Project Supernova

高效、快捷通用ML推理服务设计

从兰军

lanjunc@vmwre.com

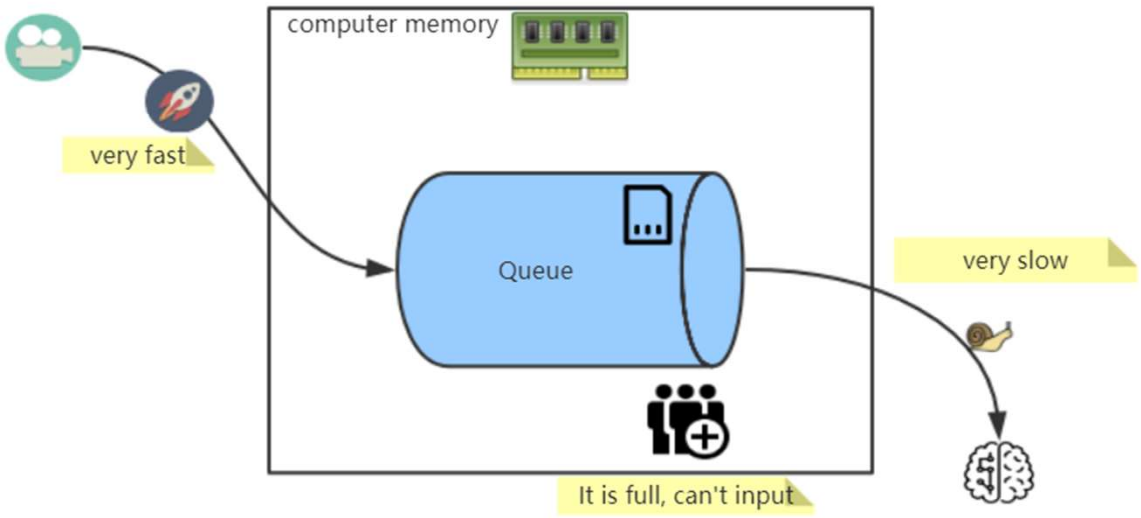
Agenda

- 面临的问题和挑战
- 解决方案
- 后续工作

面临的问题和挑战 1/2

读取视频流和模型推理的时间差距很大

- 为了保证模型推理的实时性，需要让读取视频流时间和模型推理时间尽可能相近。



面临的问题和挑战 2/2

对各类平台的支持，如arm架构

- 安装开发库面临的问题

  很多ML及图像处理开发库找不到合适arm的版本。

  开发库的安装脚本没有对arm提供支持。

- 编程语言某些第三方库不能很好的支持arm

面临的问题和挑战 2/2

对各类平台的支持，如arm架构

Q: Pip fails with `Could not find a version that satisfies the requirement ...`?

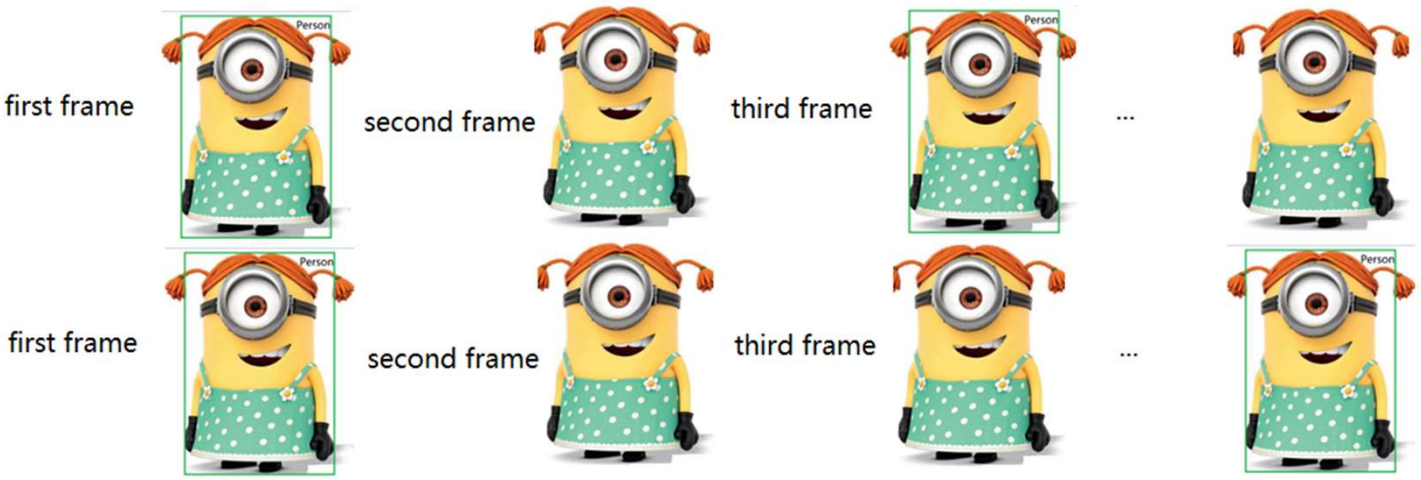
A: Most likely the issue is related to too old pip and can be fixed by running `pip install --upgrade pip`. Note that the wheel (especially manylinux) format does not currently support properly ARM architecture so there are no packages for ARM based platforms in PyPI. However, `opencv-python` packages for Raspberry Pi can be found from https://www.piwheels.org/.

具体说明了某些开发库是不支持arm安装的。

解决方案

消除视频流和模型推理的时间差

跳帧适配

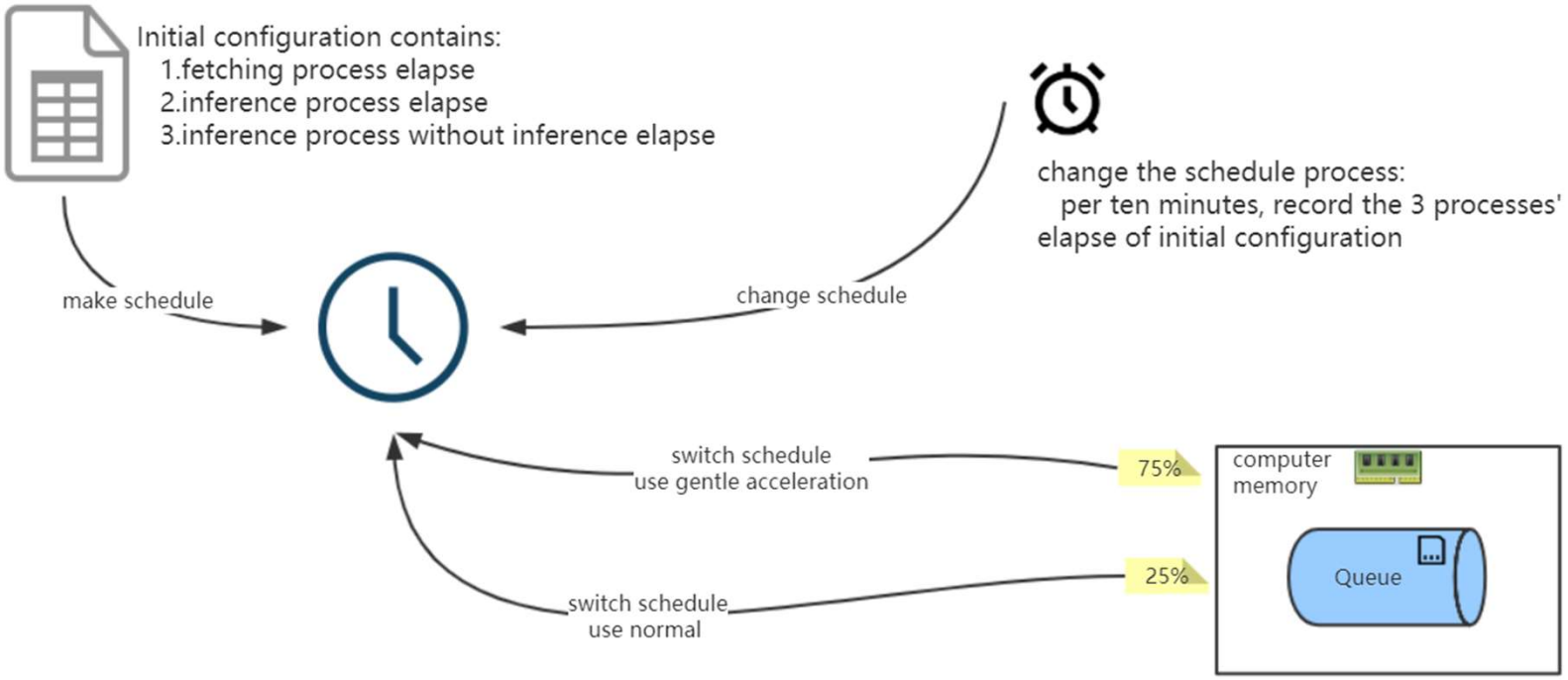


根据运行时间来计算的处理策略

如果堆积的帧太多，进行提速策略。

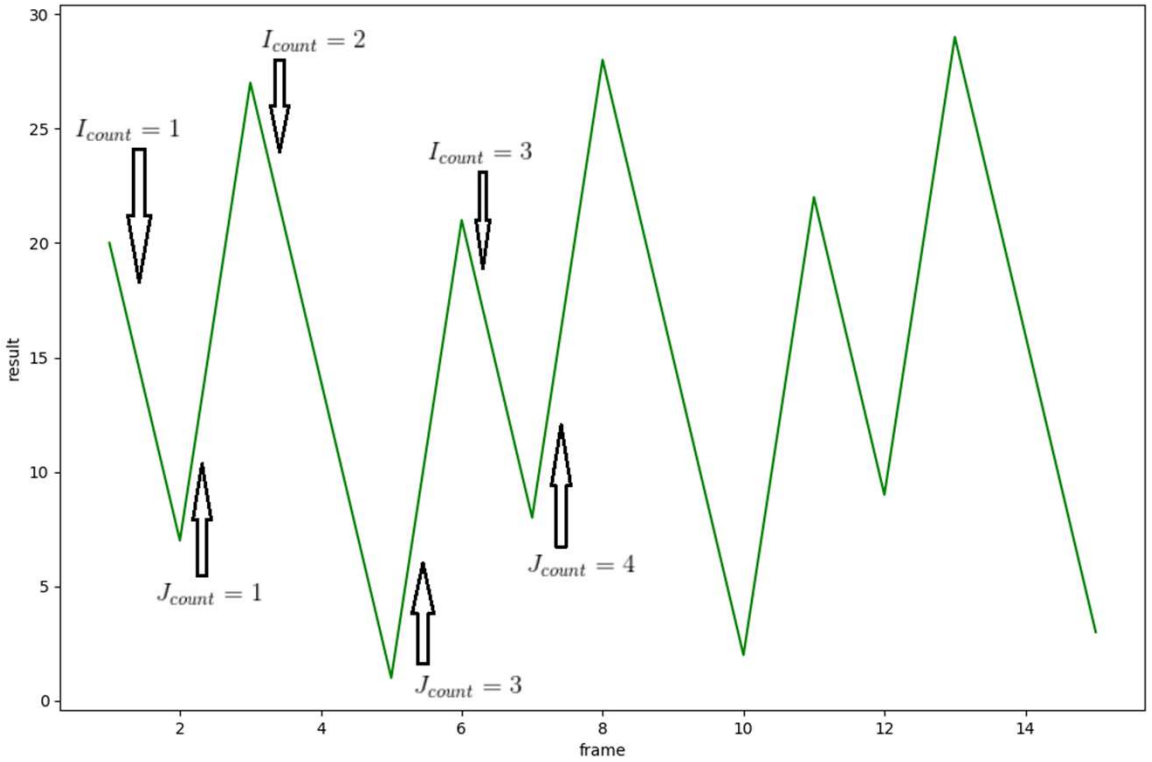
解决方案

跳帧策略


Initial configuration contains:
1.fetching process elapse
2.inference process elapse
3.inference process without inference elapse
change the schedule process:
per ten minutes, record the 3 processes' elapse of initial configuration
make schedule
change schedule
switch schedule
use gentle acceleration
75%
computer memory
Queue
25%
switch schedule
use normal

解决方案

方法实践



上图是根据实验数据画出的曲线图

$$y=(IP-FP)x-(IP-NI)I_{count}, y<FP-NI$$

$$y=(NI-FP)x+(IP-NI)J_{count}, y\geq FP-NI$$

实际运行环境

Cpu：Intel 1.6GHz 4核

内存：8G

Tpu：google coral usb edge tpu

解决方案

对各类平台的支持

- 源码编译安装ML及图像处理开发库
- 修改开发安装脚本的源码
- 修改编程语言第三方库的源码

解决方案

对各类平台的支持

解决问题后，列举出目前支持的几种平台设备

后续工作

小样本模型训练

- 在edge端可以进行迁移学习。用小样本retrain模型。
- 训练新类别也更有效。

  传统的反向传播，添加新类别需要重新训练所有类别。

  而Supernova支持快速的模型训练。

后续工作

模型压缩

- 保证模型精度前提下，降低模型复杂度。
- 方便部署。
- 提升模型训练速度。