ERNIE 4.5 技术报告解读

2025年6月30日　4 分钟阅读

#Opensource #ERNIE-4.5 #论文 #技术

> *本文介绍了百度开源的ERNIE 4.5模型，并对其技术原理、主要贡献、论文方法、评估结果和局限性进行了详细解读。*

ERNIE 4.5 技术报告简报

ERNIE 技术报告 (https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf)

GitHub: https://github.com/PaddlePaddle/ERNIE

Hugging Face: https://huggingface.co/baidu

核心亮点和主要主题

ERNIE 4.5 系列模型代表了百度在大型语言模型 (LLMs) 和视觉语言模型 (VLMs) 方面的最新进展。该报告详细介绍了其创新的架构、预训练策略、模型优化、后训练方法、训练框架、推理与部署以及全面的评估结果。核心亮点包括其异构 MoE 架构、自适应分辨率视觉编码器、高效的数据管理、以及在处理长上下文和多模态推理方面的卓越能力。

最重要的思想或事实

架构创新：异构 MoE 与多模态融合

异构 MoE 结构： ERNIE 4.5 采用了“异构模态结构”，与传统单模态 MoE 模型不同，它“支持模态间的参数共享，包括自注意力参数共享和专家参数共享，同时允许每个独立模态拥有专用参数。” 这种设计不仅通过专用视觉专家实现视觉信息的有效学习，还在训练过程中增强了语言模型原有的知识和推理能力。

视觉编码器（自适应分辨率）： 摒弃了传统 ViT 对固定分辨率输入的限制，ERNIE 4.5 使用“自适应分辨率视觉编码器”。该编码器“独立调整每个输入图像的高度和宽度到最接近 ViT 补丁大小的倍数”，从而“近似保留了原始长宽比，避免了固定大小调整引入的失真。” 它还利用 2D 旋转位置嵌入（RoPE）编码空间信息，并采用图像打包技术高效利用计算资源。

适配器（Adapter）： 为了统一视觉和文本表示，“设计了一个适配器作为视觉编码器和语言模型之间的模态桥接模块。” 该适配器通过空间和时间令牌压缩进行特征融合，并减少序列长度，利用像素洗牌（pixel

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#Opensource #ERNIE-4.5 #技术

shuffle）技术将相邻的特征重排为更紧凑的形式。所有静态图像都被视为合成的两帧视频，以实现跨模态的一致时间建模。

预训练策略

分阶段预训练： 采用三阶段预训练策略：

阶段一：纯文本训练（短上下文与长上下文）： 首先在大规模纯文本数据上进行预训练，建立核心语言能力、事实知识库和文本生成能力（8k 短上下文）。随后通过调整旋转位置嵌入（RoPE）的频率基数θ，将模型上下文长度扩展到 128k 令牌，并上采样长于 16k 令牌的文档。

阶段二：纯视觉训练： 报告中未直接详细描述纯视觉训练的具体内容，但提及其作为预训练的独立阶段。

阶段三：多模态联合训练： 在此阶段，模型进行多模态信息融合训练。

Token-Balanced Loss： 针对多模态训练中图像和提示位置被遮蔽导致的梯度不平衡问题，引入了"Token-Balanced Loss 函数"。该函数通过归一化因子 (|Ui|+ |Mi|)−1 确保"每个样本的损失贡献由其总序列长度的倒数加权，独立于具体的遮蔽配置。"

REEAO：比特级确定性预训练数据管理器： 为了应对大规模训练中数据管道的波动（如检查点恢复、节点故障、资源调整等），REEAO 旨在"防止数据重复或遗漏等重大问题"，确保数据处理的比特级确定性。

领域特定数据： 构建了跨行业、金融、医疗、消费娱乐等垂直领域的大规模数据集，通过"渐进式挖掘和条件训练"以及"音频转录和增强"等策略，解决高质量领域数据稀缺和专业性问题。

后训练与奖励系统

统一奖励系统： 针对 LLMs 和 VLMs 的后训练，设计了一个统一奖励系统，包括：

监督微调 (SFT)： 将 SFT 数据细分为十个不同主题领域，包括科学与数学、编码、逻辑、信息处理、创意写作、多语言、知识问答、多轮与角色扮演以及安全。

强化学习：推理任务：参考引导 LLM-as-a-Judge (RLLM)： 利用 LLM 作为公正评估器，将模型生成结果与参考答案语料库进行基准测试。

沙盒： 用于编程任务的受控隔离测试环境，评估代码的功能、正确性、可靠性和合规性。

参考引导判别奖励模型 (RDRM)： 类似于 RLLM，但 RDRM 在评分过程中明确由参考答案引导，使其成为"开卷考试"，确保模型输出与参考答案的内容和结构特征高度一致。

非推理任务：清单感知验证器： 定义一组明确且客观可评估的标准，确保生成响应符合既定规范。

生成式奖励模型 (GRM)： 结合多维评估标准和动态反馈机制，对每个查询进行定制评估。

判别奖励模型 (DRM)： 传统的强化学习框架，通过判别任务学习奖励函数以引导模型输出。

可验证奖励的强化学习（VLM）： 特别针对 VLMs，报告提到"通过可验证奖励的强化学习"，例如在"视觉谜题"任务中，使用两个 LLM 评估政策模型响应的正确性，一个评估内部一致性，另一个验证最终答案。

训练框架与优化

异构并行和混合并行：异构并行架构： 针对多模态模型训练，实现模型架构与并行策略的协同。

分层负载均衡策略： 确保不同模态（例如 MoE 主干和 ViT 编码器）之间的计算资源分配和负载均衡。

MoE 主干的混合并行： 包括节点内专家并行和内存高效的流水线调度。

FP8 混合精度训练： 利用 FP8 混合精度训练，以减少内存使用和加速计算。

计算优化：操作员级重计算： 优化了传统的重计算方法，仅保留必要张量进行反向计算，从而减少内存开销。

FlashMask： 用于灵活的注意力掩码和长上下文训练。

框架原生容错系统： 确保大规模训练的稳定性和可靠性。

推理与部署

量化： 提供多种量化方案，包括 FP8、INT8、INT4，甚至 2 比特权重专用量化。

2 比特量化： ERNIE-4.5-300B-A47B 可以在一台 141GB H20 GPU 上部署，"模型大小从 BF16 基线减少了 80%，且几乎无损。"

W4A8 量化： 用于提高推理吞吐量。

注意力与 KV 缓存量化： 通过 4 比特或 8 比特量化 KV 缓存以减少内存，并以 8 比特精度计算批次矩阵乘法（BMM）和 Softmax 以降低延迟。

推理加速：W4A8 内核加速： 利用 CUTLASS 内核和快速位移反量化，显著提高内存吞吐量和推理速度。

高效注意力内核： 在 Hopper 架构 GPU 上利用 FP8 计算，在 Ampere 架构 GPU 上利用 INT8 计算，并针对 Softmax 计算优化了指数函数实现。

推测解码： 用于加速解码器负载。

部署： 支持 Prefill-Decode (PD) 解耦部署与大规模专家并行，并提供多级负载均衡。

开源开发工具

ERNIEKit： 专门的开发工具包。

FastDeploy： 大型语言模型和视觉语言模型的推理与部署工具包，支持 vLLM 接口，提供 PD 解耦、全面的低比特量化支持（W8A8、W8A16、W4A8、W4A16、W2A16），以及对 NVIDIA GPU、昆仑芯 XPU、海光 DCU 和昇腾 NPU 等多种硬件的支持。

评估与结果

语言模型评估： ERNIE-4.5-Base 模型在通用任务（C-Eval, CMMLU, MMCU, AGIEval, MMLU, MMLU-Redux, MMLU-Pro）、事实知识（SimpleQA, ChineseSimpleQA）、推理（BBH, DROP, ARC, HellaSwag, PIQA, WinoGrande, CLUEWSC）、代码生成与理解（Evalplus, MultiPL-E）和数学推理（GSM8K, MATH, CM17K, MGSM, ASDIV, SVAMP, MATHQA, CMATH）等多个基准测试中，与 DeepSeek-V3-Base 和 Qwen3-30B-A3B-Base 等现有 SOTA 模型进行了系统评估。ERNIE-4.5-300B-A47B-Base 在多项基准测试中表现出色，尤其在知识、推理和数学方面取得了领先分数。

后训练语言模型评估： 后训练模型（ERNIE-4.5-300B-A47B）在通用、知识、指令遵循、数学和推理任务的评估中，与 Qwen3-235B-A22B、DeepSeek-V3-0324 和 GPT-4.1 进行了比较，显示出强大的竞争力，在多项中文基准测试中（如 C-Eval, CMMLU, ChineseSimpleQA）超越了其他模型。

多模态模型评估： ERNIE-4.5-VL-424B-A47B 在视觉知识、文档与图表理解、多模态推理、视觉感知和视频理解等多个维度上与 Qwen2.5-VL 系列模型进行了比较。结果显示，ERNIE-4.5-VL 在 OCRBench, AI2D, DocVQA, CountBench 等文档和视觉感知任务上表现尤为突出，在多模态推理的 MMMU 和 MathVista 也展现了强大的能力。

定性示例： 报告附录提供了丰富的定性示例，展示了 ERNIE-4.5-VL 在 OCR 解析（包括古汉语识别和多语言文档）、文档理解（如病理报告的表格提取、产品说明书的摘要）、视频时间定位、视觉谜题、化学和数学推理、深度语义图像理解、视觉模式识别、表情符号测验、深度排序和计数等复杂多模态任务中的能力。例如，它能根据图片中的物体深度进行排序，并准确识别图中笔的颜色数量。

结论

ERNIE 4.5 系列模型通过其创新的异构 MoE 架构、自适应分辨率视觉编码器、精细的分阶段预训练、领域数据挖掘、以及全面的后训练策略，包括统一奖励系统和强化学习，实现了强大的语言和多模态理解与生成能力。同时，通过先进的量化技术、推理加速和完善的部署工具链，显著提升了模型的部署效率和硬件兼容性。各项基准测试和定性示例均表明，ERNIE 4.5 在多项复杂任务上达到了行业领先水平，特别是在中文和多模态场景下展现出卓越的性能。

ERNIE 4.5 概览

模型家族： 了解ERNIE 4.5系列中的不同模型（例如ERNIE-4.5-300B-A47B-Base、ERNIE-4.5-VL-424B-A47B等），包括它们是否为多模态、是否经过后训练、以及是否具备"思考"模式。

核心创新： 识别ERNIE 4.5与传统单模态MoE模型的主要区别，例如异构模态结构、跨模态参数共享（自注意力、专家参数共享）、专有模态参数。

关键技术点： 理解模态隔离MoE路由技术和多模态联合预训练在提升视觉信息学习和语言模型能力方面的作用。

架构

异构MoE (Mixture-of-Experts)： 理解MoE架构如何允许模型拥有多个“专家”网络，以及异构性如何支持不同模态的特定处理。

视觉编码器 (Vision Encoder)：图像编码： 掌握ERNIE 4.5如何处理图像输入，特别是自适应分辨率视觉编码器与传统固定分辨率ViTs的区别。

关键技术： 了解2D旋转位置嵌入 (RoPE) 在编码2D空间信息中的作用，以及图像打包技术如何提高计算资源利用率。

适配器 (Adapter)：作用： 理解适配器作为视觉编码器和语言模型之间模态桥接模块的功能。

压缩机制： 了解空间和时间令牌压缩的具体方法（例如2x2块上的空间压缩，以及序列长度的时间压缩），以及像素重排 (pixel shuffle) 在此过程中的应用。

统一处理： 掌握如何将静态图像处理为两帧合成视频以实现跨模态的时间建模一致性。

多模态位置嵌入 (Multimodal Position Embedding)： 理解其在统一不同模态令牌位置信息中的作用。

预训练 (Pre-Training)

预训练数据：数据来源： 了解预训练数据来自何处（纯文本、多样化领域数据）。

领域特定数据： 理解为了增强领域特定任务能力，如何构建大型数据集，并采用何种数据来源策略（渐进式挖掘、条件训练、ASR音频转录）。

REEAO：位确定性预训练数据管理器： 理解其在解决大规模模型训练中数据管道问题（如数据重复、遗漏）方面的作用。

预训练方案 (Pre-Training Recipe)：多阶段训练： 详细了解三个阶段的训练过程：

阶段I：纯文本训练： 目标、上下文长度和RoPE频率基线。

阶段II：纯视觉训练： 视觉部分的专门训练。

阶段III：多模态联合训练： 如何结合不同模态的数据进行训练。

长上下文训练： 理解如何通过逐步增加序列长度和调整RoPE频率基线来扩展模型的上下文处理能力。

模型优化：路由器正交化损失 (Router Orthogonalization Loss)： 理解其在MoE模型训练中的作用。

令牌平衡损失 (Token-Balanced Loss)： 掌握其引入的原因（解决梯度不平衡）和计算方式。

指数移动平均 (EMA)：作用： 理解EMA如何稳定训练动态并提高泛化能力。

理论分析： 了解EMA与学习率衰减的类比关系。

有效衰减窗口： 掌握如何通过EMA衰减系数α控制有效衰减窗口的大小。

后训练 (Post-Training)

大型语言模型 (LLMs) 的后训练：监督微调 (Supervised Fine-Tuning, SFT)： 了解SFT数据的分类、涵盖的十个主题领域。

统一奖励系统 (Unified Rewarding System)： 理解其组成部分：

基于参考的LLM作为评估器 (RLLM)： 如何利用LLM评估模型输出。

沙盒 (Sandbox)： 安全隔离的编程任务评估环境。

基于参考的判别性奖励模型 (RDRM)： 与传统模型的区别，以及如何利用参考答案指导评估。

非推理任务的奖励模型：清单感知验证器 (Checklist-Aware Verifiers)：基于明确、客观标准进行评估。

生成式奖励模型 (GRM)： 多维评估标准和动态反馈机制。

判别式奖励模型 (DRM)： 传统RL框架中通过判别任务学习奖励函数。

强化学习 (Reinforcement Learning)： 理解数据过滤、奖励信号解耦、分层处理和奖励归一化在增强训练稳定性方面的作用。

视觉语言模型 (VLMs) 的后训练：监督微调： 与LLM类似。

可验证奖励强化学习 (Reinforcement Learning with Verifiable Rewards)： 理解视觉谜题的RLVR训练如何利用两个LLM评估一致性和正确性，以及其不限制响应格式的优势。

训练框架 (Training Framework)

多模态模型训练的异构并行性：架构： 理解异构并行性如何共享MoE骨干网络和ViT编码器之间的并行布局。

分层负载均衡策略： 了解粗粒度负载均衡和细粒度动态平衡分区。

MoE骨干网络的混合并行性：节点内专家并行 (Intra-Node Expert Parallelism)： 理解如何在节点内实现专家并行。

内存高效流水线调度 (Memory-Efficient Pipeline Scheduling)： 了解1F1B（一次前向，一次后向）调度与传统VPP调度相比的内存效率优势。

FP8 混合精度训练： 理解FP8数据类型在减少内存使用和加速计算中的作用。

计算优化：带最佳计算-内存权衡的重计算 (Recomputation with Best Computation-Memory Tradeoffs)： 理解操作符级别重计算如何减少内存开销，并与传统方法的区别。

FlashMask： 理解其在灵活注意力掩码和长上下文训练中的应用。

框架原生容错系统 (Framework-Native Fault Tolerance System)： 理解其在处理大规模训练中的故障和波动方面的作用（例如零成本检查点、在线探查器、SDC扫描器）。

推理与部署 (Inference and Deployment)

概述： 理解ERNIE 4.5模型（MoE和密集模型）的部署灵活性，以及提供的量化方案（FP8、INT8、INT4、2比特）。

量化 (Quantization)：多精度支持： 了解BF16、FP8以及低精度选项（W4A8、2比特）。

2比特量化： 理解其模型大小缩减比例、部署要求，以及与向量量化和非相干处理方法的对比。

灵活量化配置： 了解头粒度 (head-wise) 和通道粒度 (channel-wise) 量化，以及静态量化。

注意力与KV缓存量化： 理解RHT在消除异常值方面的应用，以及其对内存和计算的优化。

推理加速：W4A8 内核加速： 理解CUTLASS内核和位移反量化。

高效注意力内核： 理解FP8/INT8计算在不同GPU架构上的应用，以及指数函数的多项式近似和FMA指令合并。

推测解码 (Speculative Decoding)： 理解其在加速推理方面的作用。

部署： 理解Prefill-Decode (PD) 分离部署、KV缓存传输设计、多机器部署中的动态角色切换。

开源开发工具 (Open-Source Development Tools)

ERNIEKit： 了解其功能。

FastDeploy： 掌握其作为推理和部署工具包的特点（PD分离与多级负载均衡、全面低比特量化推理支持、多硬件支持）。

评估与结果 (Evaluation and Results)

语言模型评估：基准测试： 了解用于评估预训练和后训练语言模型的各类基准（通用、事实知识、推理、代码生成与理解、数学推理）。

比较对象： 与DeepSeek-V3-Base、Qwen3-30B-A3B-Base、GPT-4.1等模型的性能对比。

多模态模型评估：基准测试： 了解用于评估多模态模型的各类基准（视觉知识、文档与图表、多模态推理、视觉感知、视频理解）。

比较对象： 与Qwen2.5-VL系列等模型的性能对比。

消融研究： 了解路由器正交化损失的有效性。

测验

请用2-3句话简要回答以下问题。

ERNIE 4.5的异构MoE结构与传统单模态MoE模型有何不同?

ERNIE 4.5的视觉编码器何处理固定分辨率图像输入这一挑战?

适配器 (Adapter) 在ERNIE 4.5的架构中扮演什么角色？它如何实现特征融合和序列长度缩减?

在ERNIE 4.5的预训练中，REEAO数据管理器解决了哪些问题?

解释ERNIE 4.5长上下文训练的两个主要子阶段是如何实现上下文长度扩展的。

什么是ERNIE 4.5中引入的"令牌平衡损失"？它解决了传统损失公式的什么问题?

ERNIE 4.5如何通过强化学习优化模型性能，特别是在处理奖励信号异构性方面?

描述ERNIE 4.5在推理和部署方面提供的两种主要低比特量化策略及其应用场景。

在高效注意力内核中，ERNIE 4.5如何针对不同GPU架构进行优化以提高softmax计算效率?

FastDeploy工具包为ERNIE 4.5的部署提供了哪些关键技术特性?

测验答案

ERNIE 4.5采用新颖的异构模态结构，支持跨模态参数共享（如自注意力、专家参数共享），同时为每个模态保留专用参数。这与传统单模态MoE模型不同，传统模型通常只在一个模态内进行专家混合。

ERNIE 4.5使用自适应分辨率视觉编码器，而不是强制将图像调整为正方形。它独立调整图像的高度和宽度到ViT块大小的最近倍数，从而大致保留原始宽高比，避免了固定尺寸调整引入的失真。

适配器在ERNIE 4.5中充当视觉编码器和语言模型之间的模态桥接模块。它通过空间和时间令牌压缩实现特征融合和序列长度缩减，其中空间压缩对2x2块操作，时间压缩将序列长度减少一半，并利用像素重排进行特征重组。

REEAO（位确定性预训练数据管理器）旨在解决大规模自回归语言模型训练中常见的数据管道问题。这些问题包括因检查点恢复、节点故障、资源调整等导致的数据意外重复或遗漏，确保了数据处理的位确定性。

ERNIE 4.5的长上下文训练分为两个子阶段：首先将最大序列长度扩展到32k，通过提高RoPE频率基线；然后进一步扩展到128k，再次提高RoPE频率基线，并对长上下文文档进行上采样，以确保模型充分接触长距离依赖。

“令牌平衡损失”是ERNIE 4.5为解决传统损失公式中梯度不平衡问题而引入的。传统损失公式对未掩码令牌较少的样本产生不成比例的更大梯度，而令牌平衡损失通过除以总序列长度（|Ui|+|Mi|）进行归一化，确保每个样本的损失贡献与其总长度成反比。

为了增强训练稳定性并优化模型性能，ERNIE 4.5在强化学习中对奖励信号进行了改进。这包括排除准确率为0或1的提示、过滤掉奖励信号组内方差不足的提示，以及在每个训练迭代中对不同来源和领域解耦、分层处理后的奖励进行独立归一化。

ERNIE 4.5支持多种低比特量化策略，包括W4A8精度和2比特权重专用量化。W4A8用于低成本场景以提高推理吞吐量，而2比特权重专用量化则将模型大小减少80%，旨在资源受限场景下降低部署门槛。

在高效注意力内核中，ERNIE 4.5针对Hopper架构GPU使用FP8计算，针对Ampere架构GPU使用INT8计算。它通过设计指数函数（用于softmax）的不同多项式近似和合并反量化步骤到FMA指令，最小化CUDA核心上的计算开销，从而提高效率。

FastDeploy为ERNIE 4.5的部署提供了多项关键技术特性，包括支持PD（Prefill-Decode）分离部署与多级负载均衡，具备全面的低比特量化推理支持（如W8A8、W2A16），以及得益于PaddlePaddle的多硬件适应能力，支持NVIDIA GPU、昆仑芯XPU、海光DCU和昇腾NPU等多种芯片。

论文格式问题

请详细阐述ERNIE 4.5中异构MoE架构的设计理念及其如何通过参数共享和专用参数来优化多模态模型的学习效率和能力。 分析ERNIE 4.5在预训练阶段如何通过“短上下文”和“长上下文”两个子阶段，并结合RoPE频率基线调整和数据上采样策略，逐步扩展模型的上下文处理能力，并探讨其对模型性能的影响。 比较ERNIE 4.5中用于大型语言模型 (LLMs) 后训练的“统一奖励系统”中的RLLM、Sandbox和RDRM，重点阐述它们在评估模型输出和引导强化学习过程中的不同机制和优势。 ERNIE 4.5的训练框架中，如何通过异构并行性、混合并行性（特别是节点内专家并行）以及内存高效流水线调度来应对大规模多模态模型训练的挑战？请结合文中提供的图示进行说明。 深入探讨ERNIE 4.5在推理和部署阶段所采用的量化策略，包括W4A8、2比特量化以及注意力与KV缓存量化。请分析这些量化技术如何协同作用以减少内存占用、加速推理，并平衡性能与精度。

关键词汇表

ERNIE 4.5： 百度开发的一个大型多模态预训练模型系列，以其异构Mixture-of-Experts (MoE) 架构和强大的语言及视觉理解能力为特色。 MoE (Mixture-of-Experts)： 专家混合模型，一种神经网络架构，其中模型的不同部分（专家）专门处理输入的不同部分，由一个"路由器"网络决定输入应由哪个或哪些专家处理。 异构 MoE (Heterogeneous MoE)： ERNIE 4.5特有的一种MoE结构，支持跨模态的参数共享，同时也允许为特定模态设置专用参数，以更有效地处理多模态数据。 视觉编码器 (Vision Encoder)：模型中负责处理和编码图像或视频输入的部分，通常将视觉信息转换为模型可以理解的向量表示。 Vision Transformer (ViT)： 一种基于Transformer架构的视觉模型，将图像分割成小块（patch）并将其视为序列进行处理。 自适应分辨率视觉编码器 (Adaptive-resolution vision encoder)： ERNIE 4.5中使用的视觉编码器，能够处理不同分辨率和宽高比的图像，避免了固定分辨率输入导致的图像失真。 2D 旋转位置嵌入 (2D Rotary Position Embedding, RoPE)： 一种位置编码技术，用于在Transformer模型中编码输入序列中令牌的相对或绝对位置信息，特别是在2D RoPE中用于编码图像块的二维空间位置。 图像打包 (Image Packing)： 一种将多张图像高效地打包到一个批次中进行处理的技术，旨在提高计算资源利用率。 适配器 (Adapter)： 在ERNIE 4.5中作为视觉编码器和语言模型之间的模态桥接模块，负责对视觉特征进行压缩和对齐，使其进入统一的嵌入空间。 像素重排 (Pixel Shuffle)： 一种图像处理操作，常用于超分辨率任务，通过重新排列像素来改变特征图的空间分辨率，在ERNIE 4.5中用于令牌的压缩。 REEAO (Bitwise-Deterministic Pre-Training Data Manager)： ERNIE 4.5中用于管理预训练数据的数据管理器，确保在大规模训练中数据处理的位确定性，避免数据重复或遗漏。 上下文长度 (Context Length)： 模型在处理输入时能够考虑的最大令牌数量，影响模型理解长距离依赖的能力。 路由器正交化损失 (Router Orthogonalization Loss)： MoE模型中一种损失函数，旨在鼓励路由器网络更有效地分配负载，并避免专家之间的冗余或重叠。 令牌平衡损失 (Token-Balanced Loss)： ERNIE 4.5引入的一种损失函数，通过归一化每个样本的损失贡献，解决传统损失公式中未掩码令牌数量不均导致的梯度不平衡问题。 指数移动平均 (Exponential Moving Average, EMA)： 一种参数平滑技术，在训练过程中计算模型参数的加权平均值，有助于稳定训练动态和提高模型泛化能力。 监督微调 (Supervised Fine-Tuning, SFT)： 在预训练模型的基础上，使用带有标签的数据集进行进一步训练，以适应特定任务或领域。 统一奖励系统 (Unified Rewarding System)： ERNIE 4.5后训练阶段采用的综合奖励机制，包括RLLM、Sandbox、RDRM等组件，用于评估模型输出并指导强化学习。 LLM-as-a-Judge (RLLM)： 一种评估方法，利用大型语言模型本身作为评估器，根据参考答案来判断另一个模型生成输出的质量。 沙盒 (Sandbox)： 一个安全隔离的执行环境，用于运行和测试模型生成的代码或计算任务，以直接评估其功能和正确性。 判别性奖励模型 (Discriminative Reward Model, DRM)： 强化学习框架中的一种奖励模型，通过学习判别任务来预测给定输出的质量或偏好，从而为模型提供奖励信号。 生成式奖励模型 (Generative Reward Model, GRM)： ERNIE 4.5中一种更高级的奖励模型，整合多维评估标准和动态反馈机制，对模型输出进行更细致的评估。 清单感知验证器 (Checklist-Aware Verifiers)： 一种基于预定义明确、客观标准的评估方法，确保模型生成的响应满足特定的规范。 异构并行性

(Heterogeneous Parallelism)： 在多模态模型训练中，同时利用不同类型的并行策略（如模型并行、数据并行）来优化计算和内存效率。 混合并行性 (Hybrid Parallelism)： 结合多种并行策略，如数据并行、模型并行和专家并行，以高效训练超大规模模型。 节点内专家并行 (Intra-Node Expert Parallelism)： 在单个计算节点内部实现专家并行，以优化MoE模型的训练效率。 内存高效流水线调度 (Memory-Efficient Pipeline Scheduling)： 一种训练调度策略，旨在减少模型训练过程中的内存消耗，例如1F1B（一次前向，一次后向）调度。 FP8 混合精度训练 (FP8 Mixed Precision Training)： 在模型训练中使用8位浮点数（FP8）与其他精度（如BF16）混合进行计算，以减少内存占用和加速训练。 重计算 (Recomputation)： 在反向传播过程中，重新计算一些前向传播的激活值而不是将其存储在内存中，以减少内存消耗。ERNIE 4.5采用"操作符级别重计算"进行优化。 FlashMask： 一种用于灵活注意力掩码和长上下文训练的技术，旨在提高注意力机制的效率。 框架原生容错系统 (Framework-Native Fault Tolerance System)： 深度学习框架内置的故障处理机制，确保在大规模训练过程中面对硬件故障或软件错误时能够稳定运行并恢复。 量化 (Quantization)： 将模型参数或激活值从高精度（如FP32、BF16）转换为低精度（如INT8、FP8、INT4、2比特），以减少模型大小和加速推理。 W4A8 量化： 一种量化方案，权重使用4位（W4），激活使用8位（A8）。 2比特量化 (2-Bit Quantization)： 极低精度的量化，将模型大小显著减小，但可能带来精度挑战，ERNIE 4.5实现了"近乎无损"的2比特量化。 KV 缓存量化 (KV Cache Quantization)： 对Transformer模型中的键（Key）和值（Value）缓存进行量化，以减少推理时的内存占用，尤其对于长上下文模型。 推测解码 (Speculative Decoding)： 一种推理加速技术，使用一个小型、快速的模型来预测大部分输出，然后用一个大型、准确的模型验证并修正这些预测。 FastDeploy： 一个推理和部署工具包，为ERNIE 4.5等大型模型提供高效、便捷的部署解决方案，支持多种量化配置和硬件平台。 PD 分离部署 (Prefill-Decode Disaggregation Deployment)： 一种分布式推理部署策略，将预填充（Prefill）和解码（Decode）阶段分离，通过动态负载均衡提高吞吐量。

参考

官网blog

Tech Report

Hugging Face

GitHub

AI Studio

分享这篇文章

相关文章推荐

多智能体强化学习（MARL）在多...

本文介绍了多智能体强化学习（MARL）在多智...

Cursor AI 最佳实践：提升编码效...

Cursor AI 最佳实践：提升编码效率与代码质量...

Chain of Draft 论文解读

本文介绍了 Chain of Draft（CoD）论文，...