)**: 通过在指令数据集上微调 LLaMA, 衍生出了 Alpaca、Vicuna 等模型, 使其能够执行指令跟随和问答对话等任务。
3. **参数高效微调(Parameter-efficient fine-tuning)**: 使用 LoRA、Prefix Tuning 等参数高效微调方法在下游任务数据上微调 LLaMA,得到 Alpaca Lora 等模型。
4. **全参数微调(Full parameter fine-tuning)**: 在特定垂直领域数据上对 LLaMA 进行全参数微调, 如在医疗对话数据上微调得到的 BianTsao 模型。
5. **多模态模型(Multimodal models)**: 将 LLaMA 扩展到多模态, 如支持图像输入的 LLaVA、MinGPT-4, 语音交互的 InstructBLIP 等。

总的来说, 该图全面地展示了以 LLaMA 为基础衍生出的丰富多样的大模型生态,涵盖了主要的优化训练范式、任务类型和具体模型。

值得注意的是, 中文模型在该生态中占据了重要地位。

Alpaca 是由斯坦福大学计算机科学系博士生 Eric Wang 等人开发的一个基于 LLaMA-7B 模型的衍生大模型。

其主要特点包括:

1. 指令精调: Alpaca 在 LLaMA-7B 的基础上, 使用了一个包含 5.2 万条指令数据的数据集进行了监督微调(Supervised Fine-tuning)。这使得 Alpaca 能够很好地理解和执行自然语言指令, 具备类似 ChatGPT 的对话交互能力。
2. 开源共享: Alpaca 项目的代码和训练数据都已经在 GitHub 上完全开源, 允许研究者和开发者基于此进行二次开发。这极大地降低了构建指令跟随型对话系统的门槛。
3. 性能优异: 在标准的指令跟随任务基准如 MMLU 上, Alpaca 的表现已经接近 ChatGPT 等封闭模型, 而参数量和计算开销却小很多。这说明了在 LLaMA 基础上进行指令精调的有效性。
4. 多语言支持: 得益于 LLaMA 模型本身强大的多语言能力, Alpaca 也具备了一定的多语言处理能力, 尽管主要还是针对英文进行了优化。
5. 可控性强: 由于 Alpaca 的训练数据是人工标注的高质量指令数据, 因此其生成的内容更加可控, 在事实性、安全性方面表现出色。
6. 开源生态: Alpaca 的开源进一步推动了 LLaMA 周边生态的繁荣, 催生了一系列基于 Alpaca 的衍生模型和应用。

总的来说, Alpaca 是 LLaMA 家族中一个代表性的指令精调模型, 其开源性、可访问性和优异的性能, 使其成为了开源界 ChatGPT 的有力竞争者。

Alpaca 的成功证明了在一个强大的基础模型上, 利用高质量的指令数据进行针对性微调, 可以显著提升模型在对话交互任务上的表现, 同时还能保持较强的可控性。

Vicuna 是由加州大学伯克利分校、卡内基梅隆大学、斯坦福大学等机构的研究者联合开发的一个基于 LLaMA 的开源对话语言模型。

其主要特点包括:

1. 大规模指令精调: Vicuna 使用了一个包含 7 万多条对话数据的指令数据集对 LLaMA-13B 进行了微调。这些数据主要来自于 ShareGPT 收集的真实人类对话,质量相当高。相比 Alpaca 的 5 万条指令数据,Vicuna 的训练语料更加丰富和多样化。
2. 多轮对话能力: 得益于大规模高质量对话数据的训练, Vicuna 具备了出色的多轮对话能力。它能够很好地理解对话的上下文, 根据之前的对话内容生成连贯且相关的回复。在这一点上, Vicuna 比 Alpaca 表现得更为出色。
3. 训练计算效率: Vicuna 在训练过程中采用了一系列优化手段, 如混合精度训练、梯度累积、DeepSpeed 等, 使得在有限的计算资源下也能高效地完成大模型的训练。这为开源社区提供了一个很好的模型训练范例。
4. 开源共享: 与 Alpaca 类似, Vicuna 的代码和训练数据也已经完全开源, 供社区使用和研究。同时, 研究团队还发布了经过训练的检查点(checkpoint), 可以直接进行推理和微调。
5. 人类对齐度高: 通过在大规模人类对话数据上的训练, Vicuna 学会了更加自然、人性化的交互方式。其生成的回复在流畅度、连贯性、同理心等方面都有更好的表现,给人一种更加亲切、自然的感受。
6. 商业化应用: 进一步扩大了 Vicuna 的影响力和应用范围。

总的来说, Vicuna 代表了开源对话大模型的最新进展。其充分利用了大规模高质量的对话数据,在 LLaMA 的基础上实现了全面的性能提升,尤其是在多轮对话和人类对齐度方面表现突出。

Vicuna 的开源和商业应用,为构建更加智能、自然的对话系统提供了重要参考。

LLaVA，全称为Large Language and Vision Assistant（大型语言和视觉助手），是一种新型的大型多模态模型。

它的目标是开发一种通用视觉助手，能够遵循语言和图像指令来完成各种现实世界任务。

LLaVA 结合了自然语言处理（NLP）和计算机视觉（CV）的能力，通过理解视觉内容并根据语言指令进行操作，从而实现对图像和文本的深入理解与交互。

LLaVA 模型的核心在于其多模态架构，它将视觉编码器（如基于 Transformer 的视觉模型）与语言模型（如 LLaMA）结合起来，形成一个能够处理图文信息的集成系统。这种设计使得 LLaVA 能够执行包括图像标注、视觉问答、文本到图像生成等在内的多模态任务。

其主要特点包括:

1. 多模态融合: LLaVA 在 LLaMA 的基础上引入了视觉特征,实现了语言和视觉信息的融合。具体来说,它使用 BLIP-2 模型提取图像特征,然后将其与文本表示进行交互, 生成与图像相关的自然语言描述或回答。
2. 图像理解能力: 得益于多模态融合, LLaVA 具备了强大的图像理解和分析能力。它可以根据输入的图像生成详细的描述,回答与图像相关的问题,甚至进行开放式的图像内容分析。
3. 通用语言能力: 作为 LLaMA 的衍生物,LLaVA 继承了原模型优秀的自然语言处理能力。因此它不仅可以处理图像相关的任务,在通用的语言理解、生成等方面也有不俗的表现。
4. 零样本学习: LLaVA 支持零样本学习(zero-shot learning), 即无需在下游任务上进行微调, 直接利用预训练的知识完成推理。这使得 LLaVA 可以灵活地应对各种形式的多模态任务, 无需重新训练模型。
5. 大规模预训练: LLaVA 在大规模图文对数据上进行了预训练, 学习了丰富的视觉-语言对齐知识。这为其在下游任务上的优异表现奠定了基础。同时, 预训练也提高了模型的泛化能力和鲁棒性。
6. 开源开放: 与其他 LLaMA 衍生物类似,LLaVA 的代码和预训练模型权重都已经开源。这为进一步研究和应用多模态大模型提供了宝贵的资源和参考。

总的来说, LLaVA 代表了 LLaMA 在多模态领域的重要拓展。

通过引入视觉特征与语言表示的交互, LLaVA 实现了对图像内容的深度理解和分析。

同时, 它在通用语言任务上的出色表现, 证明了多模态学习对于提升语言模型性能的积极作用。

可以预见, LLaVA 将为多模态大模型的研究和应用开辟新的方向, 推动人工智能向更加全面、贴近人类智能的方向发展。