SSNet: An Encoder-Decoder Network for Crowd Counting

沈溯
18307130102
email: 18307130102@fudan.edu.cn

Abstract

人群计数是一个计算机视觉领域广受关注的任务。其中，基于密度图统计的人群计数方案由于其模型获取空间信息能力强、模型效果好等诸多有点，受欢迎程度越来越高。本文提出了一个基于 Encoder-Decoder 模型的网络 SSNet。SSNet 从输入的人群图片生成密度图，再在密度图上进行统计，以实现人群计数的目标。SSNet 在 ShanghaiTech PartB 数据集上进行了简单的训练和测试，获得了 MAE = 37.329, RMSE = 64.474 的成绩。

SSNet 的源码: https://github.com/bblss123/DIP2021-SSNet

1 Introduction

人群计数的任务是在给定的图像中统计出现的人的个数。近年来，随着例如外滩踩踏事件的一众公共安全事故的发生，人们对精确统计一幅图像中的人数的需求越来越大。在绝大多数情况下，利用人力，通过人眼来估计一幅图像中的人数并不实际，特别是当图像中的人数特别多的情况下。所以很多人诉诸于计算机视觉，希望能够通过数字图像处理的手段，利用计算机来自动完成人群计数。

早期人们大多使用传统的数字图像处理技术来进行人群计数。大部分方法利用检测的思路，通过统计图像中的人头或者身体的某部分出现的次数来获取图像对应的人数 (1; 2; 3)。但是在很多图片中，人像重叠的问题非常突出，此时基于检测的方法的准确率则会显著地下降。

为了克服这个问题，基于回归的人群计数方案 (4; 5) 渐渐兴起。基于回归的人群计数方案将输入的图像回归到一个表示人数的数字上。然而，这类方案无视了图像中人分布的空间信息。

随着深度学习重新回归人们的视野，人们意识到神经网络在解决人群计数问题方面存在很大的潜力。近年来也涌现出越来越多基于深度学习的人群计数方案。

Preprint. Under review.

基于密度图统计的人群计数方案 (6; 7; 8; 9; 10) 是一种典型的使用深度学习解决人群计数问题的方案。在基于密度图统计的方法中，模型利用输入的图像，生成一个人像分布的密度图，再在密度图上进行进一步的统计以获得图像行人的个数。这类方案的好处是充分利用了人群分布的空间信息。但是由于在训练过程中需要预先生成 ground truth 的密度图，这类方法对用于训练的数据集提出了更高的要求。

基于密度图统计的人群计数方案通常能在人群计数方面获得非常好的效果，虽然它对数据集的要求较高，但是随着近年来涌现出的诸多优秀数据集，基于密度图统计的模型的训练也渐渐得到了保障。人们的关注点也从基于回归的人群技术方案聚焦到基于密度图统计的人群计数方案上。

本文提出了一个基于密度图统计的方案。本文设计了一个基于 Encoder-Decoder 结构的网络 SSNet，该结构又受到了生成对抗网络 (11) 的启发，使用了一个类分辨器的网络作为 Encoder 来抽象输入图片特征，使用了一个类生成器的网络作为 Decoder 来生成密度图。由于这个网络是笔者亲手设计的第一个网络，所以笔者将其命名为 SSNet。

2 Related Work

在现如今的大多数数据集中，数据集除了给出有关图像人数的标注外，还会对图像中人头出现的像素点进行标注。我们自然能够想到利用一个只包含 0 和 1 的矩阵来描述图像中人群的分布信息，而对所有矩阵中所有元素求即可获得最终的人数。但是由于单个像素包含的信息有限，卷积神经网络很难仅通过单个点的信息习得人群分布的信息。所以密度图的概念被提出。

在密度图中，我们并不是简单的使用 0 和 1 来表示某个像素上是否出现了人头，而是利用一系列权重来表征每个像素上存在人头的个数的概率。原有的 01 矩阵上标 1 的像素权重会偏大，而其周边的权重则会随着距离的增加，渐渐减小。我们将这样处理后的图片称为密度图。

常见的密度图处理的方案是利用高斯核函数进行处理。使用高斯核函数处理的优点是其处理后的结果满足上述有关密度图的描述，而且对经过高斯核函数处理过后的像素权重求和，得到的结果等同于对原有的 01 矩阵求和的结果。

二维高斯函数的公式如下:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}$$

MCNN(10) 是一个较早使用密度图统计的方法。MCNN 吸取了多尺度网络的思想，分别使用不同大小的卷积核组成不同的网络分支进行计算，再将所有网络分支的输出组合成一个特征以生成密度图。此外，MCNN 针对高斯核函数进行了改进，提出了基于几何适应高斯核的密度图。几何适应高斯核函数如下:

$$F(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), with\ \sigma_i = \beta \overline{d_i}$$

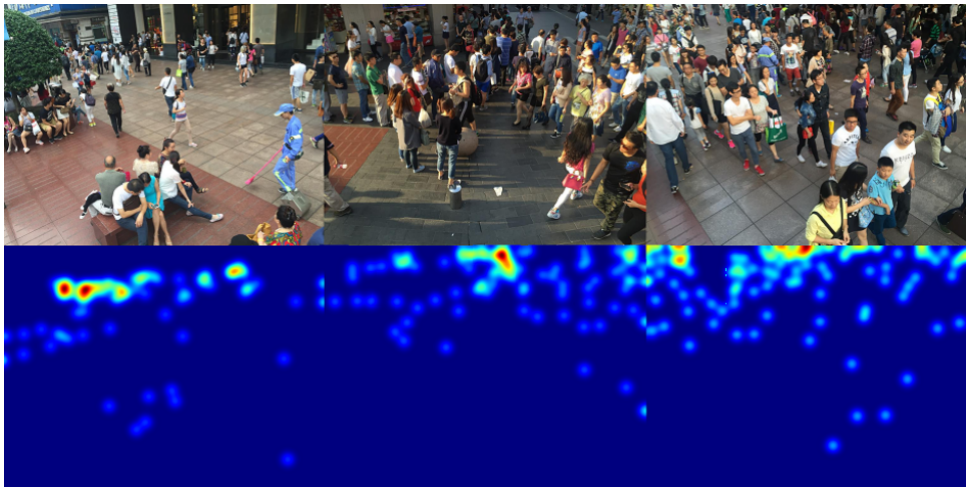

图 1: ShanghaiTechPartB 数据集中一些图像生成的密度图的展示。图中第一行的图片表示原图，第二行的图片表示对应的密度图。本图来自 CSRNet(9) 论文

CSRNet(9) 则设计了一系列空洞卷积层，并且沿用了 MCNN 中的密度图计算方法，CSRNet 在多个数据集上达到了非常理想的效果。

3 SSNet

基于密度图统计的方案的主要研究方向是如何使用输入的图像得到能够表现人群分布的密度图。其本质是利用一幅图像，生成另一幅尺寸和其相同，每个像素点上是一个 0 到 1 的浮点数的密度图。注意到这个过程与生成对抗网络的训练过程非常相似。在生成对抗网络中，我们使用一个生成器来生成图片，再使用一个判别器来对该图像进行分类。但在密度图生成的任务中，我们可以直接对比生成的密度图和 ground truth 的密度图，以替代生成对抗网络中鉴别器的工作。相应的，我们将度量生成的密度图和 ground truth 的密度图的差别作为训练的损失函数。

SSNet 的具体架构如图 1 所示。

3.1 Encoder

Encoder 由 8 层卷积神经网络构成。每层神经网络的卷积核大小为 4，步长为 2，并且填充为 1。即，每层卷积神经网络在进行卷积的过程中，会对上一层网络的输出进行 1/2 的池化。以 ShanghaiTech Part_B 数据集为例，输入图片的宽为 1024，高为 768，在经过 8 层卷积神经网络后，得到的特征张量的宽和高分别为 4 和 3。

此外，Encoder 仿照了 DCGAN(11) 中鉴别器的设计，除了第一层卷积神经网络，其后每层卷积神经网络的出口都增加了 BN 层和 LeakyRelu 层。其中 LeakyRelu 的参数设置成 0.2。

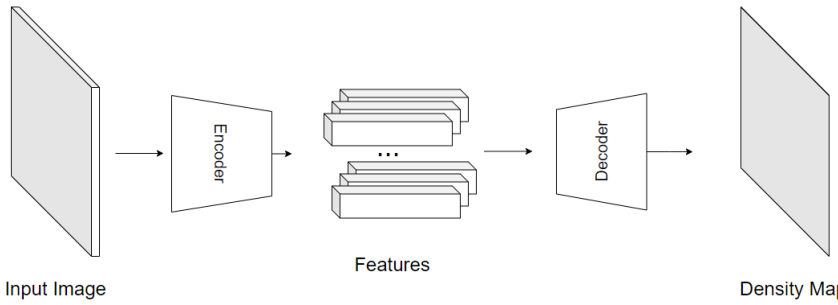

图 2: SSNet 的总体结构。其中，Encoder 是一个 8 层的卷积神经网络，输入图像每经过一层 CNN，都会进行一次 Batch Normalize 操作，并经过一层 leakyRelu 层。Decoder 一个 8 层的去卷积神经网络，其将特征向量去卷积成一个宽和高与输入图像相同，通道数为 1 的密度图。

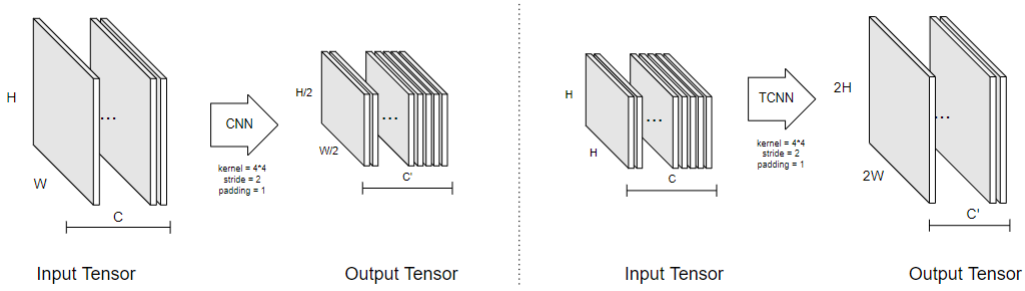

图 3: 左图展示了 Encoder 中的卷积神经网络对输入张量的处理。输入的张量为 $C \times W \times H$，输出的张量为 $C' \times W/2 \times H/2$。通常情况下，$C' = 2C$。右图展示了 Decoder 中的去卷积神经网络对输入张量的处理。输入的张量为 $C \times W \times H$，输出的张量为 $C' \times 2W \times 2H$。通常情况下，$C' = C/2$

3.2 Decoder

Decoder 由 8 层去卷积神经网络构成。每层神经网络的卷积核大小为 4，步长为 2，填充为 1。每层去卷积神经网络在进行卷积的过程中，会对输入张量进行宽和高 2 倍的放大。所以经过 8 层去卷积神经网络后，特征的宽和高都会恢复成输入 Encoder 时候的大小。同样，Decoder 仿照了 DCGAN 中生成器的设计，在除最后一层的所有去卷积神经网络的出口增加了 BN 层和 Relu 层。最后一层的去卷积神经网络使用 tanh 函数进行激活，并经过一层 relu 层。

3.3 Loss function

SSNet 的目的是最小化 Decoder 的输出和 ground truth 的密度图的区别。SSNet 选择使用几何适应高斯核函数来生成密度图，在训练时选择使用这两者的 BCELoss 来作为训练的损失函数。

4 Experiment

4.1 Experiment Details

在实际的实验中，使用 adam 作为优化器进行实验，batch size 设置成 2。一开始的学习率设置成 0.0002，观察到 100 轮后训练出现很大规模的抖动。testRMSE 在 90 150 间抖动，损失函数的差别也很大。推测是学习率设置过大。在这 100 轮训练的基础上选出最好的模型后，将 lr 设置成 0.00001 继续训练，此后模型损失函数稳步下降，并在经理了约 300 轮的训练后得到了一个较为收敛的模型

4.2 Experiment Results

由于算力和时间有限，本模型仅在 Shanghai Tech Part_B 上进行了训练和实验。

表 1: 不同方法在 ShanghaiTechPartB 上实验的对比

Method	MAE	MSE
MCNN(10)	26.4	41.2
CSRNet(9)	10.6	16.0
SSNet	37.3	64.7
Local Binary Pattern + ridge regression(10)	59.1	81.7

4.3 Ablation experiment

本文进行了一个简单的消融实验，用来验证 Encoder 和 Decoder 中的 8 层网络是有效的。本文设计了一个对照网络，其中 Encoder 和 Decoder 分别只有 5 层，每层的参数不变，以 Shanghai Tech Part_B 数据集为例，768*1024 的图像将被 Encoder 编码成 64 个 24*32 的特征张量 (在 SSNet 中，8 层 CNN 会将图像编码成 512 个 3*4 的特征向量)。Decoder 则是 Encoder 的逆运算，除了在最后一层输出的是 1*768*1024 的网络。

在实际实验过程中，这个对照网络使用 BCELoss 进行训练时，testRMSE 随着 loss 的减少不减反增，最后稳定在一个较大的数附近。推测原因是神经网络的参数量不足，导致在进行规模较大的图片生成任务时拟合能力较弱。

由于时间和算力的原因，并没有进一步增加网络深度来测试网络的效果。事实上，适当增加网络的深度可能会训练得到一个更好的模型。

5 Conclusion

本文设计了一个基于密度图统计的人群计数方法 SSNet. SSNet 使用了 Encoder-Decoder 结构，并且借鉴了生成对抗网络的实现，将 DCGAN 中的鉴别器和生成器分别作为本模型中的 Encoder 和 Decoder。从实验结果来看，SSNet 的准确率与大部分同上使用密度图统计的深度学习方法还有一些差距，但是明显要优于基于回归的方法。这并不意味着 SSNet 不具备潜能。事实上，由于时间和算力的局限性，针对 SSNet 的调参工作较为困难，而在训练过程中损失函数和每轮训练的 testRMSE 也尚在波动，即，模型可能还未完全收敛。此外，SSNet 具有参数量较小、训练速度较快的优点。将 Batch-size 设置成 2 进行训练时，每轮 epoch 仅占用 1G 的显存，GPU(2080 max-q) 的占用也仅有 29%，300 轮的训练约占用 2 小时。可能在继续优化网络结构，并且训练充分之后，SSNet 还能得到更好的效果。

References

[1] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. 2008. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In International Conference on Pattern Recognition. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/icpr/", 1–4.

[2] Zhe Lin and Larry S Davis. 2010. Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 4 (2010), 604–618.

[3] Vincent Rabaud and Serge Belongie. 2006. Counting crowded moving objects. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 1. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 705–711.

[4] Antoni B Chan and Nuno Vasconcelos. 2009. Bayesian poisson regression for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision. IEEE, IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 545–551.

[5] Antoni B Chan and Nuno Vasconcelos. 2012. Counting people with low-level features and Bayesian regression. IEEE Transactions on Image Processing 21, 4 (2012), 2160–2177.

[6] Haoyue Bai, Song Wen, and S-H Gary Chan. 2019. Crowd counting on images with scale variation and isolated clusters. In Proceedings of the IEEE International Conference on Computer Vision Workshops. IEEE, "https://dblp.uni-trier.de/db/conf/iccv/", 0–0.

[7] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. 2018. Scale aggregation network for accurate and efficient crowd counting. In Proceedings of the European Conference on Computer Vision. Springer, https://dblp.uni-trier.de/db/conf/eccv/, 734–750

[8] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. In Advances in Neural Information Processing Systems. MIT press, "https://dblp.uni-trier.de/db/conf/nips/", 1324–1332.

[9] Yuhong Li, Xiaofan Zhang, and Deming Chen. 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the

IEEE Conference on Computer Vision and Pattern Recognition. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 1091–1100.

[10] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. 2016. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, "https://dblp.uni-trier.de/db/conf/cvpr/", 589–597.

[11] Radford, A., Metz, L., and Chintala, S., "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks" , arXiv e-prints, 2015.

[12] Bai, H. and Chan, S.-H. G., "CNN-based Single Image Crowd Counting: Network Design, Loss Function and Supervisory Signal" , arXiv e-prints, 2020.