南京郵電大學
Nanjing University of Posts and Telecommunications

Deepseek

汇报人：庄智杰

2025年2月17日

报告大纲

Contents

1. 技术原理
2. 与国内外大模型的对比
3. 应用

1. 技术原理

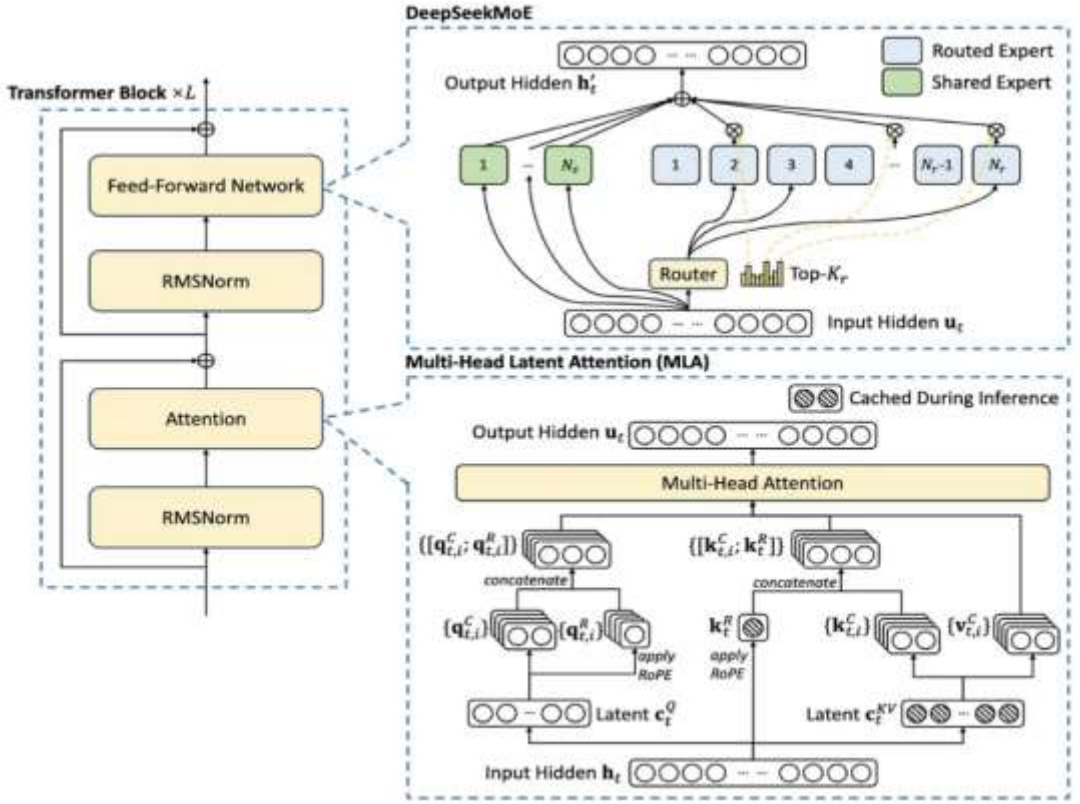
DeepSeek-V3 在推理速度上相较历史模型有了大幅提升，在目前大模型主流榜单中，DeepSeek-V3 在开源模型中位列榜首，与世界上最先进的闭源模型不分伯仲。DeepSeek-V3 主要采用了**多头潜注意力**(MLA对传统多头注意力机制的改进) 和 **DeepSeekMoE架构**（对传统MoE架构的改进）。

- **多头潜注意力（MLA）**

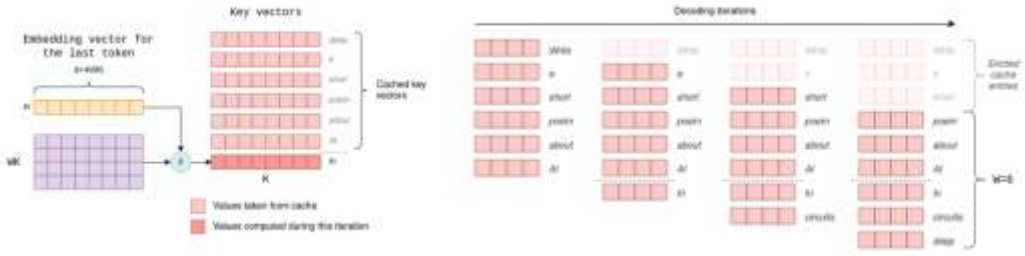
一种改进的注意力机制，旨在提高Transformer模型在处理长序列时的效率和性能。优化了**键值（KV）矩阵**，显著减少了内存消耗并提高了推理效率

- **DeepSeekMoE架构**

融合了**专家混合系统(MoE)**、**多头潜在注意力机制**和**RMSNorm**三个核心组件。通过专家共享机制、动态路由算法和潜在变量缓存技术，该模型在保持性能水平的同时，实现了相较传统MoE模型40%的计算开销降低。

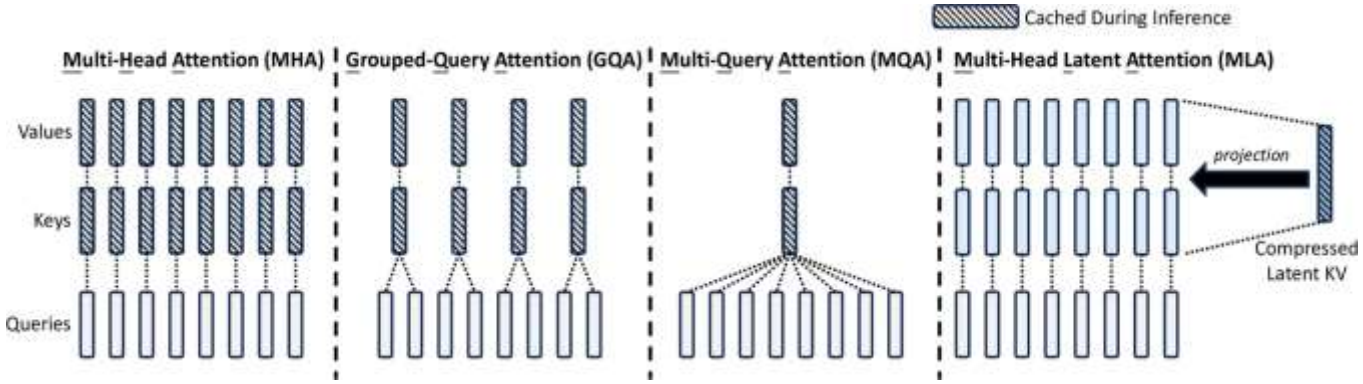


1. 技术原理-多头潜注意力 (MLA)



KV Cache：缓存所有过去 Token 的相关内部状态，主要是注意力机制中的键（Key）和值（Value）向量

DeepSeek使用的**MLA技术**大大节省KV缓存，降低了计算成本。本质是对KV的有损压缩，同时尽可能保留关键细节。与分组查询和多查询注意力等方法相比，MLA是目前开源模型里显著减小 KV 缓存大小的最佳方法。

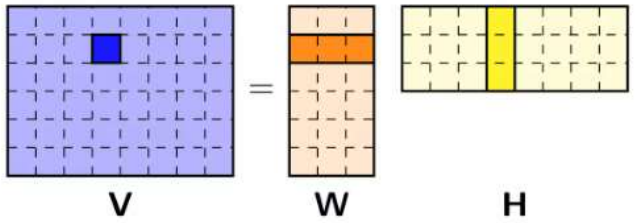


- If

$$V = WH$$

- then

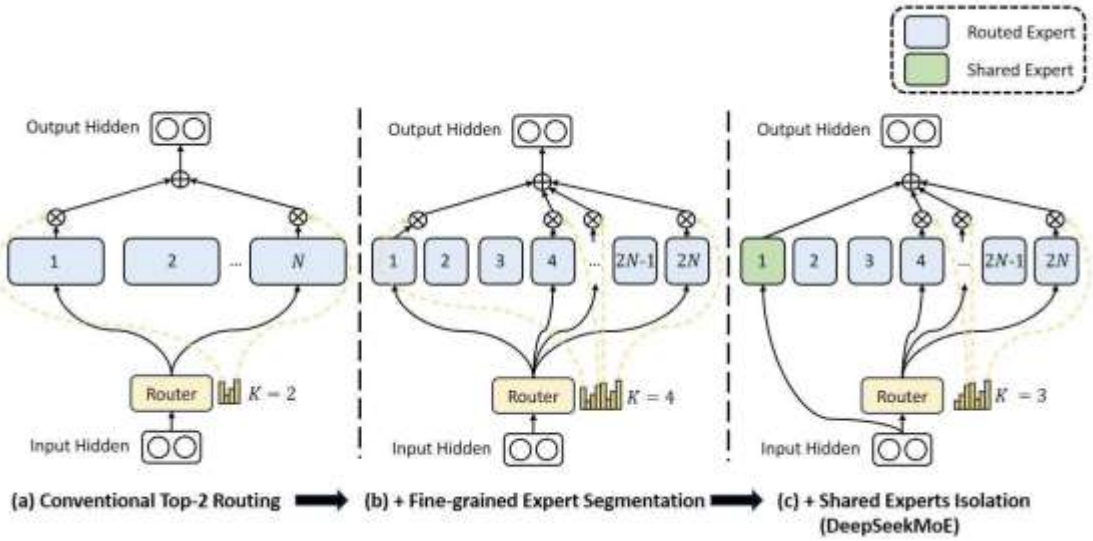
$$v_{i,j} = w_{i,*}h_{*,j} = \sum_{x=1}^{k} w_{i,x}h_{x,j}$$



MLA的方法是**将KV矩阵转换为低秩形式**：将原矩阵表示为两个较小矩阵的乘积，在推断过程中，仅缓存潜向量。规避了分组查询注意力和多查询注意力的查询的信息损失，从而在降低KV缓存的前提下获得更好的性能。

1. 技术原理-MoE架构

“增加共享专家” + “无辅助损耗负载平衡”



通过动态调整，DeepSeek-V3 在训练过程中获得了
比有辅助损失均衡模型更好的性能。

传统 VS MoE架构

特性	稠密模型（Dense）	MoE架构
参数利用率	全参数参与计算	稀疏激活（仅调用部分专家）
扩展方式	增加模型深度/宽度	横向增加专家数量
计算效率	计算成本与参数规模线性增长	计算成本与激活专家数相关
典型场景	中小规模通用任务	超大规模多任务/多模态场景

报告大纲

Contents

1. 技术原理
2. 与国内外大模型的对比
3. 应用

2. deepseek与国内外其他模型对比

模型名称	模型架构	训练数据	训练成本	生成速度	开源情况	多语言能力	推理能力	长文本处理	硬件需求
DeepSeek-R1	Transformer + **混合专家(MoE)架构**，融合**强化学习优化**	多领域高质量数据，侧重中文及多语种	极具性价比（数百万美元级别，**远低于同类顶级模型**）	**极快**，推理效率约为**GPT-4 Turbo的17%**	商用版闭源（部分技术或接口可能开放）	主要以中文为主，同时支持英文及其他少数语种	强劲，在数学、代码与自然语言推理任务上表现突出	支持长文本上下文处理，适合复杂文档与问答	高性能GPU集群，采用低成本高效优化方案
ChatGLM3 (智谱/清华)	Transformer（双语优化，侧重中文对话）	大规模中英文数据，中文占优	较低，相对平价开源大模型	快，适合**低资源设备部署**	开源（部分版本）	强于中文，支持中英文对话	优秀于中文对话及**知识问答**	支持扩展上下文（部分版本支持长文本）	适合低成本部署，**普通GPU或云服务均可**
GPT-4 (OpenAI)	Transformer解码器 + RLHF	数万亿tokens，覆盖多领域、多语种	数十亿美元级别（极高）	高（依赖专用云服务器）	闭源	非常强，支持多语种	业界领先，逻辑与创意生成均出色	**支持8K至32K上下文**	高端GPU群 A100/H100等
文心一言 (百度)	Transformer（解码器或编码器-解码器混合）	以中文为主的大规模数据，兼顾多领域	高（依托海量中文数据）	较快，针对中文场景优化	闭源	主要**擅长中文**，部分多语种支持	优秀于中文自然语言处理	支持较长上下文，适合长文生成与问答	高性能服务器，适合云端部署

2. deepseek与国内外其他模型对比

模型名称	模型架构	训练数据	训练成本	生成速度	开源情况	多语言能力	推理能力	长文本处理	硬件需求
PaLM 2 (Google)	Transformer解码器	数万亿tokens，多语种跨领域	极高	高效（推理经过优化）	闭源	强，支持多种语言	优秀，特别在**数学与编码任务**上	支持较长上下文（具体长度未公开）	高性能GPU集群
LLaMA 2 (Meta)	Transformer解码器	数十亿至万亿tokens，主要以英文为主	较GPT-4低，但总体成本仍较高	快，优化了推理效率	部分开源	较强，但**重点为英文**	表现优异，**适合学术与商业应用**	标准约4K tokens	依赖高性能GPU，相对门槛较低
通义千问 (阿里)	Transformer（针对商业场景优化）	大规模数据集，覆盖多领域，中文优势明显	高（企业级投入）	快，优化了推理流程	闭源	以中文为主，支持部分外语	强，**适合复杂商业应用**	支持较长文本输入	高性能云服务器
讯飞星火 (科大讯飞)	Transformer（语音与文本结合）	大规模语音与文本数据，中文为主	较高，依赖专业语音数据	快，响应迅速	闭源	主要支持中文，外语支持有限	强于语音识别与生成，中文理解优秀	**适合短对话与语音场景**	专业语音处理硬件+高性能服务器）

报告大纲

Contents

1. 技术原理
2. 与国内外大模型的对比
3. 应用

3. 应用能力

自然语言处理（NLP）

01

文本生成：生成文章、对话、创意内容。

信息理解与问答：解析复杂问题,提供精准答案。

多语言支持：支持多种语言的翻译和交互。

情感分析与意图识别：识别用户情绪和需求。

多模态能力

02

图像理解：解析图片内容。

跨模态生成：根据文本生成图像，或根据图像生成文本内容。

数据分析与决策支持

03

数据清洗与处理：自动清洗杂乱数据，提取关键字段。

可视化与报告生成：将数据转化为图表或总结性报告。

预测建模：基于历史数据预测趋势（如销量、用户行为）。

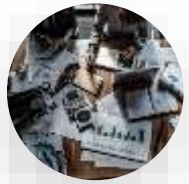
个性化推荐与交互

04

内容推荐：根据用户兴趣推荐文章、产品或服务。

对话式交互：通过聊天机器人提供全天候客服。

3. 案例示范



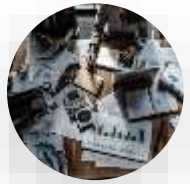
• 某电商平台引入DeepSeek-Pro，提供7x24小时全天候自动化应答，并支持多轮对话与情感智能分析，客服问题解决率飙升40%，人力成本锐减60%。

• 某营销团队使用 DeepSeek 生成创意文案并结合用户画像推荐最佳投放渠道，提升其点击率

• 某云计算公司使用 DeepSeek 实时解析日志数据，识别异常模式，简化服务器日志分析，加速故障定位。

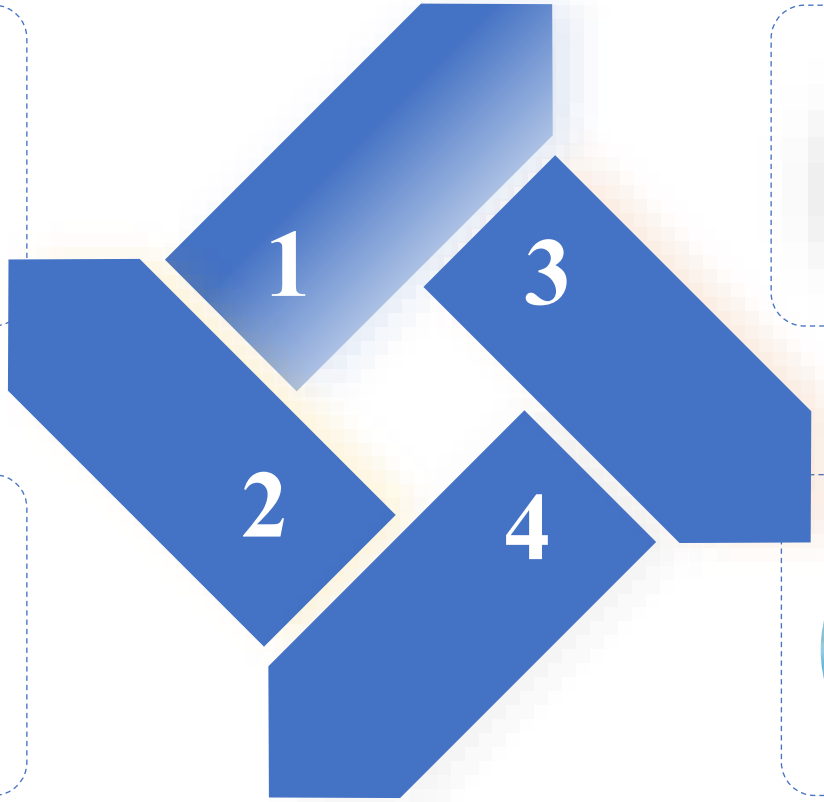
• 某医院利用Deepseek进行将医生语音记录转为结构化文本，自动生成标准化病历模板，提升医疗辅助手段。

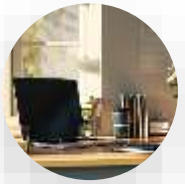
3. 案例示范



企业场景:

智能客服
金融风控
法律文档分析





创意与内容:

广告文案生成
短视频脚本设计

教育场景:

个性化学习助手
论文润色与查重

医疗场景:

病历结构化处理

Thank You !