

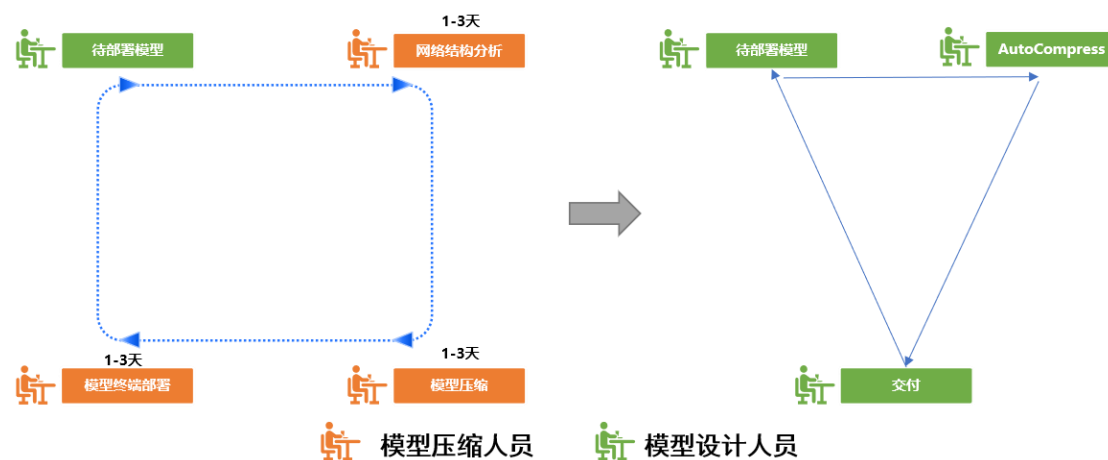
1.AutoCompress 简介

AutoCompress 是自动驾驶实验室多源融合感知组推出的全自动化的模型压缩部署工具链。该工具链旨在解决深度神经网络模型的终端部署问题，涵盖模型转换，模型优化，模型推理三大功能，可使用户一键完成深度神经网络模型的压缩部署。



2.AutoCompress 背景

深度神经网络计算复杂度高，模型参数量大的特性，使其难以在一些场景和设备上部署，尤其是在终端设备上的部署。尽管近年来各种模型优化加速技术层出不穷，但为算法工程师带来了较大的学习成本。即使可以将模型压缩部署工作交付给专业人员，也会带来相应的时间成本和人力成本。为此，我们开发了 AutoCompress 工具链。



3.AutoCompress 的优势

易用性：

- 仅需几行代码便可完成模型的转换、剪枝，量化与部署
- 无需源码编译，使用 Pip 一键安装
- 简易的使用文档，几分钟便可上手

通用性：

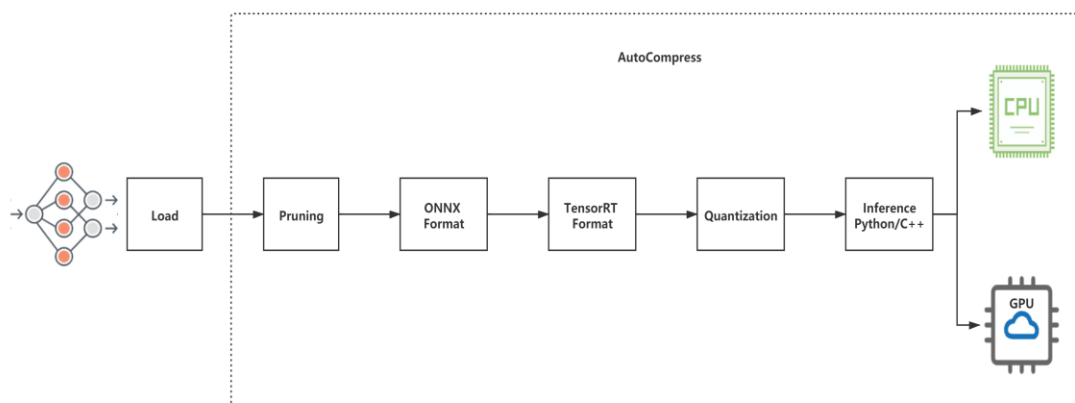
- 主要面向图像分类、目标检测，语义分割等视觉任务，几乎支持所有经典的视觉模型
- 增加了对当前 SOTA 模型的支持

拓展性：

- 支持自定义算法
- 支持自定义算子

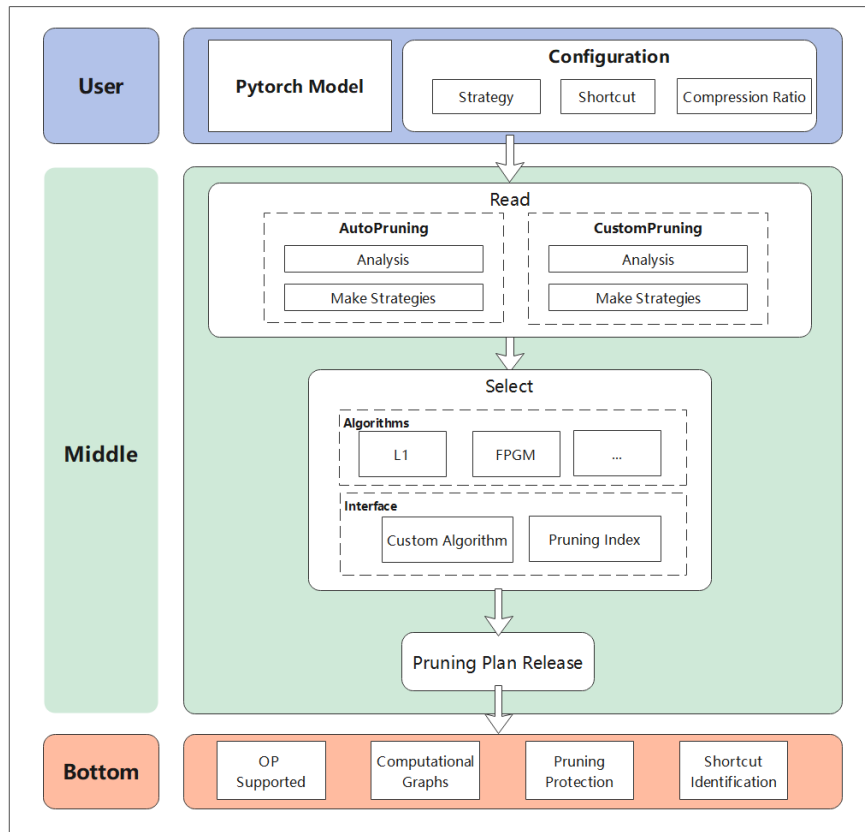
4.AutoCompress 的架构介绍

4.1 总体架构



AutoCompress 由模型剪枝、模型转换，模型量化和模型推理四个组件组成。其中，模型剪枝为团队自研，模型转换与量化推理是在 ONNX 与 TensorRT 框架下的二次开发，并在此基础上打通“Pytorch->模型剪枝->ONNX 转换->TensorRT 转换->模型量化->推理加速”流程，实现流程全自动化。

4.2 模型剪枝架构



用户可从 AutoPruning 与 CustomPruning 两种剪枝方案中任选一种，并与其他选项共同生成配置。读取配置后，AutoCompress 使用基于结构先验的快速模型剪枝方法生成 Strategy，并在此基础上限制卷积核的规整性以提高推理速度。以上方法均依赖于算子支持，计算图构建，剪枝保护，跳连层识别等底层技术的支持。

5.AutoCompress 的功能列表

模型剪枝

- 支持结构化剪枝与非结构化剪枝
- 结构化剪枝可实现真正的模型推理加速
- 支持全自动化剪枝与用户自定义剪枝
- 支持图像分类、目标检测、语义分割等任务
- 基于结构先验的快速模型剪枝

模型量化

- 支持 FP16 量化
- 支持 IN8 量化及精度一键校准

模型转换

- Pytorch 一键转换 ONNX

- ONNX 一键转换 TensorRT
- 支持模型动态输入

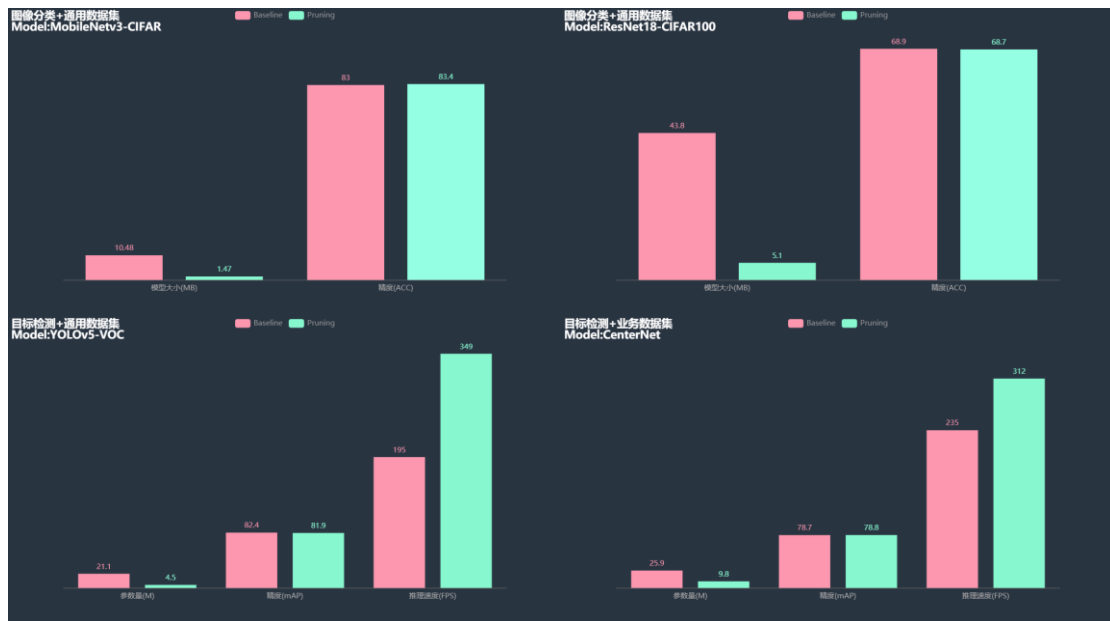
模型推理

- Python API
- C++ API (TODO)

Image classification	Object Detection	Semantic segmentation
ResNet	YOLOv1-v5	PSPNet
MobileNetv1-v3	CenterNet	U-Net
ResNext	FCOS	DeeplabV3
...

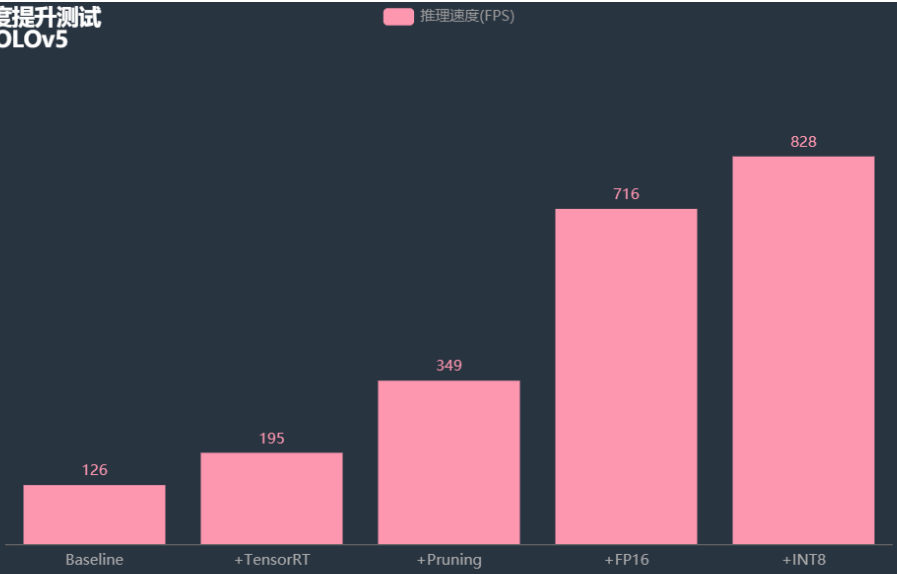
6.实验结果

6.1 剪枝



6.2 工具链整体

工具链速度提升测试
Model:YOLOv5



工具链速度提升测试
Model:CenterNet

