CSCI 1430 Final Project Report: Implementation of Dilated CNN for use in Highly Congested Scenes

Brown University

Abstract

I present my implementation of a network for Congested Scene Recognition called CSRNet intended to provide a data-driven and deep learning method that can understand highly congested scenes and perform accurate count estimation as well as present highquality density maps. The proposed CSRNet is composed of two major components: a convolutional neural network (CNN) as the front-end for 2D feature extraction and a dilated CNN for the back-end, which uses dilated kernels to deliver larger reception fields and to replace pooling operations. CSRNet is an easy-trained model because of its pure convolutional structure. I was able to achieve adequate performance in limited amount of time, with only 21.8% drop in accuracy (Mean Absolute Error, MAE) compared to the original, highly tuned state-of-the-art method.

1. Proposal Notes

1. Overriding principle: show us your effort.

2. If you wish us to consider any aspect of your project, it should be presented here.

3. Please include a problem statement, related work, your method, your results (figures! tables!), any comparison to existing techniques, and references.

4. If you made something work - great! If you didn't quite make it work - tell us about the problems, why you think it didn't work, and what you would do to fix it.

5. Length: Approximately four pages for techical description and one page for societal implications.

6. Please include an appendix to the report which details what each team member contributed to the project. One paragraph max each.

7. Any other materials should go in an appendix.

2. Introduction

Growing number of network models have been developed [3, 6] to deliver promising solutions for crowd flows monitoring. These works for congested scene analysis are mostly based on multi-scale architectures. They have achieved high performance in this field but the designs they used also introduce two significant disadvantages when networks go deeper: large amount of training time and non-effective branch structure (e.g., multi-column CNN (MCNN) in [7]). In this work, I implemented existing design of CSRNet to provide solution for crowd counting. This model uses pure convolutional layers as the backbone to support input images with flexible resolutions. By taking advantage of such innovative structure, this model (if implemented properly) potentially outperforms the state-of-the-art crowd counting solutions made previously [5, 2, 1].

3. Related Work

Following the idea proposed Li et al. [4], model concentrates on a novel approach to concentrate on encoding the deeper features in congested scenes and generating high quality density map. This paper highlights limitations of the previously existing state-of-the-art approaches such as multi-column based architecture (MCNN). Several disadvantages in these approaches consist of that Multi-column CNNs are hard to train, such bloated network structure requires more time to train. Also, Since the branch structure in MCNN is not efficient, the lack of parameters for generating density map lowers the final accuracy. So, this paper suggest model in which such disadvantages are not present.

4. Method

Following the similar ideas stated in the original design I've used VGG16 as the front-end of CSRNet because of its strong transfer learning ability and its flexible architecture for easily concatenating the back-end for density map generation. I've also remove the classification part of VGG16 (fully-connected layers) and build the proposed CSRNet with convolutional layers in VGG-16. The output size of this

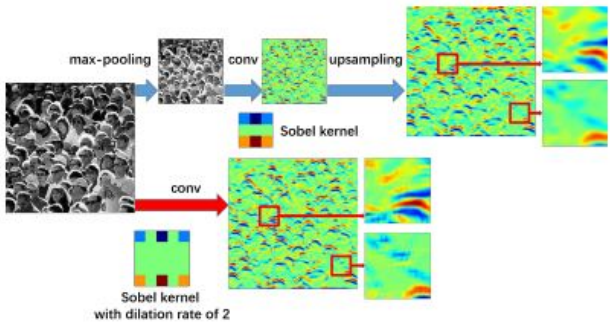

Figure 1. Comparison between dilated convolution and maxpooling, convolution, upsampling. The 3 × 3 Sobel kernel is used in both operations while the dilation rate is 2. [4]

front-end network is 1/8 of the original input size. Next, I've included one of the critical components of our design, which is dilated convolutional layer. A 2-D dilated convolution was defined as follow: (Eq. 1):

$$y(m,)(=\sum_{j=1}^{M}\sum_{j=1}^{M}x(m+r\times i,n+r\times j)w(i,j) \quad (1)$$

$y(m,n)$ is the output of dilated convolution from input $x(m,n)$ and a filter $w(i,j)$ with the length and the width of M and N respectively. The parameter$r$ is the dilation rate. If $r=1$, a dilated convolution turns into a normal convolution.

Dilated convolutional layers have been demonstrated in segmentation tasks with significant improvement of accuracy and it is a good alternative of pooling layer. For maintaining the resolution of feature map, the dilated convolution shows distinct advantages compared to the scheme of using convolution + pooling + deconvolution.

Configurations of CSRNet in Table 1. Regarding the front-end, we adopt a VGG-16 network(except fully-connected layers) and only use 3 × 3 kernels. Since the output (density maps) of CSRNet is smaller (1/8 of input size), we choose bilinear interpolation with the factor of 8 for scaling and make sure the output shares the same resolution as the input image.

We use a straightforward way to train the CSRNet as an end-to-end structure. The first 10 convolutional layers are fine-tuned from a well-trained VGG-16. For the other layers, the initial values come from a Gaussian initialization with 0.01 standard deviation. Stochastic gradient descent (SGD) is applied with fixed learning rate at 1e-6 during training. Also, we choose the Euclidean distance to measure the difference between the ground truth and the estimated density map we generated.

5. Results

We demonstrate our approach in the data set provided by original creators of the model [8]. Compared to the previous state-of-the-art methods (fig 3), our model is more accurate. But, my implementation performs worse comparing the original model implementations.

Configurations of CSRNet
input(unfixed-resolution color image)
front-end (fine-tuned from VGG-16)
conv3-64-1
conv3-64-1
max-pooling
conv3-128-1
conv3-128-1
max-pooling
conv3-256-1
conv3-256-1
conv3-256-1
max-pooling
conv3-512-1
conv3-512-1
conv3-512-1
back-end
conv3-512-2
conv3-512-2
conv3-512-2
conv3-256-2
conv3-128-2
conv3-64-2
conv1-1-1

Table 1. Configuration of CSRNet. All convolutional layers use padding to maintain the previous size. The convolutional layers' parameters are denoted as "conv-(kernel size)-(number of filters)-(dilation rate)", max-pooling layers are conducted over a 2 × 2 pixel window with stride 2.

5.1. Technical Discussion

What about your method raises interesting questions? Are there any trade-offs? What is the right way to think about the changes that you made?

Differences in my implementation from the original architecture is that in Ground truth generation they use geometry-adaptive kernels to tackle the highly congested scenes, while I simply utilize Gaussian filter to perform same operation. This decreases accuracy of my model (comparing to the orignal), but make it more manageable to implement and also greatly speeds up ground truth density map generation. My approach might have been the main reason I was not able to attain same performance as the original creators of the model. They utilize highly customized algorithm for density map generation specifically designed for their main dataset. While, my model is more generalized and less tuned for this specific dataset.

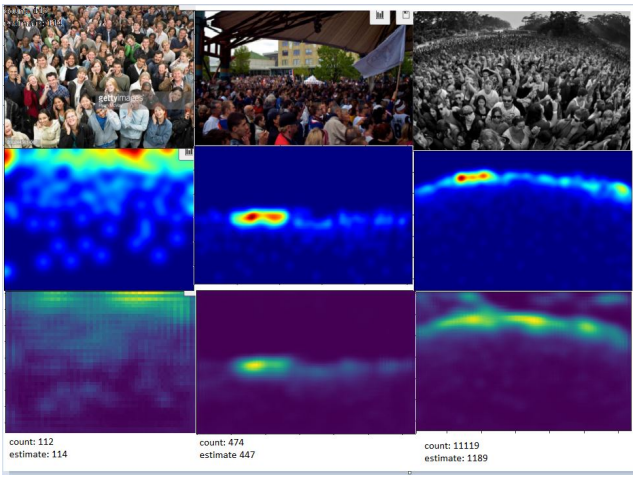

Figure 2. The first row shows the samples of the testing set in ShanghaiTech Part A dataset. The second row shows the ground truth for each sample while the third row presents the generated density map by CSRNet

My hyper-parameters and number of epoch probably highly differ from the original implementation as well. With more time I could potentially train model for more epochs and fine tune parameters to increase performance further.

5.2. Societal Discussion

Response to the Swap Critique.

1) Dataset bias. Other group had concerns regarding comprehensive quality of the project's dataset and a fact that facial recognition algorithm can be very biased towards a specific gender and race if the dataset's quality is not appropriately screened. Therefore, the concern is not valid.

My model doesn't utilize any facial recognition. It does not perform any classification at all. So, these concerns are not warranted. In fact, our model is completely gender and race blind.

2) Environmental impact. Other group had concerns regarding environmental impacts caused by training process as it may needed to be trained with a large volume of data that consumes huge energy; causing massive carbon emission problems.

My model's beauty is its efficiency and ease of training process. It's a simple combination of 2 CNNs, so the training can be done on consumer grade laptop with no great environmental concerns. Therefore, the concern is not valid.

3) Privacy concerns. Other group had concerns regarding collecting data from public places and running count and facial recognition technologies.

While these concerns are true, I am out of the agency in regards to model potential harmful use of the model, as it's simple implementation of the already existing model. We can avoid misuse of our implementation by not allowing any access to it outside of this course.

4) Data storage and security. Other group had concerns regarding secure storage of data needed for the model training.

While concern is valid, we utilize data from open sources, taken in public spaces. This data is already available to anyone. So, we are out of agency here in terms of securing the data. Additionally, kind of features present in the scenes can hardly be used to invade anyone's privacy as its extremely congested.

5) Liability. Other group had concerns regarding how I will want to deal with liabilities for unforeseen consequences that may potentially arise from use of such model.

My implementation of the model was created with no intentions of ever get used outside of this course. Therefore, there can be no liability issues. Additionally, I simply recreate architecture behind existing project and model. Therefor, my own "ideas" can't be used anywhere.

6. Conclusion

In this work, I've implemented a novel architecture called CSRNet for crowd counting and high-quality density map generation with an easy-trained end-to-end approach. We used the dilated convolutional layers to aggregate the multiscale contextual information in the congested scenes. By taking advantage of the dilated convolutional layers, CSRNet can expand the receptive field without losing resolution. We demonstrated our model on 1 crowd counting datasets and compare its performance with the orignal and with other the state-of-the-art models. I was able to achieve adequate performance in limited amount of time. This project allowed me to expand my knowledge in working with CNNs and gave me insight in developing, training, and testing a model from grounds up.

References

[1] Lokesh Boominathan, Srinivas S. S. Kruthiventi, and R. Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. *CoRR*, abs/1608.06197, 2016. 1

[2] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *in Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2547–2554. 1

[3] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, mar 2015. 1

[4] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. pages 1091–1100, 2018. 1, 2, 4

[5] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. *CoRR*, abs/1708.00953, 2017. 1

Architecture	MAE	MSE
CSRNet A	69.7	116.0
CSRNet B	**68.2**	**115.0**
CSRNet C	71.91	120.58
CSRNet D	75.81	120.82
OurCSRNet	82.7	127.2

Other state-of-the-art Methods

Method	MAE	MSE
Col. 1 of MCNN	141.2	206.8
Col. 2 of MCNN	160.5	239.0
Col. 3 of MCNN	153.7	230.2
MCNN Total	110.2	185.9
A deeper CNN	**93.0**	**142.2**

Figure 3. Results comparison of different models on the same dataset [4]. *Left:* My result are worse comparing to the different configurations of the original model. *Right:* But, still outperform other state-of-the-art approaches such as MCNN [7].

[6] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–841, 2015. 1

[7] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 4

[8] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. pages 589–597, 2016. 2