基于卷积神经网络的人群计数算法研究

向飞宇,张秀伟
(西北工业大学 计算机学院,陕西 西安 710129)

摘 要:随着城市监控网络的完善,对人群图像的计数处理正产生巨大价值。传统人群计数方法存在准确度低,无法处理高遮挡图像,受光影影响大等问题。卷积神经网络在人群计数上表现出色,但仍存在精确度较低,无法排除背景图像干扰等问题。为提高对复杂人群图像的感知能力,减少背景区域对统计的影响,并同时生成人群密度特征图像,在卷积神经网络的基础上增加空间与通道注意力模型,对不同通道和不同位置的图像赋予不同的权重以增加目标区域的影响力,同时更换全连接层为上采样层,输出与输入图像大小相同的人群密度特征图像。实验中使用 ShanghaiTech 数据集以及 NWPU-Crowd 数据集进行训练与测试,在与 MCNN、CSRNet 等网络的比较结果中显示,使用了注意力模型与全卷积神经网络的算法在平均绝对值误差与均方误差两项数据上有较好的结果,表示该算法在高密度高遮挡的人群图像计数上有着更高的精确度。

关键词:人群计数;全卷积神经网络;注意力模型;扩张卷积;特征提取
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2021)07-0042-05
doi:10.3969/j.issn.1673-629X.2021.07.007

Research on Crowd Counting Algorithm Based on Convolution Neural Network

XIANG Fei-yu, ZHANG Xiu-wei
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: As the growing of urban monitoring network, the counting of crowd image is proved to be of great value. Traditional methods of crowd counting have problems like low accuracy, inability to deal with high occlusion images and being greatly affected by light and shadow. Convolution neural network performs well in crowd counting, but problems like rather low accuracy and being influenced by background area still exist. To improve the perception of complex crowd image, decrease the influence of background area and create crowd density image in the same time, spatial-wise and channel-wise attention model are added to convolution neural network in order to give more weight to target area. Full connect layers are replaced by up-sampling layers to output crowd density image by the same size of the original input image. ShanghaiTech dataset and NWPU-Crowd dataset are used to train and validate the network, and the comparison among networks like MCNN and CSRNet shows that the proposed algorithm with attention model and fully convolutional network has better results in the mean absolute error and mean square error data, indicating that it has a higher accuracy in the high-density and high-occlusion crowd image counting.

Key words: crowd counting; fully convolutional network; attention model; expansion convolution; feature extraction

0 引 言

人群计数通过对图像中人像特征的检测来获取人群数量与密度信息,对于道路监控图像的人群计数可以快速地获取实时的人群状态信息,在分析商圈人流密度、预防疫情期间人群聚集以及踩踏事故等方面有重要意义。

传统的人群计数方法主要有两大类:基于目标检测的 Haar 小波检测法[1]、梯度方向直方图检测法[2]等尝试在对图像进行处理之后检测到图像中每一个独立的人形或者移动的兴趣点,从而同时获得人群计数与每个人的位置信息,但是在面对高遮挡和高密度的场景下效果不佳;另一类方法基于对目标轨迹信息的跟踪[3],通过对连续图像或者监控视频图像的对比分析判断出正在移动的人像,此类方法包括贝叶斯聚类方

收稿日期:2020-09-16 **修回日期**:2021-01-18
基金项目:国家级大学生创新训练项目(201910699020)
作者简介:向飞宇(1999-),男,CCF 会员(C9026G),研究方向为机器视觉、计算机应用;张秀伟,博士,副教授,研究方向为多源视觉信息融合和协同处理。

法[4]和 KLT 跟踪器方法[5]等,在高遮挡场景效果很好,但是计算耗时较长,算法的实时性能不好,同时对于图像的连续性有一定的要求,不适用于单张静态图像的人群计数。

近年的研究主要围绕卷积神经网络(convolution neural network,CNN)来展开,自 2012 年 AlexNet[6]在 ImageNet 挑战赛上获得冠军后,CNN 被广泛应用于图像识别领域当中,陆续出现了 ZFNet[7]、VGGNet[8]、GoogLeNet[9]、ResNet[10]等更加深层次也更加有效的卷积神经网络。CNN 的主要算法原理是将输入图像与不同大小的卷积核进行卷积,从而提取出不同大小的图像特征信息,相比于普通的神经网络模型将输入数据转化为一位向量的处理方式,CNN 使用卷积进行特征提取能够有效地获得像素之间的空间关系,因此在图像识别领域当中取得了极佳的效果。

具体到人群计数领域,由于 CNN 的经典网络架构会使用全连接层将像素一维化为用于分类的固定长度的特征向量,无法在获得人群计数的同时获得人群密度图像,因此全卷积神经网络(fully convolutional network,FCN)[11]被提出用于人群计数。FCN 将传统 CNN 网络结构中的全连接层改为反卷积结构,使得输出由特定长度的特征向量变成与原始输入图像相同的特征图像,同时解除了输入图像的尺寸限制。FCN 对于人群计数的适用性使其成为最新网络架构的基础。为了解决人群计数算法中特征提取的空间信息相对粗略以及缺乏对不同通道间信息的处理能力问题,文中提出了一种具有注意力模型的全卷积神经网络,通过对通道数据与空间数据的注意力模型来获取更加细化的特征图像。实验证明融合了注意力模型与全卷积神经网络的人群计数方法在多个数据集上的表现优于之前的方法。

1　人群计数网络架构综述

图 1 为文中人群计数网络架构示意图,主要包括全卷积神经网络以及注意力模型两大部分,其中 VGG-16 卷积神经网络骨干,扩展卷积层与后续的上采样层构成一个全卷积神经网络。为获取更高的精确度与更多的特征信息,在扩张卷积层之后增加了空间注意力模型(spatial-wise attention model,SAM)[12]与通道注意力模型(channel-wise attention model,SAM)[13],最终输入的人群图像通过上采样层被转化为分辨率大小相同的人群密度特征图像。

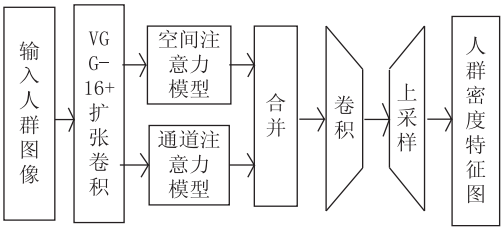

图 1　人群计数网络架构

2　全卷积神经网络

全卷积神经网络 FCN 是在传统卷积网络 CNN 的基础上,将最后的全连接层替换为反卷积层,从而将接受的输入图像尺寸从固定的尺寸扩展到任意尺寸,同时不再输出不具有空间上下文关系的特征向量,而是输出与原始图像相同尺寸的包含空间特征信息的特征图。基于这些特点,全卷积神经网络被广泛应用于图像处理领域,尤其是适用于需要在获得人群密度计数的同时获得人群密度图像的人群计数领域。标准的全卷积神经网络一般包括三个部分:卷积层、池化层和上采样层。

2.1　卷积层

卷积层是用于提取图像特征的经典结构,每一层包含多个具有权重和偏置的卷积单元,可以在前向传播的过程中提取输入图像的特征信息转化为特征映射,在反向传播中学习每个卷积单元的相关参数。图 2 为文中网络使用的卷积层示意图,由 VGG-16 卷积神经网络的前 10 个卷积层以及使用扩张卷积[14]的 6 个卷积层组成,图中[k3,3-s1-c64-d2-R]表示卷积核为 3 × 3,步长为 1,输出为 64 通道,扩张间隔为 2,使用 RELU 为激活函数的卷积层。使用 VGG-16 预训练网络可以有效地减少训练用时,同时保留很高的精确度。在整个网络架构当中没有使用全连接层,因此输入图像可以是任意尺寸的人群密度图像,从而使

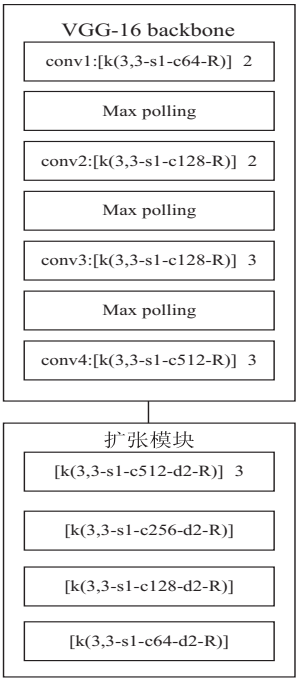

图 2　卷积层示意图

得整个网络适用于更多的应用场景。由于 VGG-16 卷积层会导致特征图像变小,使得最终经过上采样的密度图分辨率降低,为了在不降低分辨率的情况下获得高精度的特征密度图,使用了扩张卷积层在 VGG-16 骨架之后对 512 通道的输出进行扩张,最终生成 64 通道的特征图像。

2.2 池化层

池化层又称下采样层,指的是将同一通道的每个 2 × 2 非重叠区域经过处理输出为一个数据作为下一层的输入,用于降低网络中参数的数量并且降低对训练数据过拟合的风险。常用的池化层包括最大池化和平均池化,最大池化使用 2 × 2 区域的最大值作为输出,能够降低卷积层的参数误差从而保留更多的纹理信息,平均池化使用 2 × 2 区域的平均值作为输出,减少邻域大小受限导致的估计值方差增大,更多地保留图像的背景信息。在 VGG-16 卷积神经网络架构中,前 10 个卷积层内增加了三个最大池化层。

2.3 上采样层

上采样层也称为反卷积层或者转置卷积层,是将输入数据插入到更大的填充图像当中,再进行卷积,以获得与输入图像相同尺寸或者更大尺寸的输出图像的操作。在文中网络中,输入图像经过卷积层和注意力模型的特征提取后得到的较小分辨率的图像,由 1 × 1 的卷积核将通道数降至 1 后,经过 8 个上采样层恢复到与输入图像分辨率大小相同的特征密度图像[15],相当于直接输出了与输入图像分辨率大小相同的人群密度特征图。

3 注意力模型

自卷积神经网络被应用于图像识别领域以来,针对这一深度学习网络架构有着多次的改进与创新。人群计数在图像识别领域之中属于目标检测的范畴,即对于有着多个目标的输入图片,要检测出所有目标的位置及其对应的类别。传统的卷积神经网络在面对目标检测问题的时候主要是使用穷举法或者滑窗法[16]来穷举目标可能出现的所有区域,然后针对这些区域进行检测、训练和预测,这一过程有以下几个缺点:

(1)使用穷举法与设计大量不同尺寸的滑窗都需要占用大量的内存;

(2)大量非目标的区域被用于训练和预测,训练和计算的效率都比较低;

(3)对于选定的目标区域的训练放弃了区域周围的空间信息,弱化了区域与整体图像之间的空间联系。

为了解决上述的缺点,选择性搜索算法[17]被提出用于产生潜在物体候选框(region of interest,ROI),基于这一思想产生了著名的目标检测网络架构 R-CNN[18]。该算法通过对颜色相似度、纹理相似度、尺寸相似度、填充相似度等特征的加权来计算两个区域之间的相似度,依次将相邻区域中相似度最高的区域进行合并,从而获得潜在物体候选框。然而,选择性搜索算法虽然解决了内存占用以及非目标区域的问题,最终被输入进 CNN 的输入数据还是目标附近的一部分像素,不包括目标区域与整个图像之间的上下文信息,在训练和分析过程中放弃了区域与整体图像之间的空间联系。在人群计数领域,原始图像当中人群的密度分布往往是有一定规律的,例如在一张道路监控图像当中高密度的人群会集中在图像两侧的人行道区域,而一张地铁车厢人群图像当中高密度人群会聚集在图像中部,这样的特征使得在人群计数领域,目标区域与原始图像之间的空间关系是非常有价值的学习特征,而使用选择性搜索算法就会损失对这一重要特征的研究与训练,从而影响最终的预测效果。为获取人群特征与原始图像之间的空间关系,文中使用了注意力模型。

注意力模型起源于人对物体的辨识过程:人观察一幅图片的时候并非直接观察整个图片的所有信息,而是对图片的某一部分进行聚焦,例如看到一张风景图的时候人的视线会分别聚焦到远方的山,近处的水等不同的目标区域。从计算机视觉的角度,就是在识别一张图片的时候对于每个不同的像素区域有着不同的权重。对于重要的区域例如目标区域的周围,较大的权重使得关键区域对于最后的特征图像影响力较大;对于非重要的区域例如背景区域,较小的权重使得这些区域对最后的特征图像影响较小甚至没有影响。而这些权重可以通过神经网络进行训练学习,从而使得神经网络获得类似于人的注意力效果。针对人群计数对于图像空间特征与通道特征的需求,主要使用两种不同的注意力模型:空间注意力模型(spatial-wise attention model,SAM)和通道注意力模型(channel-wise attention model,CAM)。

3.1 空间注意力模型

图 3 为空间注意力模型的架构示意图,对于前置的 CNN 骨干输出的大小为 $C \times H \times W$ 的输出图像,在与三个 1 × 1 的卷积核进行卷积以减少参数和跨通道整合后,分别重整为三个不同大小的输出:S_1 与 S_3 大小为 $C \times HW$,S_2 大小为 $HW \times C$,在对 S_1 和 S_2 进行矩阵乘法和 softmax 归一化运算后,获得一张大小为 $HW \times HW$ 的空间注意力图像。为了使得最后输出的特征图像同时包含原有的图像特征信息和由注意力模型提取的大范围上下文空间信息,同时为了使最终输出与输入的大小尺寸相同,将上一步获得的 $HW \times HW$ 的图像与重整后的 S_3 进行矩阵乘法,获得大小尺寸为 $C \times H$

× W的空间注意力图像,再与最初的骨干输出进行带参数的加权和,并通过 1 × 1 的卷积核来学习这一参数。整个过程用公式表达为:

$$S^j_{\text{final}} = \lambda \sum_{i=1}^{HW} \left(\frac{\exp(S^i_1 \cdot S^j_2) \cdot S^i_3}{\sum_{i=1}^{HW} \exp(S^i_1 \cdot S^j_2)} \right) + F^j$$

式中, S^j_i 表示 S_i 在 j 处的值, F^j 表示输入的原图像在 j 处的值, λ 为可学习的参数。

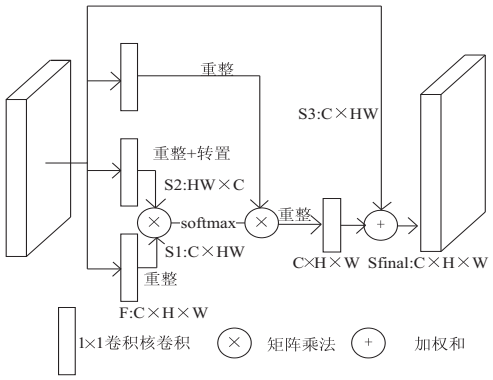

图 3　空间注意力模型示意图

3.2　通道注意力模型

空间注意力模型提取的是单通道上大范围的上下文信息,关注于人群在原始图像上的空间分布特征。除了空间分布特征以外,对于高密度的人群计数原始图像,目标区域和背景区域往往拥有较为相似的纹理特征,这使得背景区域的空间特征容易对目标特征产生额外的影响,使用通道注意力模型可以有效增加目标特征的权重,使得最终结果受到背景区域的影响减少。

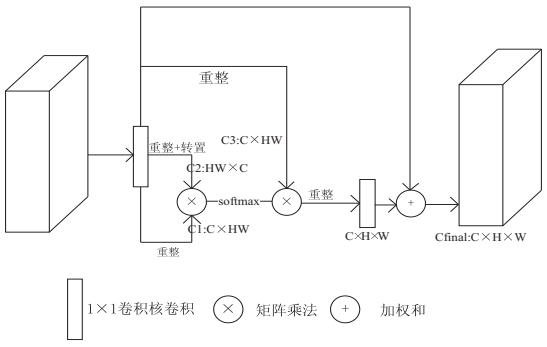

图 4　通道注意力模型示意图

通道注意力模型与空间注意模型整体上架构类似,如图 4 所示。最初的特征图仅与一个 1 × 1 的卷积核进行卷积降维后,同样重整为三个不同大小的输出: C_1 与 C_3 大小为 $C \times HW$, C_2 大小为 $HW \times C$,在对 C_1 和 C_2 进行矩阵乘法和 softmax 归一化运算后,获得一张大小为 $C \times C$ 的通道注意力图像。将上一步获得的 $C \times C$ 的图像与重整后的 C_3 进行矩阵乘法,获得大小尺寸为 $C \times H \times W$ 的通道注意力图像,最后与空间注意力模型一样与初始图像进行加权和。用公式表达为:

$$C^j_{\text{final}} = \mu \sum_{i=1}^{C} \left(\frac{\exp(C^i_1 \cdot C^j_2) \cdot C^i_3}{\sum_{i=1}^{HW} \exp(C^i_1 \cdot C^j_2)} \right) + F^j$$

式中, C^j_i 表示 C_i 在 j 处的值, F^j 表示输入的原图像在 j 处的值, μ 为可学习的参数。

4　模型训练及测试结果

4.1　实验环境及相关参数

实验采用的处理器为英特尔 i5-9600KF,运行内存为 8 GB,使用 GPU 加速,GPU 为 RTX2070super,显存为 8 G,开发环境为 pycharm;训练数据被预处理为 576 × 768 的大小,整个网络的初始学习率为 10^{-5} ,并在每个批次学习完成后下降为之前的 0.995 倍;每个批次包括 4 张图像,使用 Adam 算法进行优化,在进行 250 次迭代之后进行测试。

4.2　损失函数

在人群计数中,常用的损失函数包括平均绝对值误差(mean absolute error, MAE)和均方误差(mean squared error, MSE),分别定义为:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|^2}$$

4.3　实验结果

为了测试网络的性能,在 ShanghaiTech 数据集[19]和 NWPU-Crowd 数据集上进行了模型的训练与测试。ShanghaiTech 数据集是上海科技大学在 2016 年国际计算机视觉与模式识别会议(IEEE conference on computer vision and pattern recognition, CVPR)的会议论文中发表的人群计数数据集,包括两个部分—Part A 和 Part B,共包含 1 198 张图像。Part A 的平均分辨率为 589 × 868,平均人群计数数量为 501,Part B 的平均分辨率为 768 × 1 024,平均人群计数数量为 123。NWPU-Crowd 数据集是西北工业大学于 2020 年发表的大型人群计数图像数据库,包括 5 109 张图像,平均分辨率为 2 191 × 3 209,平均人群计数数量为 418。表 1 给出了相关算法在 ShanghaiTech 数据集上的表现,除在 Part A 的 MSE 上表现不如 CP-CNN,其余项均有一定提升。表 2 给出了相关算法在 NWPU-Crowd 数据集上的表现,文中方法在 MAE 与 MSE 上均表现最佳,较之前算法有可观的提升。

表 1　ShanghaiTech 数据集上的表现

网络模型	Part A		Part B	
	MAE	MSE	MAE	MSE
MCNN	110.2	173.2	26.4	41.3
Switching-CNN	90.4	135.0	21.6	33.4

续表 1

网络模型	Part A		Part B	
	MAE	MSE	MAE	MSE
CP-CNN	73.6	106.4	20.1	30.1
VGG-16	71.4	115.7	10.3	16.5
CSRNet	68.2	115.0	10.6	16.0
文中方法	66.4	113.8	9.6	15.1

表 2 NWPU-Crowd 数据集上的表现

网络模型	MAE	MSE
MCNN	232.5	714.6
VGG-16	105.8	504.4
CSRNet	121.3	433.5
CANNNet	93.6	489.9
文中方法	82.6	398.2

5 结束语

基于卷积神经网络的人群计数系统在传统卷积神经网络的基础上融入注意力模型与全卷积神经网络架构,在保留卷积神经网络对图像特征提取的基础上,增加了对于图像整体空间信息和通道信息的编码与学习,实现了对人群图像的准确计数与人群密度图生成,相较于之前的算法在准确度上有可观的提升。在未来的研究中,可以尝试对神经网络进行轻量化,在尽可能保证准确率的基础上将人群计数扩展到移动端和小型设备上,实现从图像获取到计数识别的一体化。

参考文献:

[1] JONES M J, SNOW D. Pedestrian detection using boosted features over many frames[C]//2008 19th international conference on pattern recognition. Tampa, FL, USA: IEEE, 2008:1-4.

[2] 常向魁,叶齐祥,刘先省,等. 基于综合色度和梯度方向直方图的运动目标跟踪算法[J]. 河南大学学报:自然科学版,2007,37(6):629-634.

[3] 伍叙励. 基于 HOG 和 Haar 联合特征的行人检测及跟踪算法研究[D]. 成都:电子科技大学,2017.

[4] BROSTOW G J, CIPOLLA R. Unsupervised bayesian detection of independent motion in crowds[C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). New York, NY, USA: IEEE, 2006:594-601.

[5] RABAUD V, BELONGIE S. Counting crowded moving objects[C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). New York, NY, USA: IEEE, 2006:705-711.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.

[7] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//European conference on computer vision. [s. l.]: Springer, 2014:818-833.

[8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//International conference on learning representations. [s. l.]: [s. n.], 2015:1-14.

[9] SZEGEDY C, LIU W, JIA Y. Going deeper with convolutions[C]//2015 IEEE conference on computer vision and pattern recognition (CVPR'15). Boston, MA, USA: IEEE, 2015:1-9.

[10] HE K, ZHANG X, REN S. Deep residual learning for image recognition[C]//2016 IEEE conference on computer vision and pattern recognition (CVPR'16). Piscataway, NJ: IEEE, 2016:770-778.

[11] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4):640-651.

[12] 张 菁,陈庆奎. 基于注意力机制的狭小空间人群拥挤度分析[J]. 计算机工程,2020,46(9):254-260.

[13] GAO Junyu, WANG Qi, YUAN Yuan. SCAR: spatial-/channel-wise attention regression networks for crowd counting[J]. Neurocomputing, 2019, 363:1-8.

[14] 刘思琦,郎丛妍,冯松鹤. 基于对抗式扩张卷积的多尺度人群密度估计[J]. 中国图象图形学报,2019,24(3):483-492.

[15] 郑 菲,孟朝晖,郭闯世. 基于反卷积特征学习的图像语义分割算法[J]. 计算机系统应用,2019,28(1):147-155.

[16] 李玲玲,刘永进,王自桦,等. 基于滑动窗口的遥感图像人造目标检测算法[J]. 厦门大学学报:自然科学版,2014,53(6):792-796.

[17] 吴素雯,战荫伟. 基于选择性搜索和卷积神经网络的人脸检测[J]. 计算机应用研究,2017,34(9):2854-2857.

[18] GIRSHICK R B, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE conference on computer vision and pattern recognition (CVPR'14). Piscataway, NJ: IEEE, 2014:580-587.

[19] ZHANG Yingying, ZHOU Desen, CHEN Siqin, et al. Single-image crowd counting via multi-column convolutional neural network[C]//IEEE conference on computer vision and pattern recognition (CVPR'16). Las Vegas, NV, USA: IEEE, 2016:589-597.