Stair-step Feature Pyramid Networks for Object Detection

Xuan-Thuy Vo, Tien-Dat Tran, Duy-Linh Nguyen and Kang-Hyun Jo*

School of Electrical Engineering, University of Ulsan, Ulsan, Korea
{xthuy,tdat}@islab.ulsan.ac.kr; ndlinh301@mail.ulsan.ac.kr;
acejo@ulsan.ac.kr

Abstract. Feature Pyramid Networks have solved scale variation problems in object detection by developing multi-level features with different scales from backbone networks. Although this network achieved promising performance without affecting model complexity, they still suffer feature-level imbalance between multi-level features, i.e., low-level features and high-level features in each stage of the backbone. Moreover, the detection head predicts classification scores and offset regression independently on each feature of multi-level features, which leads to inconsistency among the detection branch. Hence, this paper releases this problem by introducing simple but effective Stair-step Feature Pyramid Networks (SFPN) to harmonize information between multi-level features. Further, the Offset Adaption Module (OA Module) is proposed to improve feature representation by adapting the feature of the classification branch with regressed offsets of the regression branch. On the MS-COCO dataset, the proposed method increases by 1.2% Average Precision when comparing with baseline FCOS [15] without bells and whistles.

Keywords: Stair-step FPN · Offset Adaption · Object detection

1 Introduction

Object detection is one of the challenging tasks in computer vision research. This problem is decomposed into two tasks: classification task and regression task. The classification task classifies each object belonging to a specific class. The regression task identifies where each object locates. Two types of detectors are one-stage and two-stage detector, which are based on the number of networks on each detection head. The two-stage pipeline first obtains a set of region proposals and then the second stage classifies each proposal to a specific class and regresses the coordinates of each proposal by learning offsets. The one-stage object detection directly places dense anchor-boxes on each location and performs classification and regression tasks on each anchor-boxes. Although one-stage detection achieves high efficiency, the accuracy of it is far lower than two-stage detection. The structure of detection architecture includes a backbone

* Corresponding author

network to extract feature of images, neck network which connects the backbone and head part to create multi-level features and detection head to predict classification scores and offsets.

One of the most popular necks is Feature Pyramid Networks - FPN [9] selecting multi-level features from backbone network by utilizing top-down pathway and lateral connection to gather neighborhood features. Even though FPN has figured out the scale imbalance problem in which the size of objects changes in large ranges, this method still suffers feature-level imbalance among multi-level features. Since the detection head predicts scores, offsets independently on each feature, it leads to inconsistency between each head branch. To overcome this problem, EFPN [16] introduced a feature aggregation module and refinement module to obtain a uniform feature for all feature pyramid. Inspired by [16], this paper proposes Stair-step Feature Pyramid Networks (SFPN) which employs bi-linear interpolation and summation operation on adjacent features to form a uniform feature, i.e., the top-down pathway. Generally speaking, the multi-level features are converted to a single feature that contains information of all features. Without losing the novelty of FPN, SFPN also creates multi-level features by down-sampling a uniform feature, i.e., bottom-top pathway. Finally, the residual connection between the top-down pathway and down-top pathway is applied to ease hard optimization.

The detection head consists of the classification and regression branch. These two branches are trained independently to predict the class probability and offset values. Hence they ignore the correlation of class prediction and bounding box prediction. To improve feature representation, the Offset Adaption Module (OA Module) is performed by adding four offset values to the rectangular grid sampling locations in 3×3 convolution.

2 Related Work

Feature Pyramid Networks - FPN [9] presented multi-level features to solve scale variation problems in object detection. EFPN [16] enhanced multi-scale features by introducing feature aggregation module and refinement module to improve feature representation.

The popular two-stage object detection is Faster R-CNN [13] which achieves great performance. Inspired by [13], many methods are proposed such as Libra R-CNN [12], TridentNet [8], Mask R-CNN [4]. The one-stage object detection brings a trade-off between accuracy and speed. One of the most popular one-stage detection is RetinaNet [10] which densely places anchor-boxes on each location. After that, the network classifies each anchor-box and predict four offset value, e.g., coordinates of object center, width, and height. Recently, the anchor-free method - FCOS [15] balances both accuracy and speed, which avoids drawbacks of the anchor-based method. The proposed method considers FCOS as a baseline.

Many methods aims to improve object detection accuracy by inserting attention module to backbone network, such as GCNet [1], Non-local Network [18], BNLNet [17]. Different from this strategy, this paper proposes the Offset Adaption Module which correlates class prediction with bounding box prediction. This module is light-weight but boosting the accuracy of detectors.

3 The Proposed Method

The overall architecture is shown in figure 1. The backbone network extracts features from the image. Five feature maps with different scales from each stage of the backbone are selected as the input of Stair-step FPN. This architecture will describe in section 3.1.

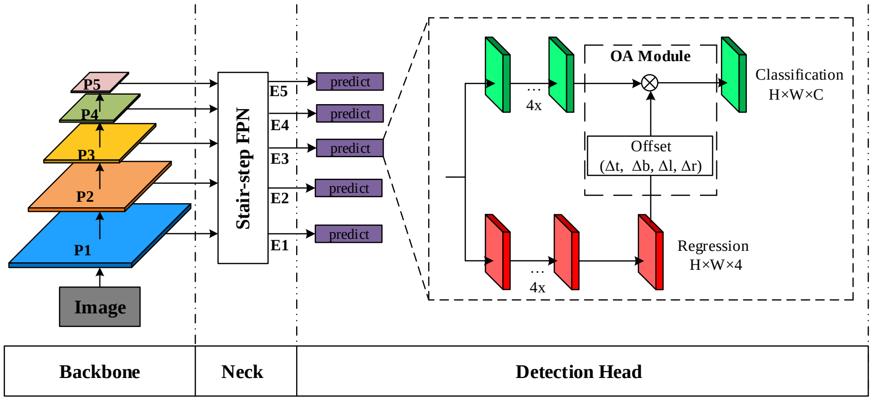

Fig. 1. The overall architecture of the proposed method includes the backbone, neck, and detection head. Stair-step FPN takes five feature maps (P_1, P_2, P_3, P_4, P_5) as input and then obtains balanced features (E_1, E_2, E_3, E_4, E_5). Offset Adaption OA Module takes four offset values and classification features as input. $\otimes$ denotes 3×3 deformable convolution.

The detection head consists of the classification branch and regression branch. The OA module considers the correlation of classification scores and bounding box prediction, described in section 3.2.

3.1 Stair-step Feature Pyramid Networks

Inspired by FPN [9], Stair-step FPN employs lateral connection using 1×1 conv and bi-linear interpolation to gather low-level and high-level features. Specifically, Stair-step FPN takes (P_1, P_2, P_3, P_4, P_5) as input. Because the number of channels and spatial resolution of each P_i is different. Therefore, 1×1 conv reduces the number of channels of the down feature suitable for summation operation with the top feature. Bi-linear interpolation up-samples the top feature to the same size as the down feature. The output of this process produces a feature pyramid (C_1, C_2, C_3, C_4, C_5). The detailed network is shown in figure 2.

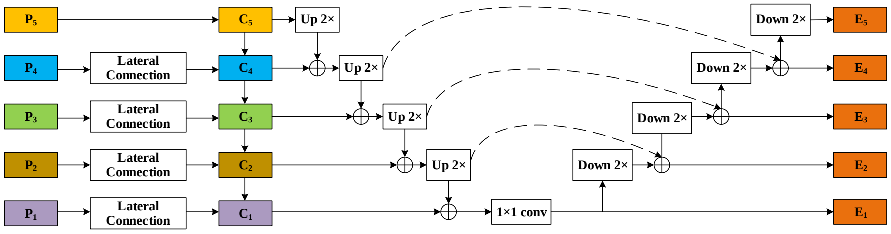

Fig. 2. The detailed architecture of the Stair-step FPN includes a top-down pathway and a down-top pathway. Up 2× denotes Up-sampling operation with scale 2. Down 2× denotes Down-sampling operation with scale 2. $\bigoplus$ denotes summation operation.

To model a uniform feature, the Stair-step FPN utilizes a top-down pathway to sum the up-sampled feature C_{i+1} with C_i steadily. Hence, multi-level features with different scales are fused to form a single feature. Then, 1×1 convolution enhances a uniform feature across all channels.

Similar to the top-down pathway, the Stair-step FPN uses the down-top pathway to create multi-level features. The enhanced feature is down-sampling two times by max-pooling operation. The residual connection is applied to improve feature representation by adding a feature in the top-down pathway with a feature in the down-top pathway correspondingly with the same spatial resolution and number of channels. Additionally, the short-cut connection is able to solve hard optimization when propagating gradient to the top-down pathway. Finally, the stair-step FPN outputs balanced multi-level features with different scales (E_1, E_2, E_3, E_4, E_5) to solve scale imbalance in object detection.

3.2 Offset Adaption

The standard convolution layers perform on fixed rectangular grid sampling. Therefore, the receptive field is not adaptive with scales or shapes of objects. Also, the feature of the classification branch can lose to adapt with the feature of the regression branch. To overcome this problem, the Offset Adaption (OA) Module is proposed to adapt the classification feature with regressed offset values, which enhances feature representation. The detailed information is shown in figure 1.

Different from standard convolution, deformable convolution [3] adds offset to rectangular grid sampling location. Hence, this strategy can change the receptive field adaptive with scales or shapes of objects. The OA module applies this operation for adapting classification features with four regressed offset values, i.e., four distances (Δt, Δb, Δl, Δr). The input of the OA module is the feature of the classification branch and distance offsets of the regression branch. The distance offsets estimate the filter offset $\Delta p_l \in \{\Delta t, \Delta b, \Delta l, \Delta r\}$ which changes rectangular grid sampling location L in deformable convolution operation. The adapted feature $y(p_0)$ at location p_0 is calculated as follows:

$$y(p_0) = \sum_{p_l \in L} w(p_l) * x(p_0 + p_l + \Delta p_l) \tag{1}$$

where x is the feature map of classification, p_l is the original location kernel weight $w(p_l)$ in grid L.

4 Experiment Setup

The Stair-step FPN and OA module are measured on challenging benchmark MS-COCO 2017 [11] for the object detection task. MS-COCO dataset consists of 115k images for training, 5k validation images for selecting the best hyper-parameters, and 20k images for testing. Because the annotation of the test set did not provide, the result is measured by the CodaLab system. To evaluate the performance, Average Precision (AP) and Average Recall (AR) are applied.

All experiments are conducted with the deep learning Pytorch framework. The parameters of the baseline FCOS [15] is set by following the standard configuration of the mmdetection [2] with 12 epochs. The integrated model is trained with a batch size of 8 on an NVIDIA Titan GPU, CUDA 10.2, and CuDNN 7.6.5. The initial learning rate is 0.00251 from 1^{st} epochs to 8^{th} epochs. It will decay by a factor of 10 at 9^{th} epochs and 10^{th} epochs. The input image is resized to 1333×800.

Table 1. Results on the validation set 2017.

Method	Backbone	Image size	Schedule	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
Faster R-CNN [13]	ResNet-50	1333×800	1×	37.4	58.1	40.4	21.1	41	48.1
Mask R-CNN [4]	ResNet-50	1333×800	1×	38.2	58.8	41.4	21.9	40.9	49.5
GC-Net [1]	ResNet-50	1333×800	1×	39.9	61.3	43.5	24.3	43.7	51.5
RetinaNet [10]	ResNet-50	1333×800	1×	36.5	55.4	39.1	20.4	40.3	48.1
FoveaBox [6]	ResNet-50	1333×800	1×	36.5	56.0	38.6	20.5	39.9	47.7
Free-Anchor [19]	ResNet-50	1333×800	1×	38.7	57.3	41.5	21.0	42.0	51.3
GHM [7]	ResNet-50	1333×800	1×	37.0	55.5	39.2	20.4	40.3	49.1
FCOS [15]	ResNet-50	1333×800	1×	36.6	55.7	38.8	20.7	40.1	47.4
Ours	**ResNet-50**	1333×800	**1×**	**37.8**	**55.9**	**38.8**	**21.0**	**40.3**	**50.1**

5 Results

Comparison with state-of-the-art. The FPN in baseline FCOS is replaced by the proposed Stair-step FPN. The pre-trained backbone ResNet [5] is trained on ImageNet [14]. The results are evaluated on MS-COCO validation set with 5k images and compared with the state-of-the-art object detectors in Tab. 1. All experiments use backbone ResNet-50 and the learning schedule is 1× denoting 12 epochs.

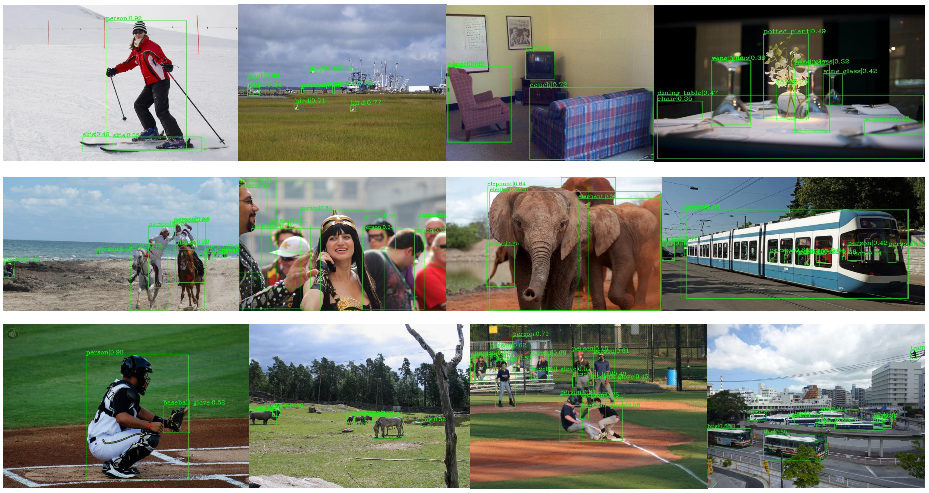

Fig. 3. The qualitative results of the proposed method on MS-COCO validation set.

The proposed method achieves 37.8 AP, which increases 1.2% higher AP than FCOS [15] with the same backbone and learning schedule without bells and whistles. Furthermore, the proposed method has surpassed most object detectors, e.g., AP of Faster R-CNN [13] with ResNet-50 is 37.4, AP of RetinaNet [10] is 36.5, AP of FoveaBox [6] is 36.5, AP of GHM [7] is 37.0. The performance on the validation set pointed out that the Stair-step FPN and OA module are boosted the accuracy of detectors by a large margin. These results demonstrate the efficiency of the proposed method. Fig. 3 visualizes the qualitative results of the proposed method on the MS-COCO validation set with different classes.

Ablation study. This work individually investigates the importance of each component, i.e., the Stair-step FPN and OA module. When the detector uses the Stair-step FPN in the neck part, the proposed method achieves 37.2 AP that obtains an absolute gain of 0.6% AP comparing with baseline. The OA module is able to improve the feature representation by adapting classification prediction

Table 2. The effect of each component in the detector. Results are measured on the validation set.

Stair-step FPN	OA Module	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
		36.6	55.7	38.8	20.7	40.1	47.4
✓		37.2	55.6	38.6	20.9	40.3	48.3
	✓	37.4	55.7	38.9	21.2	40.1	49.6
✓	✓	37.8	55.9	38.8	21.0	40.3	50.1

with regressed offsets. Hence, the results demonstrate the effectiveness of the OA module, shown in figure 2. Specifically, the OA module boosts the accuracy by 0.8% AP comparing with baseline. Finally, the detector utilizing all proposed methods increases the accuracy by 1.2% AP, compared to the baseline.

6 Conclusion

This paper proposes the simple but effective Stair-step Feature Pyramid Networks solving feature-level imbalance and scale variation problem in object detection. The Stair-step FPN employs the top-down pathway to harmonize feature-levels of multi-level features with different scales from outputs of the backbone network into a uniform feature and down-top pathway to create multi-level features. To better correlate classification prediction with regression branch, the novel Offset Adaption module is introduced to align classification features with four distance offsets by using deformable convolution. The experiments on the MS-COCO dataset confirm the improvement of the proposed methods, achieving state-of-the-art object detection.

References

1. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
2. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
5. 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. pp. 770–778 (2016)
6. Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., Shi, J.: Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing **29**, 7389–7398 (2020)

7. Li, B., Liu, Y., Wang, X.: Gradient harmonized single-stage detector. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8577–8584 (2019)
8. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6054–6063 (2019)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
10. 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. pp. 2980–2988 (2017)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
12. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra r-cnn: Towards balanced learning for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 821–830 (2019)
13. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
14. 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. International journal of computer vision **115**(3), 211–252 (2015)
15. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: A simple and strong anchor-free object detector. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
16. Vo, X.T., Jo, K.H.: Enhanced feature pyramid networks by feature aggregation module and refinement module. In: 2020 13th International Conference on Human System Interaction (HSI). pp. 63–67. IEEE (2020)
17. Vo, X.T., Wen, L., Tran, T.D., Jo, K.H.: Bidirectional non-local networks for object detection. In: International Conference on Computational Collective Intelligence. pp. 491–501. Springer (2020)
18. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
19. Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q.: Freeanchor: Learning to match anchors for visual object detection. In: Advances in neural information processing systems. pp. 147–155 (2019)