C++读取pytorch新版pth文件进展（大概1.4版本更新新版保存方式）：

1. 肉眼分辨大致信息 （已完成）
2. 代码定位信息 （已完成）
3. 代码读取解析权重（已完成，state\_dict保存形式和整个模型直接保存，可以跨版本读取解析， 1.7版本的库能解析1.12版本保存的文件）
4. C++ pickle （pytorch使用python pickle序列化非权重信息，只要实现C++版就是跨版本读取解析）
5. 最小化引用或者最小化动态库，方便迭代更新。（直接获取官方最新版本的文件替换编译，目前所有dll加起来接近一个G，实际我只需要压缩类，读取类和torch核心。）

目的：

1. 去除gen\_wts.py。这个脚本暂时存放在yolov5\_inference里面，因为yolov5仓库保存模型的方式是直接保存整个模型。所以torch.load模型时依赖整个yolov5仓库。简单点说，所有整个模型直接保存的pth文件都依赖原有代码去读取（python pickle导致），原有pt->wts的方式受限。C++版本没有这个限制。
2. 避免疑似GIL或者加载torch库造成缓慢的问题。
3. 模型转换由pt->wts->trt变成pt->trt。减少磁盘IO申请。
4. 抛弃python的一小步。

缺点：

1. 缺少模型结构信息（虽然我可以用torch.fx的脚本读取，但是这个是python环境下，做不到C++环境读取模型结构有向无环图信息。）
2. 可能比原来的方式稳定，不一定对应现在王涛和我复现BUG情况猜测的原因。