占位符

大模型技术内容

关于硬件和操作系统的要求

大模型技术内容

关于硬件和操作系统的要求

占位符

- 在实际的大模型微调技术落地应用时，硬件和软件同等重要；
- 若只使用OpenAI的在线大模型及在线微调API，则无需考虑硬件及计算资源的问题，只需要按计算量付费即可；
- 若是进行开源大模型部署及微调，则需要提前规划计算资源，购买硬件进行本地计算或者购买在线GPU算力均可；

占位符

大模型技术内容

关于硬件和操作系统的要求

- 课程对硬件的要求核心体现在开源大模型本地部署时进行推理及微调时对GPU显存的消耗。相比之下，对CPU和内存要求并不高；（尽管ChatGLM-6B支持在Intel CPU和Mac上运行，但较为繁琐且占用内存至少在32G以上，因此更推荐在GPU上运行）
- 课程对硬件的最低要求是根据ChatGLM-6B的推理和微调时显存占用进行推荐和要求；
- 除了硬件外，操作系统方面推荐使用Linux系统，相比Window或MacOS，Linux对大模型推理和微调方面功能支持更加完善，同时也是工业场景中使用最多的操作系统；

课程大模型技术阶段

GPU硬件要求及成本

占位符

- 理论上：在进行少量对话时不同精度对显卡要求

量化等级	推理时 GPU显存占用	高效微调时 GPU显存占用	最低配置 显卡	显卡显存	购买硬件 成本	租赁GPU成本（/小时）
全精度	20G	22G	3090	24G	8500	30
半精度	13G	14G	3090	24G	8500	30
INT8	8G	9G	2080ti	11G	2200	12
INT4	6G	7G	2060s	8G	1200	10

课程大模型技术阶段

GPU硬件要求及成本

占位符

- 实际上：在进行多轮对话时不同精度对显卡要求

量化等级	推理时 GPU显存占用	高效微调时 GPU显存占用	最低配置 显卡	显卡显存	购买硬件 成本	租赁GPU成本（/小时）
全精度	30G	22G	3090双卡	48G	8500*2	50
半精度	20G	14G	3090	24G	8500	30
INT8	12G	9G	3080ti	12G	4200	15
INT4	10G	7G	2080ti	11G	2200	12

复现课程微调过程
最低配置要求

占位符

- **GPU**：RTX 2060s显卡，8G显存；可以进行INT 4量化模式下高效微调；
- 其他硬件设备没有要求，GPU也可考虑租赁在线算力；
- 显卡显存越大，可以支持更高精度的微调，以及更多轮的对话；
- 若想完整本地复现课程全部内容，则需要至少双卡GPU的配置以及安装Linux操作系统；
- 相比较高的本地硬件门槛，更推荐购买在线GPU算力进行课程学习。在线GPU算力和本地硬件运行并无本质区别；

大模型部分课程教学
采用的硬件版本

占位符

同样适用于大多数消费级实验环境

- **GPU**：3090双卡，涡轮版；总共48G显存，能够适用于大多数试验和复现性质深度学习任务；同时双卡也便于模拟多卡运行的工业级环境；
- **CPU**：AMD 5900X；12核24线程，模拟普通服务器多线程设置；
- **存储**：64G内存+2T SSD数据盘；内存主要考虑机器学习任务需求；
- **电源**：1600W单电源；双卡GPU的电源在1200W-1600W均可；
- **主板**：华硕ROG X570-E；服务器级PCE，支持双卡PCIE；
- **机箱**：ROG太阳神601；atx全塔式大机箱，便于高功耗下散热；

大模型技术试验与应用的硬件理解及配置升级方案

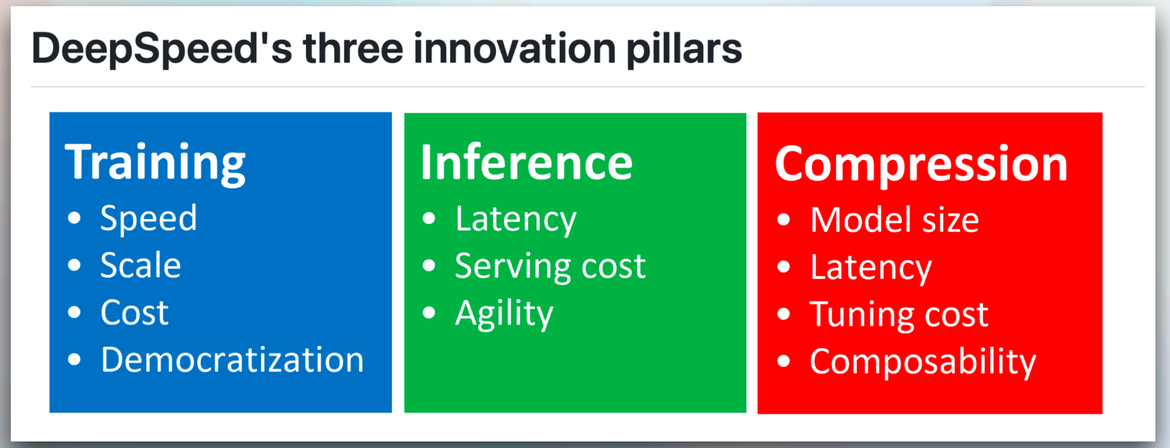
占位符

- 多显卡配置时需要注意，推荐购买涡轮版显卡（后端散热，更利于多卡散热），但如果是家用的话噪声很大，同时注意电源规格及主版是否支持双卡；
- 消费级配置最多支持双卡，4卡或者更多GPU配置，推荐购买服务器；
- 3090比4090综合性价比更高，不过4090计算速度几乎是3090的两倍，有需求亦可考虑升级，不过4090需要的机箱空间更大、电源配置也要求更高；
- 双卡GPU升级路线：3090—>4090—>A100 40G—>A100 80G；
- 大模型的工业级实践要求，大模型全量微调需要至少4张A100 80G显卡；

Linux操作系统与DeepSpeed并行训练框架

占位符

- 课程将介绍基于DeepSpeed的深度学习训练流程，前者是微软开源的深度学习优化库，旨在为大规模训练任务提供更高效的训练策略和GPU并行方法；
- RFHL方法则是由DeepSpeed的一个子项目DeepSpeed Chat提供；
- Windows下只支持部分DeepSpeed功能实现，因此课程会介绍Ubuntu操作系统的安装和使用；



- 项目地址： https://github.com/microsoft/DeepSpeed