Joint COCO and Mapillary Workshop at ICCV 2019: COCO Instance Segmentation Challenge Track

Technical Report: Context Routing for Object Detection

Zhe Chen, Jing Zhang, and Dacheng Tao
UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering
The University of Sydney, Darlington, NSW 2008, Australia
{zhe.chen1,jing.zhang1,dacheng.tao}@sydney.edu.au

Abstract

Contexts are critical for robust object detection. However, it is still challenging to model context relationships between contexts and local features effectively and efficiently. In this report, by taking advantage of the favorable expressive capacity of dynamic routing mechanism [14], we present an effective network architecture for modeling context relations in object detectors. Specifically, we derive a novel and efficient routing mechanism, called context routing algorithm, to model context relationships. In the context routing algorithm, we first employ a relation modeling network to learn effective modeling of context relationships. Then, the context routing algorithm encodes the modeled context relationships through encapsulation and transforms the encapsulated relationships to allow more dynamics for context modeling. In practice, the context routing algorithm is easy-to-implement within different modern detectors. Using the MS COCO dataset [12], our proposed context routing algorithm promisingly improves the baseline detectors. In particular, the Mask RCNN [7] detector with context routing network can achieve around 5 % higher performance on the test-dev set.

1. Introduction

Over the years, researchers have confirmed that contexts are crucial for improving object detection performance [13, 1]. Despite the critical role of contexts, it is still difficult to model complicated relationships between contexts and local features for object detection. Early studies [16, 6] only fused additional visual features as contexts without considering their relationships with respect to local features. Recent studies then attempted to further model the context relationships. For example, the study [4] introduced a spatial reasoning process to model object-object relationships for detection. The study [2] introduced non-local networks to model global spatial relationships. In the meantime, the study [8] introduced an attention module to model object relations, and the authors of [5] instead used empirical measures like intersect-over-union (IoU) scores to model object relations. Although promising, we find that these prevailing methods generally do not describe relationships explicitly. They also encode contextual relationships within weights that only suppress context features, allowing limited degrees of freedom to describe complex relationships.

We are inspired by the recent development of dynamic routing between capsules [14, 10]. In this routing algorithm, complicated attributes of an entity can be encoded within capsules, and the relations between capsules are adaptively modeled by an iterative routing-by-agreement mechanism. However, the original dynamic routing algorithm also only suppressed the capsules during routing. Besides, the required computational costs for the iterative routing procedure can be excessively large if considering all the potential pair-wise context relationships existed on the large feature maps of an object detector.

To tackle the above issues and further improve the modeling of context relationships, we devise a novel and efficient algorithm to encode context relationships better and also introduce more dynamics for context modeling. In particular, after extracting different types of context features, we first employ a relation modeling network to learn and estimate their context relationships with respect to local features. The relation modeling network takes as input both the context features and the local features and delivers the output, which aligns the dimensions of context features, to represent the corresponding context relationships. Then, we encapsulate the modeled context relationships to encode diversified attributes within a normalized representation. To allow more dynamics for context modeling, we apply an exponential transformation function to squash the modeled relationships, enabling neural networks to decide by themselves either suppress or enhance context feature when refining local features. Overall, we call our proposed context modeling algorithm as context routing (CR) algorithm.

In practice, the proposed CR algorithm is efficient and effective for improving object detection. It could learn effective modeling of context relationships without requiring the exhaustive refinement procedure in the original dynamic routing algorithm. Furthermore, it can be implemented as a complementary and easy-to-plug-in module in various modern object detectors, including FPN [11] and Mask RCNN [7]. On the MS COCO detection dataset, comprehensive evaluation results validate that the proposed CR algorithm can promisingly boost the detection performance. In particular, on the instance segmentation task, the CR delivers around 5% performance gain with respect to the baseline Mask RCNN detector, out-performing other context modeling methods on the MS COCO dataset.

2. Context Routing

2.1. Context Modeling Framework

In this study, we consider contexts as surrounding visual features that can refine the recognition of local objects. Different from current studies that generally do not model relationships explicitly and only allow limited dynamics for encoding relations, we instead devise an efficient and effective approach that can better encode relationships and can introduce more dynamics for context modeling, aiming to magnify the benefits of contexts for object detection.

Formally, we define that $\mathbf{v}_{x,y}$ is the local feature that describes an object at the location (x, y) on the image plane. We denote $N(\mathbf{v}_{x,y})$ as the collection of context features that surround the (x, y): $N(\mathbf{v}_{x,y}) = \{\mathbf{v}_{x',y'} | for\ dist((x', y'), (x, y)) \leq \tau\ and\ (x', y') \neq (x, y)\}$, where τ is a distance threshold. We use the symbol $\Omega(\mathbf{v}_{x,y}, N(\mathbf{v}_{x,y}))$ to represent the relations between a local feature $\mathbf{v}_{x,y}$ and its context features $N(\mathbf{v}_{x,y})$. For simplicity, we drop the notions (x, y) in the following sections, and thus: $\mathbf{v} = \mathbf{v}_{x,y}$, $N(\mathbf{v}) = N(\mathbf{v}_{x,y})$. We formulate the overall context modeling framework for object detection as a function f_{cm}:

$$\tilde{\mathbf{v}} = f_{cm}\Big(\mathbf{v}, N(\mathbf{v}), \Omega\big(\mathbf{v}, N(\mathbf{v})\big)\Big), \tag{1}$$

where $\tilde{\mathbf{v}}$ is the local feature refined by modeling contexts. In practice, we validate the effectiveness of the implementation of the context modeling function f_{cm} by observing the performance gain of an object detector that utilizes $\tilde{\mathbf{v}}$ for detecting objects.

2.2. Context Routing Algorithm

In this study, we implement the f_{cm} function as follows:

$$\tilde{\mathbf{v}} = \mathbf{v} + \sum_i \Omega(\mathbf{v}, \mathbf{v}_i^n) \cdot \mathbf{v}_i^n, \tag{2}$$

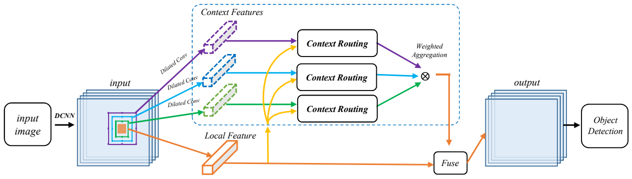

Figure 1: The proposed context modeling architecture which applies context routing to improve the modeling of relationships between context features and local features.

where $\mathbf{v}_i^n \in N(\mathbf{v})$ and i indexes over different collected context features.

To implement $\Omega(\cdot)$ with a better relation modeling result, we introduce the context routing algorithm. In general, the context routing algorithm performs the following operations subsequently. First, taking the contexts and local features as input, we employ a relation modeling network to learn the effective modeling of context relationships. Rather than encapsulate context features, we instead encode the modeled relations within capsules. Then, the encapsulated relations help route the context features. Here, we apply an exponential transformation function to squash the encapsulated relations and allow more dynamics to represent relations. Lastly, the estimated relations are used to weight context features and refine local features according to Eq. 2. To sum up, we implement $\Omega(\cdot)$ as follows:

$$\Omega(\mathbf{v}, \mathbf{v}_i^n) = T\Big(g\big(R(\mathbf{v}, \mathbf{v}_i^n)\big)\Big), \tag{3}$$

where $R(\mathbf{v}, \mathbf{v}_i^n)$ represents the relationship information extracted by the relation modeling network, $g(\cdot)$ is the function that encapsulates relations, and T is the function that transform the encapsulated relations to allow more dynamics for routing. Fig. 1 shows the architecture of our proposed context routing algorithm in an object detector.

To obtain a proper estimation of relationships, we employ a small network to learn the estimation of $R(\cdot)$. We mainly concatenate the feature vector of $\mathbf{v}$ and $\mathbf{v}_i^n$ to obtain the network input, and then we let the network compute their relationships which have the same dimensions with $\mathbf{v}_i^n$. This can be simply implemented based on convolution operations.

After the estimation of $R(\cdot)$, we then encode the extracted relationship information via an encapsulation operation $g(\cdot)$. We mainly follow the encapsulation operation as introduced in the original dynamic routing study [14], in which representations are regularized and reshaped based on L2 normalization operations to obtain capsules. As analyzed in the original study, the L2 normalization-based encapsulation operation can encode various attributes of an entity or object, including its poses, orientations, shapes, and so on. Different from the original study that encodes the appearance information about presented entities, we encode the relationship information within capsules, so that complicated relationships can be modeled properly.

To introduce more dynamics for context modeling, we tend to hypothesize that the desired relationship-based weights should be a re-scaling factor that can be larger than 1, allowing the enhancement of a feature. Besides, the weights should be neither negative nor extremely large. Negative scaling factors would change the representations of input vectors rather than weighting them, and the extremely large scaling factors would make the gradients easily blow up during training. To satisfy these properties, we tend to apply an exponential transformation operation $T(\cdot)$. Consider $\hat{\mathbf{r}}$ as the modeled relations: $\hat{\mathbf{r}} = g\big(R(\mathbf{v}, \mathbf{v}_i^n)\big)$, we have:

$$T\big(\hat{\mathbf{r}}\big) = \exp\Big(\alpha \cdot \hat{\mathbf{r}}\Big), \tag{4}$$

where α controls the value range of the weight coefficients. Using this transformation, negative relation values will be changed to values near 0, meaning that irrelevant context features will be suppressed, and *vice versa*. The hyper-parameter $\alpha(\alpha > 0)$ can control the extent of enhancement.

3. Experiments

3.1. Implementation and Training Settings

The proposed context routing algorithm is easy-to-plug-in for different object detectors. In this study, we mainly implement the context routing algorithm as a complementary network for the detection pipeline. In the detection pipeline, such as FPN [11], we first attach two subsequent context routing networks after each feature pyramid level. The channel for one context feature vector of the coarsest level is 32, and we halve the channel number at every lower level. We extract contexts with four different dilation rates ranged from 3 to 11. Also, we attach one context routing network after the RoI align for recognition heads. In these heads, the context feature channel is set to 32, and we extract contexts with dilation rates of 3 and 5. We apply group normalization on the weighted contexts to stabilize training.

We use the standard training set of MS COCO dataset and do NOT include any external datasets. We conduct ablation study on the *val* set and perform overall evaluation on the *test-dev* set of MS COCO. We mainly follow the training configurations as used in “mmdetection” [1] for overall evaluation and this challenge, while we only use a smaller image scale, *i.e.* (1000, 600), for the ablation study.

3.2. Ablation Study

In this section, we perform ablation study to evaluate different components of the proposed context routing algorithm. Note that we use ResNet50 as the backbone network, and use FPN [11] as the baseline detector.

Method	AP^{bbox}	AP$^{bbox}_{50}$	AP$^{bbox}_{75}$
R-50-FPN	36.0	57.5	38.5
R-50-FPN + CF	36.9	59.4	39.4
R-50-FPN + CF + CR	**37.9**	**59.7**	**41.8**

Table 1: Effects of contexts and context routing on the baseline detector. “CF” means introducing context features without modeling relationships. “CR” means applying the proposed context routing algorithm to model relationships. Scores on *val* are reported.

Weighting Method	AP^{bbox}	AP$^{bbox}_{50}$	AP$^{bbox}_{75}$
Softmax	37.0	58.4	40.2
Sigmoid	37.3	58.8	40.6
Context Routing	**37.9**	**59.7**	**41.8**

Table 2: Comparison of different relation-based weighting strategies. Scores on *val* are reported.

Transformation T	w/o T	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
AP^{bbox}	36.7	37.3	37.5	**37.9**	37.6

Table 3: Effects of contexts and context routing on the baseline detector. “CF” means introducing context features without modeling relationships. “CR” means applying the proposed context routing algorithm for the context features. Scores on *val* are reported.

We first evaluate the overall effects of the proposed context routing algorithm in a ResNet-50-FPN baseline detector. Table 1 shows the detailed performance on *val* set. The presented results show that including additional context features, denoted by “CF”, is beneficial for improving detection performance. By further applying the proposed context routing algorithm, denoted by “CR”, we achieve around 3% improvement *w.r.t.*“CF” and around 5% improvement *w.r.t.* the baseline model.

We also compare the performance of context routing with two other common context feature weighting strategies, including softmax-based weighting and sigmoid-based weighting. These compared weighting strategies perform softmax or sigmoid operations on the output of the relation network. Table 2 presents the detailed performance of different weighting strategies using the same ResNet-50-FPN baseline detector and the same types of extracted contexts. We can observe that the proposed context routing algorithm achieves superior performance, demonstrating its effectiveness of improving context modeling based on modeled relationships.

Besides, we study the influences of different choices of α in Eq. 3. The results are presented in Table 3. These results have illustrated that only using the encapsulated relationships could degrade performance, meaning that the negative encapsulation values are not beneficial for improving context modeling. In addition, the α brings the highest improvement when it is 3, showing that neither small α nor

[1] https://github.com/open-mmlab/mmdetection

Methods	Detector	AP^{bbox}	AP$^{bbox}_{50}$	AP$^{bbox}_{75}$	AP^{segm}	AP$^{segm}_{50}$	AP$^{segm}_{75}$
baseline	ResNet101-Mask RCNN	40.6	61.8	44.8	36.4	58.7	38.6
	ResNeXt101-Cascade Mask RCNN	45.0	63.7	49.1	38.7	60.8	41.8
Relation Net[8]	ResNet101-FPN	38.9	60.5	43.3	-	-	-
Context Refine [5]	ResNet101-Mask RCNN	42.0	62.9	46.4	-	-	-
GC-Net[15]	ResNeXt101-Cascade Mask RCNN	46.6	65.9	50.7	40.1	62.9	43.3
Context Routing	ResNet101-Mask RCNN	42.8	63.8	47.3	38.0	60.7	40.2
	ResNeXt101-Cascade Mask RCNN	**47.1**	**66.0**	**51.3**	**41.0**	**63.2**	**44.8**

Table 4: Overall evaluation of different context modeling methods on the *test-dev* set of MS COCO.

	AP^{segm}	AP$^{segm}_{50}$	AP$^{segm}_{75}$
HTC [3]	44.0	67.6	48.0
HTC & Context Routing	44.8	68.0	49.1
+ Mask Scoring [9]	45.0	68.3	49.3
+ Multi-scale Testing	46.1	70.2	50.5
+ Ensemble	**47.0**	**69.8**	**51.6**

Table 5: Effects of bells and whistles on *test-dev* of instance segmentation task of MS COCO.

large α is favorable for context routing. In practice, small α could limit the introduced dynamics, while too large α could make the training unstable.

3.3. Overall Evaluation

We present the overall evaluation of our proposed context routing algorithm on the *test-dev* set of MS COCO dataset, comparing to other cutting-edge context modeling methods. Detailed performance is listed in the Table 4. It shows that the context routing algorithm improves baseline detectors by around 5% for both Mask RCNN and Cascade Mask RCNN on bounding box scores, out-performing other context modeling methods, .

3.4. Conclusion

This report describes a novel context relationship modeling method for improving the utilization of contexts in object detectors. Promising performance gains on the MS COCO dataset demonstrate its effectiveness on object detection and instance segmentation.

References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, pages 2874–2883. IEEE, 2016. 1

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *arXiv preprint arXiv:1904.11492*, 2019. 1

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[4] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, pages 4106–4116. IEEE, 2017. 1

[5] Zhe Chen, Shaoli Huang, and Dacheng Tao. Context refinement for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–86. Springer, 2018. 1, 4

[6] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, pages 1134–1142. IEEE, 2015. 1

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *ICCV*, 2017. 1, 2

[8] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, volume 2. IEEE, 2018. 1, 4

[9] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 4

[10] Hongyang Li, Xiaoyang Guo, Bo Dai, Wanli Ouyang, and Xiaogang Wang. Neural network encapsulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 252–267, 2018. 1

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*. IEEE, 2017. 2, 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 1

[13] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898. IEEE, 2014. 1

[14] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NIPS*, pages 3856–3866, 2017. 1, 2

[15] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*. IEEE, 2018. 4

[16] Sergey Zagoruyko, Adam Lerer, Tsung-Yi Lin, Pedro O Pinheiro, Sam Gross, Soumith Chintala, and Piotr Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016. 1