Journal of KIIT. Vol. 18, No. 1, pp. 27-34, Jan. 31, 2020. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2020.18.1.27

Object Localization and Detection Using SALNet with Deformable Convolutional Network

Md Foysal Haque*, Dae-Seong Kang**

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (NO.2017R1D1A1B04030870).

Abstract

Convolution filters inherently characterized the inputs by the aggregation classification. The filters regulate simple topological composition to reconstruct the input as same as the real structure. Convolutional neural networks (CNNs) use the invariable filters arrays to classify the objects. In this paper, a unique convolutional architecture introduced to classify the source by offset learning. The network uses spatial augmentation localization to improve the object localization performance. The spatial augmentation localization network (SALNet) follows the concept of offset learning. It adopts the offset learning to classify the pixel information for feature extraction and localization. Moreover, the framework addresses the feature area by using the offset classification that calculates the pixel values of the nearest neighbor for reducing the cost computation. The process uses improved deformable convolutional architecture to augment the test sets for the precise detection task.

요 약

컨볼루션 필터는 기본적으로 집합 분류에 따라서 입력을 특징화한다. 이런 필터는 실제 구조와 동일하게 입력을 재구성하기 위해 단순한 토폴로지 구조를 조절한다. CNN(Convolutional Neural Networks)은 객체를 분류하기 위해 변하지 않는 필터 배열들을 사용한다. 본 논문에서는 오프셋 학습을 통한 객체를 분류하는 특유의 컨볼루션 필터 구조를 제안한다. 정확한 객체 분류를 위한 객체 국소화 성능을 향상하기 위해서 네트워크는 공간적 증가 국소화를 사용한다. 이런 국소화 절차는 오프셋 학습의 개념을 따르고 변하지 않는 필터 배열이 아닌 개선된 변형 가능한 컨볼루션 구조를 사용하여 테스트 데이터 셋을 증가시킵니다. 특징 추출 및 국소화를 위해 픽셀 정보를 분류하는 오프셋 학습을 채택한다. 또한, 프레임워크는 비용 계산을 줄이기 위해 가장 가까운 픽셀 값을 계산하는 오프셋 분류를 사용하여 객체 영역을 처리한다. 증가된 테스트 데이터 셋을 통해 더 정확한 검출 작업이 가능하다.

Keywords
convolutional neural network, object detection, deformable convolutional network, SALNet

* Dept. of Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0634-9783
** Professor, Dept. of Electronic Engineering, Dong-A University
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Dec. 27, 2019, Revised: Jan. 22, 2020, Accepted: Jan. 25, 2020
· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University, 37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea,
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. Introduction

Improving the detection accuracy of an object detector is the most challenging work in the computer vision task. Due to increasing interest in the visual recognition task recently, number of robust detection frameworks introduced with significant results. Image classification [1], object detection [2], and semantic segmentation [3] achieved significant accuracy. Detection task faces different challenges such as noisy background and intra-class variations in low brightness. These challenges can be defeated by offset learning and augmentation classification of source data. Most of the robust network uses improvable learning strategy and activation methods to train the network. Furthermore, even deploying improvable learning methods into the base network the high cost calculation lacking the network to achieve significant classification accuracy. The methods also use the invariant feature transformation method to classify feature information. Support vector machine (SVM) [4] and SIFT (scale-invariant feature transformation) [5] uses the invariant feature transformation classification to learn about the object and localize the objects in the source image. These networks environment are only capable of working on known and fixed work sets. These networks modeled for fixed geometric transformation tasks and limited data augmentation. Moreover, hand-crafted convolutional models performance is not satisfactory for complex convolutional transformation.

Convolutional neural networks inherently introduced different optimization and generalization methods to improve the classification performance. It deploys different convolutional models for geometric transformation to classify objects from test sets. Different pooling layers used to reduce the spatial resolution at a different fixed ratio. These methods stimulate the network to classifies the objects by various geometric transformations. Object detection method achieved significant progress in the current vision task [6]-[9], these approaches based on the enhanced bounding box method.

In this paper, we proposed an enhanced convolutional architecture to improve detection performance. It adopts the deformable convolutional layers, and deformable pooling layers insist with the base convolutional neural network to detect multiple objects. The improved deformable convolutional architecture modeled with the geometric transformation to classifies the offset of the source image. The convolutional layers address the offset areas by seeking each grid of the sampling object. Offset values calculated from the feature maps of additional convolutional layers.

Ⅱ. Related Algorithm

Deep convolutional neural network [1][8] models exhibit significant performance to handle complex tasks in current object detection. It achieved better performance than shallow models to solve complex vision tasks. The achievement was possible because of robust learning capability using simplistic feature extraction layers. Deep convolutional neural network widely used to classify objects, object detection, and object recognition tasks.

2.1 Spatial Augmentation Localization

Spatial augmentation localization network (SALNet) [10] examines the source image to discover and address the location of each single object from the test image. Localizing notation of objects from the different typical areas of the source image is one of the challenging tasks for object classification. In terms of searching for the exact location of objects, the process depends on exploiting execution of objects spatial shape. The proposed concept inspired by the geometric transformation. The spatial augmentation

localization is an initial process of locating the primary object location of a single object. In this process network firstly executes pixel classification for search a particular object area. Then the classifier executes the object area by the pixel segmentation process. After pixel segmentation, the network executes the pixel classification to find the object address with classifying the nearest neighbor pixels of the selected area. Nearest neighbor classification conduct by following eq. (1) and eq. (2).

$$f(x) = f(x_k) \tag{1}$$

$$f(x_k) = \frac{x_{k-1} + x_k}{2} < x \leq \frac{x_k + x_{k+1}}{2} \tag{2}$$

where, convolutional filters extracts pixel values of each points of the input image. The process of spatial augmentation localization shown in Fig. 1.

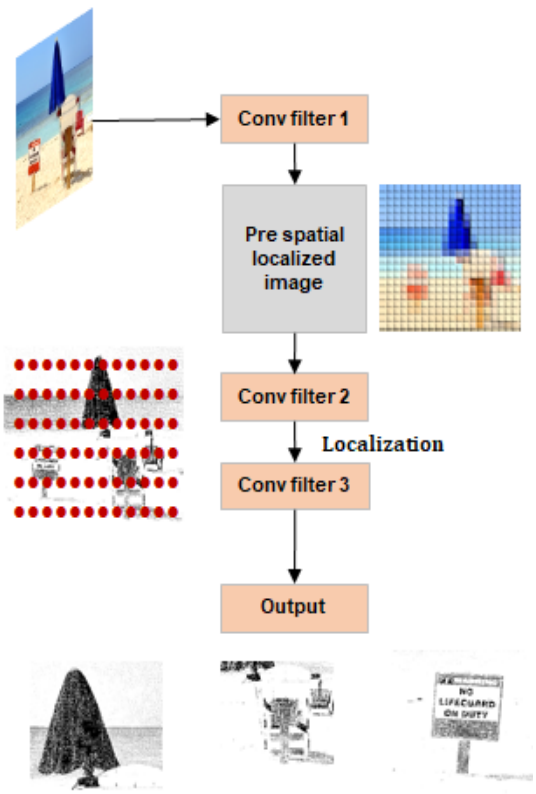


Fig. 1. Architecture of (SALNet) spatial augmentation localization

2.2 Geometric Transformation

Geometric transformation redefines the spatial relationship between points of in an image [11]. The operation manipulated an image by observing and varying the spatial layout. The method executed successful vision task on image classification, medical imaging, and remote sensing. The concept of geometric transformation represents the image by transforming mapping from one coordinate to another coordinate. The transformation operation varies depending on the application. It uses different forms to map the image. The transformation operation uses projective transformation, affine transformation, and polynomial transformation to map the image. In eq. (3) illustrate the process of geometric transformation.

$$z(x) = \sum_{k=0}^{K-1} c_k h(x - x_k) \tag{3}$$

where, h is kernel weight, c_k is coefficients of kernel, K is the data samples, and $x - x_k$ is feature mapping area.

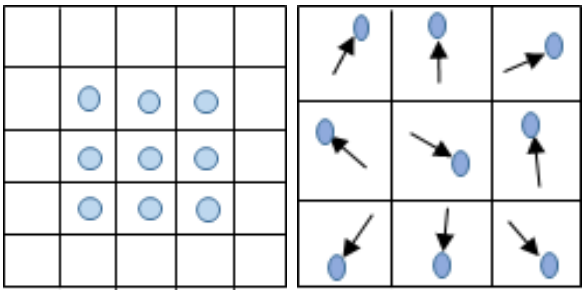
2.3 Deformable Convolution

Deformable convolution neural network [12] presents a unique convolutional approach that classifies image data by varying viewpoints, scales, and poses. It consists of two different convolution processes first one uses deformable convolution to address the object boundaries with grid sampling, and the second one uses standard convolutional layers to survey offset learning with respect to the mapped feature data. The offset learns from the bin position using deformable pooling. Combining these modules with a neural network, it analyzes the feature information to configure the specific object address by pooling and sampling patterns. Localizing and detecting objects deformable pooling layers play an important role in detection task Deformable pooling layers are similar to

average-pooling and max-pooling. The convolutional layers extract C part feature maps for detection.

$$b_c^{(x,y)} = \left\{ m_c^{z_{\delta_x,\delta_y}} - \sum_{n=1}^{N} a_{c,n} d_{c,n}^{\delta_x,\delta_y} \right\} \quad (4)$$

In eq. (4), the output of the deformable pooling layer denoted as b_c where x and y element of the output. The output map size denoted as m_c, and it confirms the visual scores for placing the c^{th} map.



(a) (b)

Fig. 2. Representation of (a) normal convolution and (b) offset representation by deformable convolution

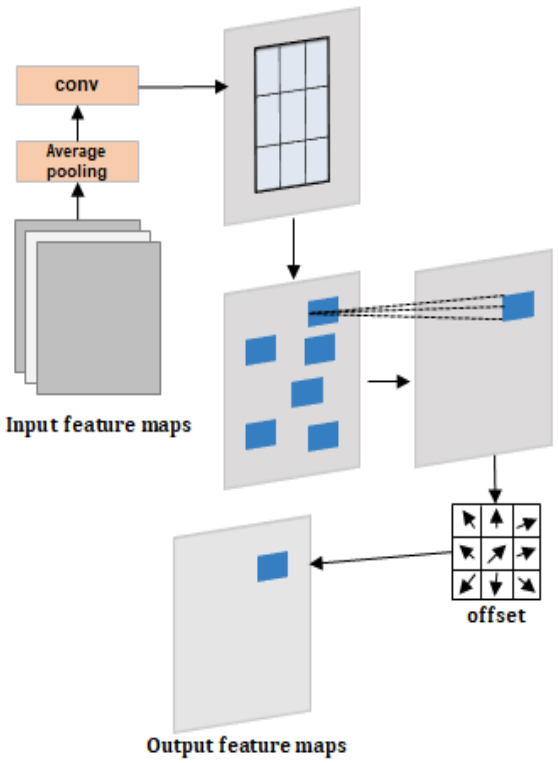


Fig. 3. Proposed architecture of 3×3 deformable convolutional neural network

In this paper, we proposed an improved deformable convolutional network that adopts the deformation strategy to classifies the object. The architecture of the proposed deformable convolution shown in Fig. 3. The deformable convolutions uses 3×3 kernels to classifies and conforms the object location with offset mapping.

$$y(P_0) = \sum_{p_n \epsilon R} w(p_n) \cdot x(p_0 + p_n) \quad (5)$$

In eq. (5), projected location P_0 carries the offset on the output feature map, R is regular grid of dilation and receptive field size. The proposed model consists of different convolutional layers including average-pooling and fully-connected layers.

III. Proposed Algorithm

Perfect organized convolutional blocks extract high-quality feature maps that carry the essential information of the objects. The main goal of robust image classifier is to train the network with high-quality feature data with less computation. Deformable convolution layers use the strategy of structure-aware convolution. The network learns the structure of the object, observing the local structure representation and geometric transformation.

In this paper, an improved deformable convolutional neural network proposed that deploys spatial augmentation localization and deformable layers classify the objects. Feature maps processes through the deformable convolution classified the feature data, and address the object location part by part. These training data help the network to localize and aligns different part of the object for obtaining enhance detection. Deformable convolutional neural network environment designed to manipulate both 2D and 3D feature data. The overview of the proposed method shown in Fig. 4. The proposed network consists of three conventional convolutional layers (conv1, conv3, and conv4), one pre-localization module (conv2), and

three deformable convolutional with deformable pooling layers. The proposed network exploited an AlexNet [7] to extract and learn the deep feature. Spatial augmentation localization network (conv2) introduced in the proposed network to improve the network localization. It pre-localized the feature data that reduce the network convolutional computation.

Moreover, the framework examines the whole image and addresses the object location. After addressing the object, the network extracts the localized feature maps. These feature data forwarded towards the next convolutional for further classification process. The convolutional layer conv3 and conv4 extracts more essential feature information to apply in deformable convolution task.

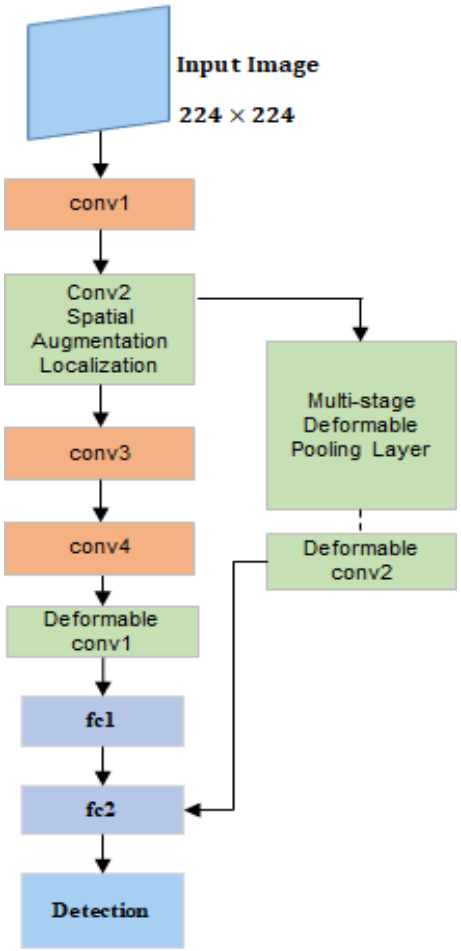


Fig. 4. Architecture of proposed SALNet with deformable convolutional network

These features pass to the improved deformable convolutional network to classify feature data to represent each of the class by using root filter. The root filter classifies the feature maps and assigned anchor boxes to point each part of an object. The anchor box moves around to address each part of the object. Furthermore, the network included regularization layers to justify and improve the abnormal displacement of the anchors. Final scores of the addressed part optimized by deformation convolution to minimize the deformation cost and localize accurate object areas.

However, to collect more deformation feature data to enhance the training process the network introduced one extra deformable convolutional network added with multi-stage deformable pooling. This layers connected with the initial layers of the base network. The multi-stage deformable pooling layers examines the initial feature data for addressing the object location with different points of view. The addressed feature maps preserve in separately in fully-connected layers. Finally, the detection block optimized all data and predicted the final detection from the confident score maps.

IV. Experiments

To evaluate the performance of proposed network, we analyzed the localization proposals. To improve the localization performance of the proposed network SALNet (spatial augmentation localization network) was added with the deformable convolutional network for object detection. The deformable convolutional network utilized 3×3 convolutional filters to proposed architecture. The network train with PASCAL VOC (VOC 2007 and 2012) [13] dataset and all test experiments conducts with the same dataset. In addition to evaluating the test proposal mAP (mean average precision) [14] sets to 0.5 as a standard, and the evaluation also examines on mAP at 0.7 & 0.9. The experimental parameters are shown in Table 1.

Table 1. Experimental parameters

Experimental parameters	Set value
Learning rate	0.001
Batch size	32
Weight decay	0.0005
Input size	224×224

V. Results

The network performed increasing number of iteration to achieve the desired detection accuracy. The spatial augmentation localization extracts pre-localized feature with less computation. The detection results are shown in Fig. 5. The proposed network achieved significant detection accuracy with less computation due to the pre-localization feature data. The test results compared with the recent novel methods reported in Table 2, 3, and the proposed network shows significant performance to locate the object parts. Also the network shows improved performance to localize objects than other robust detection methods. To evaluate network performance, the mean average precision (mAP) sets to three thresholds such as 0.5, 0.7, and 0.9. In 0.5, the proposed network achieved 81.3 percent detection accuracy that is higher than other image classifiers.

Table 2. Detection results on MS COCO test set

Model	mAP		
	0.5	0.7	0.9
SALNet with deformable convolutional network	**81.3**	**44.7**	**43.3**
Deformable convolutional net [12]	69.7	42.6	27.1
R-FCN [15]	75.8	38.3	39.8

Table 3. Analyzing Multi-scale detection results between the proposed method and robust detection frameworks test

Method	Setting	AP_S^{bbox}	AP_M^{bbox}	AP_L^{bbox}
Proposed network	multi-scale test	**28.6**	**51.2**	**61.7**
Deformable convolutional net [12]		24.6	44.3	52.7
R-FCN [15]		21.4	43.6	48.9
Faster R-CNN [16]		19.1	42.2	48.7
SSD [17]		20.6	41.7	49.8

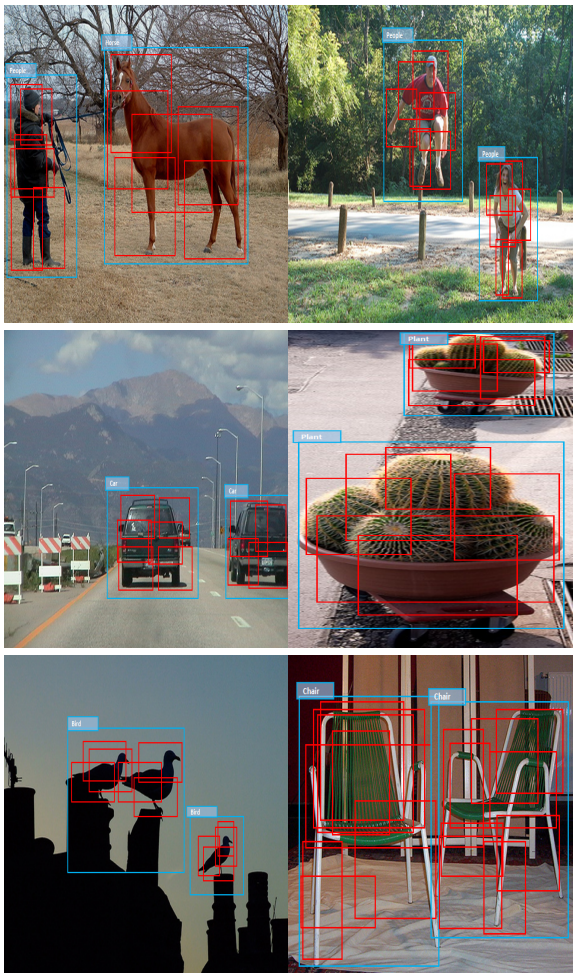


Fig. 5. Detection results of proposed network. The deformable parts are shown in red and region proposals are shown in blue

VI. Conclusions

The proposed model introduced a localization network that classifies and extracts localized informative feature data to improve the detection performance. It generated the pre-localized feature data for classification through a deep and deformable based convolutional network. The network focused on classifies feature data by discriminative element. It learns and represents feature data with a part-based representation through an effective way without any complex annotation in the training process. For this reason, the network achieved significant gains on the detection task. The proposed network achieved 81.3 percent detection accuracy that relatively notable than the conventional object detection algorithms.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", The IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, pp. 770-778, Jun. 2016.

[2] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition., Kauai, HI, USA, pp. 511-518, Dec. 2001.

[3] J. Logan, E. Shelhamer, T. Darrell, and U. Berkeley, "Fully convolutional networks for semantic segmentation", Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, pp. 3431-3440, Dec. 2015.

[4] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers", Neural Processing Letters, Vol. 9, No. 3, pp. 293-300, Jun. 1999.

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, Nov. 2004.

[6] N. Dlal and B. Triggs, "Histograms of oriented gradients for human detection", 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, Vol. 1, pp. 886-893, Jun. 2005.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems 25, pp. 1097-1105, 2012.

[8] K. Simonayan and A. Zisserman, "Very deep convolutional networks for large-Scale image recognition", arXiv preprint arXiv:1409.1556, Apr. 2015.

[9] R. Girshick, "Fast R-CNN", The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440-1448, Apr. 2015.

[10] M. F. Haque, H. Y. Lim, and D. S. Kang, "Spatial augmentation localization based deformable convolutional neural networks for object detection", The Proceedings of KIIT DCS Summer Conference, Vol. 14, No. 1, pp. 66-69, Jun. 2019.

[11] J. L. Mundy and A. Zisserman, "Geometric invariance in computer vision", Cambridge, MA: MIT Press, Vol. 92, May 1992.

[12] W. Ouyang, et al., "DeepID_Net: multi-stage and deformable deep convolutional neural networks for object detection", In Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403-2412, Sep. 2014.

[13] M. Everingham, et al., "The pascal visual object classes challenge: a retrospective", International Journal of Computer Vision, Vol. 111, No. 1, pp. 98-136, Jan. 2015.

[14] S. Nowzin, "Optimal decisions from probabilistic model: The intersection-over-union case", In Proceedings of The IEEE International Conference on Computer Vision and Pattern Recognition, pp. 548-555, Jun. 2014.

[15] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks", Advances in Neural Information Processing Systems 29 (NIPS 2016), pp. 379-387, May 2016.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", Advances in Neural Information Processing Systems 28 (NIPS 2015), pp. 91-99, Jun. 2015.

[17] W. Liu, et al., "SSD: Single shot MultiBox detector", European Conference on Computer Vision 2016, pp. 21-37. Sep. 2016.

Authors

Md Foysal Haque

2015 : B.Sc. in Electrical & Electronic Engineering, University of Information Technology & Sciences(UITS).
2018 ~ 2020 : M.Sc. in Electronic Engineering, Dong-A University.
Research Interests : Digital Image Processing, Computer Vision, and Pattern Recognition

Dae-Seong Kang

1994 : Ph.D. in Electrical Engineering, Texas A&M University.
1995 ~ Present : Professor at Dong-A University.
Research Interests : Image Processing and Compression