基于 NVIDIA JetsonTX2 的视频行为检测研究

李 龙 卿粼波* 李诗菁 何小海
(四川大学电子信息学院 四川 成都 610065)

摘 要 行为检测是计算机视觉领域的一个重要研究方向,在交通监控、人机交互等方面都有着广泛的应用。目前,基于深度学习的 C3D 行为检测网络与传统行为检测相比,其检测精度虽然有了提高,但存在网络参数量大的问题。为进一步提高检测结果的准确性以及降低网络参数量,采用改进的 SqueezeNet 与 C3D 相结合的卷积神经网络,并引入 BN 层与 short-cut 结构。将训练模型部署到 NVIDIA JetsonTX2 上,对视频行为进行分析、检测。实验结果表明,改进后的 SqueezeNet-C3D 卷积神经网络相比于 C3D 神经网络在精度上提高了 4.4%;改进后的 SqueezeNet-C3D 网络与 SqueezeNet-C3D 网络相比,参数量降低了 15%。可见该网络具有精度高、参数量少的优点。

关键词 SqueezeNet 深度学习 计算机视觉 NVIDIA JetsonTX2 行为检测

中图分类号 TP391.41 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.03.026

VIDEO BEHAVIOR DETECTION BASED ON NVIDIA JETSONTX2

Li Long Qing Linbo* Li Shijing He Xiaohai
(*College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, Sichuan, China*)

Abstract Behavior detection is an important research direction in the field of computer vision. It has a wide range of applications in traffic monitoring and human-computer interaction. At present, compared with traditional behavior detection, C3D behavior detection network based on deep learning has improved the detection accuracy, but it has a large amount of parameters. In order to further improve the accuracy of the detection results and reduce the amount of network parameters, we adopted a modified convolutional neural network combining SqueezeNet and C3D, and BN layer and short-cut structure were introduced. Finally, the training model was deployed on the NVIDIA Jetson TX2 to analyze and detect the video behavior. The experimental results show that the improved SqueezeNet-C3D convolutional neural network has an accuracy improvement of 4.4%, compared with the C3D neural network. Compared with the SqueezeNet-C3D network, the improved SqueezeNet-C3D network has a 15% reduction in parameter size. It can be seen that the proposed network had the advantages of high precision and small parameter quantity.

Keywords SqueezeNet Deep learning Computer vision NVIDIA JetsonTX2 Behavior detection

0 引 言

行为检测是计算机视觉领域和图像处理中的一个重要研究方向[1]。在传统行为检测领域,DT[2]算法是最经典的算法之一,它主要包括密集采样特征点,特征点轨迹跟踪以及基于轨迹的特征提取三个部分。2013 年由 IEAR 实验室发表的 iDT[3]算法,对 DT 算法做了改进,主要包括对光流图像的优化、特征正则化方式以及特征编码方式,大大提升了算法的效果。自深度学

收稿日期:2019-04-16。国家自然科学基金项目(61871278);四川省科技计划项目(2018HH0143);四川省教育厅科研项目(18ZB0355)。**李龙**,硕士生,主研领域:多媒体通信,信息系统。**卿粼波**,副教授。**李诗菁**,硕士生。**何小海**,教授。

习应用到行为检测领域后,使用基于深度学习的方法[4]得到的效果已经明显超过了使用传统算法。

深度学习理论提出以来,研究人员发现应用深度学习进行行为检测,可以有效提高检测效果和性能,因此深度学习在实时视频的行为检测[5-6]开始广泛应用,到现在为止,其检测效率和精度已经有了很大提高。在深度学习理论中,Two-Stream[7]是一个主流方法,它由时间、空间两个网络组成。该方法提出对视频序列中每两帧计算密集光流,得到密集光流的序列。然后对光流序列和图像序列分别训练卷积神经网络模型,再训练一个 fusion 网络进行融合图像序列和光流序列的网络。C3D[8](3-Dimensional Convolution)是另一个主流方法,在目前来看,使用 C3D 方法得到的效果要比 Two-Stream 方法略差些,但 C3D 网络结构简单,而且 C3D 运行时间短,处理速度快,所以仍然是当前研究热门。因为嵌入式平台如 NVIDIA JetsonTX2 携带方便,性能强大,所以使得更大型、更复杂的神经网络可以广泛地部署到嵌入式平台上。为提高检测精度以及减少参数量,本文以 C3D 网络为基础,结合 ResNet[9]的 short-cut 结构以及改进的 SqueezeNet[10]来进行网络结构调整,并将网络模型部署到 NVIDIA JetsonTX2 上进行行为检测,总体结构如图 1 所示。

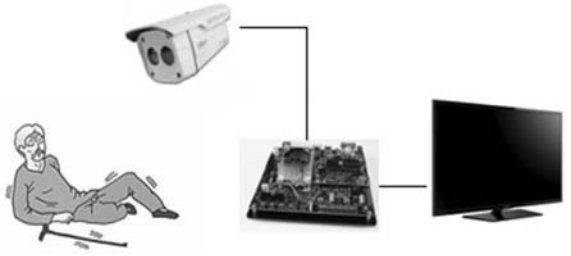

图 1　总体结构图

1　C3D 网络与网络结构改进

1.1　C3D 网络

采用 C3D 卷积神经网络来进行视频行为检测,网络结构如图 2 所示。该网络采用三维卷积对视频连续帧进行操作,相比于二维卷积更能简单有效地处理时间信息,是一种简单、高效、通用、快速的行为识别神经网络。但是 C3D 网络参数量较大,难以将它跟其他参数量较大的网络同时部署到同一个 NVIDIA JetsonTX2 上,并且准确度不高。为提高检测的准确度以及减小网络参数量,本文借鉴 SqueezeNet 网络对 C3D 网络进行改进。

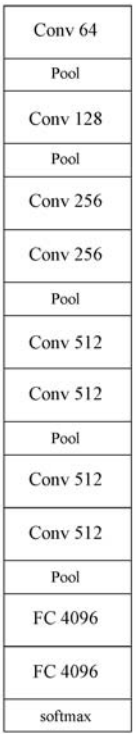

图 2　C3D 网络结构

1.2　网络结构改进

由 UC Berkeley 与 Stanford 研究人员设计完成的 SqueezeNet 网络,其设计目标并不是想得到更好的检测精度,而是希望能够简化网络复杂度。所以 SqueezeNet 主要是为了降低卷积神经网络模型参数数量而设计的,相比较直接使用 3×3 的卷积核,SqueezeNet 的参数量和理论计算量理论上都降为原来的 5/36。与传统的卷积方式不同,SqueezeNet 把原本为一层的卷积分解为两层:squeeze 层和 expand 层,每层卷积后都有一个激活层,squeeze 层里都是 1×1 的卷积,数量为 s_1;expand 层里有 1×1 和 3×3 两种卷积核,数量分别为 e_1、e_3,在数量上 $4 \times s_1 = e_1 = e_3$。expand 层之后将 1×1 和 3×3 卷积后得到的 feature map 进行拼接,然后把这两层封装为一个 Fire_Module,如图 3 所示。Fire_Module 输入的 feature map 为 $H \times W \times C$,输出的 feature map 为 $H \times W \times (e_1 + e_3)$,可以看到 feature map 的分辨率是不变的,变化的是通道的数量。

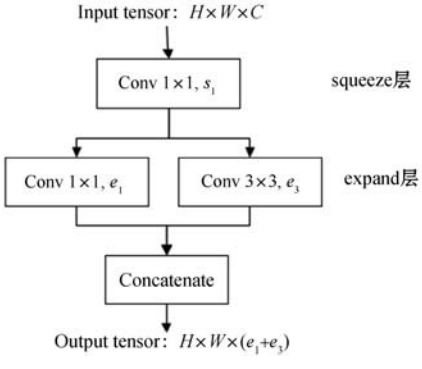

图 3　Fire_Module

综上所述,SqueezeNet 可以有效减少网络参数量,为进一步减少网络参数量并提高检测准确度,本文对 SqueezeNet 提出如下两种修订,以应用到 C3D 网络中。

(1) 因为所用网络使用的是三维卷积,所以将 Fire_Module 中的 1 ×1 和 3 ×3 卷积改为使用三维卷积 1 ×1 ×1 和 3 ×3 ×3。为提高准确度,借鉴 InceptionV1[11]结构,在 Fire_Module 的 expand 层中增加一个数量为 e_5 的 5 ×5 ×5 卷积支路,如图 4 所示,输出大小为 $H \times W \times (e_1 + e_3 + e_5)$。这样不仅增加了网络的宽度,而且改进后的 Fire_Module 包含了 3 种不同尺寸的卷积,同时也增加了网络对不同尺度的适应性,从而提高准确度。而网络越到后面,特征也越抽象,每个特征所涉及的感受野也更大,因此随着网络深度的增加,3 ×3 ×3 和 5 ×5 ×5 的卷积比例也会增加。

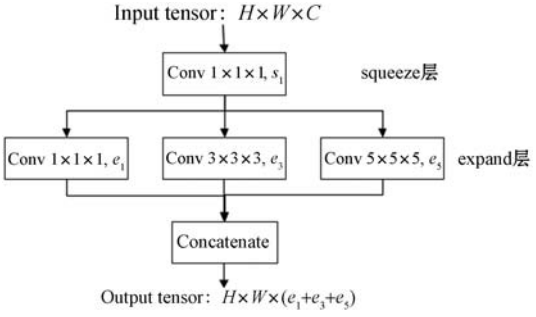

图 4 基于 Inception V1 的 Fire_Module V1

(2) 如果将网络中的卷积全部应用为上述改进后的 Fire_Module V1,因为 5 ×5 ×5 卷积所需的计算量太大,会造成特征图厚度增大,网络参数量也会随之增加。为减少网络参数量,借鉴 InceptionV3[12]结构,提出另一种方法,将 Fire_Module 中的 3 ×3 ×3 卷积替换为 3 ×1 ×3 卷积,在不影响网络性能的情况下,大大减小参数量,如图 5 所示。

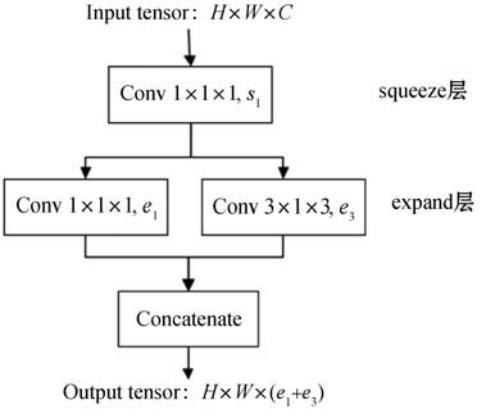

图 5 基于 InceptionV3 的 Fire_Module V2

1.3 整体网络结构

本文在 C3D 网络基础上,对网络结构进行调整,首先,紧跟着卷积层引入 BN[13](Batch Normalization)层与 short-cut 结构,如图 6 所示。BN 层在神经网络层的中间,它起到预处理的功能,也就是对上一层的输入进行归一化后,再送到网络的下一层,这样做可以有效防止梯度弥散,也可以在网络训练过程中加快网络收敛速度,加速网络训练。其次,将网络中的卷积层替换为 Fire_Module V1、Fire_Module V2,如果全部使用 Fire_Module V1,会造成网络参数量增大,经多次实验得出,Fire_Module V1、Fire_Module V2 按图 7 所示进行卷积层替换,参数量会大大减小。最后,替换后的网络深度变深,为防止训练时出现梯度退化问题以及提高精度,在 Fire_Module V2 通道数量相同的模块之间,添加 short-cut 结构。

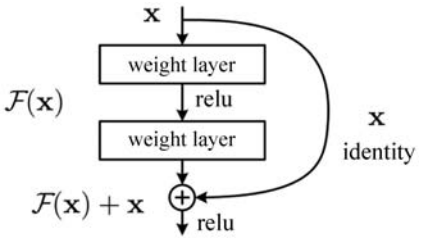

图 6 Short-cut 结构

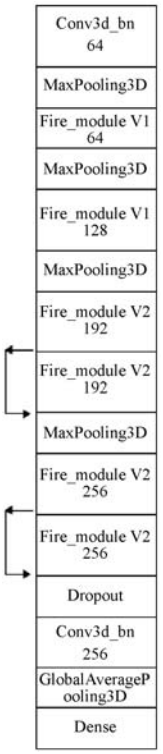

图 7 整体网络结构

2 训练及测试结果

2.1 网络训练

本文在服务器上采用 GPU 模式进行网络训练,其中训练平台配置:Intel(R) Core(TM) i7-6700 3.4 GHz 处理器;显卡为显存 12 GB 的 NVIDIA Titan X;Ubuntu 16.04 64 位操作系统;深度学习框架为 Keras。使用

UCF101 数据集,该数据集包含动作 101 类,共有 13 320 个视频,每个视频大小为 320 × 240。开始训练前,先将数据集中的视频转换为图片格式,按照 3∶1的比例将数据集分为训练集、测试集。

如图 8、图 9 所示,当训练约 15 个 epoch 后,网络收敛趋于平稳,约 20 个 epoch 后准确率达到 97.1%。

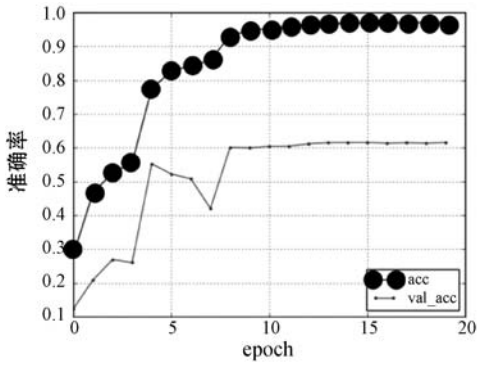

图 8　模型准确率

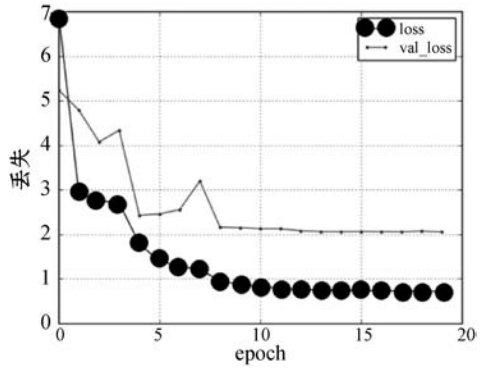

图 9　模型丢失

2.2　测试结果

将本文提出的网络所得模型进行评估并与其他文献中的行为识别网络在 UCF101 数据集上进行对比,其中 SqueezeNet-C3D 为使用 Fire_Module V1 模块的 C3D 网络,Improved SqueezeNet-C3D 为使用 Fire_Module V1 和 Fire_Module V2 结合的网络。结果如表 1 所示。

表 1　与其他网络对比

神经网络	准确率/%
C3D(1 net)[8]	82.30
LRCN[14]	71.12
Improved LRCN[15]	80.96
Temporal ConvNet[7]	83.70
SqueezeNet-C3D	84.16
Improved SqueezeNet-C3D	86.70
TSN(RGB)[16]	85.10
TSN(RGB + Flow)[16]	94.00

Temporal ConvNet 为基于深度学习的、以光流(Flow)数据作为输入的人体动作识别网络,TSN(RGB + Flow)为以光流和 RGB 数据作为输入的人体动作识别网络,其余为以 RGB 数据作为输入的人体动作识别网络。可以看到,本文提出的网络比只以光流数据作为输入的 Temporal ConvNet 高出 3%;当以 RGB 图片数据作为输入时,本文提出的网络比 C3D 高出 4.4%,比 TSN(RGB)高出 1.6%;但与 TSN(RGB + Flow)相比,本文的识别率较低。可见当 RGB 信息与光流信息融合时,能有效提高识别率,表明光流信息在提升识别率中起到重要的作用。本文只以 RGB 数据作为输入,这也是本文提出的网络识别率比 TSN(RGB + Flow)低的原因。但光流信息需要从视频帧形成,这样做会使计算量增加,所用时间也会增加,进而导致实时性变差,不利于实时检测。

本文使用改进的 SqueezeNet 与使用 SqueezeNet 网络总的参数进行对比,如表 2 所示。

表 2　网络参数对比

网络	Trainable params	Total params
C3D[8]	16 117 733	16 117 733
SqueezeNet-C3D	1 837 893	1 843 429
Improved SqueezeNet-C3D	1 564 869	1 570 789

使用本文的 ImprovedSqueezeNet-C3D 网络与使用 SqueezeNet-C3D 网络相比,参数量降低了 15%,与 C3D 网络参数量相比降低了 90.3%。由此可见,本文网络大大降低了对计算机硬件的要求。

3　结　语

本文采用改进的 SqueezeNet 与 C3D 相结合的卷积神经网络,引入 BN 层,随着网络深度的增加采用了 ResNet 的 short-cut 结构,对走路、跑步、打架、摔倒、坐、等动作进行检测识别,取得较好的检测结果,并得到以下结论:(1) 本文提出的网络提升了检测的准确度,具有较好的识别率。(2) 本文提出的网络参数量较少,降低了模型的训练及预测时间,使得网络性能在嵌入式平台(如 NVIDIA JetsonTX2)上进一步提高。

参　考　文　献

[1] 陈煜平, 邱卫根. 基于视觉的人体行为识别算法研究 综述[J]. 计算机应用研究, 2019(7):1 - 10.

[2] 朱伟,吴耀祖,刘泽祥,等. 基于部位密集轨迹的人体行为识别[J]自动化技术与应用,2018,37(9):116 - 120.

(下转第 172 页)

[14] 冯艳红,于红,孙庚,等. 基于词向量和条件随机场的领域术语识别方法[J]. 计算机应用,2016,36(11):3146-3151.
[15] 孙娟娟,于红,冯艳红,等. 基于深度学习的渔业领域命名实体识别[J]. 大连海洋大学学报,2018,33(2):265-269.
[16] 丁晟春,王莉,刘梦露. 基于规则的动物卫生事件舆情信息抽取研究[J]. 计算机应用与软件,2018,35(9):56-62.
[17] 冯蕴天,张宏军,郝文宁. 面向军事文本的命名实体识别[J]. 计算机科学,2015,42(7):15-18,47.
[18] 余晨,毛喆,高嵩. 基于规则的海事自由文本信息识别方法研究[J]. 交通信息与安全,2017,35(2):40-47.
[19] Liu H, Chen C, Zhang L, et al. The research of label-mapping-based entity attribute extraction[C]//2010 IEEE International Conference on Progress in Informatics and Computing. IEEE, 2010: 635-639.
[20] 叶正,林鸿飞,苏绥,等. 基于支持向量机的人物属性抽取[J]. 计算机研究与发展,2007,44(S2):271-275.
[21] Huang R, Riloff E. Classifying message board posts with an extracted lexicon of patient attributes[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 1557-1562.
[22] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[23] Hammerton J. Named entity recognition with long short-term memory[C]// Conference on Natural Language Learning at Hlt-naacl. 2003.
[24] 冯艳红,于红,孙庚,等. 基于 BLSTM 的命名实体识别方法[J]. 计算机科学,2018,45(2):261-268.
[25] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Netw, 2005, 18(5):602-610.
[26] Peng N, Dredze M. Named entity recognition for chinese social media with jointly trained embeddings[C] //Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 548-554.
[27] 世界事故调查跟踪[EB/OL]. Aviation Safety Information System of CAAC. [2018-09]. http://safety.caac.gov.cn/index/initpage.act.
[28] 何炎祥,罗楚威,胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件,2015,32(1):179-185,202.

(上接第 130 页)

[11] Tilkov S, Vinoski S. Node.js: using JavaScript to build high-performance network programs[J]. IEEE Internet Computing, 2010, 14(6):80-83.
[12] 牛建伟,刘洋,卢邦辉,等. 一种基于 Wi-Fi 信号指纹的楼宇内定位算法[J]. 计算机研究与发展,2013,50(3):568-577.
[13] 张勇,李飞腾,王昱. 基于 KDDA 和 SFLA-LSSVR 算法的 WLAN 室内定位算法[J]. 计算机研究与发展,2017,54(5):979-985.

(上接第 159 页)

[3] 吴心筱,伍堃. 利用深度神经网络的无监督视频表示[J]. 北京交通大学学报,2017,41(6):8-12.
[4] 朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6):848-857.
[5] 叶茂权. 基于视频的人体行为识别的理论与方法研究[D]. 成都:电子科技大学,2018.
[6] 余兴. 基于深度学习的视频行为识别技术研究[D]. 成都:电子科技大学,2018.
[7] 高阳. 基于双流卷积神经网络的监控视频中打斗行为识别研究[D]. 西安:西安理工大学,2018.
[8] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.
[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE, 2016:770-778.
[10] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[EB]. arXiv preprint arXiv: 1602.07360,2016.
[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Computer Vision and Pattern Recognition. 2014: 1-9.
[12] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Las Vegas:IEEE, 2016:2818-2826.
[13] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//International Conference on International Conference on Machine Learning. 2015:448-456.
[14] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Computer Vision and Pattern Recognition. IEEE, 2015:677.
[15] 周文. 人体行为时空特征提取与识别算法设计与实现[D]. 北京:北京交通大学, 2018.
[16] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[J]. Acm Transactions on Information Systems, 2016, 22(1): 20-36.