基于胶囊网络在复杂场景下的行人识别

程换新,刘文翰,郭占广,张志浩
(青岛科技大学 自动化与电子工程学院,山东 青岛 266061)

摘 要:大数据环境下,对行人检测的需求度不断提高,然而视频中的信息越来越丰富,视频中所获取的场景也愈加复杂。在如此背景下,目前大多使用卷积神经网络进行识别,但识别率不高。在原有的胶囊网络模型的基础上,增加了两层卷积层并将胶囊维度进行了扩展,同时使用了动态路由迭代算法,提出了一种基于改进胶囊网络的行人识别模型(PRM-ICN),该网络能够更有效地减少复杂背景中多余信息的干扰。实验在 TensorFlow 框架下使用三个国际知名且有一定难度的公开通用数据集 CUHK01、CUHK03 和 Market-1501 上进行验证,并将结果与 PRM-AlexNet 和 PRM-VGG-16 两个著名的行人识别网络相对比。实验结果表明在三个数据集上,所提出的网络模型在 CMC 曲线和 MAP 指标下都要优于其他两个网络,证明了所提模型在复杂场景下识别效果的优越性。

关键词:大数据;深度学习;胶囊网络;行人识别;TensorFlow
中图分类号:TP391.9 **文献标识码**:A **文章编号**:1673-629X(2021)02-0075-05
doi:10.3969/j.issn.1673-629X.2021.02.014

Pedestrian Recognition in Complex Scenes Based on Capsule Network

CHENG Huan-xin, LIU Wen-han, GUO Zhan-guang, ZHANG Zhi-hao
(School of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: In the context of big data, the demand for pedestrian detection is constantly increasing. However, the information in the video is getting more and more abundant, and the scenes acquired in the video are also becoming more and more complicated. Under such background, convolutional neural network is mostly used for recognition at present, but the recognition rate is not high. Based on the original capsule network model, two convolutional layers are added and the capsule dimension is extended. At the same time, according to dynamic routing iteration algorithm, a pedestrian recognition model based on the improved capsule network (PRM-ICN) is proposed, which can more effectively reduce the interference of redundant information in the complex background. The experiments are verified under the TensorFlow framework using three publicly available datasets CUHK01, CUHK03 and Market-1501, and the results are compared with two famous pedestrian recognition networks, PRM-AlexNet and PRM-VGG-16. Experiment shows that on the three data sets, the proposed network model is greater than another two networks under the CMC curve and the MAP index, which proves its superiority in complex scene recognition.

Key words: big data; deep learning; capsule network; pedestrian recognition; TensorFlow

0 引 言

随着国家的各项城市工程建设,视频监控摄像头的数量也在不断增长,随之带来的就是视频信息的大数据化。因此,现如今开发视频数据中的重要价值越来越受到人们的重视。比如公安部门可以通过视频信息实现对目标人物的跟踪、搜寻,亦可用来分析行人的种种行为是否存在异常[1],这无疑是保障社会安全的又一有力手段。

“深度学习”这一概念是在 2006 年被提出的,在这短短的十几年间已经发展出了大量算法,但是将深度学习这一技术实际应用于图片分析依然以卷积神经网络(CNN)为主。然而因为卷积神经网络的一些结构缺陷,例如无法充分地表明下层各对象之间的空间关系,而且在池化的过程中会将一些位置信息丢失[2],这使得卷积神经网络在某些场景下的图片识别并不能达到令人满意的效果。

收稿日期:2020-03-31 **修回日期**:2020-07-31
基金项目:国家海洋局重大专项项目(国海科字[2016]494 号 No.30)
作者简介:程换新(1966-),男,博士,教授,研究方向为人工智能、图像识别等;通讯作者:刘文翰(1996-),男,硕士研究生,研究方向为人工智能、计算机视觉。

胶囊(capsule)[3]的概念是由 Sabour S 等人首次提出的,他们在文中创建了一个结构简单的三层胶囊网络(capsule network),并且利用该网络实现对 Mnist 手写数字识别,识别的准确率达到了 97.5%,直接超越了 LeNet-5 模型[4]。Hinton 等人在 2018 年发表的论文中对胶囊网络中的动态路由迭代算法进行了介绍,并提出了一种新的 EM 路由算法,对胶囊网络核心路由算法进行改进[5]。改进的胶囊网络通过使用动态路由算法替代了 CNN 的池化操作[6-7],从而使得特征损失有所减小,能够在一定程度上提高图像识别的准确率。但是在目前复杂环境的应用场景下的识别研究还是少数。该文在原有 Hinton 提出的胶囊网络的基础上,对网络的结构进行一定的改进,提出了基于改进胶囊网络的行人识别算法 PRM-ICN。并使用三个公开通用的数据集 CUHK01[8]、CUHK03[9] 和 Market-1501[10]对该模型进行训练及验证。

1 改进胶囊网络模型

经过改进后的胶囊网络模型如图 1 所示,共包含三层卷积层,一层多维 Primary Capsule Layer,一层 Intermediate Capsule Layer 以及一层 Advanced Capsule Layer。

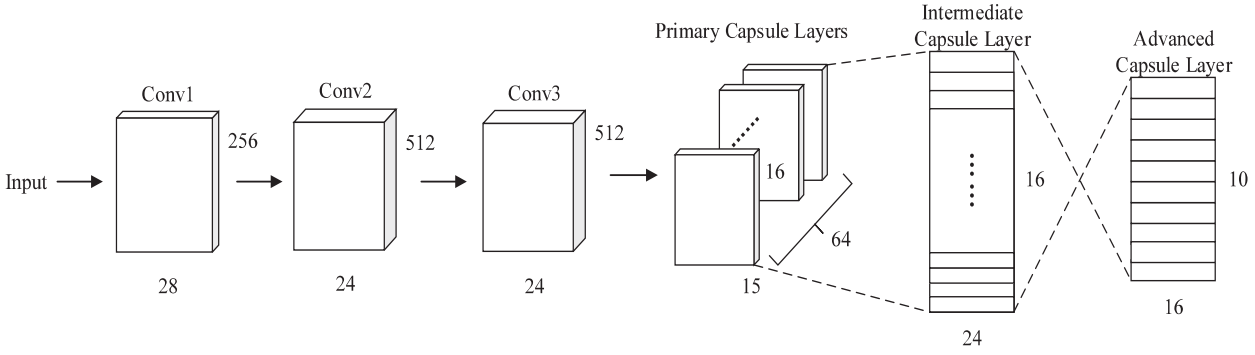

图 1 改进的胶囊网络模型结构

1.1 增加卷积层

在复杂场景下的图片所包含的信息量是巨大的,然而在图片中存在着过多的干扰信息。为了能够减少无用信息的干扰,充分联系图片中各特征的关系,并且可以在进入 Primary Capsule Layer 之前过滤一部分噪声。该网络在 Conv 1 层之后又额外增加了两层卷积层 Conv 2 和 Conv 3,从而可以减少复杂背景中多余信息对网络产生的干扰。

1.2 胶囊维度扩展

经过三层的卷积网络后,输入图片的大量有用特征被提取,经过 Primary Capsule Layer 和 Intermediate Capsule Layer 对信息进行处理后将其压缩至胶囊中。该网络中的典型结构是胶囊结构,胶囊是存储信息的单元,胶囊结构的维度越大,就有充足的存储单元对网络中的有效信息进行保存。因此,在该网络将其维度扩展至 16D。

1.3 Intermediate Capsule Layer

在胶囊层的内部,底层的特征胶囊利用姿态关系来对高层特征进行预测,之后利用动态路由算法以及筛分决策机制对高层胶囊进行选择性的激活,这就等同于筛选出了部分低层胶囊网络的预测结果,并且使高层胶囊选择性激活。

经过上述的改动之后,该网络的运行情况如下:

Conv 1:先将输入的彩色图片用 256 个 5×5 大小的卷积核进行卷积操作,其卷积步长为 1。并且在卷积操作过程中使用 ReLu 激活函数。

Conv 2:对 Conv 1 经过卷积得到的初步特征使用 512 个 5×5 的卷积核进行卷积操作,其卷积步长为 1,进而得到 Conv 2 层的输出结果。

Conv 3:进一步地将 Conv 2 层卷积得到的特征进行卷积操作,使用 512 个 5×5 大小的卷积核。

Primary Capsule Layer:对 Conv 3 层的输出结果进行向量化操作。采用 16 组不同的卷积核,而每一组卷积核中又包含 64 个不同的 15×15 的卷积核,卷积的步长设置为 1,该卷积操作的激活函数使用 ReLu。经过该步操作后,得到低级特征 U_i,该特征为 1×16 的向量。其过程如图 2 所示。

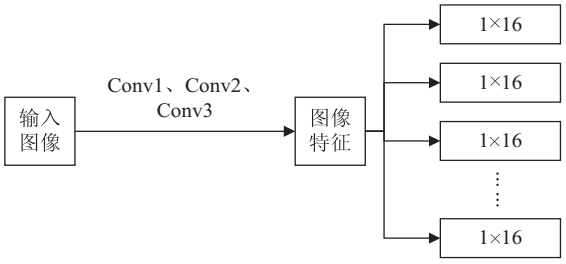

图 2 Primary Capsule Layer 结构

Intermediate Capsule Layer:通过上一层得所到的低级特征 U_i 以及各胶囊层之间的姿态关系 W_{ij} 来对高级特征 $U_{j|i}$ 进行预测。即 $U_{j|i} = U_i \cdot W_{ij}$。

Advanced Capsule Layer:利用得到的底层特征对高层特征进行预测,并利用动态路由算法以及筛分决

策机制对高层特征胶囊进行选择性激活,最终实现分类功能。

2　关键算法

2.1　向量神经元

在胶囊网络中,每个胶囊包含着众多神经元,每一个神经元存储了从图片中获取的特征。与 CNN 不同,在胶囊网络中采用向量神经元而非标量神经元,这就使得神经元可以表达的信息更丰富,从而能够提高网络的识别率。每一个向量神经元都有其自身的属性,各种各样的实例化参数都可以包含于其属性当中,比如姿态、变形、速度等。除此之外,胶囊还存在一个特殊属性,该属性描述的是图像中某一类别实例的存在与否。该属性的值为概率,其大小又取决于该向量神经元的模长,模长越大则概率越大,反之亦然。向量神经元通过 Squash( )函数进行激活,该函数能够对该向量的长度进行放大或缩小,而向量的长度又代表某一时间发生的可能性。经过该函数的激活后,能够将特征显著的向量进行放大,将特征不够明显的向量进行缩小,从而提高识别率。

2.2　姿态关系转换

胶囊层中物体各部分之间的分层姿态关系通过姿态矩阵表现出来[11-12]。为了正确地识别物体,首先应当保持分层姿态关系。而姿态无非包括旋转(rotation)、平移(translation)和缩放(scale)三种。若将某个物体姿态先逆时针旋转 60°,再将其向右平移 3 个单位,之后再缩放至原来的 50%,那么可以通过以下矩阵连乘的方式得到。其中等式右边前三个矩阵分别为 $\boldsymbol{R}$、$\boldsymbol{T}$、$\boldsymbol{S}$,而该三个矩阵相乘即为姿态矩阵 $\boldsymbol{M}$。

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \cos\frac{\pi}{3} & -\sin\frac{\pi}{3} & 0 \\ \sin\frac{\pi}{3} & \cos\frac{\pi}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}$$

2.3　动态路由迭代算法

在胶囊网络中,使用动态路由迭代算法来预测高层特征,动态路由迭代的过程如图 3 所示。该算法的输入为低层 l 中的所有胶囊及其输出 u,以及路由迭代计数 r。输出为一个高层胶囊 V_j。该算法中的 b_{ij} 是一个临时变量,它的值会在迭代过程中更新,在训练开始时,b_{ij} 的值被初始化为零。对于 c_i,$c_i = \mathrm{softmax}(b_i)$[13],而高层特征 $U_{j|i} = U_i \cdot W_{ij}$,并且 $S_j = \sum_i c_{ij} U_{j|i}$,最终 $V_j = \mathrm{squash}(S_j)$。

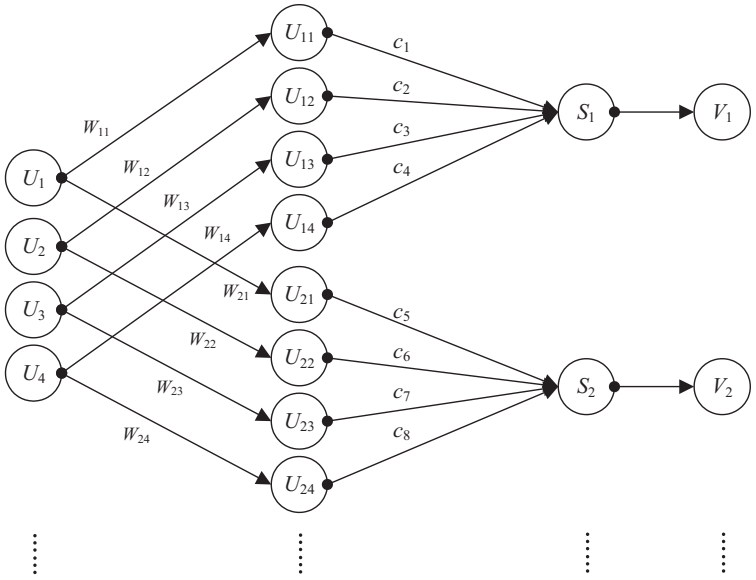

图 3　动态路由迭代算法

2.4　Squash 函数

该函数能够对该向量的长度进行放大或缩小,并且保证每个胶囊的长度都介于 0 到 1 之间。因此,每个事件发生的概率大小都可以使用每个胶囊的长度所替代,这就能提高大概率事件的可能性,而降低小概率事件的发生几率,从而提高识别的准确度。该函数的表达式为:

$$V_i = \frac{\| S_j \|^2}{1 + \| S_j \|^2} \cdot \frac{S_j}{\| S_j \|^2}$$

其中,V_i 是该函数的输出,也就是长度在 0 到 1 之间的一个向量;S_j 是输入该函数的一个胶囊,而该胶囊是经过前述三层卷积运算后的结果。对于等式右边的第

一项,该项的作用仅为放缩胶囊的长度,使得长度大的胶囊更长、长度小的胶囊进一步缩小。对于等式右边的第二项,该项的作用是保持原有的胶囊方向不变,也就是保持事件除概率外的各项特征不发生变化。

3 实验与分析

为了验证提出的基于胶囊网络的复杂场景下行人识别模型(PRM-ICN)的性能,在三个流行的公开数据集 CUHK01、CUHK03 和 Market-1501 上进行了实验,实验的结果及分析见下文。

3.1 实验设置

在本次实验中,整个网络的实现是在 TensorFlow 框架下完成的,编程语言使用的是 Python,GPU 型号为 GTX 1080Ti。对于实验结果的评价,采用常用的两个指标,即累计匹配曲线(cumulative match curve, CMC)和平均精度均值(Mmean average precision, MAP)。在行人识别领域中,一个必不可少的对模型的重要评价指标就是 CMC 曲线,该指标能够充分反映分类器的性能。MAP 曲线即为平均 AP 值,是对多次查询结果求平均 AP 值。

此外,为了验证本网络的有效性,还与另外两个在该领域性能不错的模型进行了对比,这两个模型分别为基于 AlexNet 的行人识别方法(PRM-AlexNet)[14]和基于 VGG-16 的行人识别方法(PRM-VGG-16)[15]。其中 AlexNet 包含 5 个激活函数为 ReLu 的卷积层和 3 层全连接层。VGG-16 网络包含 13 个激活函数为 ReLu 的卷积层和 3 层全连接层。

3.2 数据集

本实验采用了三个公开通用的数据集 CUHK01、CUHK03 和 Market-1501 对该模型进行训练及验证。数据集 CUHK01 中包括 971 人,其中每人都有包含正面、背面以及两侧的四张图片,在实验中选取前 750 人作为训练集,剩余作为测试集;数据集 CUHK03 包含 13 164 张行人图片,涵盖了 1 360 个人,在实验中选取前 1 100 人作为训练集,剩余作为测试集;数据集 Market-1501 包括 1 501 个行人共 32 668 张图片,在本次实验中选取前 1 300 个人的图片作为训练集,剩余作为测试集。数据集中的部分图片如图 4 所示。

图 4 数据集中的部分图片

3.3 实验结果分析

在三个数据集上的实验结果如表 1 所示。从表中可以看出,因为 CUHK01 中的图片数据相对较少,而 PRM-AlexNet 和 PRM-VGG-16 网络结构过于复杂超参数也较多,使得在训练过程中极容易发生过拟合的现象,很难利用到其深层网络的作用,从而导致两者的 CMC 较低。在三个数据集上,提出的基于胶囊网络的行人识别模型的 CMC 曲线值均高于上述两种方法。

表 1 不同方法的 CMC 比较

方法	CUHK01	CUHK03	Market-1501
PRM-AlexNet	43.0	45.0	54.2
PRM-VGG-16	51.2	59.2	72.2
PRM-ICN	**61.3**	**71.7**	**83.7**

该文创建的模型 PRM-ICN 在三个数据集上的 MAP 值如图 5 所示。从图中可以看出,在三个数据集下,所提模型在 MAP 值上的表现都优于其他两种网络。在 CUHK01 数据集上,所提模型较其他两种方法的 MAP 值分别提高了 0.18 和 0.16;在 CUHK03 数据集上,所提模型较其他两种方法的 MAP 值分别提高了 0.22 和 0.20;在 Market-1501 数据集上,所提模型较其他两种方法的 MAP 值分别提高了 0.17 和 0.10。

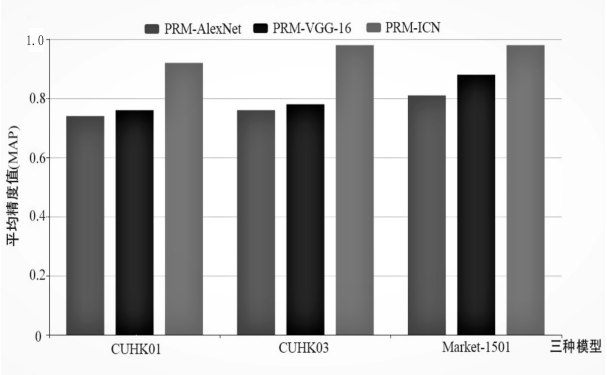

图 5 不同模型的 MAP 值

通过上述实验表明,提出的基于胶囊网络的复杂

场景下行人识别模型的性能较好。因为其在对不同姿态和空间关系的识别对象进行处理时,可以通过姿态转换来对图像进行变换,提高了识别准确率。

4　结束语

针对复杂场景下行人识别难度大的问题,为了让胶囊网络能够在复杂场景下对行人图像进行更高准确度的识别,在原始的胶囊网络基础之上对网络结构进行了优化,使之提高了在处理复杂场景中消除无用信息的能力,提出了基于改进胶囊网络的行人识别模型,并与在该领域识别效果不错的模型 PRM-AlexNet 和 PRM-VGG-16 进行了对比。所提出的模型在 CMC 曲线及 MAP 值上都优于前两者。但是该模型的识别率还未达到预期,并且还存在一定的局限性,比如并未详细考量光线变化较大的情况。未来会将更多的环境因素考虑在内,不断地修改网络结构、完善模型,使之识别准确率进一步提高。

参考文献:

[1] YUAN Y, FANG J, WANG Q. Online anomaly detection in crowd scenes via structure analysis[J]. IEEE Transactions on Cybernetics, 2015, 45(3): 548-561.

[2] GEIRHOS R, RUBISCH P, MICHAELIS C, et al. Image net-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness[EB/OL]. (2019-01-14)[2019-07-08]. https://arxiv.org/pdf/1811.12231.pdf.

[3] SABOUR S, FROSST N, HINTON G E. Dynamic rou-ting between capsules[C]//Advances in neural information processing systems 30. Long Beach: NIPS, 2017: 3856.

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[5] HINTON G E, SABOUR S, FROSST N. Matrix capsules with EM routing[C]//Sixth international conference on learning representations. Vancouver: ICLR, 2018.

[6] TOLIIAS G, SICRE R, JEGOU H. Particular object retrieval with integral max-pooling of CNN activations[EB/OL]. (2015-11-18)[2019-07-08]. https://arxiv.org/pdf/1511.05879v1.pdf.

[7] MUKHOMETZIANOV R, CARRILLO J. Caps net comparative performance evaluation for image classification[EB/OL]. (2018-05-28)[2019-07-08]. https://arxiv.org/ftp/arxiv/papers/1805/1805.11195.pdf.

[8] BEDAGKAR-GALA A, SHAH S K. A survey of approaches and trends in person re-identification[J]. Image and Vision Computing, 2014, 32(4): 270-286.

[9] ZHONG Zhun, ZHENG Liang, ZHENG Zhedong, et al. CamStyle: a novel data augmentation method for person re-identification[J]. IEEE Transactions on Image Processing, 2019, 28(3): 1176-1190.

[10] MCLAUGHLIN N, DEL RINCON J M, MILLER P C. Person reidentification using deep convnets with multitask learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 525-539.

[11] 曲铁军,任顺清,陈希军. 姿态变换矩阵性质的研究[J]. 宇航计测技术,2004,24(4):58-63.

[12] LIU W Y, WEN Y D, YU Z D, et al. Large-margin softmax loss for convolutional neural networks[C]//Proceeding of the 33rd international conference on machine learning. New York: ACM, 2016: 507-516.

[13] ELLIOTT D L. A better activation function for artificial neural networks[R]. City of College Park: Institute for Systems Research, University of Maryland, College Park, 1993.

[14] WRIGHT R E. American property: a history of how, why, and what we own (review)[J]. Register of the Kentucky Historical Society, 2011, 109(2): 276-278.

[15] LIU Luoqi, ZHANG Li, LIU Hairong, et al. Toward large-population face identification in unconstrained videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(11): 1874-1884.