Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting

Pongpisit Thanasutives[(1)], Ken-ichi Fukui[(2)], Masayuki Numao[(2)], Boonserm Kijsirikul[(3)]

(1) Graduate School of Information Science and Technology Osaka University
(2) Institute of Scientific and Industrial Research Osaka University
(3) Department of Computer Engineering Faculty of Engineering Chulalongkorn University

Abstract

- In this paper, we propose two modified neural networks based on dual path multi-scale fusion networks (SFANet) and SegNet for accurate and efficient crowd counting. Inspired by SFANet, the first model, which is named M-SFANet, is attached with atrous spatial pyramid pooling (ASPP) and context-aware module (CAN).
- The encoder of M-SFANet is enhanced with ASPP containing parallel atrous convolutional layers with different sampling rates and hence able to extract multi-scale features of the target object and incorporate larger context.
- To further deal with scale variation throughout an input image, we leverage the CAN module which adaptively encodes the scales of the contextual information. The combination yields an effective model for counting in both dense and sparse crowd scenes.
- The second model is called M-SegNet, which is produced by replacing the bilinear upsampling in SFANet with max unpooling that is used in SegNet. This change provides a faster model while providing competitive counting performance.

Introduction

Crowd counting task: To count a number of people in a given image for public safety, surveillance monitoring, etc.

Problem [2]: Heavy occlusion (noisy image), perspective distortion (different camera angles), scale variation (different sizes of head and surrounding context), etc. These problems are later solved by the multi-scale-aware modules and the dual-path decoder [3].

Data preprocessing [1]: To generate the density map ground truth $D(x)$, we follow the Gaussian method with a fixed standard deviation kernel. Assuming that there is a head annotation at pixel x_i represented as $\delta(x - x_i)$, the density map can be constructed by convolution with Gaussian kernel. Attention map is generated based on the threshold applied to the corresponding density map.

$$D(x) = \sum_{i=1}^{C} \delta(x - x_i) * G_\sigma(x)$$

$$A(\mathbf{i}) = \begin{cases} 0 & 0.001 > D(\mathbf{i}) \\ 1 & 0.001 \leq D(\mathbf{i}) \end{cases}$$

Goal: To tackle the scale variation caused by perspective distortion and improve the crowd counting accuracy using a combination of multi-scale-aware modules with encoder-decoder networks.

Methodology

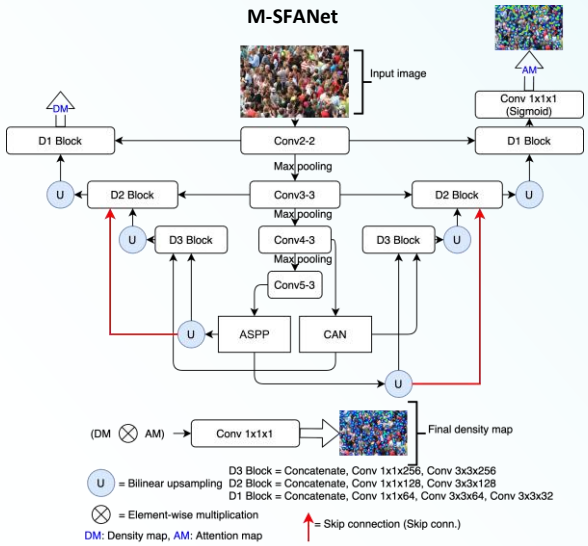

Figure 1: The architecture of M-SFANet. The convolutional layers' parameters are denoted as Conv (kernel size)-(kernel size)-(number of filters).

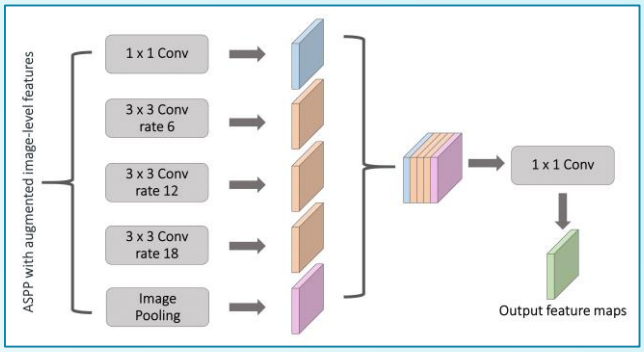

Figure 2: Atrous spatial pyramid pooling (ASPP) with augmented image-level features

- Atrous spatial pyramid pooling (ASPP) [4] module applies several effective fields-of-view of atrous convolution and image pooling to the incoming features, thus capturing multi-scale information. The atrous rates are 1, 6, 12, 18. Thanks to atrous convolution, loss of information related to object boundaries (between human heads and background) throughout convolutional layers in the encoder is alleviated.

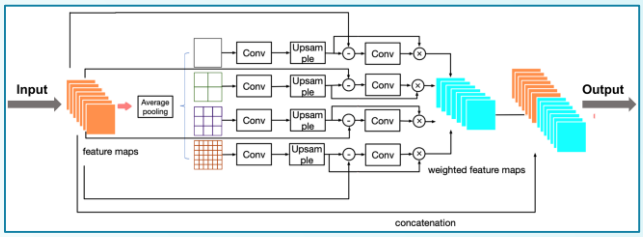

Figure 3: Context-aware module (CAN)

- Context-aware module (CAN) [5] module is capable of producing scale-aware contextual features using multiple receptive fields of average pooling operation. The pooling output scales are 1, 2, 3, 6. The module extracts those features and learns the importance of each such feature at every image location by measuring the difference from their neighbors, thus accounting for potentially rapid scale changes within an image..
- Dual-path decoder [3] for highlighting crowd regions in images before producing the final high-resolution density maps estimation.
- Design heuristics: (1) Apply ASPP to the most shrunken features in order to learn various scale features statically (with equal importance). Here, ASPP is more preferable than CAN, because, at the lowest resolution features, the weights of each extracted features are not diverse across image locations. (2) Apply CAN at the high-level feature maps, ⅛ in size of the original input, to adaptively encode the scales of the contextual information. By doing (1) and (2), we encode multi-scale features at different levels of learned semantic information, and hence; the scale variations problem is alleviated.

Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting

Pongpisit Thanasutives[(1)], Ken-ichi Fukui[(2)], Masayuki Numao[(2)], Boonserm Kijsirikul[(3)]

(1) Graduate School of Information Science and Technology Osaka University
(2) Institute of Scientific and Industrial Research Osaka University
(3) Department of Computer Engineering Faculty of Engineering Chulalongkorn University

M-SegNet

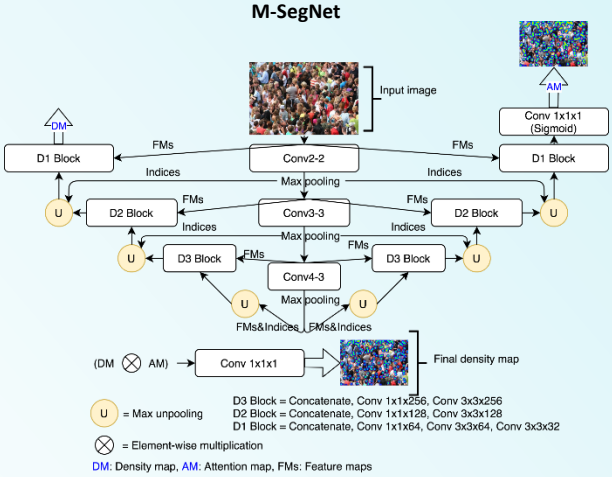

Figure 4: The architecture of the proposed M-SegNet.

- There are no CAN and ASPP to additionally emphasize multi-scale information and the bilinear upsampling is replaced with max unpooling operation using the memorized max-pooling indices [6].
- Less computational resources than M-SFANet with competitive performance. More suitable for speed-constrained applications.

Comparison with state-of-the-art methods on ShanghaiTech and UCF_CC_50 dataset

Method	Part A		Part B		UCF_CC_50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
CAN	62.3	100.0	7.8	12.2	212.2	**243.7**
SFANet	59.8	99.3	6.9	10.9	219.6	316.2
S-DCNet	58.3	95.0	6.7	10.7	204.2	301.3
SANet + SPANet	59.4	**92.5**	6.5	**9.9**	232.6	311.7
M-SegNet	60.55	100.80	6.80	10.41	188.40	262.21
M-SFANet	59.69	95.66	6.76	11.89	**162.33**	276.76
M-SFANet + M-SegNet	**57.55**	94.48	**6.32**	10.06	167.51	256.26

- We outperform previous methods in terms of MAE in ShanghaiTech and UCF_CC_50 datasets (More results on the paper).
- According to the ablation study (See Table II in the paper.) in ShanghaiTech, M-SFANet w/o CAN attain a higher MAE than M-SFANet w/o ASPP on SHA while the result is converse on SHB. This empirically shows that CAN and ASPP are effective for crowded scenes and sparse scenes respectively. Subsequently, integrating both modules successfully decreases MAE/RMSE on both SHA and SHB.
- M-SFANet and M-SegNet also outperform previous works on vehicle counting task, TRASCOS, indicating the generalizability of our approaches (See Table VII in the paper).

Results and Discussion

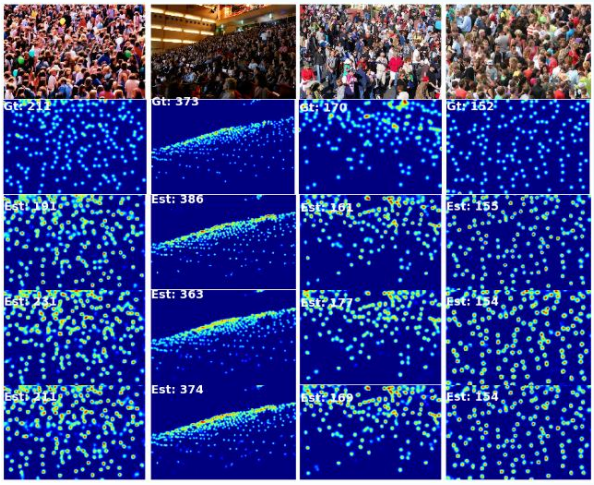

Figure 5: Visualization of estimated density maps. The first row is sample images from ShanghaiTech Part A. The second row is the ground truth. The 3rd to 5th rows correspond to the estimated density maps from M-SegNet, M-SFANet and M-SegNet+M-SFANet respectively.

Conclusion

- For M-SFANet, we add the multi-scale-aware modules to SFANet [3] architecture for better tackling drastic scale changes of target objects.
- Furthermore, the decoder structure of M-SFANet is adjusted to have more residual connections in order to ensure that the learned multi-scale features of high-level semantic information will impact how the model regress for the final density map.
- For M-SegNet, we change the up-sampling algorithm from bilinear to max unpooling using the memorized indices employed in SegNet. This yields the cheaper computation model while providing competitive counting performance applicable to real-world applications.

Selected references

- [1] Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 589-597).
- [2] Gao, G., Gao, J., Liu, Q., Wang, Q., & Wang, Y. (2020). CNN-based Density Estimation and Crowd Counting: A Survey. arXiv preprint arXiv:2003.12783.
- [3] Zhu, L., Zhao, Z., Lu, C., Lin, Y., Peng, Y., & Yao, T. (2019). Dual path multi-scale fusion networks with attention for crowd counting. arXiv preprint arXiv:1902.01115.
- [4] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).
- [5] Liu, W., Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5099-5108).
- [6] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2481-2495.