Bidirectional Non-local Networks for Object Detection

Xuan-Thuy Vo, Lihua Wen, Tien-Dat Tran, and Kang-Hyun Jo(✉)

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

Abstract. The convolutional neural networks have reached great achievements in solving the challenging problems of computer vision tasks, such as image classification, object detection, semantic segmentation. The core element of CNNs is the convolution operation, which gathers important features by constructing pixel relationships in a local region. Even though CNNs are universally exploited in visual feature understanding, they still have drawbacks due to that the receptive field is restrained inside local neighborhoods by the physical construction of the convolution layer. The Non-local Network introduces a novel method for modeling long-range dependencies to remedy the local neighborhood problem, which computing the correlations between the query position and all positions to capture global context features and then performing a weighted sum of the features at all positions. As a complementary part of the Non-Local Network, the proposed method called Bidirectional Non-local operation designs the bidirectional relationship, which the informative feature at a specific-query position is gathered and distributed to all positions. Notably, this work relaxes the Bidirectional Non-local complexity by simplifying the network based on the same attention maps for different query positions. To evaluate the effectiveness of the proposed method, the Bidirectional Non-local block is embedded into the backbone network of the detector Mask R-CNN. Without bells and whistles, the integrated network achieves 0.9 points higher Average Precision than Global Context Network on the major baseline.

Keywords: Bidirectional non-local networks · Non-local networks · Object detection

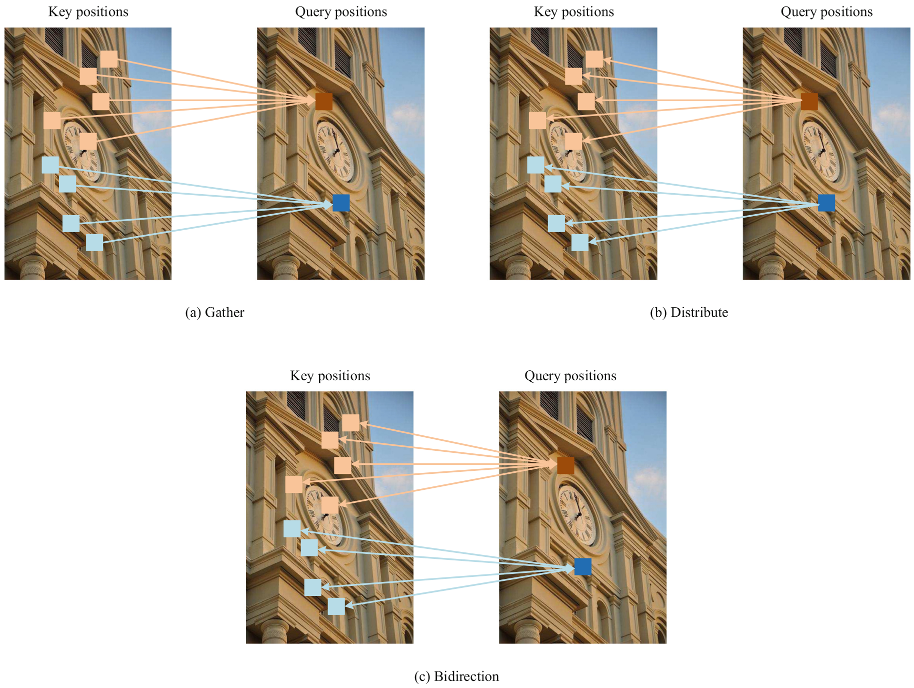
1 Introduction

Object detection is one of the most vital and challenging problems in understanding the visual world. Object detection consists of two tasks. The first task is to classify what objects inside the given images. The second task is to determine where objects locate. Object detection has been widely utilized in many applications, such as autonomous cars, robot vision, intelligent surveillance systems.

N. T. Nguyen et al. (Eds.): ICCCI 2020, LNAI 12496, pp. 491–501, 2020.
https://doi.org/10.1007/978-3-030-63007-2_38

Currently, the rapid development of deep learning techniques, remarkably Convolutional Neural Networks (CNNs), has brought a bright future in the computer vision task (e.g., image classification, object detection, instance segmentation), as an efficient approach for automatically extracting feature representations from the visual world (e.g., images, videos).

Due to the physical design of the convolution operation, the receptive field is constrained to the local regions. To remedy this problem, capturing long-range dependency is helpful for extracting the global contextual feature of visual data. In CNNs, long-range dependencies are modeled by deeply stacking many convolutional layers to enlarge the receptive field. Nonetheless, repeating many convolutional layers is not an effective way, increasing the computational cost. Furthermore, this strategy leads to the difficult optimization that is time-consuming to converge to global points.



(a) Gather

(b) Distribute

(c) Bidirection

Fig. 1. Illustration of bidirectional information paths. They gather the features of all key positions to form an attention map. Another information flow distributes the information of query positions to all key locations.

The non-local network NLNet [16] introduces the non-local block inspired by non-local operation [1], modeling the long-range dependency. A non-local operation calculates the correlation between a query position and all positions and then gathers the feature of all positions by weighted average (Fig. 1(a)).

Therefore, the relationship between each pairwise is not bidirectional. As a complementary of the non-local block, the proposed method, named BNL, presents the bidirectional information paths (Fig. 1(c)) for understanding complex visual world, not only aggregates the feature of all query position to model global context but also distributes the important information at each position to key positions globally.

The global context network GCNet [2] investigates the simplified version of the non-local block based on the query-independent attention map for all positions. This simplified network dramatically reduces the number of parameters when compared with the non-local network but still maintaining accuracy. Based on the study of the global context network, this paper also relaxes the computation cost of the non-local block but surpasses the efficiency of the global context network and non-local network with only a slight increase of the computational cost.

The Bidirectional Non-local (BNL) block applies to any existing architecture of the backbone networks. To perform the improvement of the proposed method, this block is inserted into the residual block of the ResNet [5]. This work conducts the experiment on the MS-COCO dataset [10] for the object detection task. As mentioned, the proposed method achieves significant improvement, outperforming the Mask R-CNN [4] + GC block, Mask R-CNN + NL block by 0.9% in Average Precision (AP).

2 Related Work

CNNs have been one of the most popular research in the computer vision field since acceptably designing networks guarantee a significant improvement in image classification [5,6,14], object detection [4,7,9–13], segmentation [4,17].

With the accelerated development of deep learning, object detection has improved both accuracy and speed. The goal of object detection is to classify what objects inside the given images and localize where objects on. Based on the number of networks, object detection task consists of two types, the two-stage method, and the one-stage method. The two-stage method [4,12,13] first creates a set of proposals by region proposal networks (i.e., uses the anchor generator on each center of sliding window) and assigns each ground-truth to each proposal identifying negative proposals or positive proposal. Then the second network classifies each anchor by classification networks and refines coordinates of the proposals by learning offset. Whereas the two-stage method, one-stage method instead of region proposal network creating anchors, they densely place anchors with different size and aspect ratio on each position. Then the classification network and localization network form the final detection with a specific class and bounding boxes. Faster R-CNN [13] is the two-stage method, one of the most popular architectures in the computer vision task related to detection. Inspired by the Faster R-CNN method, many architectures such as Mask R-CNN [4], Libra R-CNN [12], TridentNet [7] have introduced. Mask R-CNN adds one branch into Faster R-CNN to predict the mask for the segmentation task. In this paper, the BNL inserts into the backbone ResNet [5] for object detection based on the detector Mask R-CNN method.

Owning to the restricted receptive field of convolution operation, Non-local network [16] proposes a new method capturing long-range dependencies based on the attention mechanism [15] and the non-local filter [1] to extract the global understanding of the visual world. The relationship between key-query position and query position is not bidirectional, gathering the information at each query position from all key-query position. Although this operation is effective, but still a high computational cost. Global context [2] studies the query-independent attention map for all positions. They only use one query position to form one attention map (i.e., corresponding to one output channel of the kernel) that represents all attention maps for other positions. From this characteristic and inspired by SENet [6], GCNet drastically reduces the number of parameters of the non-local block but still maintains the accuracy of non-local networks. Different from the non-local block and global context block, the proposed method introduces a bidirectional non-local (BNL) block that aggregates the feature at each position from all positions vice-versus distributes the information path at each query position to all positions (Fig. 1). Moreover, this work, inheriting the observation of GCNet, relaxes the high computational cost but surpasses the accuracy of the non-local network and global context network with a slight increase of the computational cost.

3 The Proposed Method

3.1 Non-local Network

To design the BNL, this section visits the non-local block [16]. As mentioned in Sects. 1 and 2, the non-local block gathers the feature information at each query position from all key positions. Equation 1 expresses this relationship as

$$\mathbf{z_i} = \mathbf{x_i} + \mathbf{W_z} \sum_{j=1}^{H*W} \omega_{\mathbf{i,j}} \left(\mathbf{W_v}\, \mathbf{x_j}\right) \tag{1}$$

where $\mathbf{x_i}$ denotes the query position, $\mathbf{x_j}$ is the key query positions. H*W is the number of key positions in the input feature map. $\mathbf{W_z}$ and $\mathbf{W_v}$ are a 1×1 convolution operation. $\mathbf{z_i}$ presents the output of this block. $\omega_{\mathbf{i,j}}$ is the correlation between the query position $\mathbf{x_i}$ and $\mathbf{x_j}$, presents four types, namely Gaussian function, Embedded Gaussian, Dot product, and Concatenation. In this paper, BNL inherits the advantage of Embedded Gaussian that forms the attention map (i.e., highlights the important regions and suppresses the unnecessary parts). The Embedded Gaussian is calculated as Eq. 2.

$$\omega_{\mathbf{i,j}} = \frac{exp(\langle \mathbf{W_q x_i},\, \mathbf{W_k x_j} \rangle)}{\sum_m exp(\langle \mathbf{W_q x_i},\, \mathbf{W_k x_m} \rangle)} \tag{2}$$

where $\mathbf{W_q}$, $\mathbf{W_k}$ is 1×1 convolution operation. The overall computation of non-local block, as shown in Fig. 2(a). The non-local block models the global contextual feature, which performs a weighted sum from all key positions based on query attention maps to each query position. Hence, the relationship at pairwise positions is not bidirectional. From this observation, the proposed network constructs the bidirectional relationship between query positions and key positions.

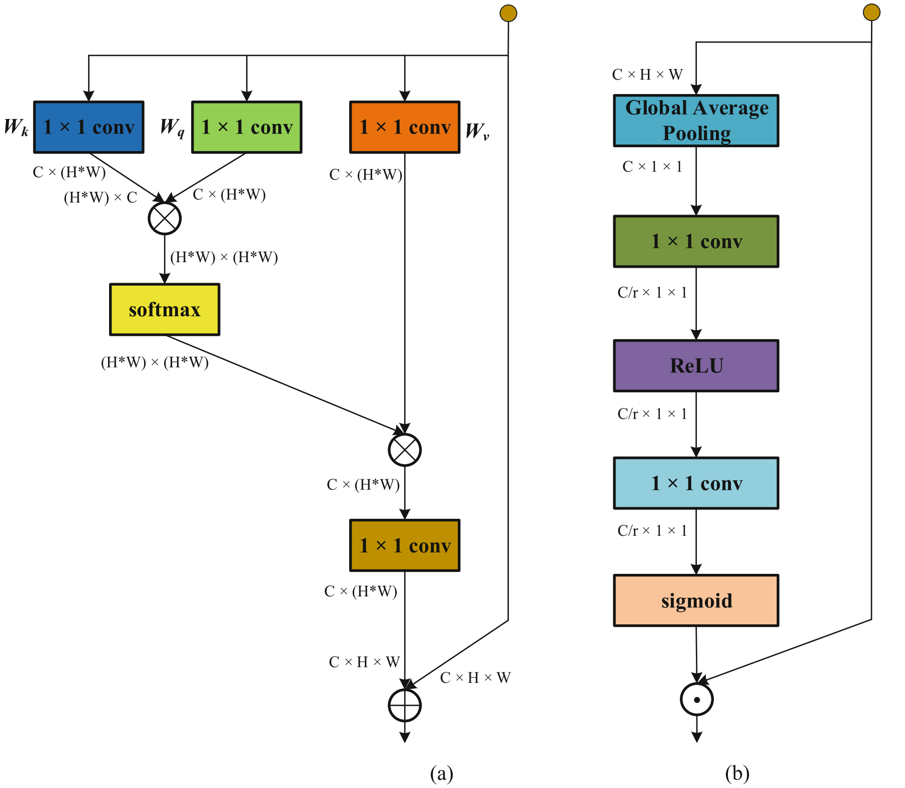


Fig. 2. (a) shows the non-local block. (b) expresses the squeeze-excitation network [6]. $\otimes$ denotes matrix multiplication, $\oplus$ denotes matrix summation and $\odot$ denotes broadcast element-wise multiplication.

3.2 Bidirectional Non-local Network

In the Eq. 2, there are many key query positions and query positions in the feature map. It leads to a large number of the computation between $\mathbf{x_i}$ and $\mathbf{x_j}$. Hence, the non-local block is simplified by approximation strategy. The first, the function $\omega_{\mathbf{i,j}}$ is converted as

$$\omega_{\mathbf{i,j}} \approx \omega_{\mathbf{i}} = \frac{exp(\mathbf{W_q x_i})}{\sum_m exp(\langle \mathbf{W_q x_i},\ \mathbf{W_k x_m} \rangle)} \tag{3}$$

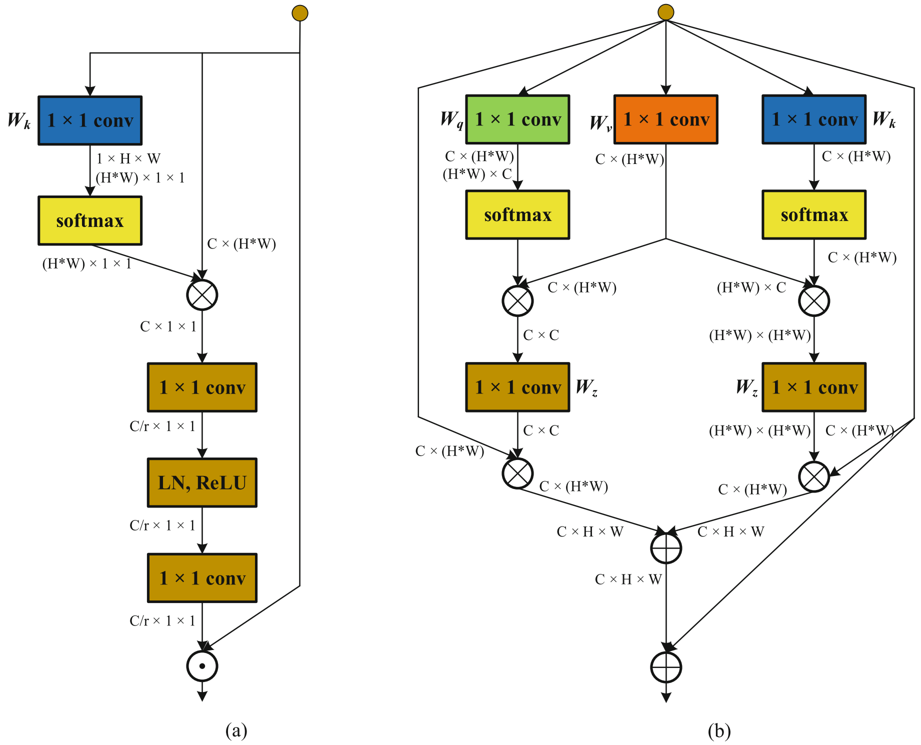


Fig. 3. (a) shows the global context block. (b) presents the bidirectional non-local block, left part is $\omega_{\mathbf{i}}$, right part is $\omega_{\mathbf{j}}$. $\bigotimes$ denotes matrix multiplication, $\bigoplus$ denotes matrix summation and $\bigodot$ denotes broadcast element-wise multiplication.

In the Eq. 3, the information path from key query position j to query position i is only related to attention map at the query location i and the correlation of i and j. To model the attention map, the function $\omega_{\mathbf{i}}$ is transformed to softmax function by ignoring $\mathbf{W_q x_i}$ in the denominator.

$$\omega_{\mathbf{i,j}} \approx \omega_{\mathbf{i}} = \frac{exp(\mathbf{W_q x_i})}{\sum_m exp(\mathbf{W_k x_m})} \tag{4}$$

Correspondingly, the function $\omega_{\mathbf{i,j}}$ is simplified as

$$\omega_{\mathbf{i,j}} \approx \omega_{\mathbf{j}} = \frac{exp(\mathbf{W_k x_j})}{\sum_m exp(\mathbf{W_k x_m})} \tag{5}$$

In the Eq. 5, the information path from key query position j to query position i is only related to attention map at the key query location j and the correlation of i and j. Especially, the function $\omega_{\mathbf{j}}$ is the same formula of the global context block [2] (Fig. 3(a)). It means that the global context block is a special case of the proposed method named bidirectional non-local network (BNL).

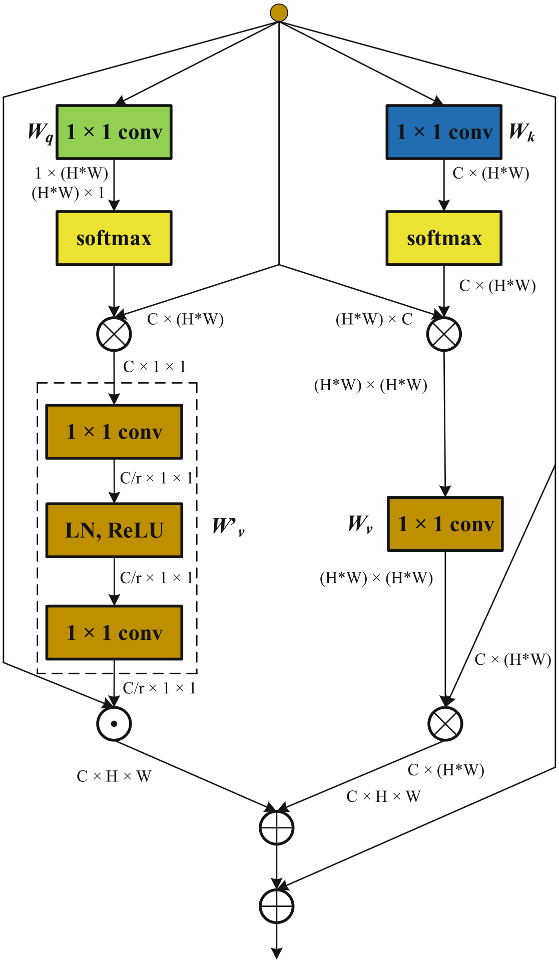


Fig. 4. Illustration of simplified bidirectional non-local block. $\otimes$ denotes matrix multiplication, $\oplus$ denotes matrix summation and $\odot$ denotes broadcast element-wise multiplication.

Finally, the bidirectional information path is formed as

$$\mathbf{z_i} = \mathbf{x_i} + \mathbf{W_z} \sum_{j=1}^{H*W} \omega_\mathbf{i} \left(\mathbf{W_v}\, \mathbf{x_j}\right) + \mathbf{W_z} \sum_{j=1}^{H*W} \omega_\mathbf{j} \left(\mathbf{W_v}\, \mathbf{x_j}\right) \tag{6}$$

The second term shows that each query position gathers the feature from other positions. The third term presents that each query position distributes the feature to other positions. Figure 3(b) shows the bidirectional non-local block.

Notably, the global context network [2] proposed the query-independent attention map for all positions. The proposed method relaxes the computational cost by inheriting this observation of GCNet. Furthermore, the $\mathbf{W_z}$ is removed, and the $\mathbf{W_v}$ is moved out, as showed in Fig. 4.

4 Experiment Setup

The proposed method conducts the experiments on challenging MS-COCO 2017 [10] for the object detection task. This dataset includes 115k images (80k images of training set + 35k images of validation subset) for training, 5k validation images for selecting the best hyper-parameters, and 20k images for testing. Because the ground-truth annotation of the test set did not publish, the result is submitted to the protocol system. The metrics are used through standard Average Precision (AP) and Average Recall (AR).

All experiments are implemented with the Pytorch framework. The BNL block is applied to stage c3, c4, c5 of the backbone network ResNet-50 [5]. The object detector is Mask R-CNN [4] with the neck FPN [8].

The Mask R-CNN is configured by following the standard setting of the mmdetection [3] with 12 epochs. The integrated model is trained with a batch size of 8 on one NVIDIA Titan GPU, CUDA 10.2, and CuDNN 7.6.5. The initial learning rate is 0.01 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.

5 Results

Comparison with State-of-the-Art. The BNL block is inserted into the architecture Mask R-CNN with the backbone ResNet-50 and the neck FPN. The integrated model with the stronger backbone ResNet-50 + BNL evaluates on MS-COCO test-dev set and compares the experimental results with the state-of-the-art object detectors in Table 1. The learning schedule is 1× for training the model with 12 epochs and 2× for training the model with 24 epochs. The ResNet-50, ResNet-50(a) denotes pytorch-style and caffe-style backbone, respectively.

The proposed method, BNL block embedded into the backbone network ResNet-50 (i.e., ResNet-50+BNL) of the object detectors Mask R-CNN achieves 39.3 AP, which increases 0.9% higher AP than GCNet with the backbone ResNet-50+GC achieves 38.4 AP without bells and whistles. Especially, the integrated method outperforms the baseline Mask R-CNN with an improvement rate of 2%. Furthermore, the proposed method has surpassed most object detectors with the same backbone, neck FPN, and learning schedule, e.g., AP of Faster R-CNN [13] with ResNet-50 is 36.4, AP of RetinaNet [9] is 35.6, AP of [4] is 37.3. The performance on test-dev set pointed out that the strong baselines are boosted by a large margin when applying the GNL block to stage c3, c4, c5 of the backbone ResNet-50. These results prove the efficiency of the proposed method.

Figure 5 visualizes the qualitative results of the proposed method on the MS-COCO validation set with three levels of the dataset.

Table 1. Results on test-dev set 2017.

Method	Backbone	Schedule	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}
Faster R-CNN [13]	ResNet-50	1×	36.4	58.4	39.1	21.5	40.0	46.6
Faster R-CNN	ResNet-50(a)	1×	36.6	58.5	39.2	20.7	40.5	47.9
Faster R-CNN	ResNet-50	2×	37.7	59.2	41.1	21.9	41.4	48.7
RetinaNet [9]	ResNet-50	1×	35.6	55.5	38.3	20.0	39.6	46.8
RetinaNet	ResNet-50(a)	1×	35.8	55.5	38.3	20.1	39.5	47.7
RetinaNet	ResNet-50	2×	36.4	56.3	38.7	19.3	39.9	48.9
Mask R-CNN [4]	ResNet-50	1×	37.3	59.0	40.2	21.9	40.9	48.1
Mask R-CNN	ResNet-50(a)	1×	37.4	58.9	40.4	21.7	41.0	49.1
Mask R-CNN	ResNet-50	2×	38.5	59.9	41.8	22.6	42.0	50.5
GCNet [2]	ResNet-50+GC	1×	38.4	59.3	41.8	21.6	41.7	49.8
Ours	**ResNet-50+BNL**	1×	**39.3**	**61.6**	**42.7**	**22.8**	**42.0**	**49.2**

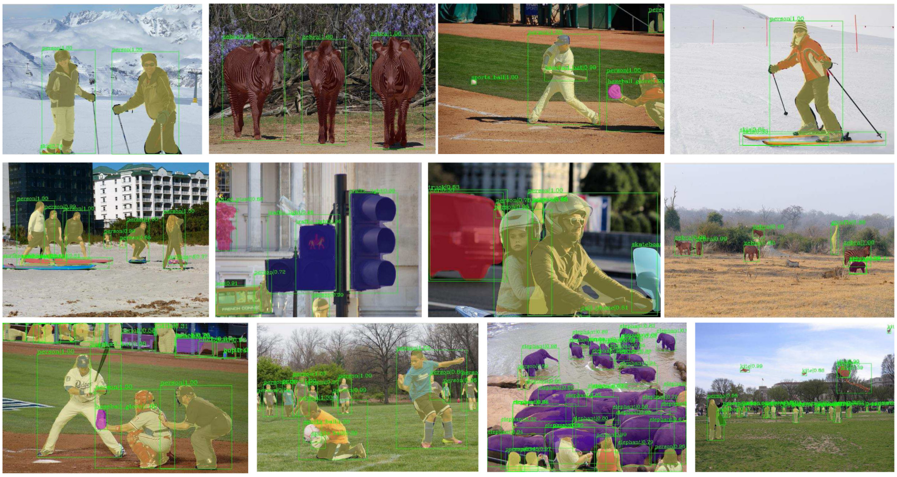


Fig. 5. The qualitative results of the proposed method on MS-COCO validation set.

Ablation Study. This work studies the importance of each component in the BNL block. The proposed method consists of gathered information block $\omega_{\mathbf{i}}$ and distributed information block $\omega_{\mathbf{j}}$. The first, the BNL module investigates the efficiency of the gathered information block by removing the distributed information block and otherwise.

Table 2 analyzes the impacts of each component in the BNL module. This experiment gradually inserts the gathered block and distributed block on the ResNet-50 Mask R-CNN baseline. The gathered component inheriting the advantage of the global context block improves 1.1% higher AP than the ResNet-50 Mask R-CNN baseline. The distributed component increases 0.5% from 37.2 AP to 37.7 AP. This block is lightweight due to that this component only used two 1 × 1 convs with a little of parameter. When combining gathered block and distributed components into the BNL block, the accuracy of the integrated model is gained by a large margin of 1.9% over the baseline.

Table 2. The impacts of each component in the BNL block. The result reports on the validation set.

Gathered	Distributed	AP^{bbox}	AP^{50}	AP^{75}	AP^{mask}	AP^{50}	AP^{75}
		37.2	59.0	40.1	34.8	55.4	35.9
✓		38.1	60.0	41.2	34.9	56.5	37.2
	✓	37.7	59.4	40.6	34.5	56.0	36.1
✓	✓	39.1	61.4	42.3	35.3	57.7	37.3

6 Conclusion

In this paper, the proposed Bidirectional Non-local (BNL) block studies the effectiveness of the gathered information block and the distributed information block. The gathered information block gathers the feature of all query position to capture long-range dependencies. The distributed block distributes important information at each position to key positions globally. By fusing two information propagation flows, the proposed methods not only encodes long-range dependencies by computing the correlation between each pair of query position but also considers the relative location of it. The experimental results demonstrate the significant improvement of the BNL block when applying to the backbone ResNet-50 of detectors Mask R-CNN baseline. Without bells and whistles, the integrated model brings 0.9 points higher AP than the GCNet and 2.0 points higher AP than the Mask R-CNN baseline.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government. (MSIT) (2020R1A2C2008972).

References

1. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 2, pp. 60–65. IEEE (2005)
2. 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 (2019)
3. Chen, K., et al.: MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
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. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
7. 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)
8. 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)
9. 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)
10. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
11. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
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., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
16. 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)
17. Zhao, H., et al.: PSANet: point-wise spatial attention network for scene parsing. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11213, pp. 270–286. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_17