NVIDIA TAO Toolkit 学习摘要

TAO（Train, Adapt, Optimize）Toolkit 是 NVIDIA 推出的低代码 AI 工具链，基于迁移学习和模型优化技术，帮助开发者快速构建、微调并优化适用于边缘设备的轻量级模型。

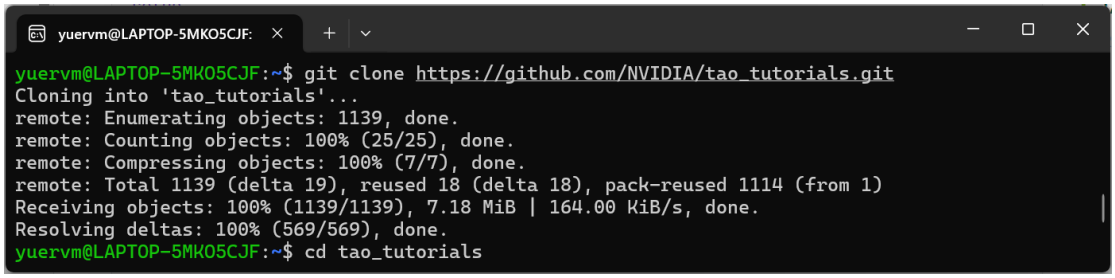

图 1：手动下载 TAO 包

获取 NGC API KEY：

$ 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

图 2：登录成功

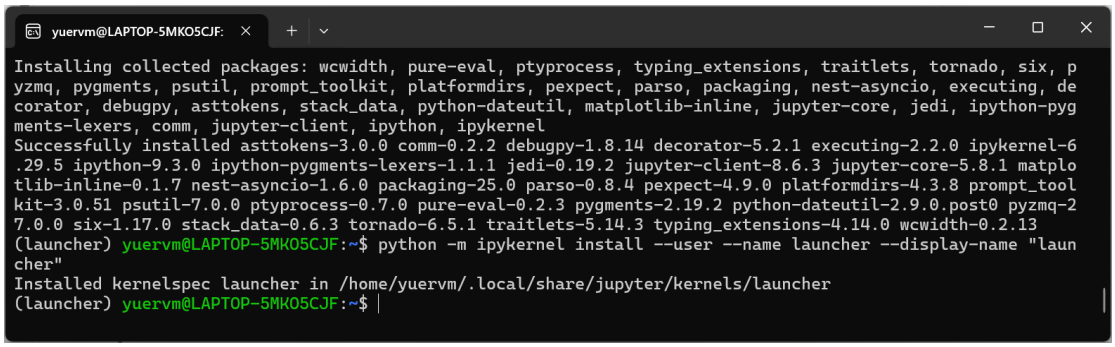

图 3：设置 Python 环境

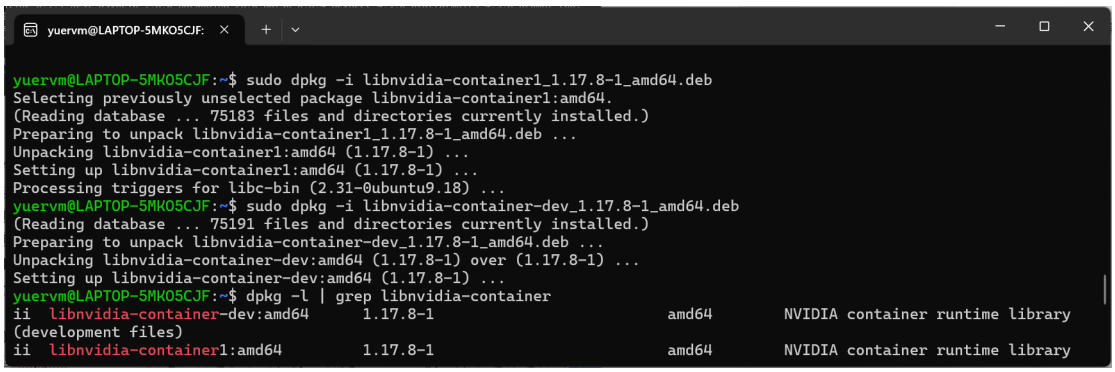

图 4：手动下载和安装 nvidia container

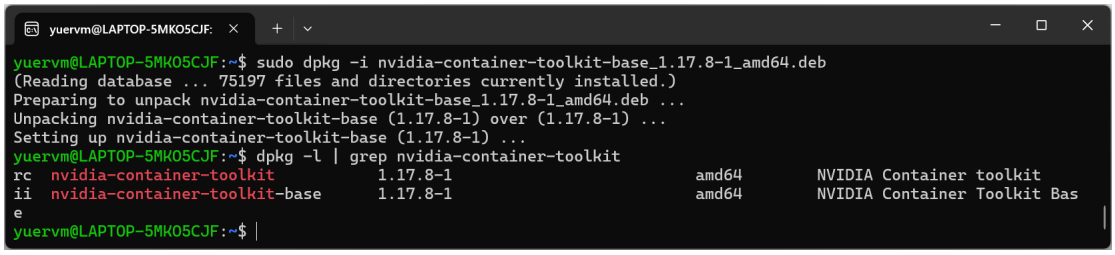

图 5：手动下载和安装 nvidia container toolkit

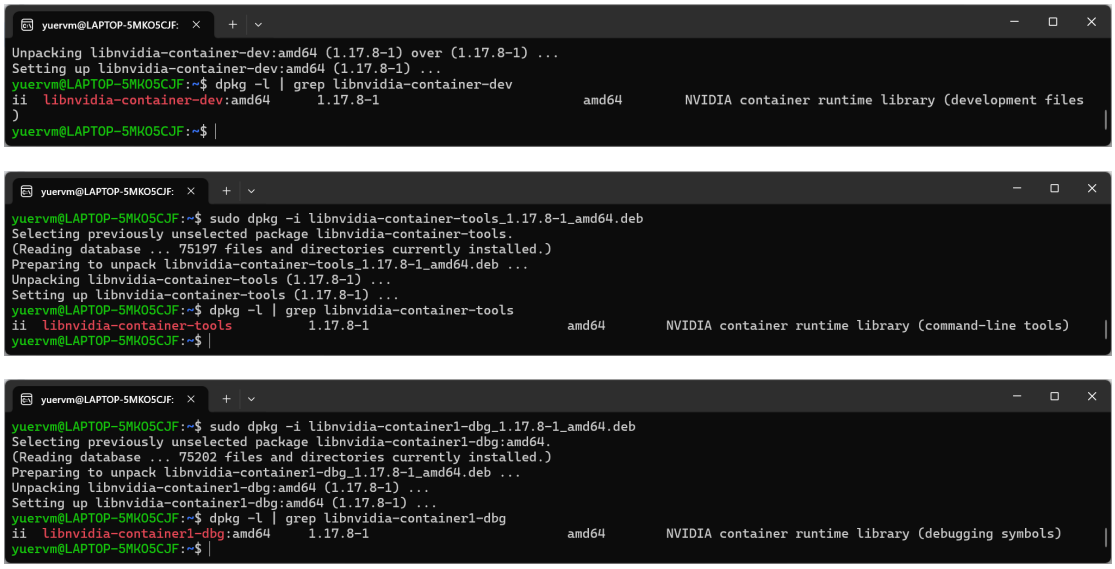

图 6：安装 nvidia container 包

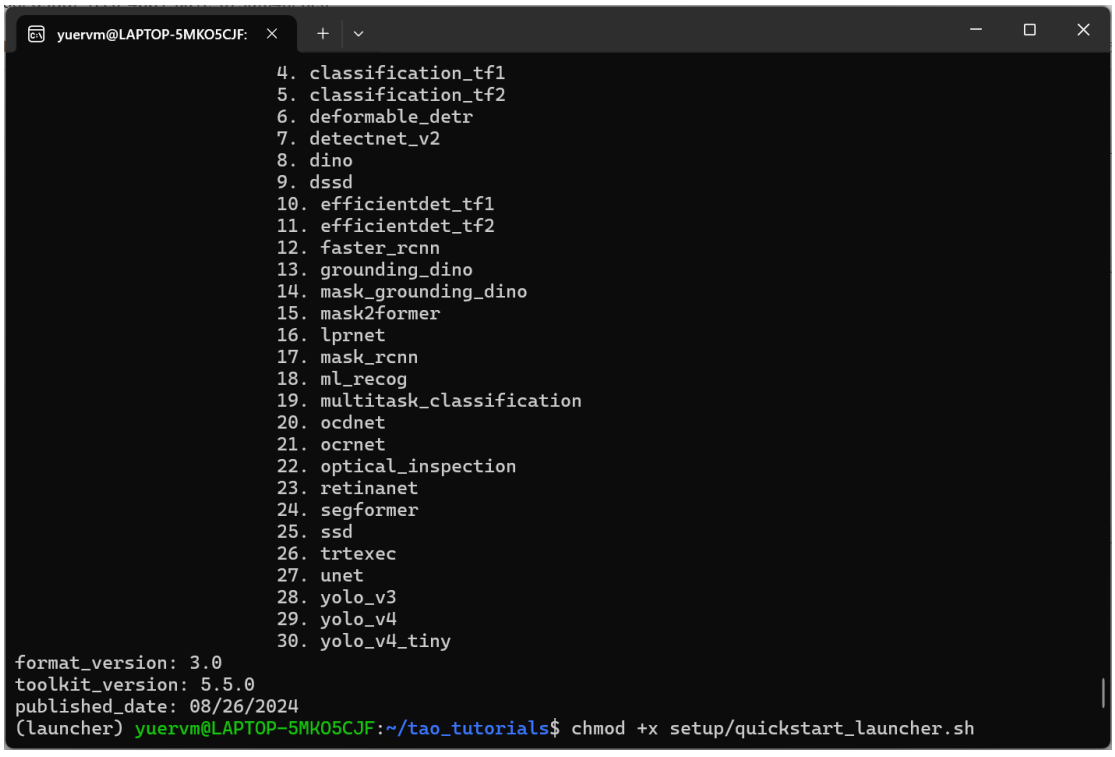

图 6：成功安装 TAO Launcher 并添加权限

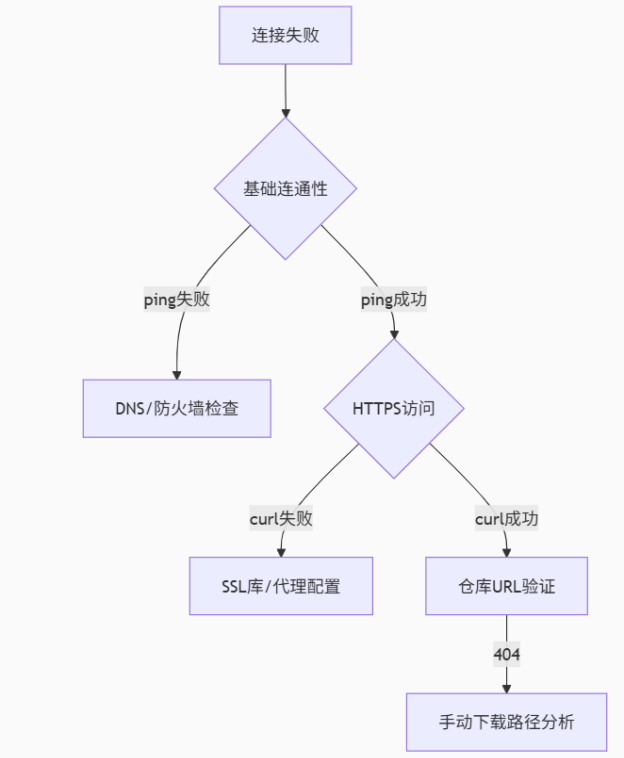

图 7：网络问题诊断框架

TAO 需要特定的 CUDA、TensorRT 和 Python 依赖组合，Docker 容器可确保环境一致，避免与主机系统冲突。TAO Toolkit 依赖 CUDA 进行 GPU 加速计算，所有训练和推理任务均通过 CUDA 调用 GPU 算力，必须安装与 TAO 版本匹配的 CUDA。TAO 生成的模型可通过 TensorRT（基于 CUDA）优化，显著提升边缘设备（如 Jetson）的推理速度。TAO Launcher 是一个基于 Python 的轻量级命令行界面，充当基于 PyTorch、TensorFlow 和 TensorRT 构建的 TAO 容器。

NVIDIA TAO Toolkit 的核心流程始于从 NGC (NVIDIA GPU Cloud) 下载预训练模型，用户可通过 Jupyter Notebook 快速启动迁移学习任务。首先在 NGC 模型库中选择适合的预训练模型（如 ResNet 或 YOLOv4），然后使用 TAO 提供的 Jupyter Notebook 模板加载模型并配置训练参数，包括数据路径、学习率和训练轮数等。在 Notebook 中，用户可以用少量标注数据对模型进行微调，TAO 会自动处理数据增强和分布式训练等复杂过程。训练完成后，通过内置工具对模型进行剪枝和量化优化，最终导出为 TensorRT 格式，直接部署到 Jetson 等边缘设备。整个过程在 Docker 容器中完成，确保环境一致性，同时支持实时监控训练指标和可视化模型性能，显著降低了从模型开发到边缘部署的技术门槛。