一大型型部署

拉供工程设计 部署城市格型栅发空车的较级 被 被 使用的过程

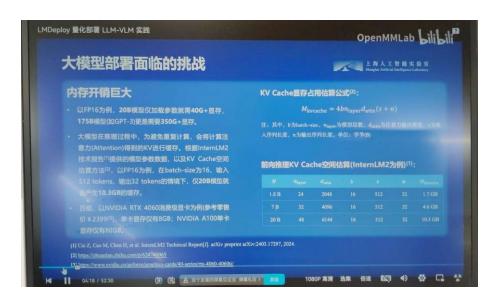
人之智能够域:模型部署是实现浑度学习 算法落地应用的关键步骤。简和言义 就是特洲练好的深度学习模型在特定新境 中运行的过程

场景: GU,草GPU/TPU/MPU,多卡集群 移动器/西缘路:移名的机器人,手机...

批学:计算量巨大



创新新校区大





3次有研究三、安置方法:

罗拉型 整枝:移除模型中不少安的剪多系的 一种,比较多数,从使模型更高效。——) 对少存储器 长,提高计算效率

级松松菜枝

②然识蒸缩:一种经典的模型压缩方法,核心思想是漏过对导转量化的净生模型"模仿"性能更好,结构更复杂的教师模型,在不改变年生模型结构的情况下提高模型性能。

②量化:将代统的线数的表示方法转化为整数式型记客额方式,以减轻深度常及模型的存储和计算负担



12. LM Deploy

