显存与模型大小对应关系

表格

描述已自动生成

显存需求计算方法

* **全精度（float32）**：每10亿个参数大约需要4GB显存。
* **半精度（FP16/BF16）**：每10亿个参数大约需要2GB显存。
* **低精度（int8）**：每10亿个参数大约需要1GB显存。
* **极低精度（int4）**：每10亿个参数大约需要0.5GB显存。
* 例如，DeepSeek的7B模型在使用FP16时，显存需求约为14GB，而使用int8时则降至7GB[1](https://www.zhihu.com/question/612046818/answer/3117938195)[2](https://segmentfault.com/a/1190000046068130)[4](https://finance.sina.com.cn/chanjing/gsnews/2025-01-28/doc-inehnfyp0625492.shtml)。
* 注意事项
* 在实际应用中，除了模型本身的显存需求外，还需考虑运行时的临时变量和其他内存占用，因此建议留有一定的余量以避免OOM（内存溢出）问题。
* 分享
* 改写
* 安装Ollama

curl -fsSL https://ollama.com/install.sh | sh

运行 默认7B

ollama run deepseek-r1

docker 安装webui

docker run -d -p 3000:8080 --add-host=host.docker.internal:host-gateway -e OLLAMA\_BASE\_URL=http://host.docker.internal:11434 -v open-webui:/app/backend/data --name open-webui --restart always ghcr.io/open-webui/open-webui:main

执行完成

以后启动使用

docker start open-webui

访问地址http://localhost:3000/

使用webui 开启Ollama的远程地址

1. 编辑Ollama的服务文件（通常位于 /etc/systemd/system/ollama.service）：

bash

sudo nano /etc/systemd/system/ollama.service

1. 在 [Service] 部分添加或修改以下行：

text

Environment="OLLAMA\_HOST=0.0.0.0"

1. 保存文件后，重新加载systemd并重启Ollama服务：

bash

sudo systemctl daemon-reload

sudo systemctl restart ollama

删除容器

docker stop open-webui

docker rm open-webui

docker volume rm open-webui