


RESEARCH

Open Access



Using transfer learning for smart building management system

Bens Pardamean^{1,5†}, Hery Harjono Muljo^{2,5†}, Tjeng Wawan Cenggoro^{3,5*†} , Bloomest Jansen Chandra^{4,5} and Reza Rahutomo⁵

*Correspondence:

wcenggoro@binus.edu

†Bens Pardamean, Hery

Harjono Muljo, and

Tjeng Wawan Cenggoro

contributed equally to this

work

³ Computer Science

Department, School

of Computer Science,

Bina Nusantara University,

Jakarta 11480, Indonesia

Full list of author information

is available at the end of the

article

Abstract

In building management, energy optimization is one of the main concern that needs to be automated. For automation, an intelligent system needs to be developed. However, an intelligent system needs to be trained in a large dataset before it can be used reliably. In this paper, we present a transfer learning scheme to develop an intelligent system for smart building management system. Specifically, the intelligent system is able to count human inside a room, which can be utilized to adaptively adjust energy usage in a room. The transfer learning scheme employs a deep learning model that is pretrained on ImageNet dataset. To enable the human counting capability, the model is trained on a dataset specifically collected for human counting case.

Keywords: Transfer learning, Deep learning, Human counting, Smart building, Building management system

Introduction

Currently, the concept of smart city is starting to be applied across the world [1–4]. Although the concept is typically implemented in a city-scale, it can also naturally be adapted in a more granular context such as in a building [5]. With this scale, the concept can be named as smart building. The implementation of this concept promises a more effective and efficient building management. Unfortunately, applying this concept requires a considerable amount of cost for procuring various type of Internet of Things (IoT) devices. Therefore, typical prototypes of smart building use only closed circuit television (CCTV) cameras as the IoT devices, which usually are already available in the building.

Using only CCTV poses a significant challenge for smart building. In case of energy management, the straightforward implementation of smart building is by using heat sensors to detect activity level in a room, which can be used to adjust the power usage of electric devices in the room. If the only available IoT devices are CCTV, a robust intelligent system with computer vision technology is needed.

To build such a robust computer vision system, a deep learning algorithm needs to be embedded within. Deep learning has been proved to have powerful performance in computer vision case such as image classification [6–11], object detection [12–15], and crowd counting [16–23]. Deep learning is also applicable for analysis of data from CCTV, which streams a big data that is difficult for other machine learning model to extract

valuable information from. However, deep learning requires a big dataset for a reliable performance. As the large dataset is not always available for every problem, training a deep learning model from scratch is considered to be impractical. To overcome the challenge, transfer learning has been broadly applied in many deep learning model developments (cite). This study introduces a transfer learning scheme that can be used to develop an intelligent system for smart building management. We focus on the development of intelligent system for counting human in a room, which can be employed for adjusting appliances for energy usage optimization. In addition, we also collected and shared a dataset that can be used in the proposed transfer learning scheme.

Literature study

The advancement of computer vision nowadays grows astonishingly fast. This growth was initiated by the use of deep learning in the ImageNet Large Scale Visual Recognition Challenge [24, 25]. At glance, it seems that the impressive performance of deep learning is the main cause of the huge growth in computer vision. However, it should be noted that the huge size of ImageNet dataset also contributes significantly to the deep learning performance. ImageNet has about 1.2 million of labeled images, which is currently one of the largest computer vision datasets. Only after it was trained on ImageNet that deep learning finally showed its extraordinary performance [6]. That particular deep learning model for computer vision, namely convolutional neural networks (CNN), was not a first choice for computer vision research since its invention in 1989 [26].

Unfortunately, a massive dataset such as ImageNet requires a laborious effort to be collected. As the consequence, it is impractical for many problems which has no large dataset available. To cope with the problem, recent research that utilize deep learning employs a concept called as transfer learning. This concept is defined as using a model that was previously trained on data from a task as a base to develop new model for other task. By using transfer learning, it is possible to use a deep learning model that has been pretrained on large dataset to learn from relatively smaller dataset. The use of this concept in deep learning was first initiated by Girshick et al. [27] to transfer utilize a CNN model pretrained on ImageNet to develop a model for object detection problem. In the following year, Yosinski et al. [28] exhaustively studied and proved the benefit of transfer learning for deep learning model. Since then, it is a standard to use an ImageNet-pretrained model in many computer vision problems. Even after the development of large dataset for object detection [29], the use of transfer learning is still widely adopted for the problem.

The benefit of transfer learning is mostly apparent in crowd counting, one of the most extensively studied computer vision problem. The most popularly used dataset in crowd counting, ShanghaiTech dataset [30], consists of only 1198 images. The other popular dataset, WorldExpo'10 [31], contains only 3980 images. The smallest dataset for crowd counting, UCF_CC_50 [32], even contains only 50 images. Despite that, the performance of crowd counting models are consistently growing fast since the use of deep learning in 2013. The fast advancement is possible by the extensive use of transfer learning. Consequently, the state-of-the-art crowd counting models within the last 6 years were always a variant of deep learning. Following this trend, Wang et al. even developed a large simulated dataset for pretraining purpose in crowd counting [33]. The dataset, named as

GTA Crowd Counting (GCC), was generated by using Grand Theft Auto (GTA) V game to obtain 15,212 synthetic crowd images.

Transfer learning scheme for intelligent human counting system

For a comprehensive understanding, we depict the whole intelligent system framework in Fig. 1. The proposed transfer learning scheme is part of the framework which is highlighted in green. The transfer learning scheme starts by acquiring a deep learning model that has been pretrained on ImageNet dataset [24, 25]. To convert the pretrained model to an intelligent human counting system, the model needs to be trained with a dataset crafted for human counting task. Therefore, we collected the required dataset, which we call as RHC (Room Human Counting) dataset. After the training, the trained intelligent human counting system is ready to process video streams from a CCTV to output the human count. It is worth noting that the CCTV stream injects a massive data to the intelligent

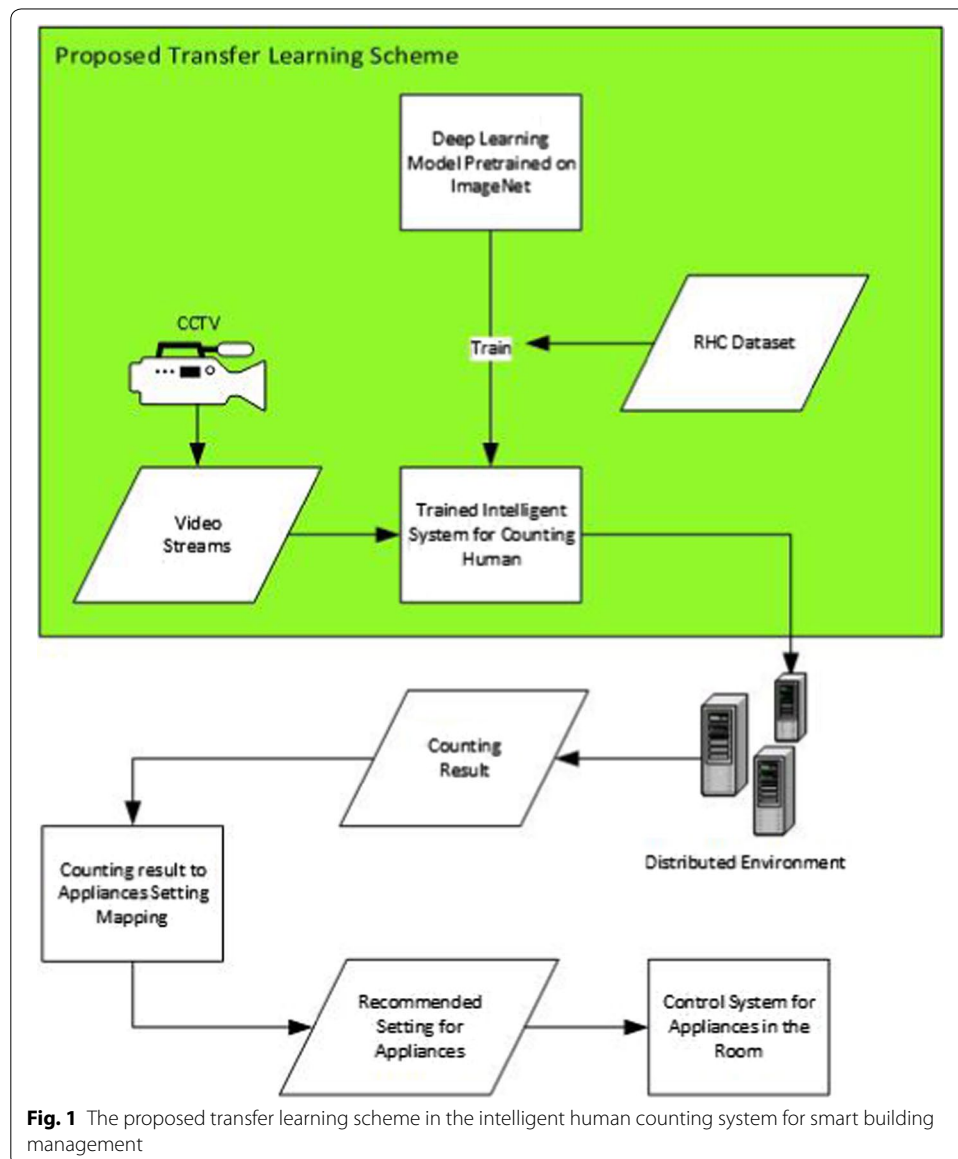


Fig. 1 The proposed transfer learning scheme in the intelligent human counting system for smart building management

system. For the system to run in real-time, it needs to be implemented using proper Big Data technology. Therefore, the intelligent system should be developed using deep learning libraries that can be implemented on apache spark. Based on the recent survey [34], Tensorflow [35] or Caffe [36] are excellent options as both libraries are supported by most deep learning frameworks for apache spark. Afterward, the predicted human count from the system is mapped to appliances adjustment setting in a control system.

Dataset collection

The images of RHC dataset were extracted from the videos captured by a CCTV in NVIDIA-BINUS AI R&D Center room. The dataset is collected only for one room to introduce a challenge for the future AI model to learn from one room only. This is necessary for developing a system that can adapt to different specification of CCTV in different room. If the model is able to robustly learn from this dataset, then it can be easily retrained using videos with different resolution from different room as long as the resolution of the new dataset is homogeneous.

In this dataset, the videos have a resolution of 640×360 pixels with a frame rate of 20 frames/s. There are 44 videos used for this dataset. The total duration of all videos is 206 h 24 min and 23 s. Figure 2 shows sample of images from the dataset.

Dataset annotation

Annotating a huge amount of data manually requires laborious work, thus it usually is infeasible. One solution that can be used to annotate a massive dataset is by developing an information system specially crafted for annotation task [37]. Therefore, we built an information system to ease the annotation process. This system takes videos from the previous acquisition process and displays them for the annotation process. The detailed explanation of this annotation system is described by Pardamean et al. [38]. In this system, the annotator decided which frame to be annotated from all videos, resulting 1217 annotated images.



Fig. 2 Sample images in RHC dataset

The dataset is annotated with the total count of human per image. We do not use the location of each human as annotation like what is typically done in crowd counting research. Training a deep learning model with the location introduces unnecessary complexity as the location information is not needed for controlling appliances usage in a room. The capability of localizing human in the model also reduces the speed of the system, which is vital for a real-time CCTV stream processing.

Dataset statistics

The human counts in RHC dataset are ranged from 0 to 13 with distributions as shown in Fig. 3. The mean human count in this dataset is 4.1249 with a standard deviation of 2.6206. We can see that the distribution is not uniform. Thus, this dataset can be considered as imbalance, which typically needs special treatment for any machine learning models to learn well from the dataset.

For a typical training procedure of machine learning, we split the dataset into three different sets: training, validation, and test set. The splitting process was done randomly with stratification to the human count. The split ratio between training, validation, and test set is 60:20:20. After the splitting process, we got a dataset with distribution as shown in Table 1.

To understand whether the current size of RHC dataset is enough for transfer learning, we compared the size with public datasets crowd counting. The crowd counting datasets is the most similar dataset to our case, which are also used for counting human. However, crowd counting differs from our case that the images contains huge number of human in outdoor setting. The dataset in crowd counting is typically much smaller than other popular computer vision cases such as image classification and object detection. Consequently, research in crowd counting usually utilize transfer learning. Therefore, the crowd counting datasets are suitable for comparison to RHC dataset. Table 2 lists popular crowd counting datasets as well as RHC dataset together with their size. We omitted GCC dataset in the list since it is a synthetic dataset and typically used only for the pretraining phase of transfer learning scheme. From the comparison, we can infer

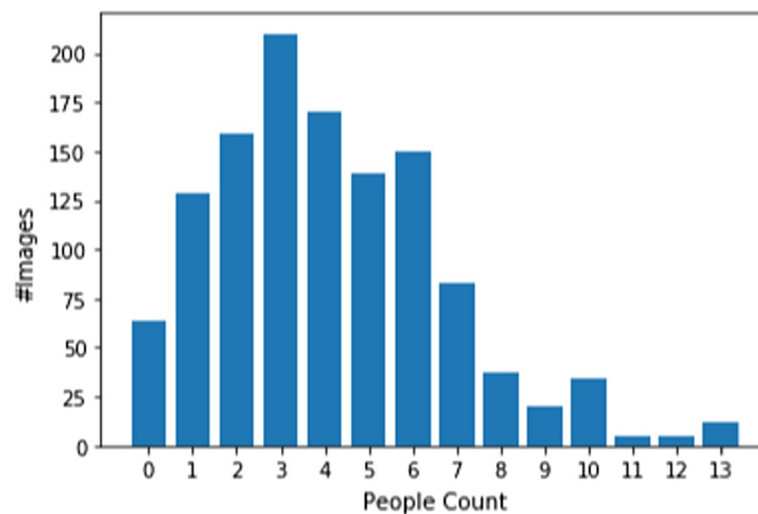


Fig. 3 Data count distribution

Table 1 Data distribution

Human count	Training data	Validation data	Test data	Total
0	38	13	13	64
1	78	25	26	129
2	95	32	32	159
3	127	41	42	210
4	102	34	34	170
5	83	28	28	139
6	90	31	29	150
7	50	16	17	83
8	22	7	8	37
9	12	4	4	20
10	21	7	6	34
11	3	1	1	5
12	3	1	1	5
13	7	3	2	12
Total	731	243	243	1217

Table 2 Comparison of datasets size

Dataset	Number of images
UCF_CC_50	50
ShanghaiTech Part A	482
ShanghaiTech Part B	716
UCF_QNRF	1535
WorldExpo'10	3980
RHC (our dataset)	1217

that RHC dataset size should be enough for deep learning. The size of RHC dataset is the third biggest dataset among the popular crowd counting datasets.

Possible challenges

We identified six possible challenges to be solved for a successful model training on RHC dataset. The first challenge is whether the trained model can count persons whose hair is covered. We see this as a challenge since most of the persons in this dataset let their hair uncovered. The second challenge is whether the model can successfully count human with overlapping heads. This challenge is common in crowd counting as the number of human captured in the images is massive. We see that a small portion in the dataset has overlapping heads, mostly for images with a large actual count.

The third challenge is introducing the trained model to exclude human outside of the room when predicting the count. The room in this dataset has a transparent glass wall on the left side, which outside can be clearly seen. Therefore, to produce a correct count prediction, the model needs to be able to exclude the persons outside of the room. The glass wall also causes the fourth challenge. When the outside of the room is darker, it turns into a mirror that reflects the persons inside the room. The model should be able to differentiate between the actual persons and their reflected figure. The fifth challenge

is related to the lighting of the room. Part of the room sometime can be darker if there is a presentation session in the room. Therefore, the model should be robust against a different light setting of the room.

The last possible challenge we identified corresponds to the distribution of this dataset. As given in Fig. 3, this dataset is not balanced to all possible count. The larger the difference between labeled counts to its mean, the smaller the number of images they have. This condition generally leads to poor performance for the labels with fewer images. This problem is called imbalanced data problem and is known to cause diminishing performance for machine learning models as well as deep learning models [39–41]. In counting case, one of the possible solutions to this problem is to create a model that is capable to extrapolate its count prediction to count labels with fewer data.

Experimental

We conducted an experiment to measure the performance of developed intelligent human counting system. In the experiment, we consider five popular CNN models as the pretrained model: AlexNet [6], VGGNet [7], GoogLeNet [8], ResNet [9], and DenseNet [11]. To enable all models to learn from RHC dataset, we changed the prediction layers with a fully connected layer consisting of one neuron. The layer outputs a single number as a predicted human count. Because the input image size of these networks is 224×224 , we resized the images in the dataset to the size before feeding them to the networks. All models are trained using Adam optimization algorithm [42] with learning rate 0.001. The performance of each model is measured using Mean Squared Errors (MSE) of the difference between predicted count and actual count.

Results and discussions

Quantitative analysis

Table 3 lists all models MSE for the test split of RHC dataset. The best MSE is achieved by AlexNet, which has the smallest number of layers. We can see a trend that the more layers the model has, the MSE is declining. We suspect that this is caused by overfitting that is suffered by the more complex models.

To check our assumption of overfitting, we tabulate the MSE for each actual count in Table 4. We also plot the MSE in Fig. 4. We can see that the complex models tend to perform worse in the actual count with less training data. Thus, we can confirm that the poor performance from the complex models is caused by overfitting.

Table 3 Model performance on test split

Model	#Layers	Test MSE
AlexNet	8	0.6240
VGG16	16	1.3762
GoogLeNet	22	3.0069
ResNet18	18	2.2546
ResNet50	50	2.1044
ResNet101	101	2.0629
ResNet152	152	1.8185
DenseNet121	121	2.0378

Table 4 Test MSE of all models for each actual count

Act. Cnt.	#Train data	AlexNet	VGG16	Goog LeNet	ResNet				Dense Net121
					18	50	101	152	
0	38	0.898	0.094	7.151	6.038	5.039	6.157	5.200	5.776
1	78	0.384	0.642	4.136	3.774	3.288	3.037	2.147	2.220
2	95	0.311	0.903	1.048	1.071	0.836	0.833	0.680	0.527
3	127	0.476	1.115	0.563	0.794	1.072	0.544	1.020	0.720
4	102	0.479	1.533	0.510	0.810	0.806	1.138	0.999	0.686
5	83	0.684	1.589	0.865	0.949	1.249	1.371	1.430	1.364
6	90	0.417	1.025	1.915	1.100	0.895	0.838	0.951	1.561
7	50	1.008	2.141	1.832	1.254	1.076	0.845	0.737	1.103
8	22	0.975	2.970	4.657	1.987	1.367	2.512	0.669	2.239
9	12	4.457	1.143	10.545	6.572	5.168	5.326	4.988	4.070
10	21	0.530	2.589	15.581	7.764	9.067	6.615	6.066	6.546
11	3	2.014	7.326	33.227	14.997	15.220	18.457	15.053	18.188
12	3	0.517	0.917	34.812	29.127	28.506	28.229	20.813	31.236
13	7	1.750	12.371	51.475	31.875	26.061	24.514	20.675	32.732

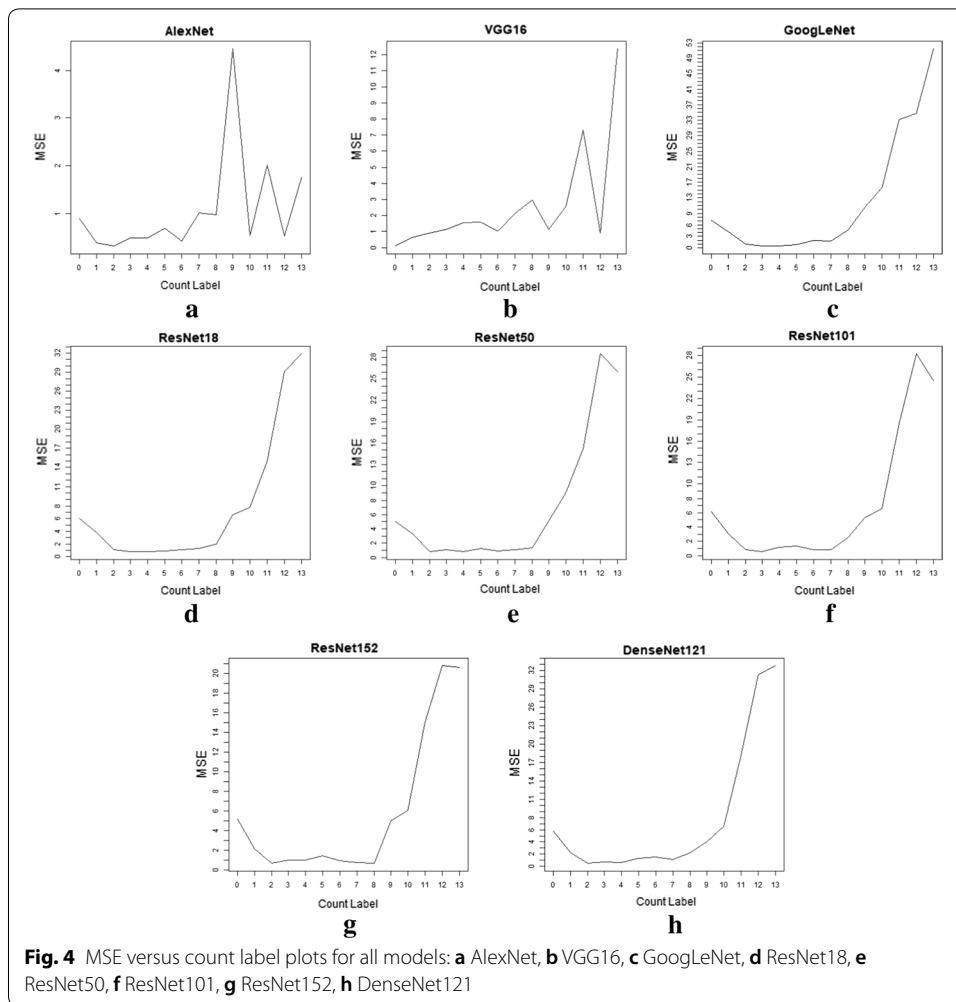
Qualitative analysis

We picked up several cases that correspond to the challenges we addressed before. These cases are tabulated in Table 5. In image (a) in the table, we see a person with a veil. Because in most training data the hair of each person is seen, we suspect that the models might be unable to count a person whose hair is covered. However, that seems to be not the case as the models count for 5.77 persons in average for this image, with the actual count is 5 persons. The problem instead is the failure of the model to exclude a person that is actually outside the room. The average count prediction is approaching 6 which indicates that the models tend to count an excessive person, which is likely the person near the most left person in the room. This failure is supported with other picture with a similar case as depicted in (b).

Although the models seem to be unable to exclude humans outside the room, it is not the case if there are more than one person outside. As seen in (c), the models do not suffer over-counting problem caused by outside persons. This fact is proved by similar predictions by all models in image (d) which outside room is relatively clear.

In fact, all models instead suffer under-counting in predicting image (c) and (d). The average count prediction is 7.23 persons compared to 11 persons in the actual count. This under-counting might be caused by several persons with an overlapping head as seen in image (c). This problem is also the possible cause of the poor performances of all models in Table 4 for large count number. However, this under-counting does not appear in images with fewer human such as image (e). In this image, there are 2 persons with overlapping head. The average count in this image is 2.83 persons, approaching the actual count of 3 persons. Therefore, the models are able to predict this case without notable problem.

In addition to image with large human count, we also checked the opposite extreme, which are images with a small human count. The average prediction of image (f), which contains only 1 person, is 2.84. This indicates that the models are











over-counting. However, in this case, it seems that the over-counting is not caused by the persons outside the room, as there are more than 2 persons clearly seen outside. Thus, we expect the over-counting probably caused by imbalance data instead.

Trained with RHC dataset, all models seem to have a robust performance against different lighting. For instance, image (g) is slightly darker than most of images in RHC dataset. However, the performances of all models are still reliable, with a slight under-counting that might be caused by overlapping instead. The models are also robust against the case where the outside room is dark, which makes the glass that separates inside and outside reflective. an example of this case is provided in image (h). It can be seen that all models do not suffer over-counting caused by the reflected figure of the persons in the room.

Conclusion and future works

In this paper, we showed that transfer learning can be used to develop an intelligent human counting system, which can be utilized for energy optimization in smart building management. To enable the development, RHC dataset is collected to train a pretrained deep learning model for counting human in a room. The result of this study shows that

Table 5 Cases with possible challenge

Image:		
	(a)	(b)
Actual Count	5.0000	5.0000
Average Pred.	5.7785	5.6457
AlexNet	4.4196	4.8554
VGG16	1.8978	5.3963
GoogLeNet	6.8223	5.9367
ResNet18	5.7509	5.3723
ResNet50	6.3831	6.2260
ResNet101	7.2255	6.4082
ResNet152	7.5966	5.5926
DenseNet121	6.1321	5.3777
Image:		
	(c)	(d)
Actual Count	11.0000	11.0000
Average Pred.	7.2369	7.5115
AlexNet	10.6564	9.5808
VGG16	8.2521	8.2934
GoogLeNet	5.6034	5.2358
ResNet18	6.9386	7.1274
ResNet50	7.3383	7.0987
ResNet101	6.9407	6.7039
ResNet152	8.0313	7.1202
DenseNet121	6.3313	6.7353
Image:		
	(e)	(f)
Actual Count	3.0000	1.0000
Average Pred.	2.8381	1.4057
AlexNet	2.3192	0.1637
VGG16	3.1656	0.0000
GoogLeNet	2.8151	2.0548
ResNet18	3.0930	2.0310
ResNet50	3.0904	2.2029
ResNet101	2.8374	1.5425
ResNet152	2.3662	1.2929
DenseNet121	3.0177	1.9574
Image:		
	(g)	(h)
Actual Count	8.0000	8.0000
Average Pred.	6.9386	6.4698
AlexNet	8.2365	5.7812
VGG16	6.6404	5.9270
GoogLeNet	6.0342	5.6323
ResNet18	6.5031	5.8712
ResNet50	7.5616	7.4109
ResNet101	6.1297	6.5895
ResNet152	7.1560	7.0117
DenseNet121	7.2469	7.5343

AlexNet is the best model for the pretrained model in the proposed transfer learning scheme. However, the size of this dataset seems insufficient to train more complex networks than AlexNet. This indicates that the dataset should be appended with more data in the future. Additionally, it is interesting to extend this dataset with additional annotations for the coordinate of each human. We believe that this additional annotation can help a complex model to improve its performance.

Abbreviations

AI: artificial intelligence; CCTV: closed circuit television; CNN: convolutional neural networks; GCC: GTA Crowd Counting; GTA: Grand Theft Auto; IoT: Internet of Things; MS COCO: microsoft common object in context; RHC: room human counting.

Acknowledgements

The raw videos was captured using CCTV in NVIDIA-BINUS AI R&D Center room. The experiments was run using NVIDIA Tesla P100 and P4 from NVIDIA-BINUS AI R&D Center.

Authors' contributions

BP, HHM, TWC designed the study; TWC and RR built the annotation system; TWC and BJC processed the data; BJC ran experiments; BP, HHM, and TWC analysed the results; BP and HHM supervised the study; BP, HHM, TWC, BJC, and RR wrote the paper. All authors read and approved the final manuscript.

Funding

This study is funded by Directorate of Research and Community Service, Directorate General of Research and Development, Indonesian Ministry of Research, Technology and Higher Education (Grant No. 23/AKM/MONOPNT/2019) as a part of 2019 Penelitian Terapan Unggulan Perguruan Tinggi Research Grant.

Availability of data and materials

The RHC dataset is available at <http://bdsrsrc.binus.ac.id/~wawan/rhc/>.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Computer Science Department, BINUS Graduate Program-Master of Computer Science Program, Bina Nusantara University, Jakarta 11480, Indonesia. ² Accounting Information Systems Program, Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta 11480, Indonesia. ³ Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia. ⁴ Computer Science Department, College of Letters and Science, University of Wisconsin, Madison, WI 53706, USA. ⁵ Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia.

Received: 18 June 2019 Accepted: 18 November 2019

Published online: 07 December 2019

References

1. Hao L, Lei X, Yan Z, ChunLi Y. The application and implementation research of smart city in china. In: 2012 international conference on system science and engineering (ICSSE). New York: IEEE; 2012. p. 288–92.
2. Dameri RP. Searching for smart city definition: a comprehensive proposal. *Int J Comput Technol*. 2013;11(5):2544–51.
3. Van den Bergh J, Viaene S. Unveiling smart city implementation challenges: the case of Ghent. *Inf Polity*. 2016;21(1):5–19.
4. Muchtar K, Rahman F, Cenggoro TW, Budiarto A, Pardamean B. An improved version of texture-based foreground segmentation: block-based adaptive segmenter. *Procedia Comput Sci*. 2018;135(September):579–86.
5. Minoli D, Sohraby K, Occhiogrosso B. Iot considerations, requirements, and architectures for smart buildings-energy optimization and next-generation building management systems. *IEEE Internet Things J*. 2017;4(1):269–83.
6. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*. 2012. p. 1097–105.
7. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556).
8. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. p. 1–9.
9. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770–8.
10. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-first AAAI conference on artificial intelligence*. 2017.
11. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 4700–8.

12. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2980–8.
13. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–49.
14. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). 2016. p. 779–88.
15. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE international conference on computer vision (ICCV). 2017. p. 2980–8.
16. Li Y, Zhang X, Chen D. Csrnet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). 2018.
17. Liu L, Wang H, Li G, Ouyang W, Lin L. Crowd counting using deep recurrent spatial-aware network. In: Proceedings of international joint conferences on artificial intelligence organization (IJCAI). 2018.
18. Shi Z, Zhang L, Liu Y, Cao X, Ye Y, Cheng M-M, Zheng G. Crowd counting with deep negative correlation learning. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), 2018. p. 5382–90.
19. Cenggoro TW, Aslamiah AH, Yunanto A. Feature pyramid networks for crowd counting. In: To appear: 2019 international conference of computer science and computational intelligence. 2019.
20. Chen X, Bin Y, Sang N, Gao C. Scale pyramid network for crowd counting. In: 2019 IEEE winter conference on applications of computer vision (WACV). New York: IEEE; 2019. p. 1941–50.
21. Liu N, Long Y, Zou C, Niu Q, Pan L, Wu H. Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). 2019.
22. Liu W, Salzmann M, Fua P. Context-aware crowd counting. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). 2019.
23. Shi M, Yang Z, Xu C, Chen Q. Revisiting perspective information for efficient crowd counting. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). 2019.
24. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. New York: IEEE; 2009. p. 248–55.
25. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211–52.
26. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten zip code recognition. *Neural Comput*. 1989;1(4):541–51.
27. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 580–7.
28. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in neural information processing systems. 2014. p. 3320–8.
29. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: common objects in context. In: European conference on computer vision. Berlin: Springer; 2014. p. 740–55.
30. Zhang Y, Zhou D, Chen S, Gao S, Ma Y. Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 589–97.
31. Zhang C, Li H, Wang X, Yang X. Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 833–41.
32. Idrees H, Saleemi I, Seibert C, Shah M. Multi-source multi-scale counting in extremely dense crowd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2013. p. 2547–54.
33. Wang Q, Gao J, Lin W, Yuan Y. Learning from synthetic data for crowd counting in the wild. 2019. arXiv preprint [arXiv:1903.03303](https://arxiv.org/abs/1903.03303).
34. Johnsirani Venkatesan N, Nam C, Shin DR. Deep learning frameworks on apache spark: a review. *IETE Tech Rev*. 2019;36(2):164–77.
35. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. Tensorflow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016. p. 265–83.
36. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia. New York: ACM; 2014. p. 675–8.
37. Cenggoro TW, Tanzil F, Aslamiah AH, Karupiah EK, Pardamean B. Crowdsourcing annotation system of object counting dataset for deep learning algorithm. In: IOP conference series: earth and environmental science, vol. 195. Bristol: IOP Publishing. 2018. p. 012063.
38. Pardamean B, Cenggoro TW, Chandra BJ. Rahutomo: a user interface for rapid data annotation of room activity level detection system. In: To appear: 2019 international conference on eco engineering development (ICEED). Bristol: IOP Publishing; 2019.
39. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. 2019;6(1):27.
40. Cenggoro TW, Isa SM, Kusuma GP, Pardamean B. Classification of imbalanced land-use/land-cover data using variational semi-supervised learning. In: 2017 international conference on innovative and creative information technology (ICITech). New York: IEEE; 2017. p. 1–6.
41. Cenggoro TW. Deep learning for imbalance data classification using class expert generative adversarial network. *Procedia Comput Sci*. 2018;135:60–7.
42. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: The international conference on learning representations 2015, San Diego, CA. 2015.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.