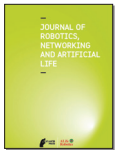
Journal of Robotics, Networking and Artificial Life
Vol. 7(2); September (2020), pp. 116–120
DOI: https://doi.org/10.2991/jrnal.k.200528.009; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal



Research Article

Crowd Counting Network with Self-attention Distillation

Yaoyao Li, Li Wang, Huailin Zhao*, Zhen Nie

School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai, China

ARTICLE INFO

Article History

Received 05 September 2019
Accepted 11 May 2020

Keywords

Self-attention distillation
dilated convolution
crowd counting

ABSTRACT

Context information is essential for crowd counting network to estimate crowd numbers, especially in the congested scene accurately. However, shallow layers of common crowd counting networks (i.e., congested scene recognition network) do not own large receptive filed so that they can't efficiently utilize context information from the crowd scene. To solve this problem, in this paper, we propose a crowd counting network with self-attention distillation. Each input image is first sent to the visual geometry group (VGG)-16 network for feature extracting. Then, the extracted features are processed by the dilated convolutional part for the final crowd density estimation. Specially, we apply self-attention distillation strategy at different locations of the dilated convolutional part to use the global context information from the deeper layers to guide the shallower layers to learn. We compare our method with the other state-of-the-art works on the UCF-QNRF dataset, and the experiment results demonstrate the superiority of our method.

1. INTRODUCTION

With the rapid growth of urban population, large-scale high-density assembly scenes are increasing, and the crowd gathering behavior is becoming more frequent and larger, which brings great difficulties and challenges to urban security systems. In order to deal with a large number of crowd monitoring data in a timely and effective manner, to prevent accidents and reduce hidden dangers in public places, crowd density estimation technology has become a research focus in the field of intelligent security.

The goal of the crowd density estimation algorithm is to estimate the number of individuals in the crowd in the entire image range through certain technical means. However, due to occlusion, perspective distortion, scale changes, and the diversity of group distribution, accurate crowd counting has always been a challenging problem in computer vision.

Traditional crowd counting algorithms are mostly based on detection and regression models. The crowd counting algorithm based on the detection model is more suitable for low-density crowd counting and has little effect on high-density scenes, similar to the pedestrian detection work. Pedestrians in the picture are detected by a pedestrian detector, and the number of detected persons is counted to calculate the total number of persons. In the crowd counting work based on the regression model, the main factors affecting the counting accuracy are the feature extraction method and the selection of the regression model. A regression algorithm is used on the extracted crowd features to establish a mapping relationship between the features and the number of people, so that the trained regressor has the ability to calculate the number of crowds. Although this method has made great progress in crowd counting, it cannot fully utilize the spatial information of the crowd, and it is still difficult to meet the accuracy requirements in dense scenes.

In recent years, due to the powerful feature expression capability and flexible architecture of neural networks, its application in the field of population density estimation has become more and more mature. By training a convolutional neural network, extracting crowd picture features, generating a corresponding crowd density map, and summing all pixels in the density map to calculate the total number of people. In order to solve the problem of population scale change, most previous counting networks used a multi-column network structure. For example, Zhang et al. [1] used multi-column convolutional neural network to capture different scales of crowd heads. Since then, in this way, Onoro-Rubio and López-Sastre [2] proposed a hydra network structure, which used multi-column convolutional neural networks to extract population characteristics of different size tiles to obtain scale information. Wang et al. [3] adds density pre-classification network on the basis of Onoro-Rubio and López-Sastre [2]. Although multi-column structures improved the counting performance, they usually had a large amount parameters, which needs much time to train. Therefore, Li et al. [4] improved the convolution operation, and proposed a Crowded Scene Recognition Network (CSRNet), which uses dilated convolution to expand the receptive field range, thereby extracting deeper feature information without increasing the amount of data. Wang et al. [5] integrates attention mechanism on the basis of CSRNet to increase context information.

Contextual information is the key to accurately estimating the number of people in a crowd counting network, especially in crowded scenarios. However, the shallow layer of the common crowd counting network (such as a CSRNet [4]) does not have a

*Corresponding author. Email: zhao_huailin@yahoo.com; www.sit.edu.cn

large receiving field, so it cannot effectively use the context information in the crowd scene. To solve this problem, this paper proposes a crowd counting network with Self-attention Knowledge Distillation (SADNet). Each input image is first sent to the VGG-16 network for feature extraction, and then the dilated convolution is used to process the extracted features. Three different attention generators have been added at different positions in the part of the dilated convolution. Use self-attention distillation strategy to obtain deeper global context information to guide shallow learning and obtain higher quality crowd density maps.

2. PROPOSED METHOD

The method of knowledge distillation was originally proposed by Hinton et al. [6], with the purpose of transferring knowledge from large networks to small networks. By introducing soft targets related to large teacher networks as part of the total loss, the training of small student networks is induced to achieve knowledge transfer. Recent research has gradually expanded knowledge distillation to attention distillation. Based on the self-attention distillation method proposed in Hou et al. [7], this paper applies attention distillation to crowd density estimation and designs a new crowd counting network with self-attention distillation. The network model can learn from itself, without any additional supervision or label annotation, to obtain deeper global context information, make the population density map estimated by the network more similar to the true value, and improve counting accuracy. The overall network structure is shown in Figure 1.

As shown in Figure 1, we apply the first 10 layers of VGG-16 to extract features. The details of VGG-16 are shown in Figure 2.

After shallow feature extraction, a six-layer hole convolution operation is used on the feature map, and then a 1 × 1 convolution is used to generate the final crowd density map, as shown in pink in Figure 1. The dilated convolution uses a 3 × 3 convolution kernel with an expansion ratio of 2, which is equivalent to six layers of 5 × 5 convolution layers. The number of channels in each layer is set to {512, 512, 512, 256, 128, 64}. The difference is that the dilated convolutional layer does not increase the amount of network calculations and avoids the loss of resolution caused by continuing the pooling operation.

The biggest feature of this paper is the application of the self-attention distillation strategy at different positions of the hollow convolutional layer, as shown in the purple part in Figure 1. After the second, fourth, and sixth dilated convolution operations, the feature map is transformed into three one-channel attention maps by the attention generator. Using the second attention map as the true value and the first attention map as the output estimate,

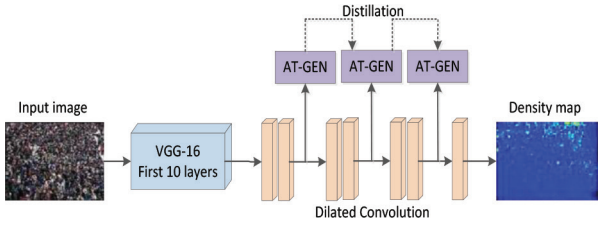


Figure 1 | The overall flowchart of the algorithm.

Figure 2 | The first 10 convolution layers of the VGG-16 network. Rectified linear unit (ReLU).

calculate the L2 loss between the two. Similarly, the third attention map is used as the true value, and the second attention map is used as the output estimation value. The L2 loss between the two is calculated. The calculated losses are fed back to the network to refine the low-level feature maps. At the same time, the feature information learned by the lower layers further improves the deeper performance of the network. Different context information is captured through attention mapping from different layers, which improves the similarity between the network-generated density map and the true-value density map. Using the network's own attention map as a distillation target does not require additional external supervision and does not increase the training time of the basic model.

Finally, the network outputs an estimated population density map. Sum all pixels in the density map to calculate the total number of people in the crowd picture.

3. EXPERIMENTS

3.1. Dataset

This experiment uses the latest dataset UCF-QNRF [8] for crowd counting to train and test the network. The UCF-QNRF data set contains a total of 1535 crowd pictures, which are divided into two parts, the training set and the test set. There are 1201 pictures in the training set and 334 pictures in the test set. These pictures not only contain a large number of people, but also complex background information such as buildings, vegetation, sky, and roads, making the data set closer to reality, and increasing the difficulty of counting people.

3.2. Density Map Generation

The crowd head annotations in the dataset are converted to true-value density maps, using the most popular adaptive Gaussian model currently available. Gaussian kernel blurs each head annotation, normalizes the sum to 1, so that all pixels in the final density map can be summed to obtain the total number of people. The formula is as follows:

$$F(x) = \sum_i^N \delta(x - x_i) * G_{\sigma_i}(x), \sigma_i = \beta \bar{d}^l \quad (1)$$

where $G_{\sigma_i}(x)$ represents the 2D-Gaussian kernel, x_i is the position coordinates of the human head in the image, $\delta(x - x_i)$ is the Dirac function of the human head, N is the total number of people included in the image, and $\bar{d}^l = \frac{1}{m}\sum_{j=1}^{m} d^i{}_j$ represents the average distance of the m heads closest to the head. In denser cases, it is approximately equal to the head size. β is a hyperparameter, here it takes 0.3.

The use of this density map makes counting network regression easier because it no longer needs to get accurate head-annotated points.

We chose the C3 framework [9,10] as the basic network to train and test the method proposed in this article.

In this paper, we uses learning rate adaptive optimization algorithm Adam to optimize the network training. At the same time, the Euclidean distance between the network estimated density map and the true value is used as the loss of the training network regression.

$$L(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\|\hat{y}(x_i;\theta) - y_i\right\|_2^2 \tag{2}$$

where θ represents the parameters to be optimized by the network, N is the number of image in the training set, x_i represents the input picture, $\hat{y}(x_i;\theta)$ represents the crowd density map estimated by the network, and y_i represents the ground truth to the input image.

The loss function can capture counting errors between the estimated and true values. The network counting error is reduced by minimizing this loss function.

3.3. Evaluation Metric

We use the Mean Absolute Error (MAE) and the Mean Square Error (MSE) to evaluate our method. MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_{X_i} - C_{X_i}^{\mathrm{GT}}\right| \tag{3}$$

where N is the total number of pictures in the test set, $C_{X_i}^{\mathrm{GT}}$ represents the true number of people corresponding to the input picture X_i, and C_{X_i} represents the number of people estimated by the network.

Mean absolute error is representative of model accuracy. In addition, to calculate the estimated variance, MSE is calculated as follows:

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{X_i} - C_{X_i}^{\mathrm{GT}})^2} \tag{4}$$

Mean square error is often used to indicate the robustness of counting predictions.

3.4. Experimental Results

The results of the self-attention distillation-based crowd counting network on the UCF-QNRF dataset are shown in Table 1.

Table 1 | The experimental results on the UCF-QNRF dataset

Method	MAE	MSE
Idrees 2013 [11]	315	508
MCNN [2]	277	426
Encoder–Decoder [8,12]	270	478
CMTL [13]	252	514
Switching CNN [14]	228	445
Resnet101 [8,15]	190	277
Densenet201 [8,16]	163	226
CL [8]	132	**191**
Our proposed	**111.7**	198.2

MAE, mean absolute error; MSE, mean square error.

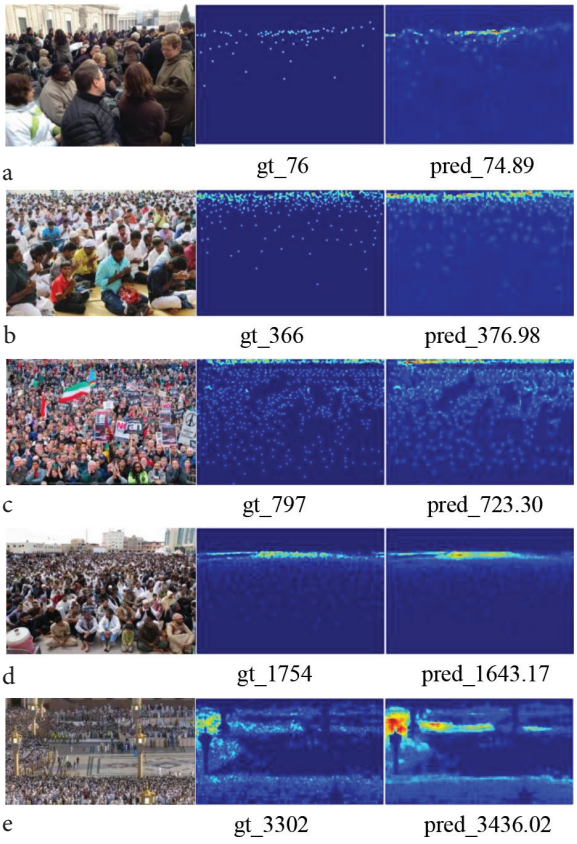


Figure 3 | Visualization of density map. a–e, Serial number of the picture; gt, ground truth, represents the actual number of people in the picture; pred, prediction, represents the number of people estimated by the network for pictures.

As can be seen from the table, our proposed method has obtained relatively small errors on the UCF-QNRF dataset. The MAE is 111.7 and the MSE is 198.2. Compared with other more advanced crowd counting networks in recent years, MAE has improved significantly. Although MSE is not the lowest, it is similar to the MSE obtained by the method proposed in Idrees et al. [8], and it is better than the other networks above. The experimental results demonstrate the effectiveness of self-attention distillation-based population counting networks.

Figure 3 demonstrates the qualitative results of our method on different crowd congested scenes. From left to right, there are input images, ground truth crowd densities and the results of our method.

It can be seen from Figure 3 that the density map estimated by the network can well reflect the distribution of the crowd, but compared with the true density map, there are still some obvious differences. In this regard, further research is needed.

4. CONCLUSION

In this paper, we propose a crowd counting network with self-attention distillation, which is improved based on the CSRNet network. The network front end uses VGG-16 to perform basic feature extraction on the input image. In the dilated convolution part of the back end of the network, we selected three different locations to join the attention generator. Use self-attention distillation strategy to obtain deeper global context information to guide

shallow learning and obtain higher quality crowd density maps. Experiments on the UCF-QNRF dataset show that the method has superior performance and higher robustness than other advanced technologies.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, Single-image crowd counting via multi-column convolutional neural network, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 589–597.

[2] D. Oñoro-Rubio, R.J. López-Sastre, Towards perspective-free object counting with deep learning, European Conference on Computer Vision (ECCV), Springer, Cham, 2016, pp. 615–629.

[3] S. Wang, H. Zhao, W. Wang, H. Di, X. Shu, Improving deep crowd density estimation via pre-classification of density, Proceedings of the IEEE International Conference on Neural Information Processing, Springer, Cham, 2017, pp. 260–269.

[4] Y. Li, X. Zhang, D. Chen, CSRNet: dilated convolutional neural networks for understanding the highly congested scenes, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Salt Lake City, USA, 2018, pp. 1091–1100.

[5] L. Wang, H. Zhao, Y. Li, Research on the multi-scale network crowd density estimation algorithm based on the attention mechanism, 2019 International Conference on Intelligent Informatics and BioMedical Sciences (ICIIBMS), IEEE, Shanghai, China, 2019, pp. 272–278.

[6] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531, 2015.

[7] Y. Hou, Z. Ma, C. Liu, C.C. Loy, Learning lightweight lane detection CNNs by self attention distillation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, South Korea, 2019, pp. 1013–1021.

[8] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, et al. Composition loss for counting, density map estimation and localization in dense crowds, Proceedings of the European Conference on Computer Vision (ECCV), 2018, Springer, Cham, pp. 544–559.

[9] J. Gao, W. Lin, B. Zhao, D. Wang, C. Gao, J. Wen, C^3 framework: an open-source PyTorch code for crowd counting, arXiv preprint arXiv:1907.02724, 2019.

[10] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 8190–8199.

[11] H. Idrees, I. Saleemi, C. Seibert, M. Shah, Multi-source multi-scale counting in extremely dense crowd images, 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Portland, OR, USA, 2013, pp. 2547–2554.

[12] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, arXiv preprint arXiv:1511.00561, 2015.

[13] V.A. Sindagi, V.M. Patel, CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, Lecce, Italy, 2017, pp. 1–6.

[14] D.B. Sam, S. Surya, R.V. Babu, Switching convolutional neural network for crowd counting, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, USA, 2017, pp. 4031–4039.

[15] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.

[16] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, arXiv preprint arXiv:1608.06993, 2016.

AUTHORS INTRODUCTION

Ms. Yaoyao Li

She received the B.S. degree from Shanghai Institute of Technology, Shanghai, China, in 2017. She is a master course student at Shanghai Institute of Technology, and her major is bionic equipment and control engineering. Her main research interests are deep learning and cross-modal research based on Generative Adversarial Network.

Ms. Li Wang

She is a master course student at Shanghai Institute of Technology, China. Her main research interests are deep learning and intelligent information processing.

Dr. Huailin Zhao

He received his PhD from Oita University, Japan in 2008. He is a professor in School of Electrical & Electronic Engineering, Shanghai Institute of Technology, China. His main research interests are robotics, multi-agent system and artificial intelligence. He is the member of both IEEE and Sigma Xi.

Mr. Zhen Nie

He received the B.S. degree from Shanghai Institute of Technology, Shanghai, China, in 2017. He is a master course student at Shanghai Institute of Technology, and his major is bionic equipment and control engineering. His main research interests are robotics and multi-agent system.