ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0061

Crowd Counting Algorithm Based on Hybrid Attention Mechanism

Pei Chen

College of Computer Science, Chongqing University, Chongqing, China.

Abstract

In order to solve the problem of poor performance in crowd counting task in different dense scenes caused by severe scale change and occlusion, a multi-scale spatial attention feature fusion network is proposed based on dense scene recognition network (CSRNet) by adding multi-scale feature fusion structure and introducing hybrid attention mechanism. Shuffle attention (SA) module combines two types of attention mechanism effectively by using shuffle unit. Specifically, SA first groups channel dimensions into sub features, and then processes them in parallel. Then, for each sub feature, SA uses shuffle units to describe feature dependencies in spatial and channel dimensions. Experiments show that the method proposed in this paper has more advantages than other methods.

Keywords

Crowd Counting; Attention Mechanism; Multiscale Features.

1. Introduction

The task of dense crowd counting is to estimate the number of people in the image or video. With the increase of the global population and the increase of human social activities, a large number of people often gather in public places, such as transportation hubs and entertainment places, which brings great hidden danger to public safety. Dense crowd counting task is widely used in video surveillance, traffic control and metropolitan security, and researchers in various countries have carried out a lot of research[1,2,3].

In the early days, the crowd counting network using CNN was a single branch network structure with only one data path. Wang et al first introduced CNN into the field of crowd counting, and proposed an end-to-end CNN regression model suitable for dense crowd scenes. In order to predict the number of people directly, the last full connection layer is replaced by a single neuron layer, Although the model automatically learns effective counting features through CNN, due to the narrow width and shallow depth of alexnet, the robustness of features is not strong enough, the counting effect is poor in crowded scenes, and the effect is not ideal in cross scene counting, lacking sufficient generalization. In order to solve the cross scene problem, Zhang et al. Proposed a cross scene counting model crowd CNN based on alexnet, and tried to output the crowd density map for the first time.

In order to solve the multi-scale problem, boominathan et al. Proposed a dual branch counting network crowdne based on CNN[4]. Through a shallow network and a deep network This combination can capture high-level and low-level semantic information at the same time to adapt to the non-uniform scaling of the crowd and the change of perspective, so it is conducive to the crowd counting of different scenes and different scales. MCNN establishes the nonlinear relationship between the image and the crowd density map. By replacing the full connection layer with the full convolution layer, the model can process the input image of any size. In order to further correct the influence of the change of view angle, MCNN does not use the fixed Gaussian kernel to generate the density map, but uses the adaptive Gaussian kernel to calculate the density map, which improves the quality of the density map[5].

ISSN: 2414-1895 DOI: 10.6919/ICJE.202105_7(5).0061

Attention mechanism is an important means to improve network performance. In this paper, shuffle unit effectively combines two types of attention mechanism[6]. Specifically, SA first groups channel dimensions into sub features, and then processes them in parallel. Then, for each sub feature, SA uses shuffle units to describe feature dependencies in spatial and channel dimensions[7]. After that, all the sub features are summarized together, and the "channel shuffling" operator is used to enable information communication between different sub features[8,9].

2. Method

2.1 Overall network structure

The main goal of this model is to learn a mapping f from the original image to the density map.

$$D_i^{est}(I_i) = F(I_i, \theta), \qquad 1 \leq i \leq N$$

Where: I_i is the input image; D_i^{est} is the predicted density map; and θ is the learned network parameter. The specific population count results can be obtained by integrating the predicted density map D_i^{est}. Based on the above mathematical model, this paper proposes a multi-scale hybrid attention feature fusion network, and the architecture is shown in Figure 1.

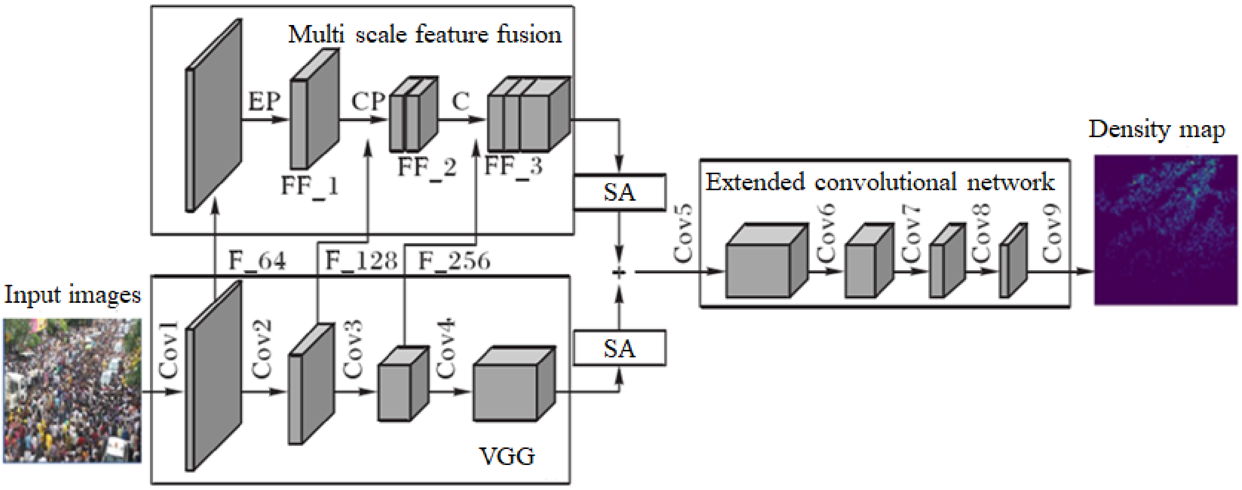

Figure 1. The network framework

The first 10 layers of VGG16 network[11,12] are selected as the front-end network, and only three pooling layers are reserved. Its powerful feature extraction ability and adjustable structure are convenient for feature fusion. The 7-layer extended convolution layer is used as the back-end network to extract more important information in a wide range of receiving fields and maintain the resolution of the output density map.

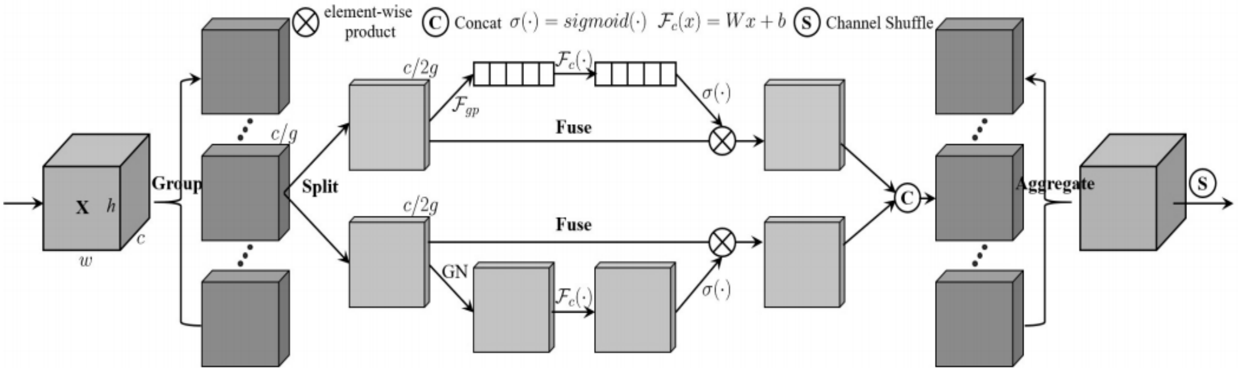

Figure 2. Network structure of SA module

2.2 SA module

Shuffle attention (SA) module combines two types of attention mechanism effectively by using shuffle unit. Specifically, SA first groups channel dimensions into sub features, and then processes them in parallel. Then, for each sub feature, SA uses shuffle units to describe feature dependencies in spatial and channel dimensions. After that, all the sub features are summarized together, and the "channel shuffling" operator is used to enable information communication between different sub features[13]. The details of the SA module are shown in Figure 2.

2.3 Comprehensive loss function

In crowd scenes, the local patterns and texture features of high-density regions are quite different from those of other regions (low-density regions or backgrounds), but the Euclidean loss is based on the assumption of pixel independence and ignores them. The local correlation of density maps is not considered

Euclidean loss is used to measure the difference between the output density map and the corresponding true value at the pixel level, which is defined as follows:

$$L_2(\Theta) = \frac{1}{N}\sum_{i-1}^{N} \|F_d(I_i;\ \Theta) - D_i\|$$

Where, $\Theta$is a set of parameters during network training, N is the number of training samples, and F_d represents the estimated density map of the network input image with parameter Θ after eight, while D_i is the corresponding true value density map.

Most counting algorithms based on density estimation optimize their counting model by measuring the error per pixel between the predicted density map and the actual density map. However, this method is not directly related to Mae and MSE, which are used to measure the accuracy, nor does it take into account the global counting error of the input image It directly measures the difference between the estimated population and the real population. The network will generate features suitable for the overall density level of the input image, which helps to produce more accurate density values. Its definition is as follows.

$$L_c = \left\|\hat{C} - C\right\|^2$$

Where $\hat{C}$ and C are the number of trained people and the real number of people respectively.

The final comprehensive loss function is as follows:

$$L = L_2 + \alpha \cdot L_c$$

α is the super parameter to adjust the proportion of different items, $\alpha = 0.02$.

3. Experimental analysis

3.1 Laboratory equipment and environment

Our experiment is on 4 Titan XP GPU. The network is based on the Pytorch framework. We use Adam optimizer to optimize the parameters and set the original learning rate to le-5. The parameters are randomly initialized by Gaussian distribution. The average value is zero and the standard deviation is 0.01. In addition to the output layer, we also use batch standardization layer and RELU layer after each convolution layer In order to improve the training speed and effectively avoid the disappearance and explosion of gradient.

3.2 Experimental analysis

The UCF_CC_50 data set proposed by Idrees et al, includes 50 images with different perspectives and carrying rates. It is a very crowded data set, with an average number of 1280 people, and the largest image contains 4543 people. Due to the limited number of scenes containing various crowd, it is a very challenging data set. Shanghai tech data set is a diverse and crowded data set, which was proposed by Zhang et al. The set consists of PartA and PartB. Part a collected 482 pictures from the Internet; Part B collected 716 pictures from the busy streets of Shanghai.

Table 1. Cross validation of UCF_ CC_ 50 data set

Test set number	MAE	MSE
1	381.69	581.23
2	143.58	182.37
3	291.39	342.63
4	257.12	339.71
5	167.41	189.76
Avg	248.24	327.14

We perform five times cross validation according to the accepted standard setting, and make the best use of the samples: divide the data set into five equal parts randomly, take four of them as the training set, and the remaining one as the test set, and conduct five times of training and testing. The results of the five experiments are shown in Table 1. Finally, the average value of error index is taken as the final result of the experiment.

Table 2. Estimation error of UCF_ CC_ 50 data set

Methods	MAE	MSE
MCNN[13]	378.21	521.13
CMTL[14]	334.76	401.97
Switch-CNN[15]	318.29	435.11
SaCNN[16]	310.53	424.74
CSRNet[17]	267.81	397.63
ours	248.24	327.14

We compare the results with the most advanced methods. The results of MAF and MSE are listed in Table 2. The estimation errors of Mae and MSE of our method are the smallest in all models, which shows that we have obtained the best estimation of the count of UCF_CC_50 data set, which is the best compared with other methods.

4. Conclusion

In this paper, a multi-scale hybrid attention feature fusion network model is proposed. Based on CSRNet, a multi-scale feature fusion structure is added and a hybrid attention mechanism is introduced. Among them, VGG-16 network structure, extended convolution, multi-scale structure and attention mechanism expand the diversity of scale perception and the acceptance range of features, enhance the ability of the model to suppress the background and retain details, and can solve the crowd counting problem in various complex scenes, which performs well in the method of calculating the number of people in the image.

References

[1] Peter J. Phillips,Gabriela Pohl. Crowd counting: a behavioural economics perspective[J]. Quality &amp; Quantity, 2021(prepublish).

[2] Chen Jingyu, Xiu Shengjie, Chen Xiang, Guo Hao, Xie Xiaohua. Flounder-Net: An efficient CNN for crowd counting by aerial photography[J]. Neurocomputing, 2021,420.

[3] Salma Kammoun Jarraya, 12, Maha Hamdan Alotibi, 13, Manar Salamah Ali. A Deep-CNN Crowd Counting Model for Enforcing Social Distancing during COVID19 Pandemic: Application to Saudi Arabia's Public Places[J]. 1 Department of Computer Science, FCIT, King Abdulaziz University, Jeddah,

Saudi Arabia; 2 MIRACL-Laboratory, Sfax University, Sfax, Tunisia; 3 Department of Computer Science, King Khalid University, Abha, Saudi Arabia; Corresponding Author: Salma Kammoun Jarraya., 2021, 66(2).

[4] Guo Qiang, Zeng Xin, Hu Shizhe, Phoummixay Sonephet, Ye Yangdong. Learning a deep network with cross-hierarchy aggregation for crowd counting[J]. Knowledge-Based Systems, 2021,213.

[5] Song Beibei, Sheng Rui, Jiang Yi-Zhang. Crowd Counting and Abnormal Behavior Detection via Multiscale GAN Network Combined with Deep Optical Flow[J]. Mathematical Problems in Engineering, 2020,2020.

[6] Yang Yifan, Li Guorong, Du Dawei, Huang Qingming, Sebe Nicu. Embedding Perspective Analysis into Multi-Column Convolutional Neural Network for Crowd Counting. [J]. IEEE transactions on image processing: a publication of the IEEE Signal Processing Society, 2020,PP.

[7] Deepesh Deshmukh, Chaitanya Deo, Ishwar Nigam, Mr. Sourabh Dave. A LOOK AT ADVANCES IN CNN-BASED CROWD COUNTING APPROACHES[J]. Ethics And Information Technology, 2020,2(2).

[8] Cao Zhijie, Shamsolmoali Pourya, Yang Jie. Synthetic guided domain adaptive and edge aware network for crowd counting[J]. Image and Vision Computing, 2020,104.

[9] Gao Junyu, Yuan Ieee Please Verify Yuan, Wang Qi. Feature-Aware Adaptation and Density Alignment for Crowd Counting in Video Surveillance.[J]. IEEE transactions on cybernetics, 2020,PP.

[10] Engineering; Investigators at Shanghai University Discuss Findings in Engineering (Mask Guided Gan for Density Estimation and Crowd Counting) [J]. Journal of Technology &amp; Science,2020.

[11] Biao Yang, Weiqin Zhan, Nan Wang, Xiaofeng Liu, Jidong Lv. Counting crowds using a scale-distribution-aware network and adaptive human-shaped kernel[J]. Neurocomputing,2020,390.

[12] Computers; Findings on Computers Reported by Investigators at Yanshan University (Crowd Counting Using a Self-attention Multi-scale Cascaded Network)[J]. Computer Weekly News, 2020.

[13] Engineering; Researchers from Civil Aviation University of China Report on Findings in Engineering (An Enhanced Scale Robust Network for Crowd Counting)[J]. Journal of Engineering,2020.

[14] Sorn Sooksatra, Toshiaki Kondo, Pished Bunnun, Atsuo Yoshitaka. Redesigned Skip-Network for Crowd Counting with Dilated Convolution and Backward Connection[J]. Journal of Imaging, 2020,6(5).

[15] Yan-Bo Liu, Rui-Sheng Jia, Qing-Ming Liu, Zhi-Feng Xu, Hong-Mei Sun. Crowd counting via an inverse attention residual network[J]. Journal of Electronic Imaging,2020,29(3).

[16] Chuanrui Hu, Kai Cheng, Yixiang Xie, Teng Li. Arbitrary perspective crowd counting via local to global algorithm[J]. Multimedia Tools and Applications,2020(prepublish).

[17] Computers; New Findings from Jiangnan University in the Area of Computers Described (Crowd Counting By the Dual-branch Scale-aware Network With Ranking Loss Constraints) [J]. Computer Weekly News, 2020.