DOI:10.14188/j.1671-8836.2018.03.008

基于多尺度全卷积网络特征融合的人群计数

彭山珍,方志军†,高永彬,黄 勃,吴晨谋
(上海工程技术大学 电子电气工程学院,上海 201620)

摘 要:图像中的人群计数在公共安全领域具有重要价值.为了解决由于摄像机透视效果、人群密度分布不均匀和严重遮挡等导致人群计数准确率低的问题,本文提出一种多尺度全卷积网络架构,用于准确地估计任意摄像头视角的静态图片的人群密度。通过利用不同尺度的卷积核,使分支网络能更好地学习图像中头部特征变化.同时,由于每个分支网络设计的网络层数量不同,因此这种多尺度的网络组合能够有效地捕捉高层的语义信息和低层的细节信息.实验结果显示,本方法在 Shanghai-tech 标准数据集上具有较高的人群计数准确率.

关 键 词:人群计数;全卷积网络;语义信息;多尺度

中图分类号:TP 391 **文献标识码**:A **文章编号**:1671-8836(2018)03-0249-06

Crowd Counting Based on Feature Fusion of Multi-Scale Fully Convolutional Networks

PENG Shanzhen , FANG Zhijun† , GAO Yongbin , HUANG Bo , WU Chenmou
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: The crowd counting in the image has an important value in public safety field. In order to solve the problem of low accuracy of population counting due to camera perspective effects, uneven population density distribution, and severe occlusion, this paper presents a multi-scale full convolutional network architecture to accurately estimate the crowd density of static images from arbitrary camera perspectives. Different scales of convolution kernels helps the branch network to learn the head feature changes in the image well. At the same time, there are differences in the number of network layers per branch network, so the combination of multi-level networks can effectively capture high-level semantic information and low-level detailed information. The results of experiments prove that this method has achieved high accuracy of crowd counting based on the Shanghai-tech standard data set.

Key words: crowd counting; full convolutional networks; semantic information; multi-scale

0 引 言

人群数量或密度是视频监控、流量控制、应急管理等许多实际应用中重要的参考信息.通过对目标区域的人群数量和密度信息进行分析,可以对很多社会安全问题起到一定的预警作用,从而实现资源的合理分配和调度.因此,人群计数和人群密度估计问题已成为计算机视觉和智能视频监控领域的热点研究内容.

人群计数就是从图像或者视频帧中获得行人数量,而人群密度估计则是获得人群在一定时间和一定空间内的分布情况.目前,人群计数的方法总体可以分为直接计数和间接计数两类.直接计数法主要包括两种方法:一种是基于行人检测的方法,它通常分两个阶段进行操作,首先生成一个实值的置信图,然后从地图上找出与个人相对应的峰值,一旦所有个人的位置被估计,就很容易完成计数;另一种是基于跟踪的方法,主要针对的是视频中的人群计数.Rabaud 等[1]使用 KLT 跟踪器的高度并行版本和

收稿日期:2017-10-11 †通信联系人 E-mail:zjfang@sues.edu.cn
基金项目:国家自然科学基金(61461021);上海市科委地方能力建设项目(15590501300);上海高校青年教师培养资助计划专项基金(ZZGCD15088)资助项目.
作者简介:彭山珍,男,硕士生,主要研究方向为计算机视觉,图像视频分析. E-mail:m020216144@sues.edu.cn

聚类方法来估计移动人数. Brostow 等[2]跟踪简单的图像特征,并将它们按照一定的概率划分成能代表独立移动实体的簇. 直接计数法的缺点是在人群密度高且有遮挡情况存在时,识别率不高,而且基于跟踪的方法只能用来估计视频中的人群,不能用于估计静止图像的人群. 间接计数法将人群视为一个整体,利用图像特征和人群个数之间的回归关系来实现行人计数,这种基于回归的方法[3~11]能够有效地解决人群遮挡问题,具有大规模人群计数的能力. 回归对象通常有人群总数和人群密度图,密度图包含更多的位置信息,在人群计数的实际应用中提供的帮助更大,因此越来越多的学者[3,5,12,13]选择密度图作为回归对象. 本方法先通过回归得出精确的人群密度图,再通过对密度图积分求出人群总数. 常用的回归模型有线性回归、高斯过程回归和神经网络等. 随着卷积神经网络(convolutional neural network, CNN)的流行,近两年,基于 CNN 的一些方法[3~5,14~17]被应用于人群计数,虽然这些方法在现有的人群计数数据集上获得了不错的效果,但还存在一些局限,如文献[5,15~17]的工作是针对特定场景的,针对特定场景学习的人群统计模型只能应用于相同场景. 文献[3]提出的多列卷积神经网络(multi-column convolutional neural network, MCNN)是由三列全卷积网络组成,每一列分支网络包含 4 个卷积层和 2 个池化层,而由于每一列分支网络的网络层数相同,使得 MCNN 学习图像多尺度特征的能力不足.

本文对 MCNN 网络进行改进,提出一种将人群密度图作为回归目标的多尺度全卷积网络架构,并将其应用于人群规模显著变化,并且场景多样化的 Shanghai-tech[3]标准数据集上.

1 人群密度图的计算

CNN 需要接受监督学习训练,训练样本的好坏对模型效果影响较大[18,19]. 本文采用目前主流的基于几何自适应内核的人群密度图计算方法[3,12]进行人群计数. 首先对图片中所有人的头部位置进行标注,并保存头部位置坐标,然后将带有头部位置坐标的图片转换成人群密度图,这个人群密度图就是我们训练时需要的数据标签. 如果有一个头部位置在像素点 x_i,将其表示为 $\delta(x-x_i)$,则可将有 N 个人的头部位置标记的图像表示为

$$H(x)=\sum_{i=1}^{N}\delta(x-x_i) \tag{1}$$

将(1)式与高斯核[15] G_σ 进行卷积,得到密度估计函数

$$F(x)=H(x)*G_\sigma(x) \tag{2}$$

每个 x_i 是 3D 场景中地面人群密度的样本,并且与不同的样本 x_i 相关联的像素对应场景中不同大小的区域. 因此,为了准确地估计人群密度 F,需要考虑地平面与图像平面之间的透视失真. 假设以头部位置代替人的位置,人群是均匀分布的,根据每人和其临近的 k 个人之间的平均距离可以得出几何失真的合理估计,而拥挤场景中几乎不可能准确地获得被遮挡的头部尺寸,只能根据与其临近的人的平均距离数据来自动确定每个人的传播参数 σ. 对于给定图像中的每个头部位置,其到 k 个最近邻的人的距离为 $\{d_1^i,d_2^i,\cdots,d_k^i\}$,因此平均距离可表示为:

$$d^a=\frac{1}{k}\sum_{j=1}^{k}d_j^i \tag{3}$$

则密度 F 可表示为

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)*G_{\sigma_i}(x),\quad \sigma_i=\beta d^a \tag{4}$$

对于参数 β,通过实验发现 $\beta=0.3$ 时,计算得到的人群密度图最准确.

本文融合所有分支网络的输出特征图并将其映射到密度图. 为了将特征图映射到密度图,采用尺寸为 1×1 的卷积核[20]. 利用欧氏距离计算估计密度图和真实密度图之间的差异. 损失函数定义如下

$$L(\theta)=\frac{1}{2N}\sum_{i=1}^{N}\|F(X_i,\theta)-F_i\|_2^2 \tag{5}$$

其中,θ 是一组可学习的参数,N 是训练图像的数量,X_i 代表输入图像,F_i 是图像 X_i 的真实密度图,$F(X_i,\theta)$是估计密度图,L 是估计密度图和真实密度图之间的损失.

2 多尺度全卷积网络架构

Shanghai-tech 数据集中人群图像来自不同的拍摄视角,由于透视失真或不同分辨率的影响,头部特征变化很大. 本文的解决方法是通过设计多尺度网络架构,使用不同尺度的卷积核,从而自适应地学习图像中头部特征变化. Shanghai-tech 标准数据集中 A 部分单张图片中最多标注人数达到 3 139 个,而 B 部分单张图片中最多标注人数仅为 578 个,两部分数据集在人群规模上存在比较大的差异. 人群规模的显著变化要求将不同尺度的特征组合使用. MCNN[3]网络的每一列分支网络都只使用了 4 层卷

积,导致网络对图像多尺度特征的学习能力不足.本文通过设计多尺度的分支网络,来进一步提升模型的泛化能力.深全卷积网络(deep fully convolutional networks)能够很好地捕捉高层语义信息[21~23],浅全卷积网络(shallow fully convolutional networks)能够学习到更加充分的低层的细节信息[21],避免丢失图像中远离摄像头的头部坐标.语义信息和细节信息的融合对于准确地估计人群数量至关重要.

图 1 是人群统计总体的网络架构,它包含三列多层次的全卷积网络,第一列 Deep Network(13 层)使用了类似于 VGG -16[24]的体系结构去捕捉人群计数所需的高层语义信息[21~23].为了减少模型参数量,本文将最后三个卷积层的通道数从 512 依次降为 256、128 和 64.VGG 网络有 5 个最大池化层(Max Pool),每层的步幅为 2,因此得到的输出特征的空间分辨率仅为输入图像的 1/32.在对 VGG-16[24]模型的改变中,我们将第三和第四个最大池化层的步幅设置为 1,并且完全移除第五个池化层,这使得网络能够以 1/4 倍的输入分辨率进行预测.这样改进是因为考虑了数据集的具体情况,由于进行预处理之后的 Shanghai-tech 数据集本身的分辨率较小,因此图像分辨率缩小太多倍数会严重影响人数统计准确性.第二列 Medium Network(6 层)是一个简单的 6 层卷积网络,卷积核大小都为 3×3,是以较小的感受野去自适应比较小的头部尺寸.同样将第三、四个最大池化层的步幅设置为 1,这样可以保持每一列网络的密度图大小一致.第三列 Shallow Network(3 层)是用来识别图像中远离摄像头的头部特征.因为 Deep Network(13 层)能够很好的捕捉高层语义的信息,所以 Shallow Network 也就不需要很多网络层数.本文提出的网络架构有 3 个卷积层,每层有 24 个 5×5 的卷积核.为确保不会因为使用最大池化而损失计数,Shallow Network 使用 3 个平均池化层,第三个池化层步幅设置为 1.文献[25]证明在卷积神经网络中选择修正线性单元[26](rectified linear unit, ReLU)作为激活函数有着很好的表现,因此,本文的三列分支网络中均选用 ReLU 作为卷积层后的激活函数.

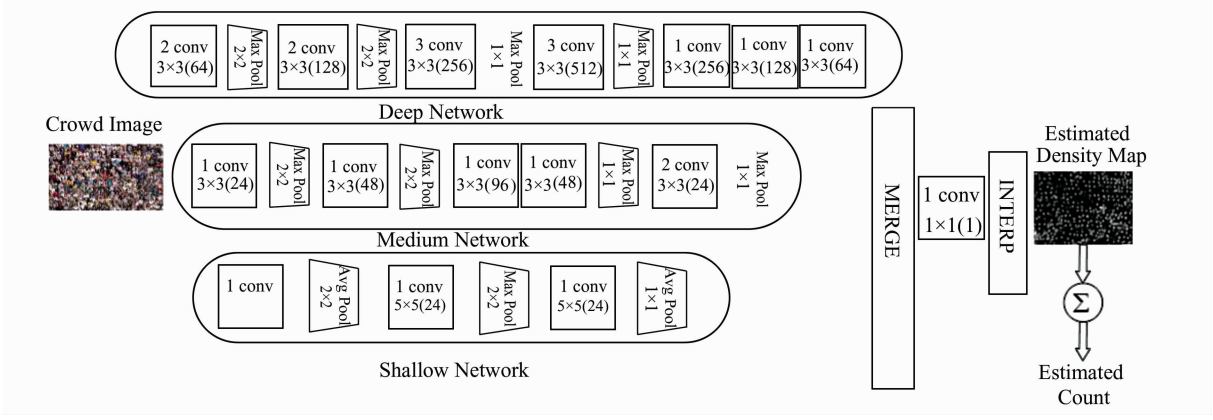

图 1　人群统计总体的网络架构

Fig. 1　Overview of the proposed architecture for crowd counting

3　实　验

3.1　实验环境

实验基于 Ubuntu 14.04,64 位操作系统,训练平台为开源深度学习框架 PyTorch,硬件配置为 GTX 1080 Ti.数据来源为上海科技大学公开的 Shanghai -tech 数据集,其中包含 1 198 个带注释的图像,总共有 330 165 人被标注了中心头像,这是目前被标注人数最多的数据集[3].该数据集由两部分组成:Part_A 由随机从互联网上爬取下来的 482 幅图像组成;Part_B 由从上海的街道拍摄的 716 幅图像组成.两部分数据集在人群密度上差异很大,因此,相比较大多数现有数据集,在 Shanghai-tech 标准数据集中进行人群计数更具挑战性.选取 Part_A 中的 300 幅图像用于训练,其余 182 幅图像用于测试;Part_B 中的 400 幅图像用于训练,其余 316 幅图像用于测试.

3.2　实验评估指标

根据现有研究工作[3,5,12],本文用平均绝对误差(MAE)和均方误差(MSE)评估不同人群计数方法,定义如下:

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|z_i-\hat{z}_i|$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i-\hat{z}_i)}$$

其中 N 是测试图像的数量，z_i 是第 i 张图像中的实际人数，$\hat{z}_i$ 是第 i 张图像中的估计人数. 一般而言，MAE 越小表示方法的估计值的准确性越高，MSE 越小表示方法的鲁棒性越高.

3.3 实验结果与分析

本文与其他方法在 Shanghai-tech 数据集上的表现如表 1 所示. 本文的方法与 MCNN[3] 相比，在 Shanghai-tech 数据集的 Part_A 测试，MAE 减少近 13 个点，MSE 减少多达 25 个点；用 Part_B 测试，MAE 减少近 1 个点，MSE 减少近 2 个点. 用 Part_A 分别将基于单个分支网络和总体网络架构的方法进行测试，结果如表 2 所示. 可以看到基于总体网络的方法在 MAE 和 MSEP 这两个评估参数上都优于基于分支网络的方法，这也证明了本文设计的总体网络架构的有效性. 图 2 和图 3 分别为 Part_A 和 Part_B 中四张测试图像及其密度图. 从图 2 和图 3 可以看出，尽管原始图像在背景和人群密度方面有很大的变化，但总体网络架构在生成合理的密度图以及估计整体人群数量方面表现出相当高的鲁棒性.

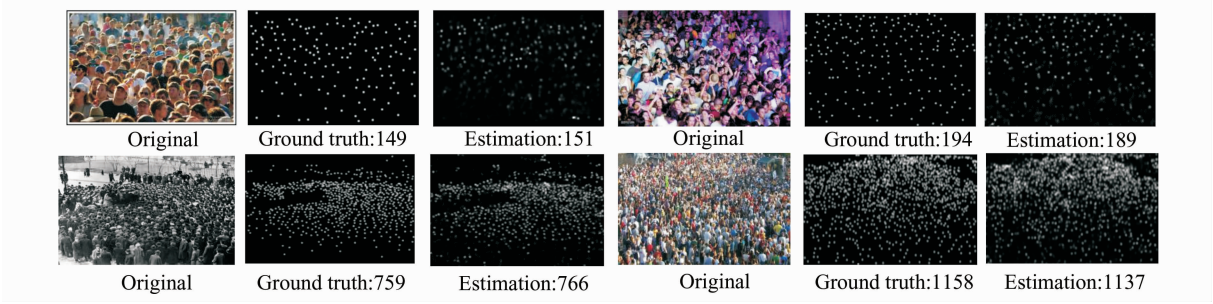

图 2 Shanghai-tech 数据集的 Part_A 中的四张测试图像及其密度图

Fig. 2 Four test images and their density maps from Part_A of Shanghai-tech dataset

图 3 Shanghai-tech 数据集的 Part_B 中的四张测试图像及其密度图

Fig. 3 Four test images and their density maps from Part_B of Shanghai-tech dataset

表 1 不同方法在 Shanghai-tech 数据集上的表现

Table 1 The performances of different methods on Shanghai-tech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
LBP+RR	303.2	371.0	59.1	81.7
Zhang *et al*.	181.8	277.7	32.0	49.8
MCNN-CCR	245.0	336.1	70.9	95.9
MCNN	110.2	173.2	26.4	41.3
Ours	97.27	147.81	25.49	40.19

注：表中数据除本文的方法外都来自于文献[3].

表 2 单分支网络和总体网络在 Shanghai-tech 数据集中 Part_A 上的表现

Table 2 The performances of single column fully convolutional networks and the whole network on Part_A of Shanghai-tech dataset

Network	MAE	MSE
Deep Network	114.09	163.41
Medium Network	119.09	191.22
Shallow Network	139.48	218.83
Whole Network	97.27	147.81

4 结 论

本文提出了一种基于多尺度全卷积网络的方法来估计静止图像的人群密度，将估计的人群密度图进行简单求和，得到最终对人群数量的估计值. 与一般基于卷积神经网络计算不同场景的人群数量的方法相比，本文的方法不需要在训练场景和测试场景上使用透视图，有效避免了实际应用中透视图难以获得的问题，极大地提高了适用性. 实验表明，使用多尺度全卷积网络的组合可以有效地解决对不同人

群规模以及高密度人群计数的困难.

本文的方法是针对单张静止图像的人群计数,接下来将考虑利用相邻视频帧之间的强时间相关性去辅助计数,以进一步提高人群计数准确性.

参考文献:

[1] RABAUD V, BELONGIE S. Counting crowded moving objects [C]//*Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2006: 705-711. DOI: 10.1109/CVPR.2006.92.

[2] BROSTOW G J, CIPOLLA R. Unsupervised bayesian detection of independent motion in crowds [C]// *Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2006: 594-601. DOI: 10.1109/CVPR.2006.320.

[3] ZHANG Y Y, ZHOU D S, CHEN S Q, *et al*. Single-image crowd counting via multi-column convolutional neural network [C]// *IEEE Conference on Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2016: 589-597. DOI: 10.1109/CVPR.2016.70.

[4] WALACH E, WOLF L. Learning to Count with CNN Boosting [DB/OL]. [2017-10-12]. *https://www.cs.tau.ac.il/~wolf/papers/learning-count-cnn.pdf*. DOI: 10.1007/978-3-319-46475-6_41.

[5] ZHANG C, LI H, WANG X, *et al*. Cross-scene crowd counting via deep convolutional neural networks [C]// *Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2015: 833-841. DOI: 10.1109/CVPR.2015.7298684.

[6] CHAN A B, LIANG Z S J, VASCONCELOS N. Privacy preserving crowd monitoring: Counting people without people models or tracking [C]// *Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2008: 1-7. DOI: 10.1109/CVPR.2008.4587569.

[7] CHEN K, CHEN C L, GONG S G, *et al*. Feature mining for localised crowd counting [C]// *British Machine Vision Conference*. Dundee: BMVA Press, 2012: 1-11. DOI: 10.5244/C.26.21.

[8] ARTETA C, LEMPITSKY V, NOBLE J A, *et al*. Interactive Object Counting [DB/OL]. [2017-02-03]. *https://www.robots.ox.ac.uk/~vgg/publications/2014/Arteta14/arteta14.pdf*.

[9] 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*. Washington, D C: IEEE Computer Society, 2015: 3253-3261. DOI: 10.1109/ICCV.2015.372.

[10] CHEN K, GONG S, XIANG T, *et al*. Cumulative attribute space for age and crowd density estimation [C]// *IEEE Conference on Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2013: 2467-2474. DOI: 10.1109/CVPR.2013.319.

[11] CHEN C L, GONG S, XIANG T. From semi-supervised to transfer counting of crowds [C]// *IEEE International Conference on Computer Vision*. Washington, D C: IEEE Computer Society, 2013: 2256-2263. DOI: 10.1109/ICCV.2013.270.

[12] SAM D B, SURYA S, BABU R V. Switching convolutional neural network for crowd counting [C]// *IEEE Conference on Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2017: 4031-4039.

[13] BOOMINATHAN L, KRUTHIVENTI S S S, BABU R V. CrowdNet: A deep convolutional network for dense crowd counting [C]//*ACM on Multimedia Conference*. New York: ACM, 2016: 640-644.

[14] SOURTZINOS P, VELASTIN S A, JARA M, *et al*. People counting in videos by fusing temporal cues from spatial context-aware convolutional neural networks [C]// *Computer Vision-ECCV 2016 Workshops*. Cham: Springer, 2016: 655-667. DOI: 10.1007/978-3-319-48881-3_46.

[15] WANG C, ZHANG H, YANG L, *et al*. Deep people counting in extremely dense crowds [C]// *ACM International Conference on Multimedia*. New York: ACM, 2015: 1299-1302. DOI: 10.1145/2733373.2806337.

[16] PARAGIOS N, RAMESH V. A MRF-based approach for real-time subway monitoring [C]//*Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2001: 1034-1040. DOI: 10.1109/CVPR.2001.990644.

[17] WANG M, WANG X. Automatic adaptation of a generic pedestrian detector toa specific traffic scene [C]//*IEEE Conference on Computer Vision and Pattern Recognition*. Washington, D C: IEEE Computer Society, 2011: 3401-3408. DOI: 10.1109/CVPR.2011.5995698

[18] FUKUSHIMA K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. *Biological Cybernetics*, 1980, **36**(4): 193-202. DOI: 10.1007/978-3-642-46466-9_18.

[19] LECUN Y, BOTTOU L, BENGIO Y, *et al*. Gradi-

ent-based learning applied to document recognition [J]. *Proceedings of the IEEE*, 1998, **86**(11): 2278-2324. DOI: 10.1109/5.726791.

[20] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014, **39**(4): 640-651. DOI: 10.1109/TPAMI.2016.2572683.

[21] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation [C] // *IEEE International Conference on Computer Vision*. Washington, D C: IEEE Computer Society, 2015: 1520-1528. DOI: 10.1109/ICCV.2015.178.

[22] 王鹏, 方志军, 赵晓丽, 等. 基于深度学习的人体图像分割算法[J]. 武汉大学学报(理学版), 2017, **63**(5): 466-470.
WAMG P, FANG Z J, ZHAO X L, *et al*. Human segmentation based on deep learning [J]. *Journal of Wuhan University (Natural Science Edition)*, 2017, **63**(5): 466-470. DOI: 10.14188/j.1671-8836.2017.05.01(Ch)

[23] GHIASI G, FOWLKES C C. Laplacian pyramid reconstruction and refinement for semantic segmentation [C] //*Computer Vision-ECCV* 2016. Cham: Springer, 2016: : 519-534. DOI: 10.1007/978-3-319-46487-9_32.

[24] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [DB/OL]. [2017-12-03]. *http://x-algo.cn/wp-content/uploads/2017/01/very-deep-convolutional-networks-for-large-scale-image-recognition.pdf*.

[25] ZEILER M D, RANZATO M, MONGA R, *et al*. On rectified linear units for speech processing [C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, D C: IEEE Computer Society, 2013: 3517-3521. DOI: 10.1109/ICASSP.2013.6638312.

[26] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines [C]// *International Conference on International Conference on Machine Learning*. New York: ACM, 2010: 807-814.

□