TESLA® CASE STUDY

Tesla P100 加速大规模视频分析与智能监控

现状

西安交大智能监控与传感研究室属于国家工程实验室，在模式识别与智能系统领域有着长期的研究积累，并在图像及视频内容自动分类与识别、目标检测与追踪、人脸检测与验证、人体身份识别等多项计算机视觉技术上拥有国际领先的成果。研究室获得国家“863”计划重点基金和国家自然科学基金的支持，并承担国家“973”重点基础研究发展计划，在国内视觉研究领域处于领先的地位。

近年来，随着深度学习相关领域的快速发展，人工智能技术在制造智能家居、构建智慧社区和建设智慧城市等诸多方面有着极为广阔的应用前景。实验室紧跟时代的潮流，与海尔集团、华为公司、蚂蚁金服等知名企业，在智能家电、高性能云服务与智能监控等多个应用领域开展了深入的合作。近期，实验室同公安部门与交通部门的研究所展开了对大规模监控视频的智能分析合作研究，基于深度神经网路提供人脸检测与验证、行人检测与检索、车辆检测与车型识别，以及聚集人口和异常事件检测等解决方案。 依托于 NVIDIA 的 Tesla 系列 GPU 加速卡，实验室构建了多 GPU 视频处理加速平台，能支持对大规模视频数据的高效处理和对多路视频流的实时监控与智能分析，从而为智能监控领域的合作研究提供了坚实的技术保障。

挑战

智能监控旨在对监控视频中的物体、行为、事件等对象，通过检测、识别、跟踪等视觉模式识别技术进行智能分析和判断，从而减少或取代人力的干预。智能监控是建设智慧城市的一项重要内容，所涵盖的技术包括对人脸、行人、车辆、标识等视觉对象识别和行为分析等，涉及无人驾驶、异常事件检测、辅助刑侦等诸多应用场景。

目前，随着监控设施的在各场所的普及，每时每刻都会产生大量的监控视频数据。一方面，人为的监视、分析这些监控视频会耗费大量的人力资源（如图 1 所示），且受限于人的注意力局限，容易遗漏可疑目标和紧急事件，耽误了宝贵的办案时间；另一方面，由于视频数据规模大，传统的高性能计算平台已难以满足视频大数据的处理需求，更难胜任实时的监控与分析。

随着基于深度学习的大规模视频分析技术的发展，采用 GPU 加速视频分析系统成为业内逐渐采用的技术方案。以对监控视频的行人跟踪、检索为例，基于深度学习的业务流程采用如图 2 所示技术框架。

图1 每时每刻都会产生大量的实时监控视频，单个人难以兼顾所有的监控录像，而增加人员数则将人力浪费在了对“正常”的视频时间段的监控上。

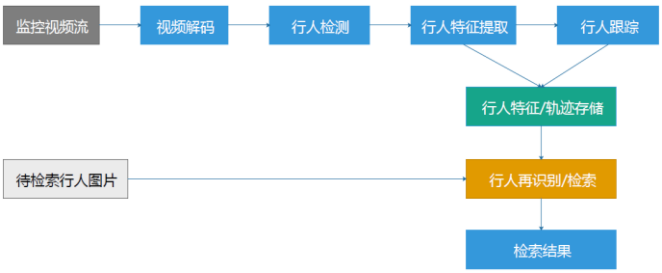

图2 监控视频行人跟踪、检索流程。从监控终端获取的行人监控视频经过解码后，输入到基于深度神经网络的行人检测算法模块，既而进行行人的检测、特征提取、轨迹跟踪，并将特征和轨迹进行结构化存储。对于待检索的行人图片，在行人再识别模块判别行人的ID，并检索行人图片、轨迹等。

其中，视频解码是视频处理的第一步，依赖于低层解码库对视频帧的解算。传统的视频解码通常采用 CPU 软解码，或基于特定的解码硬件设备进行硬解码。软解码的计算时间取决于用于解码的核心线程数，为达到较高的解码速率，通常需要把 CPU

核心占满。当视频路数增多时，为保证实时性，解码所需的计算资源也会随之提升。旧版本的 NVIDIA GPU 提供了基于 CUDA 的解码加速，但加速效果不明显。

视频解析流程中所用到的检测、特征提取和识别算法（如图中行人检测、行人特征提取和行人再识别）目前主要基于深度卷积网络实现。当视频流增加或视频对象数量增多时，单 GPU 受限于浮点核心数（FP 核），算法模型的计算处于满载的状态，往往需要采用具有更强计算能力、数量更多的 GPU 加速器来并行加速。因而，如何选取合适型号的 GPU，以及如何设计并行处理框架，成为系统设计人员需要考虑的问题。

方案

目前，本研究团队基于双路 Tesla P100 GPU 加速器搭建了视频分析并行加速平台，能支持 8 路视频流的实时解析。Tesla P100 加速器为 NVIDIA Pascal 架构的核心产品之一，采用了 HBM2 的 CoWos 技术使得内存带宽达到了 Maxwell 架构的 3 倍，为大数据处理任务提供了更高的吞吐率。P100 的单精度浮点数（FP32）计算能力也达到了 9.3TFLOPS（采用 Nvlink 更高，可以到 10.6TFlops），相比前代 Kelpler 架构的 K80 产品有了显著的提升。更值得注意的是，P100 提供了对视频硬解码的支持，仅使用很小的 GPU 资源就能达到极快的视频解码速度，能用更短的时间、实时的完成大规模视频解析任务。

本研究团队设计了一款多 GPU 进程并行框架来实现对多录视频流解析的加速功能，如图 3 所示。硬件平台搭建上，采用监控设备+双路 P100+双 E5-2620v4 的方案，其中双路 P100 用于视频硬解码、视频对象（如行人）检测和特征提取，而双 CPU 用于视频对象跟踪和特征/轨迹的结构化存储。采用此方案的是为了符合各算法模块的计算特性：检测和特征提取模块均基于大规模深度神经网络实现，适合采用 P100 GPU 加速计算；行人跟踪模块基于概率图模型等算法实现，这些算法更适合在 CPU 上使用多线程实现。通过将不同任务划分到适合的硬件上，可以最大化的实现资源的利用率，并得到较好的并行加速效果。

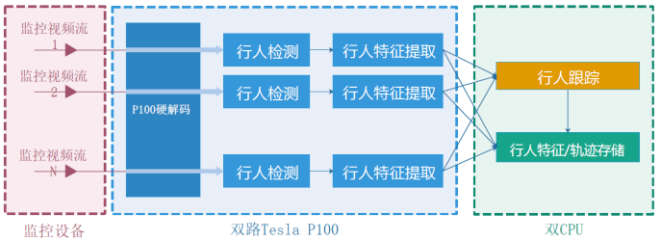

图3 行人监控视频分析系统的并行化实现

本研究团队对P100的解码器性能进行了评测，并和CPU方案进行了对比分析。对一段帧率为25fps，分辨率为540P的H264编码的行人监控录像进行解码测评，结果如表1所示。可以看出，采用P100硬解码可以取得高达170x的加速率，显著高于使用16个CPU核心的解码加速效果。

测试模块	平均解码速率	加速率
E5-2620 v4 单线程	604 fps	24x
E5-2620 v4 线程数=4	1742 fps	70x
E5-2620 v4 线程数=16	2674 fps	107x
P100 硬解码	4240 fps	170x

表1 行人视频解码性能对比评测

采用双路P100的并行方案后，可以并行处理高达8路的视频数据率，且加速比近似线性。以基于SSD的行人检测模块为例，对于540P的视频，E5-2620v4 CPU的单帧检测时间达150ms，6.6fps的帧率难以满足实时性的要求；相比而言，单个P100可同时检测4帧，平均每帧的检测时间低至20ms，50fps的帧率完全可以满足实时处理的需要。

影响

随着以 Tesla P100 为代表的 NVIDIA Pascal 架构的出现，基于 NVIDIA GPU 加速器的 HPC 平台对大规模视频数据的处理速度得到了质的飞跃，原本受限于硬件速度而耗时巨大的计算任务，现在在短期内就可以完成。本研究团队基于双路 P100 实现的视频分析并行加速平台显著的提升了智能监控系统的运行效率（如图 4 所示），使得原本数以日记的视频处理时间缩短到几个小时，并能实时地进行监控数据分析。此外，可通过扩展 GPU 设备的方法近似线性地提升视频分析的速度，从而进一步增强系统的并行处理能力。

近年来，NVIDIA GPU 硬件性能的快速升级极大的提升了算法研究人员的成果产出率，推动了人工智能技术的迅速发展。基于 NVIDIA GPU 构建的高性能计算平台为大数据分析提供了硬件支撑，使原本需要消耗大量人力的重复性工作现在可以智能系统来完成，为建设智慧城市、开启智能生活提供了提供了坚实的保障。

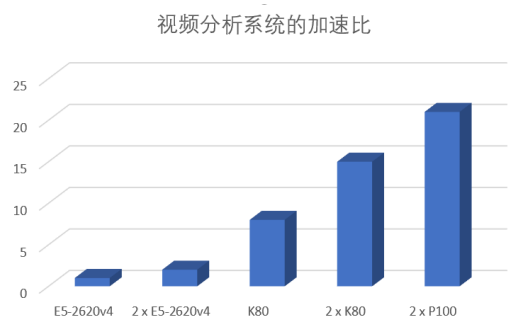

图 4 视频分析系统在各硬件设备上的加速