• 深度

嵌入式视觉的优化与前端设备智能化的趋势

2017-04-21 15:32:08 • 深度学习 • 计算机视觉

深度学习算法的突破性进展，硬件计算能力的提升，以及移动互联网带来的海量数据，成就了这次人工智能浪潮。2013 年，Facebook 平均每天上传 3 亿 5000 万张照片。现在研究人工智能的公司，必须要有先进的算法，强大的 GPU 训练集群，以及源源不断的海量标注数据。

深度学习的爆发，给计算机带来了全新的认知能力，特别是在计算机视觉感知方面，在一些领域计算机的检测和识别能力已经超过人类。本期阅面科技资深研究员宋向明将以「嵌入式视觉与前端设备智能化」为主题给大家分享更多干货内容。

人工智能成为移动互联网后又一个创业浪潮，从互联网演进到移动互联网，用户数和数据量都产生爆炸性增长。人工智能也会经历一个类似的过程，从云端智能迈向移动设备端的智能，今后的大部分算法将会运行在设备端，完全基于本地的计算能力或者部分基于本地的计算能力来完成任务。

嵌入式视觉是计算机视觉的一个方向，随着深度学习的兴起，在算法层面的准确度也相应提高，与之前相比，嵌入式视觉的视频分析就是把云端或本地服务器的处理能力放到嵌入式系统上，使得它能够非常低功耗和实时的进行本地处理。

嵌入式视觉的广阔应用场景

计算机视觉的视频分析主要应用在人机交互，环境感知，智慧商业分析，自动驾驶以及安防监控等应用场景，一般要求实时处理，有些功能可以离线分析。

人机交互：顾名思义就是机器和人的交互。它可以准确的判断访客的属性信息，以及年龄性别的分析。同时基于一些手势的控制，人机的互动性也将会上一个层次，但是这通常需要实时信息的反馈。

环境感知：举个例子，当机器人到一个陌生的环境时，通过环境感知技术，能快速准确地知道现在所处的环境位置，为导航或者其他功能的决策做个判断。

智慧商业分析：商业场所的视频分析有两个方面，一方面是当顾客到达一个商业场所，对她的属性进行识别，比如年龄、喜好等，另一方面是商场的人流数据统计，这对于营销决策起着不可或缺的作用。

自动驾驶：目前自动驾驶市场非常广大，包括摄像头在内的多种传感器是必备部件。相关的视觉技术要求包括汽车检测/跟踪、路标行人检测等。

安防监控： 公共场所日益关注的安全及监控是全球范围内推动智能摄像头需求量增长的重要因素之一。通过人脸检测、跟踪与识别、人的属性和动作行为检测、车的检测与跟踪、物体标注等技术，可以非常实时的找出安防缺陷与问题。

前端设备智能化的必要性

目前来看，我们身边许多智能设备，如摄像头、机器人等，它们都需要强大的本地实时交互、计算的能力，这也意味着前端设备上需要有智能化的能力。

随着智慧城市，智慧商业，智能家庭的发展，越来越多的摄像头产品上线，传统的视频监控存储，人工查看的方式，已经完全无法满足现在对视频分析的需求。如果使用大量服务器进行实时视频分析，那么视频的传输，存储，分析的成本非常高，只能在某些特定领域使用，限制了应用场景和规模。

人机交互，环境感知方面，需要实时的理解和响应，即使网络条件差，或者没有网络，也需要能够正常工作。为了解决这些问题，深度学习必须在前端有限的计算资源和功耗下运行。

然而深度学习算法，计算量非常大，通常需要运行在高性能的服务器上，对于在前端运行提出了非常高的要求。这些正是嵌入式视觉的机遇与挑战。Nvidia 预计，到 2020 年，全球预计将会有 10 亿台监控摄像头投入使用。将传统的前端带摄像头的设备，升级为具有一定智能的设备，继而在前端本地就能实时的完成特定的任务，比如检测到感兴趣的目标，并进行下一步的追踪或者识别，对环境实时建模，自动导航，极具应用和商业价值，开创一个新的时代。

深度学习针对嵌入式的优化方式

前端设备智能化，前景广阔。硬件+算法一体化的解决方案，以最优的性价比提供给客户，才是嵌入式视觉解决方案的核心竞争力。嵌入式视觉中需要做非常多的优化工作。深度学习针对嵌入式方面的优化，主要有网络结构优化，模型压缩，定点化，二值化，结合 SIMD，缓存，多线程，异构计算的优化。

算法软件优化

1. 网络结构优化

基于一个初始版本，对网络结构进行调整，某些层的修改，参数的调整，使得它能够在不降低精度的情况下速度更快。

2. 模型压缩剪枝

把一些不必要的分支给砍掉，在进行一个预测的时候，计算量相对会减少一些，速度变快。

3. 定点化，二值化

深度学习模型的参数都是浮点数，相对来说它的计算比整数要复杂一些，特别在一些低端的芯片，乘法器都不够多的情况下，浮点的性能就会比较差。如果把它转成定点整数运算，那么在精度下降 1% 的情况下，它的速度将会带来几倍的提升。

二值化比定点化更进一步，一个权重值只占用一个比特，并且可以将乘法运行转换为异或操作，在特定硬件上并行性会更高，执行速度会更快，非常适合在低端芯片上使用。

4. SIMD，缓存，多线程

SIMD, 单指令多数据，一次一条指令做多个操作，增加缓存命中，减少内存访问。一些不同的算法如果放在不同的线程中去跑，对外提供的整体组合的效果会非常的快。

5. 异构计算

与硬件相关比较大，根据我选择的不同的硬件、不同的方案、定制化指令的不同，硬件选择都会接触到异构计算。

将高性能服务器上运行的算法，迁移到嵌入式平台实时运行，其难度非常大，除了算法软件层面的优化，还需要充分利用硬件提供的计算能力。在硬件选择方面，更是需要选取最适合的方案才能搭配出最优的性价比。

硬件选择

1. ASIC 专用芯片

> 谷歌 TPU，中科院 DIANNAO 系列，（需要找到大量使用的客户，才能降低成本）

在人工智能早期，只有少数公司用到这个方案，所以它的受众并不会特别大。

2. 基于 GPU 的方案

> GPU，英伟达公司推出的 Jetson TX1，Jetsong TX2 等嵌入式 GPU 方案

GPU 中有多核并发的优势，在上面运行深度学习的复杂运算时，可以进行并行运算。同时，GPU 本身支持定点、浮点的操作，用 GPU 方案，相对来说能达到一个几倍的加速。

3. 基于 FPGA 的方案

> FPGA，主要供应商是赛灵思公司，各家算法公司在其上进行开发

FPGA 对开发人员的要求非常高，首先要对软件很熟悉，又要非常熟悉硬件，现在有些公司提供的一些解决方案，相当于能够直接将深度学习的模型导到他们做的 FPGA 方案上去，然而，他们并不知道内部是如何优化的，整个 FPGA 方案的成本会非常高。

4. 基于 DSP 的方案

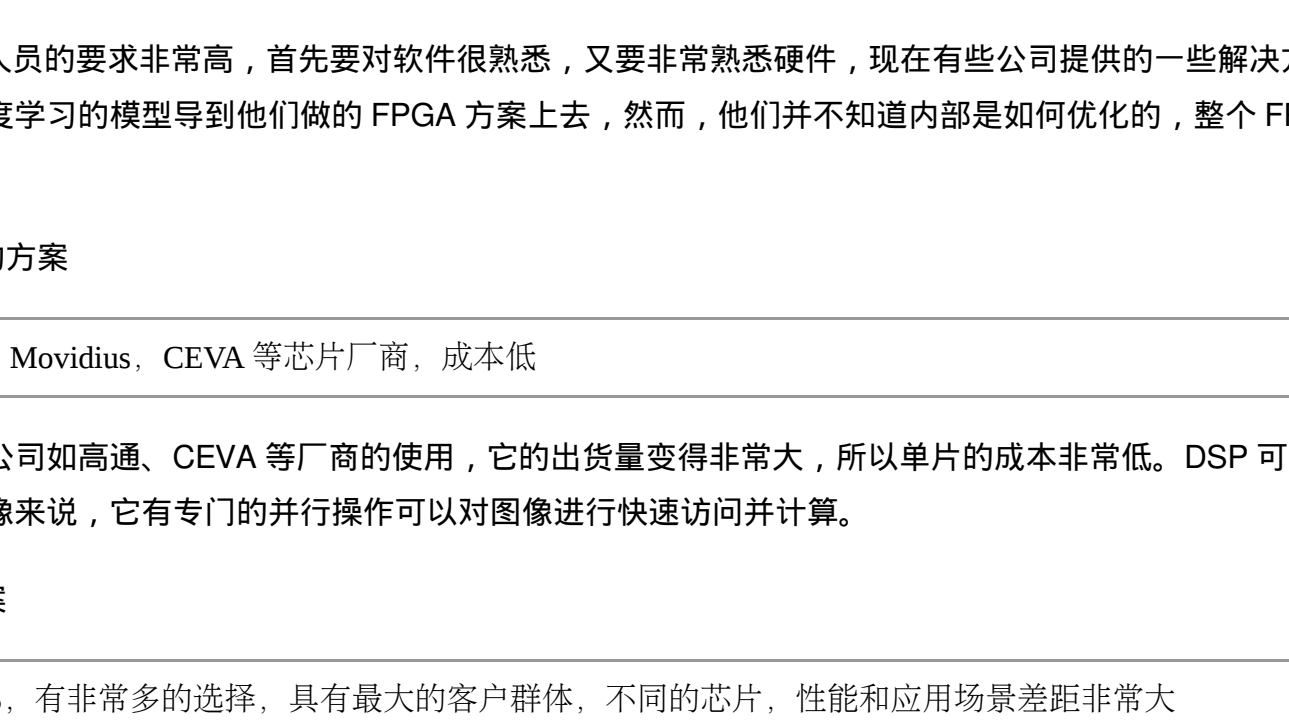
> DSP，高通，Movidius，CEVA 等芯片厂商，成本低

伴随着一些大公司如高通、CEVA 等厂商的使用，它的出货量变得非常大，所以单片的成本非常低。DSP 可以进行数字信号处理，以图像来说，它有专门的并行操作可以对图像进行快速访问并计算。

5. 纯 CPU 方案

> ARM，MIPS，有非常多的选择，具有最大的客户群体，不同的芯片，性能和应用场景差距非常大

与前面说的四种优化方式相比，它的场景非常大，对于我们来说更多注重的是它优化的方向。因为所有方案里面都是有 CPU 的，它是一个必不可少的方案，如果在 CPU 上做的很快, 那么在一些硬件、协处理器的情况下，速度会更加提升。



在 Android TV（ARM CPU）上本地运行的实时人脸追踪+人脸识别

从硬件成本考量，GPU 和 FPGA 由于价格及功耗过高，限制了应用的场景，基本上在普通消费级产品中不会使用。在高端的一些方案当中，可能会选择的是 GPU 和 FPGA 方案。因为它现在能提供一个比较大的计算量和计算能力，对于并不是非常注重成本的情况下，这两个方案在目前是比较合适的选择。

另外，纯 CPU 的方式对算法优化提出了非常高的要求。如果要使用它的话，可以进行两个方面的优化。一个是云端方式，在云端去做，但是这可能并不是我们所看好的方案。另外一个在 CPU 上经过非常高的优化之后，使得算法全部或者部分的在 CPU 上运行。人工智能是一个趋势，低成本的 CPU 方案目前能做的事情不多，所以芯片厂商们都会在这块进行升级，而在升级完成的 2-3 年时间窗口内，纯 CPU 的方式还是需要做一些软件上的优化。

未来几年内比较看好 ARM+DSP 的方案，它有非常高的性价比，既可以提供比较低的成本，又具有很高的计算能力，技术演进路线比较自然，产品面向市场的时间也比较快。

如果再长远一些看，ASIC 专用芯片将会变成主流。届时人工智能领域将会产生一些规范，软件算法也已然定型，一个专用芯片只能专门去做一个服务，比如自动驾驶专用芯片，它整合了行业中自动驾驶所有的复杂场景，只要使用这块专门的芯片再加上一些传感器就可以获得自动驾驶的能力，是一种面向行业的专业应用了。SYNCED

宋向明：阅面科技首席架构师，2012 年上海交大计算机体系结构硕士毕业。百度高级研发工程师，从事大数据分析，大型网站性能优化。目前主要负责人脸识别，智慧视频分析等系统的架构和开发工作。

阅面科技专栏文章：

专栏 | SLAM算法解析：抓住视觉SLAM难点，了解技术发展大趋势 (http://mp.weixin.qq.com/s?__biz=MzA3MzI4MjgzMw==&mid=2650725468&idx=2&sn=e4d2758ba82733d00f0142b8a7869b17&chksm=871b1822b06c913456efbd857639afdf0548852997199b56a48ea14b30dcc9415272ba60b5b0&scene=21#wechat_redirect)

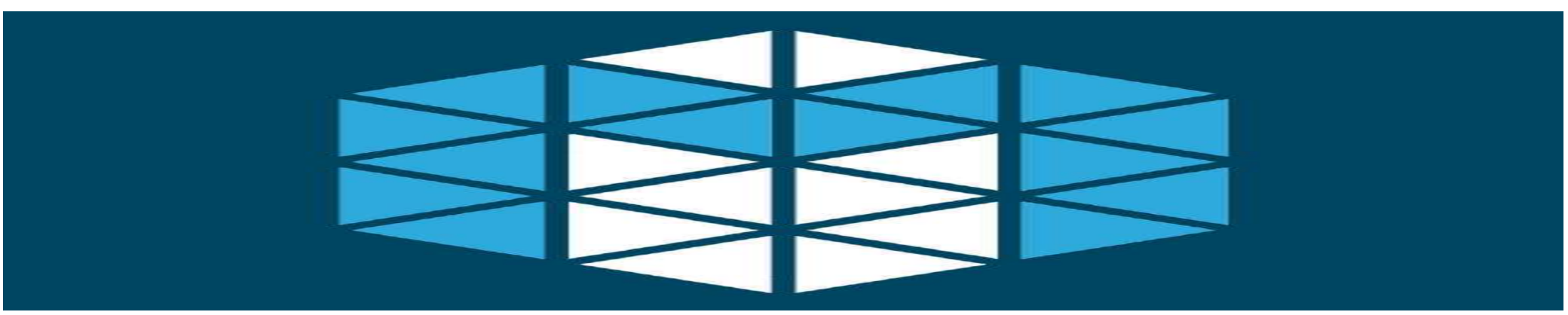
干货 | 物体检测算法全概述：从传统检测方法到深度神经网络框架 (http://mp.weixin.qq.com/s?__biz=MzA3MzI4MjgzMw==&mid=2650725146&idx=3&sn=453e29cb6179e8e06df2133269c20812&chksm=871b1f64b06c967276274cee90f5a036c09b08cec9e52111af6744c7f0c19d5a8c7e0eb06e0b&scene=21#wechat_redirect)

参与成员：

REPOST (/user/author/index/pid/oadnmfo0o0o6)

转载

相关文章



无需反向传播的深度学习：DeepMind的合成梯度

(/article/2695)

英特尔深度学习产品综述：如何占领人工智能市场

(/article/2693)

深度学习在NLP领域成绩斐然，计算语言学家该不该惊慌？

(/article/2659)

评论

共有0条评论，点击展开

(/user/author/index/pid/OaDnMFO0O0O6)

转载

35

文章 (/user/author/index/pid/OaDnMFO0O0O6)

0

评论 (/user/author/comment/pid/OaDnMFO0O0O6)

0

收藏 (/user/author/favorite/pid/OaDnMFO0O0O6)

24小时最热 一周最热

微软**RobustFill**：无需编程语言，让神...

2017年04月22日

(/article/2691)

用人工智能做金融风控？这里是一位实践者的思考

约23时前

(/article/2715)

三张图读懂机器学习：基本概念、五大流派与九种...

2017年04月22日

(/article/2690)

谷歌**TPU**之后还有高通，人工智能芯片竞赛已经...

约23时前

(/article/2713)

绿公司年会，依图创始人朱珑的AI梦与伦比...

2017年04月24日

(/article/2710)