Advanced Engineering Technology and Application

ISSN: 2090-9543 VOLUME 7, ISSUES 3, 2018 www.egyptfuture.org/ojs/

Deep Convolutional Neural Network based Person Detection and People Counting System

Maksat Kanatov[1,*] *and Lyazzat Atymtayeva*[2]

[1] Kazakh-National Research Technical University after K.I.Satbayev, Almaty, Kazakhstan
[2] Suleyman Demirel University, Kaskelen, Kazakhstan

Received: 3 Jul. 2018, Revised: 12 Aug. 2018, Accepted: 27 Aug. 2018
Published online: 1 Sep. 2018

Abstract: Nowadays, computer vision is an actively developing and one of the most important part of Artificial Intelligence. There are a lot of works in this area. Computer Vision has a wide range of uses. For example, in medicine it can be used for diagnosing a Magnetic Resonance Imaging or X-ray image, in security systems - for detecting intruders, for driverless cars or robotics to navigate in space, etc. However, often image recognition works together with object detection. Usually required object for recognition takes only the small part of original image whereas the rest part of the image does not carry useful information for recognizing. By this reason to optimize computation time, we need to find this object, before recognizing. There are various methods and technologies for object detection in the image. This work demonstrates one of the actual method in computer vision which named Deep Convolutional Neural Networks and shows the advantages of this method in person detection and people counting.

Keywords: object detection, neural networks, deep cnn, machine learning, computer vision

1 Introduction

Computer Vision began to develop in parallel with Artificial Intelligence. In the begin of this developing there were hand-craft methods [1], [2], [3], [4], [5]. However, the best works are done in last 20 years. One of the best methods is Viola-Jones method [6], which successully used for Face detection. They proposed model for object detection, based on the rectangle filters which can be evaluated extremely rapidly at any scale [10]. The examples can be shown in the Figure 1. In this algorithm, firstly the sum of all regions are calculated (in the light and dark regions). After that, the difference between the two sums is calculated. In the situation like Figure 1C, dark region multiply to 2, to make it equal to light region. This method had very good accuracy. In 2001 for face detection system, accuracy was about 80% with false detection less than 2%, and about 95% with false detection about 33%. Four years later, Viola and Jones used this algorithm for pedestrian detection [10]. For static objects it had a lot of false detection, as it is shown in Figure 2. However, authors used this method only for moving objects. The results were significantly improved. Figure 3 shows the results of Viola-Jones' method for dynamic pedestrian detection. For that time, it was a perfect result. This method was used in several good works [7], [8], [9].Similar options like methods based on Local Binary Patterns [13], [12] and Histograms of Oriented Gradients [11] were appeared and

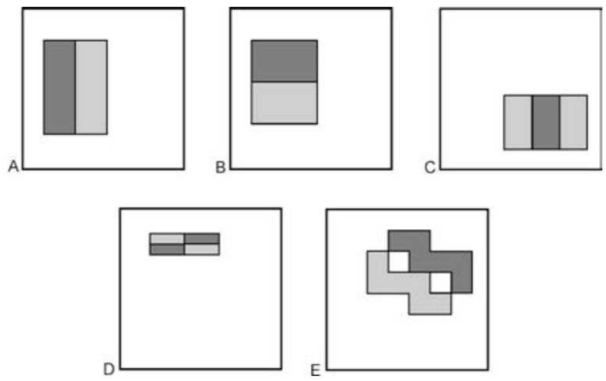

Fig. 1: The rectangle filters used in the method of Viola-Jones [10].

successfully used for object detection. However, in the last ten years Deep Convolutional Neural Networks make very good results in computer vision. In this article we describe Deep Convolutional Neural Networks (CNN) for object detection tasks with implementation phases and results. Our work is based on architecture of Deep CNN proposed by Joseph Redmon and Anelia Angelova [16]. We used this model for person detection task. Also, in this article we describe the dataset which used for training and testing, and environments used in this work.

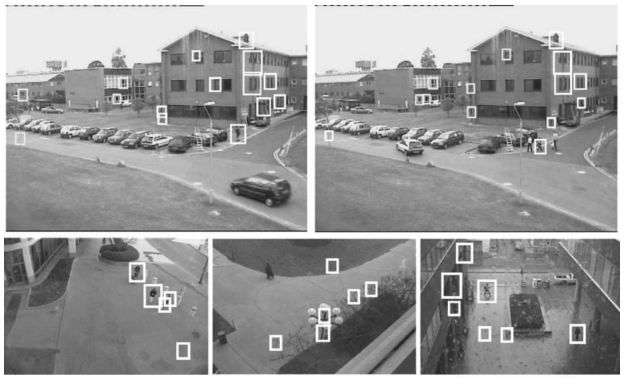

Fig. 2: Pedestrian detection for static object.

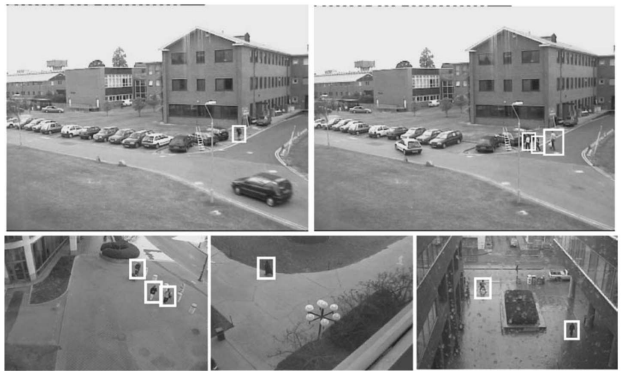

Fig. 3: Pedestrian detection for dynamic object.

2 Deep ConvNets

The concept of artificial neural networks appeared more than fifty years ago. However, a great interest in artificial neural networks began in the late 80's. The active use of neural networks for image recognition started at the beginning of the 21st century. Currently, neural networks occupy a large share in the field of image recognition. Convolutional neural networks (CNN) are a special kind of artificial neural networks, which was proposed by Yann Lecun [14]. This artificial neural network model is a part of deep learning [17]. First great job was done by Alex Krizhevsky [18]. He won ILSVRC-2012 competition using CNNs for image recognition task using ImageNet dataset [19] with a large margin of results. This caused great interest in CNN, and now, CNN is the most popular and actively used in image recognition tasks. Visual

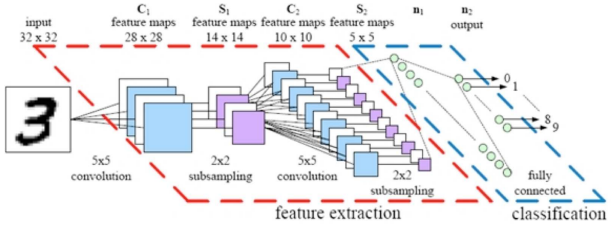

Fig. 4: The standard architecture of CNNs for image recognition task.

object recognition task in computer vision is to splite into two steps [20] feature extraction and classification. Any architecture of Artificial Neural Networks [21] is prepared for special features, which are extracted from the image. Extraction can be done by some methods and algorithms, or can consists of preprocessing procedures. Also, vectorized pixels of the image can be used as a features for CNNs. Important, that features must be extracted the same for training and recognizing. Generally, CNNs are based on Artificial Neural Networks. However, CNNs have a structure, which is similar to the visual perception of the human brain (Figure 4). The recognition procedure starts from feature extraction from the visual objects. Then, data goes to fully connected Artificial Neural Networks for classification. Usually, CNNs have a different layers [22], as a convolution layers, subsampling layers and neuron layers. In the convolutional layer, using the feature maps (Figure 4), data is extracted from the some local receptive field. Subsampling layer is needed to reduce data at the output of convolutional layer. Neuron layer is a output of all CNNs architecture(n_1 and n_2 in the Figure 4). It is like classical neuron model. In this case, neuron layer works for classification task. The convolutional and subsumpling layers make feature extraction tasks. Our architecture based on Deep CNNs, proposed by J.Redmon et al. [16] and it is called You Only Look Once (YOLO). In the previous work, authors trained 224 x 224 network classifier and increased to 448 for detection [15]. In this new architecture, they used fine tuned classification networks at the full 448 x 448 resolution. It is used for 10 epochs on ImageNet architecture [23]. Network model is became faster and works better with large scale images. Also, this model can predicts coordinates of the

recognized object using fully connected layers on top of the convolutional feature extractor. On the other hand, this architecture works very fast. For example, popular VGG-16 [24] is very powerful and has a very good accuracy in classification. However, computational cost of VGG-16 is very high. That is why, YOLO is based on faster Googlenet architecture [25]. This model 3 times faster in the same accuracy, comparing VGG-16. That is why, we chose YOLO architecture for training and testing our visual object detection system.

3 Dataset

Dataset is the most important part of any Deep Neural Networks. Because, deep nets require a lot of training samples for good recognition, and data must be labeled very accurately. We tested our system for person detection task. It means, we needed a lot of labeled images with the persons. For this task, PASCAL-Visual Object Classes 2012 (PASCAL-VOC 2012) dataset [26] was the most optimal choice. This dataset consists of 4 087 RGB images with person from the different situation, and 10 129 labels of person. It means, that for each images authors of this dataset prepared txt file with coordinates of persons in the images. Some txt files can have more than one labels. It depends how many persons in the image. Figure 5shows some examples of PASCAL-VOC dataset. As you can see, images of person are very different. Training on this database makes Deep CNNs model very universal, rather than Viola-Jones' Pedestrian detection method [10]. Because, Viola-Jones' method detects only full body of the person, when Deep CNNs can detect person from the any angle, even if not the full part of the person is on the image.

Fig. 5: Samples from the PASCAL-VOC 2012 dataset.

4 Training and testing

We divided dataset into 3 groups: train (80%), validation (10%) and test (10%) data. Training was done on two NVIDIA Quadro 2000 GPU. It tooks 54 hours. To exclude overfitting, we must take the model with minimum of validation error, as it is shown in the Figure 6. We registered average loss and saved network model every 1000 iteration. On the 2200th iteration we had got the minimum average loss (0.043) and after this iteration error began to grow. That is why we stopped training, and tested this model on test set. We have got accuracy about 91%, and with 1% false positive and 8% false negative on training set. In the next step, we used this network model

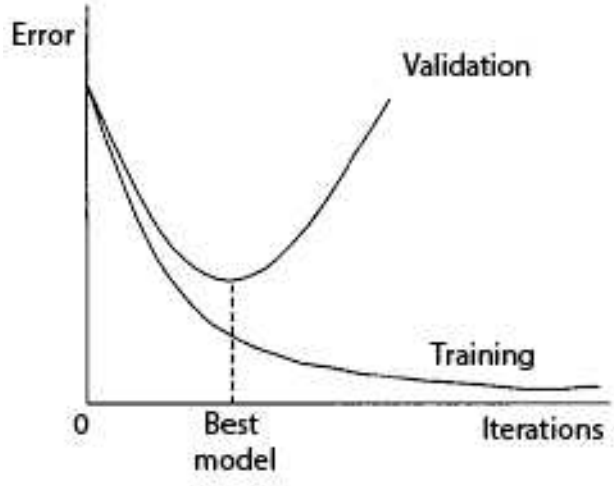

Fig. 6: A graph for determining the best iteration for a network model.

for real time person detection. In this case we have an almost perfect results. We used RTSP from RGB IP camera with 1280 x 720 pixels resolution. We recorded 1 hour of video (12 fps HD) with the drawing rectangle around detected person, where 158 people passed through the corridor of our university during this time. Figure 7 illustrates images from this experiment. The results were amazing. All these persons were detected. However, we noticed that in some sequences of the movement of the persons, in the some frames there were false negatives. In total were 2780 false negative from 43200 frames (12fps x 3600seconds). That is less than 10%. Later, we tested this model on spontaneous images (Figure 8). Results also were perfect. Network model detects person very well. It detects the person, even if only head is presented in the image. In some cases, reflection and shadow can be detected as a person(¿0.3 threshold). For this experiments we used OpenCV [27], Tensorflow [28], CUDA, cuDNN [29] and other libraries. For programming it was used Python v3.6. On our PC HP Z240 (CPU Xeon E3-1230, 32GB RAM, GPU NVIDIA Quadro M200), we were able to get a Real-Time Visual Objec Detection System with 12FPS using HD camera with resolution 120x720 of RGB frame. Running without GPU on this PC gives about 2 FPS.

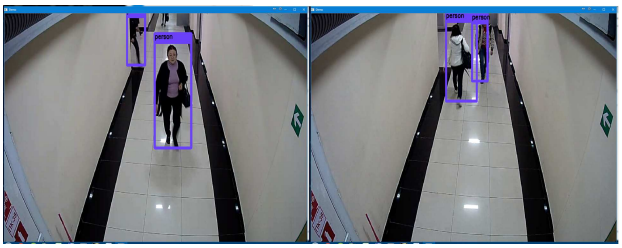

Fig. 7: Real time person detection in the Satpayev University using Deep CNN.

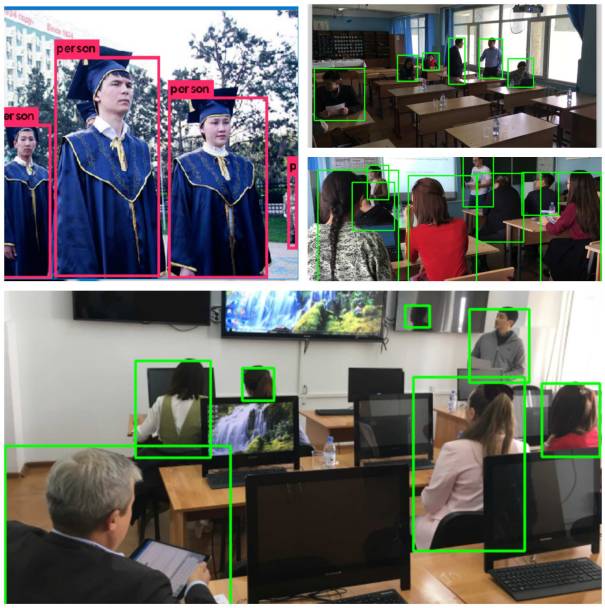

Fig. 8: Person detection using spontaneous images.

5 Conclusion

This work shows that Deep Convolutional Neural Networks are very productive in Visual Object Detection. Some researchers say, that Deep NNs need so many computation and this is not reasonable. However, we live in the epoch, where in the 10 year the most powerful computers, which needed server rack turned into pocket computers as smart-phones. In our opinion, it is a normal to use powerful machines to take the results as demonstrated in this article. Because, Deep Nets are very muck superior to all other method, like a Viola-Jones'. To draw the conclusion, on can say that Deep CNN the best way to solve the problems of Visual Object Recognition. Further, we plan to develop and test a system of counting people, relying on this system.

6 Acknowledgement

This research was supported by grant of the program of Ministry of Education of the Republic of Kazakhstan BR05236699 Development of a digital adaptive educational environment using Big Data analytics. We thank our colleagues from Suleyman Demirel University(Kazakhstan) who provided insight and expertise that greatly assisted the research. We express our hopes that they will agree with the conclusions and findings of this paper.

References

[1] Papageorgiou, C. P., Oren, M., and Poggio, T. (1998, January). A general framework for object detection. In Computer vision, 1998. sixth international conference on (pp. 555-562). IEEE.

[2] Jain, A. K., Ratha, N. K., and Lakshmanan, S. (1997). Object detection using Gabor filters. Pattern recognition, 30(2), 295-309.

[3] Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., and Poggio, T. (1997, June). Pedestrian detection using wavelet templates. In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on (pp. 193-199). IEEE.

[4] Yakimovsky, Y. (1976). Boundary and object detection in real world images. Journal of the ACM (JACM), 23(4), 599-618.

[5] Gennery, D. B. (1979, August). Object detection and measurement using stereo vision. In Proceedings of the 6th international joint conference on Artificial intelligence-Volume 1 (pp. 320-327). Morgan Kaufmann Publishers Inc..

[6] Viola, P., and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-I). IEEE.

[7] Lienhart, R., and Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In Image Processing. 2002. Proceedings. 2002 International Conference on (Vol. 1, pp. I-I). IEEE.

[8] Fergus, R., Perona, P., and Zisserman, A. (2003, June). Object class recognition by unsupervised scale-invariant learning. In Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on (Vol. 2, pp. II-II). IEEE.

[9] Viola, P., Jones, M. J., and Snow, D. (2003, October). Detecting pedestrians using patterns of motion and appearance. In null (p. 734). IEEE.

[10] Viola, P., Jones, M. J., and Snow, D. (2003, October). Detecting pedestrians using patterns of motion and appearance. In null (p. 734). IEEE.

[11] Dalal, N., and Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[12] Pietikinen, M. (2005, June). Image analysis with local binary patterns. In Scandinavian Conference on Image Analysis (pp. 115-118). Springer, Berlin, Heidelberg.

[13] Jin, H., Liu, Q., Lu, H., and Tong, X. (2004, December). Face detection using improved LBP under Bayesian framework. In Image and Graphics (ICIG'04), Third International Conference on (pp. 306-309). IEEE.
[14] LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995.
[15] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
[16] Redmon, J., and Farhadi, A. (2017). YOLO9000: better, faster, stronger. arXiv preprint.
[17] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. nature, 521(7553), 436.
[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009
[20] Nixon, M., Aguado, A.S.: Feature Extraction and Image Processing, Second Edition. Academic Press, 2nd edn. (2008)
[21] Haykin, S.: Neural Networks and Learning Machines. Prentice Hall, 3 edn. (2008)
[22] Peemen, M., Mesman, B., and Corporaal, H. (2011, August). Efficiency optimization of trainable feature extractors for a consumer platform. In International Conference on Advanced Concepts for Intelligent Vision Systems (pp. 293-304). Springer, Berlin, Heidelberg.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 2015. 2
[24] K. Simonyan and A. Zisserman., Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
[25] C. Szegedy, S. Ioffe, and V. Vanhoucke., Inception-v4, inception-resnet and the impact of residual connections on learning. CoRR, abs/1602.07261, 2016.
[26] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. International journal of computer vision, 111(1), 98-136.
[27] Bradski, G., and Kaehler, A. (2000). OpenCV. Dr. Dobbs journal of software tools, 3.
[28] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., and Kudlur, M. (2016, November). TensorFlow: A System for Large-Scale Machine Learning. In OSDI (Vol. 16, pp. 265-283).
[29] Sanders, J., and Kandrot, E. (2010). CUDA by example: an introduction to general-purpose GPU programming. Addison-Wesley Professional.

Maksat Kanatov master of Computer science, researcher at Satbayev University in Almaty, Kazakhstan. His research interest includes Deep Convolutional and Recurrent Neural Networks, Automated Facial Expression Recognition systems using Deep CNNs.

Lyazzat Atymtayeva received the Ph.D and Doctor of Science degree in Mechanics, Mathematics and Computer Science at Suleyman Demirel University, Kazakhstan. Her research interests are in the areas of mechanics, applied mathematics and computer science including the numerical and rigorous mathematical methods and models for mechanical engineering and computer science, intelligent and expert systems in Information Security, Project Management and HCI. She has published research papers in reputed international journals of mathematical and computer sciences. She is reviewer and editor of international journals in mathematics and information sciences.