Lightweight Deep Neural Networks for Dense Crowd Counting Estimation

Stanislav Sholtanyuk
Belarusian State University
Minsk, Belarus
ssholtanyuk@bsu.by
ORCID 0000-0003-0266-7135

Aliaksandr Leunikau
Belarusian State University
Minsk, Belarus
Alex.levnikov@gmail.com

Abstract. **In this paper, productiveness problems of deep neural networks for dense crowd counting prediction have been explored. Deep neural network CSRNet has been considered, and its shallow modifications (named CSRShNet-1 and CSRShNet-2) have been designed and researched. It has been shown that for relatively small crowds (up to 500 people) it is possible to reduce training time by using shallow networks with keeping an appropriate prediction accuracy.**

Keywords: **crowd counting, deep neural networks, convolutional neural networks, supervised learning, neural network performance, neural network accuracy**

I. Introduction

Nowadays crowd size estimation and prediction of various crowd characteristics are important tasks for such activities as industry, traffic organization, social services, security systems and many others. By knowing crowd characteristics, it is possible to make immediate decisions for safety and preventing emergencies.

Using deep convolutional neural networks is the most common approach for crowd counting. Most of them use supervised learning algorithms. State-of-the-art methods include many implementations, e.g., CSRNet [1], D-ConvNet [2], MCNN [6], MRCNet [3], SANet [5], SPANet [4].

For our research, the CSRNet is chosen as the basic model because it has best results for counting a highly congested crowd, and it has simple realisation. This neural network demonstrates accurate results, but the training process is characterized by long time. We constructed and investigated shallow networks based on CSRNet which predict crowd counting with a decent accuracy and not consume too much time.

II. Description of Models and Their Features

A. CSRNet and shallow networks

CSRNet has sequential architecture and consists of convolutional and pooling layers, as pictured on Fig. 1. The input for this neural network is an image with arbitrary resolution, and the output is the estimation of crowd density map represented with a matrix with 1/8 size of the initial image. The neural network consists of two parts: pre-trained layers sequence from VGG-16 network [7] and successive trainable layers. All convolutional layers use kernels with size 3x3, but trainable layers use dilated kernels (Fig. 2).

Two shallow modifications, denoted as CSRShNet-1 and CSRShNet-2, have been designed for reducing the training time. They also consist of some pre-trained VGG-16 layers, and trainable layers use dilated kernels.

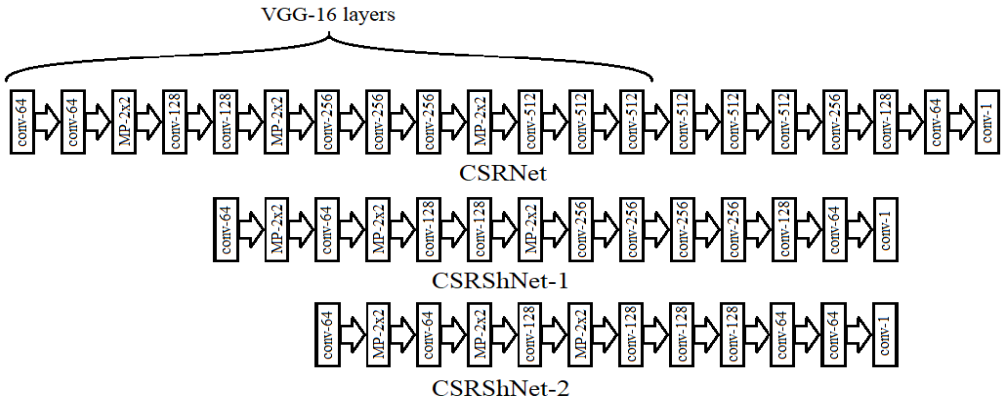


Fig. 1. Architectures of three considered neural networks, from top to bottom: CSRNet (the original one), CSRShNet-1, CSRShNet-2. Convolutional layers are denoted as conv-X, where X stands for number of output channels, and max pooling layers (which use pooling kernels with size of 2x2) are denoted as MP-2x2. In each neural network, there are pre-trained layers from VGG-16 network. Convolutional layers from VGG-16 have dilated rate 1, and other have dilated rate 2

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

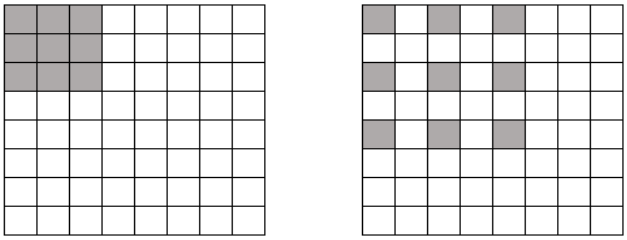


Fig. 2. 3x3 convolutional kernels with dilation rates 1 (left) and 2 (right)

B. ShanghaiTech dataset

In this research, ShanghaiTech dataset [3] of crowd images with different resolution was used for training and testing the neural networks. This dataset is divided into two parts, A and B. In the A part, there are as many images as 300 for training and 182 for testing. In the B part, there are 400 and 316 images in training and testing parts, respectively. Statistical characteristics of both samples are represented in Table I. As shown in this table, images from the part A represent larger crowds.

All neural network was trained on both parts with supervised learning. Ground truth was calculated with specific MAT-files that contain information about counting of crowd. Those files are included in ShanghaiTech dataset. It contains label positions for all people's heads for all images.

C. Software implementation

For realizations networks we use Google Colab with GPU. PyTorch framework on Python was used for neural network implementations [8].

III. Methodology

Program implementation includes the following steps:

Step 1: *Construction of ground truth density maps for each image*. For this, from the corresponding MAT-file, matrix with zeros and ones is constructed (one stands for a label, corresponding to an individual, and zero means no label). Then, gaussian blurring is applied to this matrix. Thus, the density map is created (Fig. 3b).

TABLE I. Numerical characteristics of crowd sizes on images from ShanghaiTech dataset

Characteristic	Part A	Part B
Minimal value	33	9
1st quartile	217	54
Median	359	95.5
3rd quartile	600	165.5
Maximal value	3138	576

Step 2: *Training*. For training the neural networks on a given training set (from either part A or part B) four copies of each image are taken, and the quadrupled sample is shuffled. Validation sample is composed of 1/5 randomly picked images from the initial dataset. Mean squared error (MSE) has been used as a loss function:

$$MSE = \frac{1}{n}\sum_{i=1}^{N}\left(\hat{n}_i - n_i\right)^2,$$

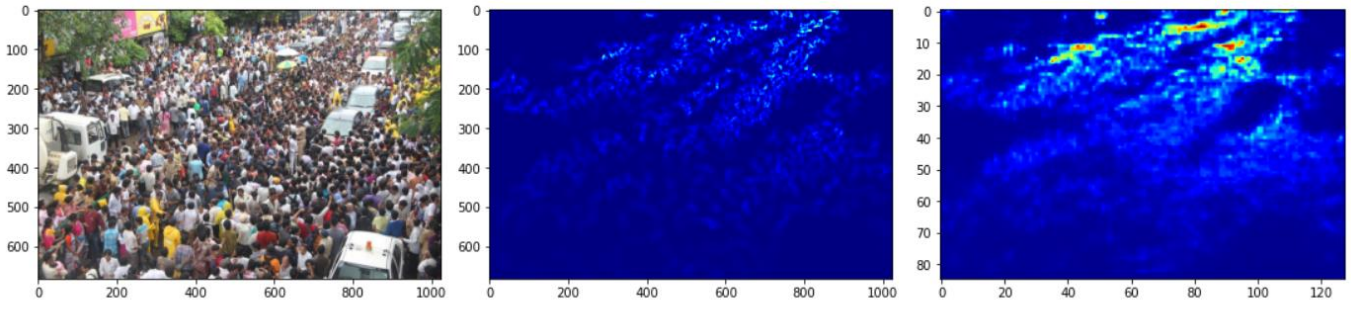
where N is sample size, $\hat{n}_i$ is estimation of the crowd size on i-th image, and n_i is the ground truth for the crowd counting. Stochastic gradient descent is used as a network optimizer. Neural network is saved after each epoch. Besides, saving the best iteration is taking place.

Step 3: *Testing*. After training, neural networks have been tested on both part A and B. Mean average error (MAE) and mean average percentage error (MAPE) was used as testing results:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{n}_i - n_i\right|,$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\hat{n}_i - n_i\right|}{n_i}.$$

An example of results is shown on Fig. 3.



a) b) c)

Fig. 3. An initial image (a). Ground truth density map (b). Density map estimated with CSRShNet-2 network (c)

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

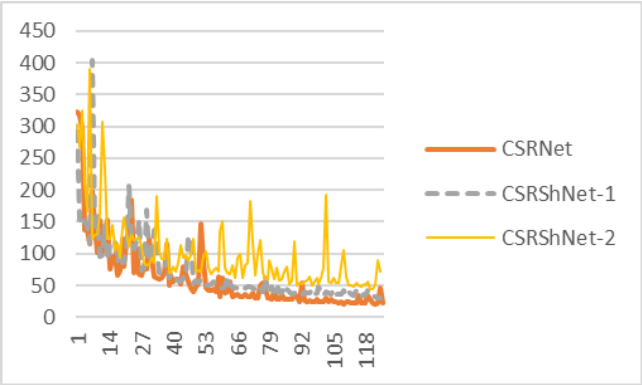


Fig. 4. Time performance results when training on part A of ShanghaiTech dataset. Horizontal axis is number of epochs, and vertical axis is MAE for the validating dataset

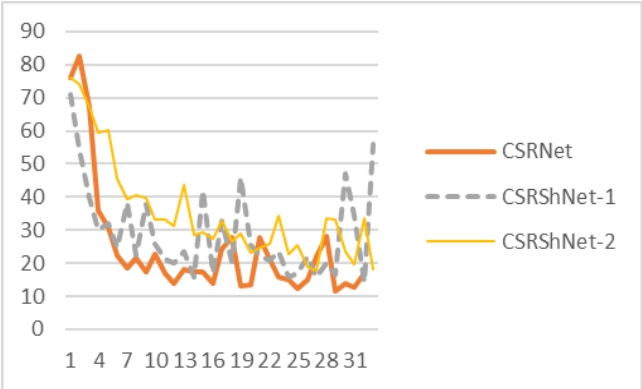


Fig. 5. Time performance results when training on part B of ShanghaiTech dataset. Horizontal axis is number of epochs, and vertical axis is MAE for the validating dataset

TABLE II. TIME AND ACCURACY RESULTS AFTER TRAINING

Name	Training Dataset[a]	Training Time, sec[b]	MAE for part A	MRE for part A	MAE for part B	MRE for part B
CSRNet	A	1413.93	71.64	0.187	22.80	0.208
CSRShNet-1	A	316.98	88.42	0.253	22.51	0.282
CSRShNet-2	A	199.61	97.69	0.301	31.14	0.426
CSRNet	B	719.75	126.55	0.299	13.87	0.129
CSRShNet-1	B	173.01	142.16	0.336	17.95	0.193
CSRShNet-2	B	99.30	171.12	0.440	21.17	0.204

[a] A stands for part A of ShanghaiTech dataset, and B stands for part B of the same dataset.

[b] Time for only epoch themselves is shown, all other activities have not been considered.

IV. RESULTS

Results for accuracy during training the neural networks on part A for 124 epochs and on part B for 33 epochs are represented on Fig. –Fig. .

From the chart on Fig. , it can be concluded that neural networks accuracy improved significantly after 15th epoch, and approximately after 55th epoch, CSRNet and CSRShNet-1 demonstrate stable results, and results of CSRShNet-2 are non-stable. According to chart on Fig. , CSRShNet-2 has greater MAEs than CSRNet, and CSRShNet-1 on different epochs has accuracy comparable with both other neural nets, so an instability is taking place.

Results of neural networks training are shown in Table II. For shallow networks training it was faster by 4-7 times than for the original network. MAE became 25-40% greater when training on the part A, and 30-60% greater when training on the part B. In terms of MAPE, performance of the shallow networks isn't very good when training on the part A since they demonstrate 25–30% error for the same sample, but there are decent results after training on part B which are about 20%. However, when testing the shallow networks on another sample rather than one they have been trained on, unsatisfactory results are obtained (25-45% when training on part A and testing on part B, and 30-45% when training on part B and testing on part A).

V. CONCLUSIONS

Proposed shallow neural networks have been designed which can be used in predicting crowd characteristics for dense crowds up to 500 people. Such network architectures for convolutional neural networks don't require too much time to train and have good performance in crowd counting.

The number of epochs for training CSRShNet-1 and CSRShNet-2 can be reduced to 30-50 without significant productiveness loss (MAPE is up to 20% for CSRNet and both shallow modifications which are trained on part B of ShanghaiTech dataset), therefore the training time can also be saved.

ACKNOWLEDGMENT

The reported study was funded by RFBR and BRFBR, project number 20-57-00025 / BRFFI F20R-134.

REFERENCES

[1] Y. Li, X. Zhang, D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1091-1100.

[2] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, G. Zheng, "Crowd Counting with Deep Negative Correlation Learning," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5382-5390.

[3] Y. Zhang, D. Zhou, S. Chen, S. Gao, Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 589-597.

[4] R. Bahmanyar, E. Vig, P. Reinartz, "MRCNet: Crowd Counting and Density Map Estimation in Aerial and Ground

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

Imagery," British Machine Vision Conference; Workshop on Object Detection and Recognition for Security Screenin (BMVC-ODRSS), 2019, pp. 1-12.

[5] X. Cao, Z. Wang, Y. Zhao, F. Su, "Scale Aggregation Network for Accurate and Efficient Crowd Counting," European Conference on Computer Vision (ECCV), 2018, pp. 734-750.

[6] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A. G. Hauptmann, "Learning Spatial Awareness to Improve Crowd Counting," IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6152-6161.

[7] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition" arXiv preprint arXiv:1409.1556, 2014.

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," Advances in Neural Information Processing Systems 32, 2019, pp. 8024-8035.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.