Deepseek V3 Technical Report

(部分一)

姓名：王宏博

年级：研究生三年级

参考文献

github - DeepSeek-V3 Technical Report
https://github.com/deepseek-ai/DeepSeek-V3/blob/main/DeepSeek_V3.pdf

bilibili deepseek v3 全网最硬核解读
https://www.bilibili.com/video/BV1XocqepEsv?spm_id_from=333.788.player.switch&vd_source=6374cb49b8a20f29e86ab9de12471afa

知乎 - Deepseek v3 技术报告万字硬核解读
https://zhuanlan.zhihu.com/p/16323685381

目录

- 性能
- 模型架构
- 训练机制创新
- 部署测试

性能

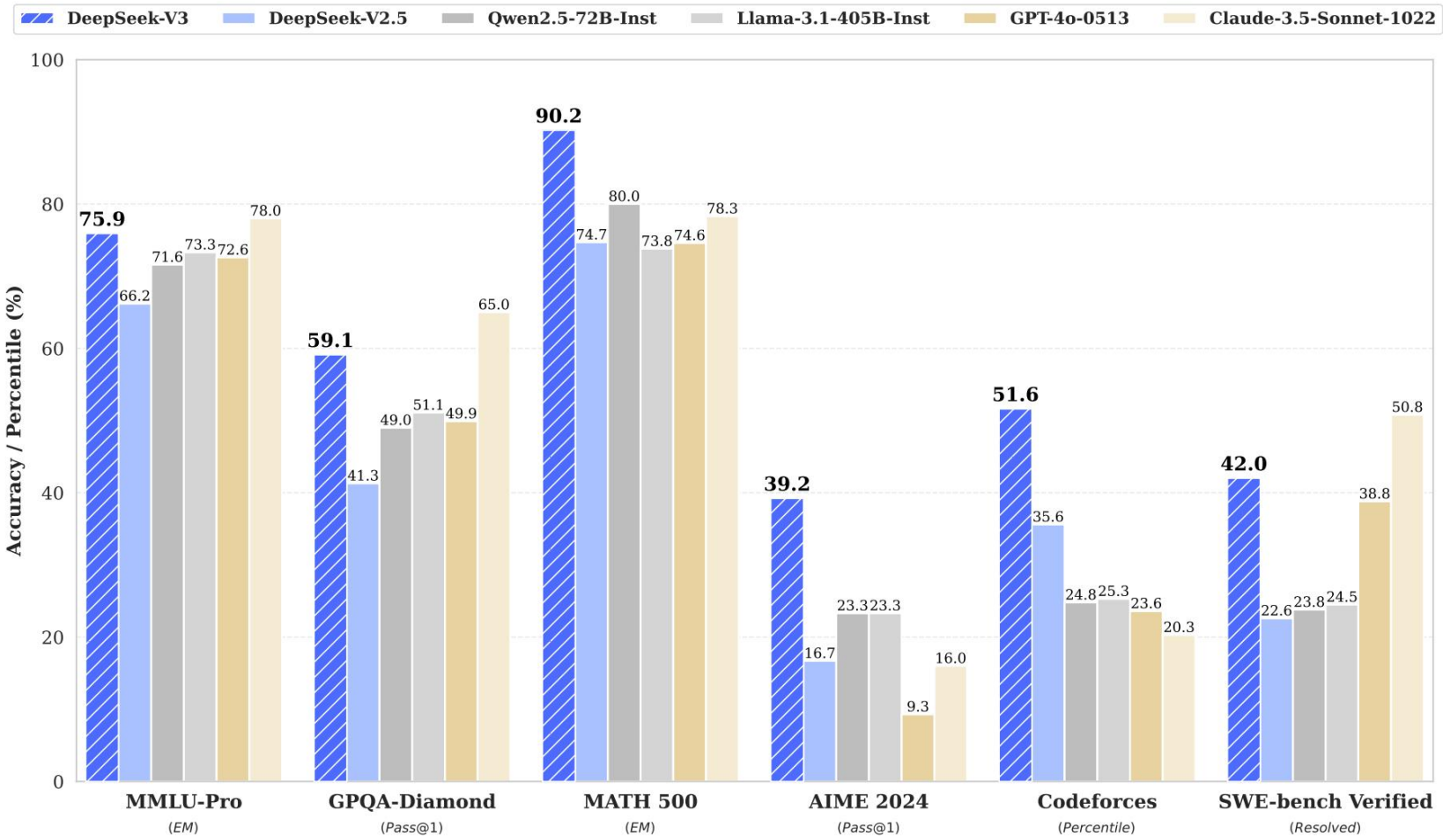
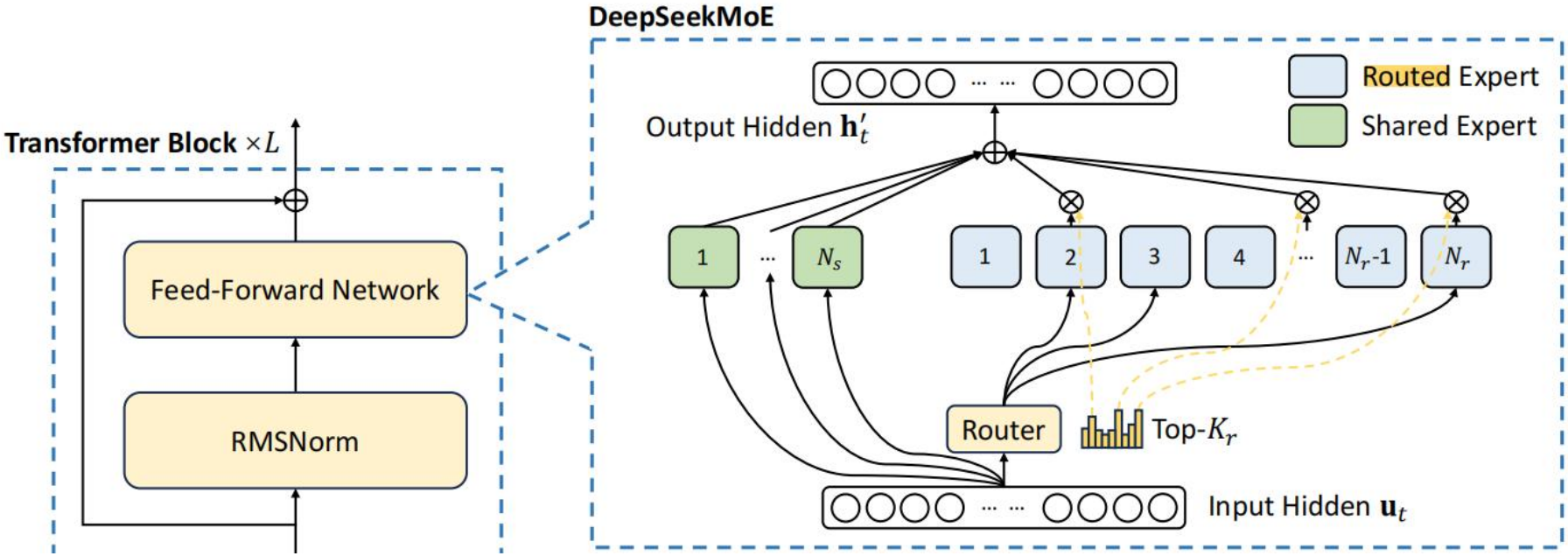


Figure 1 | Benchmark performance of DeepSeek-V3 and its counterparts.

MoE (Mixture of Experts) 架构



- MoE是FFN层的有效替代，使用专家系统将任务细分
- 可以精准分配任务，同时有效减少计算量
- 计算量可减少10的数量级

MoE (Mixture of Experts) 架构

- 计算量可减少10的数量级(案例)

3.2.1原始 MLP 的计算量

- **MLP 结构**:
  - 两个矩阵:
    - 第一个矩阵: `[h, 2.5h]`。
    - 第二个矩阵: `[2.5h, h]`。
  - 每个 token 向量的计算量为:

$$\text{计算量} = h \times 2.5h + 2.5h \times h = 5h^2$$

3.2.2 MoE 的计算量

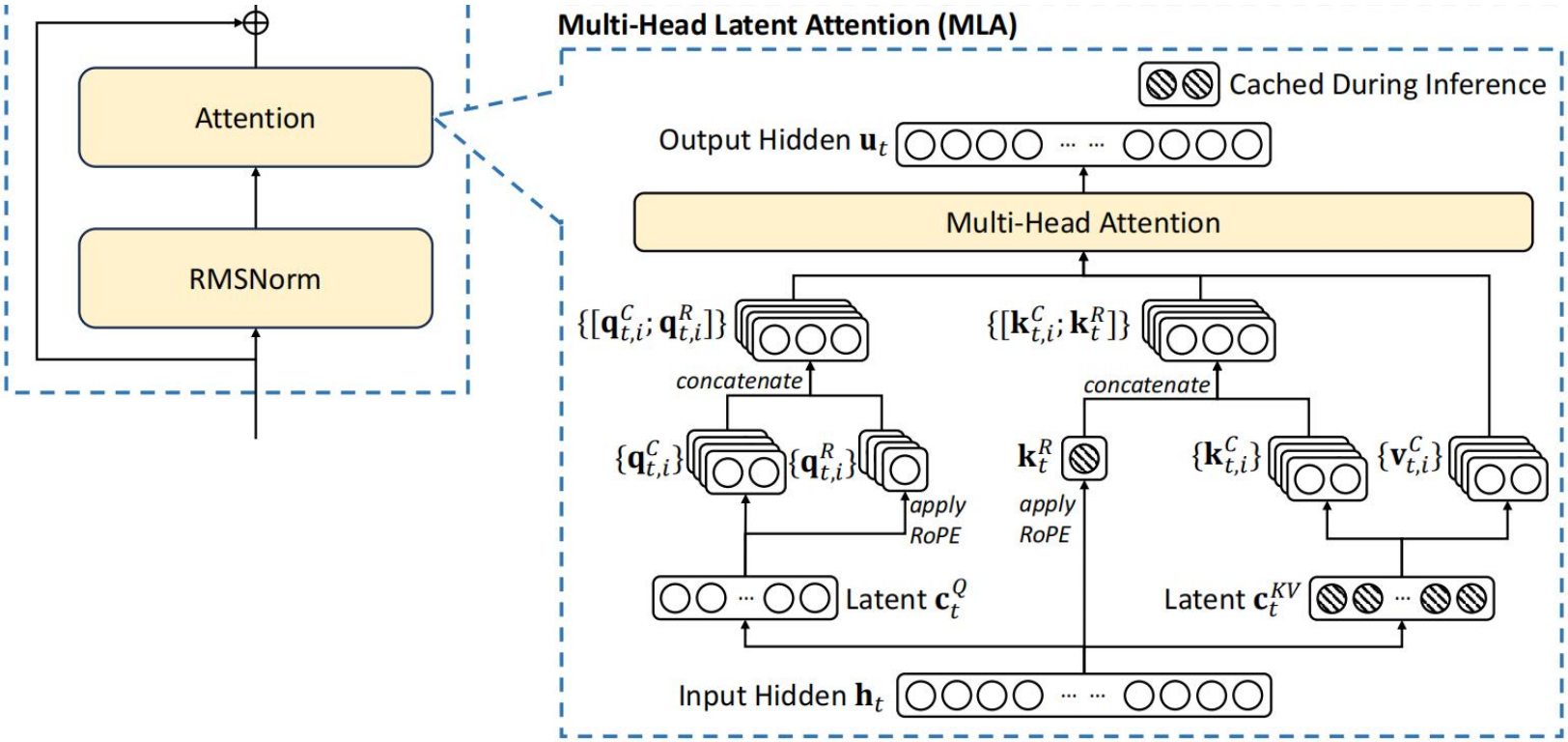
- **MoE 结构**:
  - 假设有 `n` 各专家，每次选用 `k` 个专家。
  - 每个专家的两个矩阵:
    - 第一个矩阵: $[h/n^{1/2}, 2.5 * h/n^{1/2}]$。
    - 第二个矩阵: $[2.5 * h/n^{1/2}, h/n^{1/2}]$。
  - 每个 token 每个专家的计算量为:

$$\text{计算量} = h/n^{1/2} \times 2.5h/n^{1/2} + 2.5h/n^{1/2} \times h/n^{1/2} = 2.5h^2/n + 2.5h^2/n = 5h^2/n$$

  - `k` 各专家的平均每 token 计算量为:

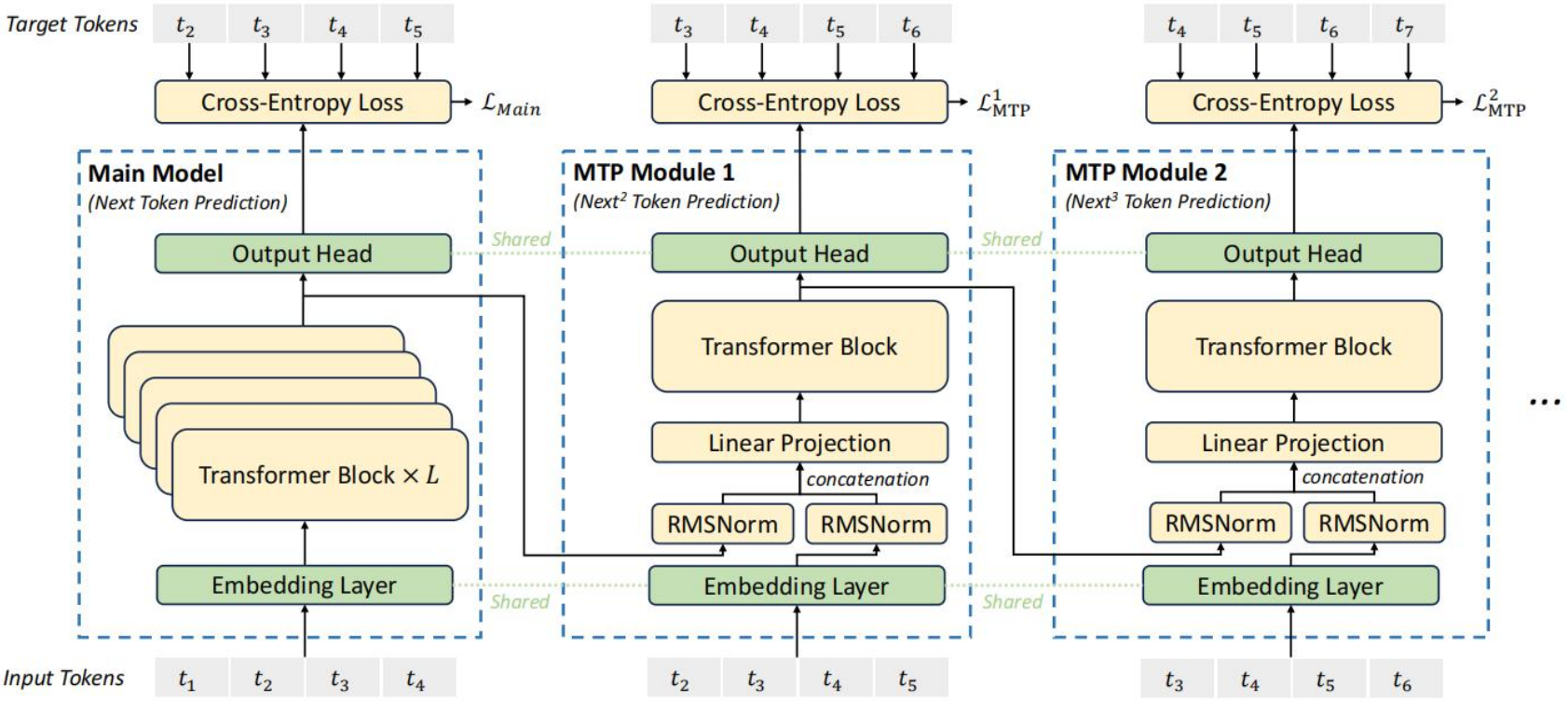
$$\text{计算量} = k \times 5h^2/n = 5kh^2/n$$

MLA架构（Multihead Latent Attention）



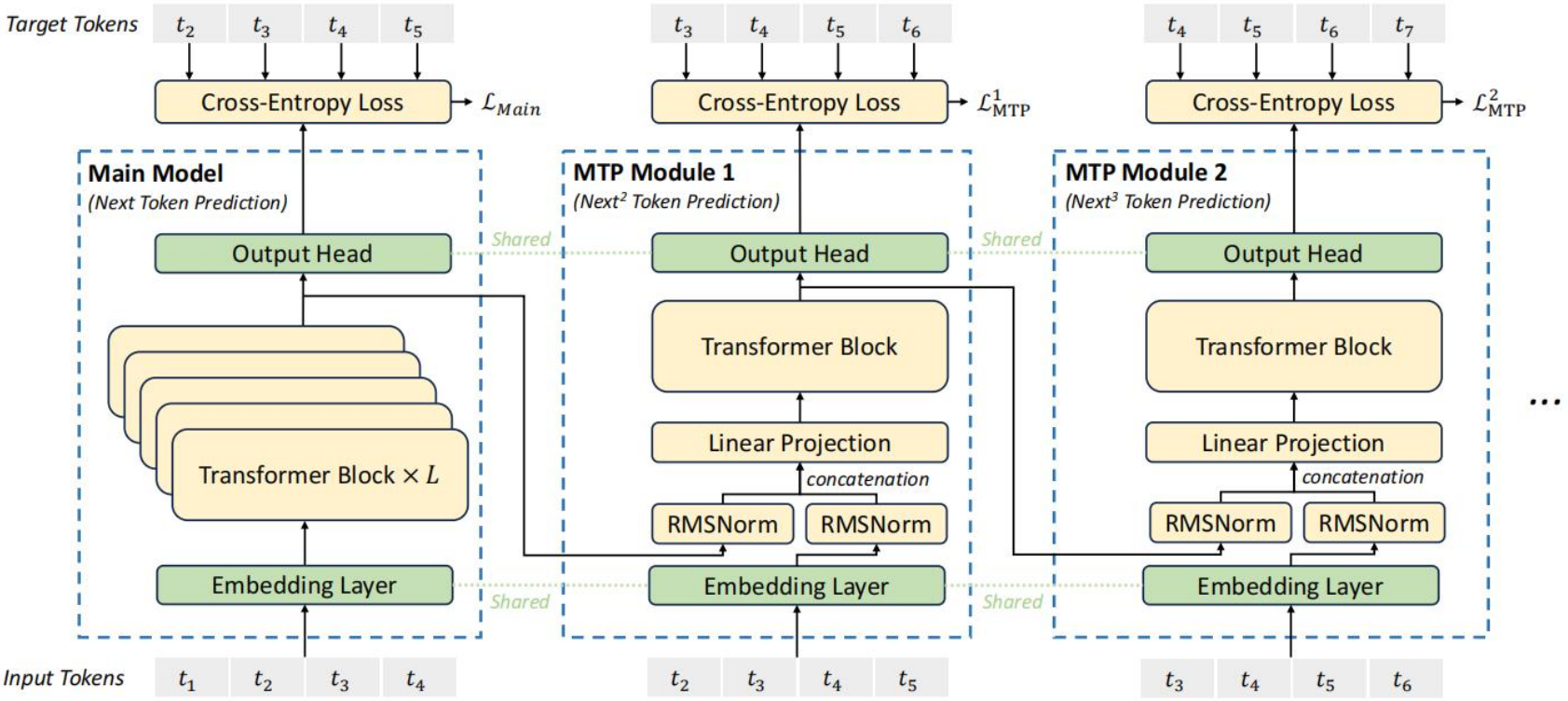
- MoE是通过低秩变换，减少KV cache 期间的显存占用
- Q,K,V均进行压缩操作

MTP训练（Multi Token Prediction）



- 训练过程为预测一个token t1 后的多个token，而不是单一的next-prediction

部署测试（Multi Token Prediction）



- 训练过程为预测一个token t1 后的多个token，而不是单一的next-prediction

谢谢！请多提意见！

http://ir.dlut.edu.cn