文章编号：2095-4980(2020)03-0515-05

基于深度学习的视频人群计数系统

向 东，卿粼波*，何小海，吴晓红

(四川大学 电子信息学院，四川 成都 610065)

摘 要：人群自动计数问题在视频监控领域引起了广泛关注。近年来，卷积神经网络(CNN)模型在人群计数方面取得了良好效果。然而，当前对于基于深度学习的人群计数的研究主要停留在 PC 端上对单幅静止图片的人群计数，网络模型参数量巨大，网络结构复杂，消耗的计算资源巨大，难以部署于实际的监控视频人群计数系统。因此，本文采用深度学习的方法，通过对网络模型进行裁剪压缩，同时使用 TensorRT 对模型进行加速，在嵌入式平台上实现了接近实时的人群计数。提出的人群计数平均绝对误差(MAE)为 21.6 且平均每秒帧数(FPS)为 22，在精确度和速度方面达到了一个很好的平衡，在嵌入式平台上运行速度较快，能达到实时的效果。

关键词：人群计数；深度学习；模型压缩；NVIDIA Jeston TX2 平台

中图分类号：TN919.82 **文献标志码：**A **doi：**10.11805/TKYDA2019234

Video crowd counting system based on deep learning

XIANG Dong，QING Linbo*，HE Xiaohai，WU Xiaohong

(College of Electronics and Information Engineering，Sichuan University，Chengdu Sichuan 610065，China)

Abstract：Automatic crowd counting has attracted widespread concern in the field of video surveillance. In recent years, the Convolutional Neural Network(CNN) has achieved miraculous results in crowd counting. However, current research based on deep learning mainly concentrates on high-performance PC to count the people with a single still picture. The network model has huge computational resources consuming due to its large amount of parameters and complex network structure, which is difficult to deploy in actual surveillance video crowd counting system. Therefore, the deep learning method is adopted to realize the real-time crowd counting on the embedded platform by pruning and compressing the network model and using TensorRT to accelerate the model inference. The proposed crowd counting algorithm achieves a balance between accuracy and speed with Mean Absoulte Error(MAE) of 21.6 and average Frames Per Second(FPS) of 22. Its performance on the embedded platform can approach the real-time result.

Keywords：crowd counting；deep learning；model compression；NVIDIA Jeston TX2

视频监控中的人群计数一直是视频监控领域的研究热点，在实际的场景中具有广泛的应用价值，对于人群检测和场景理解具有重要意义，如通过分析道路人数来动态控制交通信号灯，优化管理行人流量、监控地铁站等交通枢纽的乘客数量、统计旅游景点客流量、监控重点公共场所安全等。因此，研究自动化、智能化的视频图像人群数量计算方法对于社会的有效分析和管理意义重大。人群计数的方法可分为传统的人群计数和深度学习驱动的人群计数两大类。传统的人群计数主要基于检测和回归方法[1]，这类方法在处理密集场景时准确度较低；基于深度学习的人群计数[2-3]一般生成密度图来计数，但存在网络模型复杂，参数量巨大和计算量大的缺点，难以部署于实际应用场景。并且，现有人群计数算法绝大多数基于高性能的 PC 平台，无法实际部署在计算资源有限的嵌入式平台。同时，随着边缘计算的兴起，更快速的服务响应需求使得算法部署在数据源端成为

收稿日期：2019-06-30；**修回日期：**2019-08-14

基金项目：国家自然科学基金资助项目(61871278)；四川省科技计划资助项目(2018HH0143)；成都市产业集群协同创新资助项目(2016-XT00-00015-GX)

作者简介：向 东(1995–)，男，在读硕士研究生，主要研究方向为通信与信息系统。email:scu-xiangdong@foxmail.com

*通信作者：卿粼波 email:qing_lb@scu.edu.cn

必然趋势。但目前基于嵌入式平台的深度学习人群计数算法的研究较少，主要受限于嵌入式平台的资源有限和网络模型的庞大。因此，对网络模型进行裁剪压缩和加速[4−15]尤为重要。本文基于深度学习的方法对视频图像进行人群计数，对网络结构进行优化，并对网络模型进行压缩剪枝，最后结合 TensorRT 优化后部署在 NVIDIA Jetson TX2 嵌入式平台上，基本达到了实时人群计数的效果。

1 网络结构和网络模型优化

1.1 网络结构

基于回归的人群计数关键在于生成高质量的人群密度图。本文的网络模型结构基于 MCNN[8]的一个多分支的网络模型结构，如图 1 所示。采用 3 列卷积核尺寸分别为 7×7、5×5 和 3×3 的卷积神经模块，使训练出来的模型对于拍摄角度和分辨率更具有鲁棒性，可以应对实际应用场景中拍摄角度变化和分辨率不同的问题；采用卷积核大小为 1×1 的卷积层代替全连接层，使模型可以接受任意大小的图像作为输入。本文设计的系统可以对任意分辨率的监控摄像头视频[14]进行实时人群计数，输入一个视频序列，模型输出相应的人群密度图，然后通过对密度图积分得到最终的人群计数结果。因此，高效的、高质量的人群密度图是深度学习应用在视频人群计数的关键。

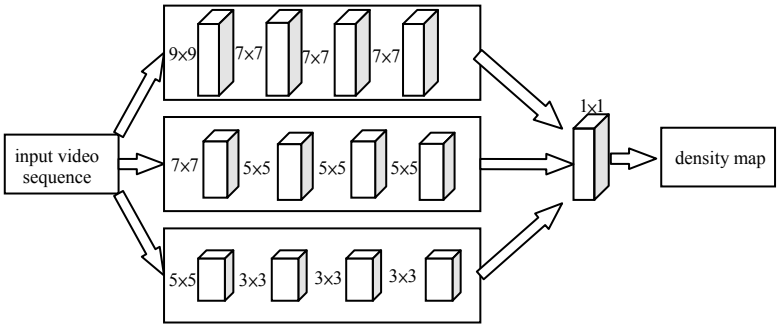

Fig.1 Structure of network model
图 1 网络模型结构示意图

1.2 网络模型压缩

随着深度学习算法研究的不断深入，网络的层数不断增加，网络模型参数数量也在不断增加。虽然准确率在逐渐提升，但模型参数消耗的内存和计算资源也在提升，庞大而冗余的模型只能在高性能 PC 平台下运行，难以移植到计算资源有限的嵌入式平台上。因此，设计紧凑的模型结构，并对模型进行压缩和加速，成为亟待解决的问题。

目前针对单幅图像的人群计算网络在不考虑推理时间和消耗资源的情况下，在数据集上可以达到较高的准确度，但模型参数量巨大，只能止步于理论研究，很难具有实际应用价值。针对已有的网络模型存在较大的冗余，某些参数或卷积核对最终的预测结果作用不大或不明显，本文在提出简洁的网络结构基础上，对预先训练好的网络模型进行裁剪压缩。

剪枝的主要过程为：首先加载预训练好的模型，其结构如图 1 所示，其准确度和可用性在 2.1 节中进行了测试和验证；然后获取准备剪枝的卷积层信息，计算准备剪枝的卷积核滤波器个数，每个滤波器对应生成一个特征图，这些特征图包含了相应的特征信息；由卷积核滤波器的总数和准备剪枝的卷积核滤波器得到需要迭代剪枝的次数。对网络进行一次前向传播，取得特定的卷积层的每一个通道经过激活层的输出作为对卷积层的评价标准，对每一层的输出的 L_2 范数进行计算，如式(1)所示，x 表示卷积层通道经过激活层后的输出，根据 L2 范数大小进行排序，选取 N 个最小值作为被剪枝的滤波器核，然后生成掩模矩阵将选中的滤波器核剪去，再对模型进行微调训练。通过这种迭代裁剪和 Retrain 的方式恢复网络模型的部分性能，避免裁剪导致过多精确度损失。通过在测试集上进行测试，保证裁剪后模型的泛化能力。这样在减小网络模型大小的同时，可以保证网络模型精确度不至于下降太多，保持其良好的泛化能力。

$$\|X\|_2=\sqrt{\sum_{i=1}^{n}x_i^2} \tag{1}$$

1.3 TensorRT 加速

TensorRT 是一个高性能深度学习推理平台，为深度学习推理应用提供低延迟和高吞吐量的服务，支持 INT8 和 Float16 数据类型优化。

TensorRT 加速推理包含两个阶段：构建(Build)和部署(Deployment)。在构建阶段，TensorRT 对网络配置进行优化，并生成一个优化过的 Plan 用于计算深度神经网络的前向传播。这个 Plan 是一个优化后的目标代码，可以序列化地存储在内存或磁盘上。在部署阶段，通常采用长时间运行的服务或用户应用程序的形式，该服务

或用户应用程序接受批量输入数据，通过对输入数据执行 Plan 来进行推理，并返回批量输出数据。针对实时人群计数在嵌入式平台 TX2 上的应用需求，同时考虑 TensorRT 对于本文所用框架的支持性，本文采用 TensorRT 对模型部署进行优化加速。

优化推理过程如图 2 所示，TensorRT 优化对模型进行 INT8 量化之前，需对精确度为 FP32 的模型进行校验。在创建 TensorRT 优化推理图之前，首先调用 create_inference_graph 函数并设置 precision_mode 参数为 INT8，以此对模型进行校准，输出一个精确度校准后的冻结 TensorFlow 图；接着使用与实际数据集分布相同或接近的校准数据运行校准图；最后将优化后的模型部署在 TX2 嵌入式平台上。

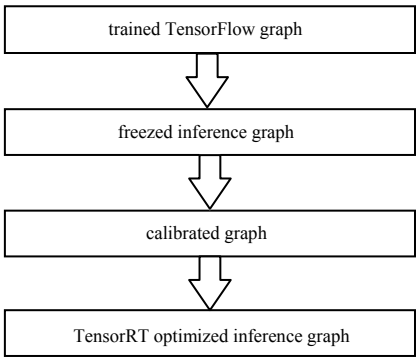

Fig.2 Reasoning process of TensorRT optimization
图 2 TensorRT 优化推理过程

2　实验结果与分析

2.1 网络训练和部署

首先对 3 个单列神经网络进行预训练，然后再合并训练。训练时对原始数据集进行 9 次随机裁剪，得到 9 张图像子块，每个图像子块是原图的 1/4，通过这样的数据增广方式使训练图像尺寸变小，以加快网络训练速度。网络采用 Adam 为优化算法进行训练，先在 GCC Dataset[9]上预训练，然后在 ShanghaiTech 数据集上进行微调，模型实现基于 Tensorflow 为后端的 Keras 框架。本文采用 GPU 模式进行网络训练，硬件为基于 Pascal 架构的 NVIDIA Titan X 显卡，其显存为 12 GB。训练完成后，进行网络裁剪压缩，并利用 TensorRT 进行推理加速，最后将实际模型部署在 TX2 嵌入式平台，通过接入监控摄像头进行实际场景的测试。对于视频序列图像，本文模型基于每一帧单独进行前向推理输出密度图，从而实现人群计数，计数结果在控制终端中实时更新，并且在显示监控图像时标记其对应的计数结果。本实验采用的平台基于 Tegra Parker 处理器，拥有 256 个 CUDA 核，浮点计算能力为 1.5TeraFLOPS，对于终端化人工智能和深度学习开发具有独特的优势和潜力。网络卷积层裁剪前后参数对比见表 1。

表 1 网络卷积层裁剪前后参数对比

Table1 Parametric comparison of network convolution layer before and after pruning

convolution layer	kernel size	number of kernels	number of kernels after pruning
Conv2D_1	9×9	16	14
Conv2D_2	7×7	32	29
Conv2D_3	7×7	16	16
Conv2D_4	7×7	8	7
Conv2D_5	7×7	20	18
Conv2D_6	5×5	40	37
Conv2D_7	5×5	20	18
Conv2D_8	5×5	10	10
Conv2D_9	5×5	24	23
Conv2D_10	3×3	48	39
Conv2D_11	3×3	24	21
Conv2D_12	3×3	12	12

2.2 测试结果

本文在已训练的模型上应用了一种具有相对较低权重的滤波器修剪方法，以减少模型的参数量，且不会引入不规则的稀疏性[10]。如表 2 所示，在 ShanghaiTech B 测试集上的测试表明，使用的迭代修剪和再训练策略使原模型大小降低约 30%，但裁剪后的模型的 MAE(计数平均绝对误差)与原模型相比并没有降低太多，对于一个实时的应用系统，在可接受范围。在 TX2 嵌入式平台上测试结果显示，TensorRT 加速后的裁剪模型的 MAE 与未加速模型的 MAE 基本相当。但从表 2 可以看出，模型在通过 TensorRT 加速后，在 TX2 嵌入式平台上，对于 360×640 大小的视频帧推理时间，相比于原模型缩短了 64.5%。综上所述，采用的裁剪压缩方法和 TensorRT 加速方法，对本系统的实时效果起到了良好的作用。

表 2 裁剪前后模型性能对比

Table2 Performance comparison of models before and after tailoring

models	MAE	parameters	FLOPS/(frames/s)	average inference time/(s/frame)	average inference power/W
MCNN[8]	19.3	127 958	255 378	0.127	7.2
pruned MCNN	21.2	89 565	176 210	0.092	6.1
pruned MCNN+TensorRT	21.6	89 565	176 210	0.045	5.7

为验证模型对视频的人群计数的能力，本文采用 2 种方式进行测试：一是将 ShanghaiTech B 测试集的图片合成视频作为输入，部分样例测试结果如图 3 所示；二是在实际场景中进行测试。在进行实际场景的测试前，先在一个自制的小型数据集进行微调训练，部分测试结果如图 4 所示。图 5 为 BeijingBRT 测试集的图片合成视频作为模型输入的部分测试结果。其中 GT 表示 Ground Truth，即视频帧中真实的人群数目，estimated 表示通过本文系统测试输出的人群数目。

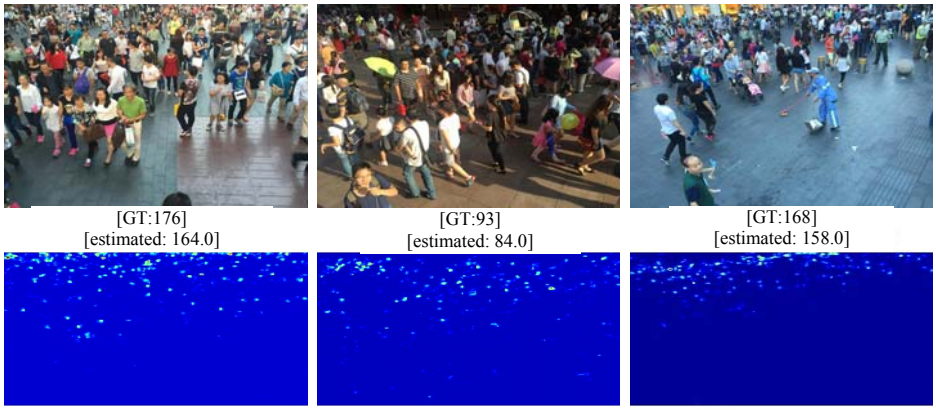

Fig.3 Crowd counting results of ShanghaiTech B testset synthetic video
图 3 ShanghaiTech B 测试集合成视频人群计数

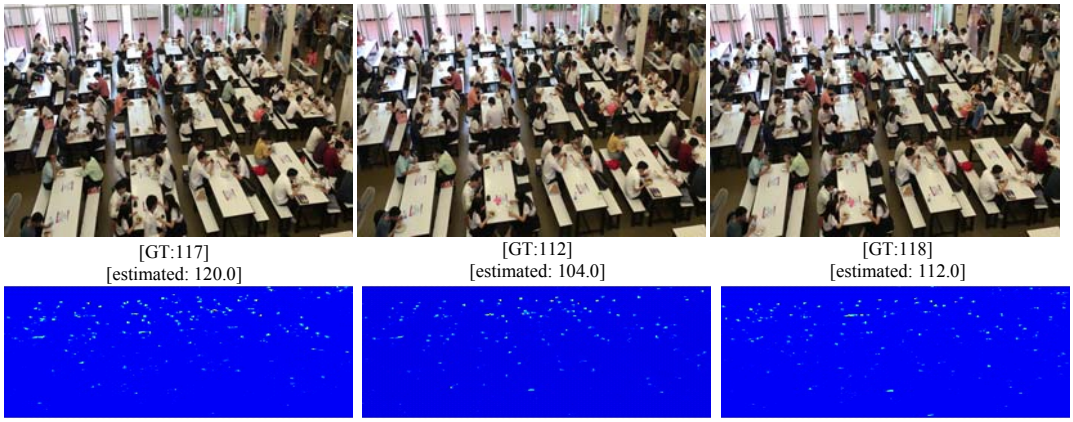

Fig.4 Crowd counting results of campus canteen surveillance video
图 4 校园食堂监控摄像视频人群计数

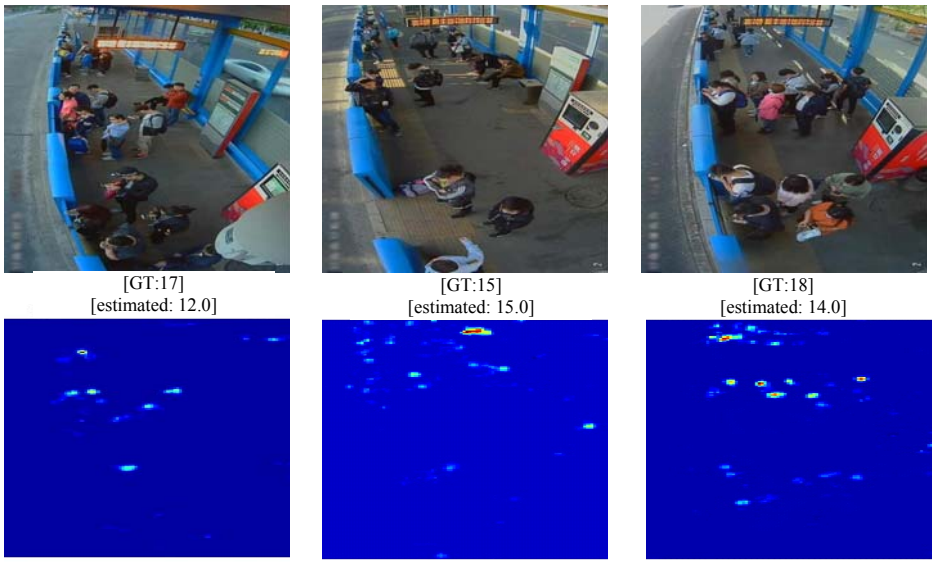

Fig.5 Crowd counting results of BeijingBRT surveillance video
图 5 BeijingBRT 监控摄像视频人群计数

3　结论

本文针对视频中人群计数的应用背景，采用了一个易于部署的深度学习网络模型，为了实现对视频序列进行接近实时的人群计数，应用了迭代裁剪滤波器的方法，通过裁剪对输出精确度贡献小的滤波器，再通过 TensorRT 加速，显著减少了前向推理所消耗的运算资源和运算时间，同时保留了模型的预测性能。在实际场景的测试中，实现了接近实时的人群计数。

参考文献：

[1] PHAM V Q,KOZAKAYA T,YAMAGUCHI O,et al. COUNT forest: co-voting uncertain number of targets using random forest for crowd density estimation[C]// IEEE International Conference on Computer Vision. Santiago,Chile:IEEE, 2016: 3253–3261.

[2] SAM D B,SAJJAN N N,BABU R V,et al. Divide and grow:capturing huge diversity in crowd images with incrementally growing CNN[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT, USA:IEEE, 2018:3618–3626.

[3] CAO X,WANG Z,ZHAO Y,et al. Scale aggregation network for accurate and efficient crowd counting[C]// Proceedings of the European Conference on Computer Vision. Munich,German:Springer, 2018:734–750.

[4] ANWAR S,HWANG K,SUNG W. Fixed point optimization of deep convolutional neural networks for object recognition[C]// Proceedings of the IEEE Conference on Speech and Signal Processing. Brisbane,QLD,Australia:IEEE, 2015:1131–1135.

[5] MIKHAIL Figurnov,AIZHAN Ibraimova,DMITRY P Vetrov,et al. Perforated CNNs: acceleration through elimination of redundant convolutions[C]// 30th Conference on Neural Information Processing System. Barcelona,Spain:[s.n.], 2016:947–955.

[6] COURBARIAUX M,BENGIO Y,DAVID J P. Binaryconnect: training deep neural networks with binary weights during propagations[C]// Neural Information Processing Systems. Montreal,Quebec,Canada:[s,n,], 2015:3105–3113.

[7] HAN Song,POOL Jeff,TRAN John,et al. Learning both weights and connections for efficient neural network[C]// Advances in Neural Information Processing Systems. Montreal, Quebec, Canada:[s.n.], 2015:1135–1143.

[8] ZHANG Y,ZHOU D,CHEN S,et al. Single-image crowd counting via multi-column convolutional neural network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,NV,USA:IEEE, 2016:589–597.

[9] WANG Q,GAO J,LIN W,et al. Learning from synthetic data for crowd counting in the wild[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,CA,USA:IEEE, 2019:8198–8207.

[10] HAN Song,LIU Xingyu,MAO Huizi,et al. EIE: efficient inference engine on compressed deep neural network[C]// Proceedings of the 43rd International Symposium on Computer Architecture. Seoul,Republic of Korea:IEEE, 2016:243–254.

[11] 覃勋辉,王修飞,周曦,等. 多种人群密度场景下的人群计数[J]. 中国图象图形学报, 2013,18(4):392–398. (QIN Xunhui,WANG Xiufei,ZHOU Xi,et al. Counting people in various crowed density scenes using support vector regression[J]. Journal of Image and Graphics, 2013,18(4):392–398.)

[12] 徐超,高梦珠,查宇锋,等. 基于 HOG 和 SVM 的公交乘客人流量统计算法[J]. 仪器仪表学报, 2015,36(2):446–452. (XU Chao,GAO Mengzhu,ZHA Yufeng. Bus passenger flow calculation algorithm based on HOG and SVM[J]. Chinese Journal of Scientific Instrument, 2015,36(2):446–452.)

[13] 周治平,许伶俐,李文慧. 特征回归与检测结合的人数统计方法[J]. 计算机辅助设计与图形学学报, 2015,27(3):425–432. (ZHOU Zhiping,XU Lingli,LI Wenhui. People counting based on feature-regression and detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2015,27(3):425–432.)

[14] 聂勇,张鹏,冯辉,等. 基于动作标准序列的 3D 视频人体动作识别[J]. 太赫兹科学与电子信息学报, 2017,15(5):841–848. (NIE Yong,ZHANG Peng,FENG Hui,et al. 3D video human motion recognition based on motion standard sequence[J]. Journal of Terahertz Science and Electronic Information Technology, 2017,15(5):841–848.)

[15] WANG Chuan,ZHANG Hua,YANG Liang,et al. Deep people counting in extremely dense crowds[C]// Proceedings of the 23rd ACM International Conference on Multimedia. New York,NY,USA:ACM, 2015:1299–302.