GATED ADDITIVE SKIP CONTEXT CONNECTION FOR OBJECT DETECTION

Haoran Li,Hongxun Yao,Yuxin Hou,Xiaoshuai Sun

School of Computer Science and Technology,
Harbin Institute of Technology, China.

ABSTRACT

Context information plays an important role in object detection. DeepID concatenates the global context for classification, while Yolo, SSD, and Crafting use local context information for detection. In this paper, we propose a straightforward method to plug the global context information into the Faster RCNN Framework, namely the Skip Context Connection (SCC). We use SCC to inject the global context into the object representation which skips the RoiPooling layer rather than drops it. Therefore, it can not only leverage the context information but also keep the location accuracy from the RCNN framework. We proposed three principles to construct the SCC blocks: **effectiveness** means fewer parameters, **additivity** means the features possess the same meaning, and **selectable** means soft gated addition. We also evaluate several different SCC blocks. The Gated Additive SCC(GA-SCC) which satisfy the three principles get the best performance. Our experiment results on PASCAL VOC 2007 show that GA-SCC can get the steady 1% improvement over the traditional RCNN method.

Index Terms— context information, object detection

1. INTRODUCTION

In recent years, the deep convolutional neural network demonstrates an excellent ability on feature representation for computer vision tasks such as object recognition, object detection etc. Especially, in object detection field, the RCNN approach makes the researchers rethink the deep learning method. By using the impressive deep models like Alexnet[1], VGG[2], inception[3] and reset[4] , the CNN based object recognition models have already bypass average human performance. In object detection task, the RCNN[5] based models, the family of Fast RCNN[6] and Faster RCNN[7] models have made a tremendous improvement over traditional approaches. However, these methods still suffered from the barriers such as scale limitation, low resolution, and potential occlusion. Although the inter-media feature map connection and deeper model have been proposed to solve these problems, object detection in complex scenes is still very difficult. As a result, recent works tried another path which takes the context into account.

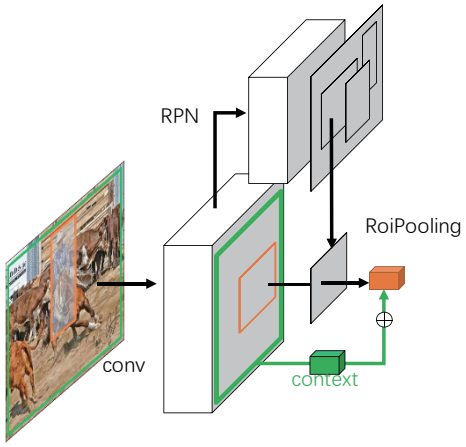

Fig. 1. $\mathcal{C}\mathbf{ontext}(\mathbf{x}) + \mathbf{x} = \mathbf{representation}$, a presentation of object should be the combination of the object itself and the scene description or here we use context.

Previous works like YOLO[8] and SSD[9] discard the RoiPooling to make the model FCN[10] end to end. The advantage of FCN[10] is that the model can cover the local context information by stack convolutional operation. However, the FCN-like model impairs the transition sensitivity without the RoiPooling. Recent works bring context into Faster RCNN like models. We divide them into two classes: (1) global context: such as DeepID[11], they put the whole image feature and the part image feature together for recognition. In a general way, they encode the whole image and the object into the same dimension features and concatenate them. (2) local context: such as convolutional pose machine[12] and many iterated models[13] in pose estimation, they use pre-supervised information and iterative optimization. These iterated blocks are built with several stacked convolutional operations in order to pass the local context information[14]. Latest work[15, 16] put the local context into the Faster R-CNN framework by adding the local around context feature, which is similar to the iterated model setting with the explicit local context region. From our point of view, we believe that global context is a kind of scene-object co-occurrence, and local context is the structure co-occurrence. The local context are generally used in pose estimation because of the strong

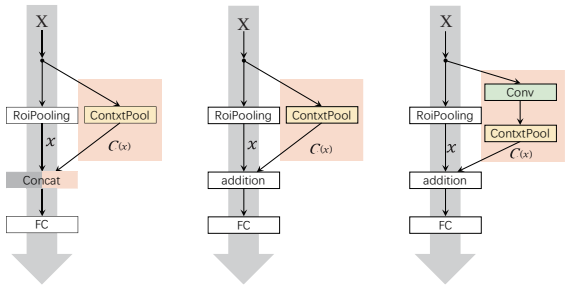

Fig. 2. (a) is Concat, (b) is addition directly, (c) is convolutional then addition

structure-level co-occurrence.

In this paper, we propose the Gated Additive SCC across the RoiPooling that can inject the global context while keeping the transition sensitivity. We give a paradigm on how to design a simple block with less modification on the base model. This block consists of a global pooling, several convolutional operations, and a gated function. We have two primary contributions in this paper: (1) we give a paradigm to design a global context block: **effective**, **selectable**, and **additivity**. (2) we test the effectiveness of our model in the VOC2007 dataset, and the results show that our SCC framework is more effective for object detection comparing to traditional methods.

2. SKIP CONTEXT CONNECTION

Here we give some notation for convenient, the final shared feature map in Faster RCNN is $\mathcal{X}$, and the feature of object extracted from this $\mathcal{X}$ using RoiPooling is x, and the context of x is $\mathcal{C}(x)$.

2.1. network design

The previous works shows that the deep model has already become a plug-pull process. The backbone of specific task model generally modified by the classic classification models. We follow this way by using the VGG16[2] and Fastr RCNN[7], and plug our SCC block into the model. Here, we proposed three basic principles to design the SCC block:

effective: As described in DeepID[11], SharpMask[17] and some other feature fusion models, they concatenate features from different branches in CNN backbone model. The concatenation is $y = [w, w_c] \cdot [x, \mathcal{C}(x)]^T$, and the extra weight w_c sometimes is very large. The large amount extra weight will lead much more computation cost. As described in Fig.2 (a), we see the concatenation[11] in Faster RCNN[7] backbone model, the extra weight is the same size with the original weight of the following fully-connected layer, which is $7 * 7 * 512 * 4096$. It is very large for the original VGG-16[2] model since the number of original parameters is only $138M$ and the extra weight is almost $102.7M$. The latest works begin to use addition to replace the concatenation like Feature Pyramid Network(FQN)[18], the addition $y = w \cdot (x + \mathcal{C}(x))$ will not lead the large extra weight. However, the result shows that this simple addition is not proper for this model because the features added are not additive.

additivity: From the SharpMask[17] to FQN[18], we can see that the features fusion method from concatenation to addition. In resnet[4], they also use the addition. The addition operation can leverage all the features ability, and also average the features representation, while the features should have the same semantic means. In deep CNN model, the backbone of the model is fixed so that the hierarchical representation is fixed. The skip connected features from different branches should describe the same things. Here the SCC in Faster RCNN[7] as showing in Fig.2 (b), the addition can be described as $y = w \cdot (x + \mathcal{C}(x))$. In this equation, the x is the representation of the object in the scene, however, the $\mathcal{C}$ is the representation of the scene. Therefore, the two features are not additive. Thus, we should use a new transfer function to make the $\mathcal{C}(x)$ additive. As shown in Fig.2 (c), the conv-add can be described as $y = w \cdot (x + w_q * \mathcal{C}(x))$, the w_q is the semantic transferring weight. The $w_q * \mathcal{C}(x)$ convert the $\mathcal{C}(x)$ to describe the co-occurrence of the object so that the two features are additive. However, after experiments, this conv-add block can only improve the detection precisions of objects only in several specific scenes, others which appear in random scenes are degraded.

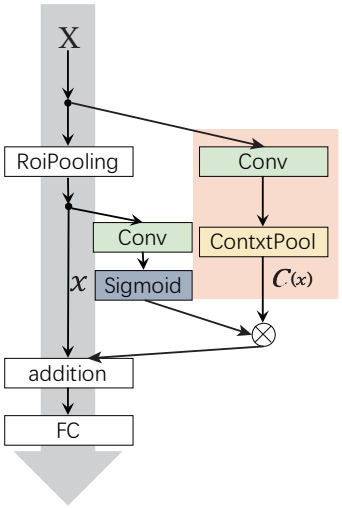

Fig. 3. $\mathbf{Gate(x) * \mathcal{C}(x) + x = representation}$

selectable: Not all the object need the context information. In other words, the object-scene co-occurrence is not ubiquitous. As described in LSTM[19] and Crafting[15], they use gated function to ensure the selectable. Here, we construct an object representation x and the additive global context information $w_q * \mathcal{C}(x)$. We just add these two features directly as shown in Fig.2 (c). That means we average the

feature ability for any object even lack of the object-scene co-occurrence. We do several experiments to test that add directly is helpful for the majority of particular object classes, but harmful for other object classes. We imputed the harm to the reason that the conv-add SCC lacks selectable ability. Therefore, we bring in the $sigmoid$ function like LSTM[19] and Crafting[15] to choose whether add the global context or not. The output value of $sigmoid(x) = 1/(1+e^{-x})$ function is $(0, 1)$, so it can be treated as a gate which allows the context adding or not.

Here, we design a GA-SCC that satisfy all the three principles above:

$$y = w \cdot (Gate(w_p * x) \otimes w_q * \mathcal{C}(x) + x) \tag{1}$$

As shown in Fig.3, x is the object representation, and it can derive the gated ability feature $w_p * x$. The $sigmoid$ function change the value into a real gated value. In the CNN model, we use share weight convolutional operation to transfer the feature ability to another. In the Eq.1, the $\otimes$ means element-wise multiplying, and the $*$ means convolutional operation. The $Gate(x) = sigmoid(x) = 1/(1 + e^{-x})$ is selectable, the w_q is for additive. We use addition to replace the concatenation for effectiveness

2.2. optimization

We plug the SCC into the Faster RCNN[7], and the model still can be end-to-end trained by back-propagation and stochastic gradient descent(SGD)[20]. The weights of extra convolutional layers in SCC block are drawn from the zero-mean Gaussian distribution with standard deviation 0.01, and all layers in Faster RCNN[7] are initialized following the original settings. We use a learning rate of 0.001 for first 50k mini-batch and multiply 0.1 for following each 50k mini-batch. The momentum is 0.9, weight decay is 0.0005. We implement the SCC use MXNet[21] and thanks to Jian Guo for original Faster RCNN[7]. The global pooling that we use Roipooling with the whole image size. We use broadcast-sum to expand the global context, which makes the context feature and the original feature same dimension during training.

3. EXPERIMENT

We evaluate our model in the PASCAL VOC 2007 detection benchmark. This dataset consists of about 5K trainval images and 5K test images over 20 object categories. We test four type SCC blocks in this dataset.

3.1. computation cost

As showing in Tab.1, we can see our GA-SCC only use $2.6M$ extra parameters. The concatenation uses almost $102.7M$ extra parameters. Although the add does not bring in the extra parameters, the result is much worse than the base model. The conv-add use one convolutional layer for additivity, the result is comparable to the base model but less than concatenation. The GA-SCC blocks not only bring the small amount of parameter but also surpass the concatenation.

methods	#. of parameters(Millions)
FRCNN[7]	$138M$
concat	$138M + 102.7M$
add	$138M$
conv-add	$138M + 2.3M$
GA-SCC	$138M + 2.6M$

Table 1. #. of parameters for the four SCC blocks.

3.2. detection result

We test our proposed four SCC blocks in VOC07. As shown in Tab.2, all the models train on the PASCAL VOC 2007 trainval 5k images. During training, we just flipped the images without using other data-augmented method, and no augmented methods like multi-scale and multi-crop are used in the test process, just one pass for the results. The concatenation gets the comparable result to the Fast RCNN[7] mAP and surpass in several categories, however, it brings in large amount of parameters that we talk above. The add SCC block collapses the results, due to the add SCC block does not satisfy the additivity. Therefore, the mAP result collapse. The conv-add SCC block has got comparable results to the Faster RCNN[7], which means the additivity is important for global context plugging. The conv-add SCC make great improvements in several categories like "aeroplane" and "bike" .etc, and collapse in several categories like "bus" and "train". We think this due to the selectable property, which means not all the object have strong object-scene co-occurrence. The final GA-SCC we proposed is a relax the conv-add SCC context information and turn the hard add into soft gated add. It can be regard as the trade-off between the original model and the straightforward conv-add model. The mAP result shows that the GA-SCC is effective.

4. CONCLUSION

Global context information is very import for object detection, especially for some object categories that appear in some fixed scenes. In this paper, we test several skip context connection blocks, and we proposed three principles for building a global context block that can embed context information more effectively. We draw the following contributions: (1) categorize the existed context information plugged in object detection models, (2) propose three principles for building a skip context connection blocks, which includes **effective**, **selectable**, and **additivity**. We test our GA-SCC method in PASCAL VOC 2007, and the results show the GA-SCC can make 1% improvement to the Faster RCNN[7].

methods	data	MAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
FRCNN[7]	07	69.9	70.0	80.6	**70.1**	**57.3**	49.9	**78.2**	80.4	82.0	52.2	75.3	67.2	80.3	79.8	75.0	76.3	39.1	68.3	67.3	**81.1**	67.6
concat	07	70.6	73.0	79.0	66.2	54.7	51.7	77.4	82.4	**84.4**	**53.5**	73.5	67.3	80.0	83.0	76.6	76.8	43.6	**70.4**	**67.7**	78.6	71.8
add	07	63.3	68.7	72.8	63.0	50.9	42.6	67.7	77.6	75.3	41.8	74.2	59.2	64.8	78.5	71.0	70.4	32.3	63.9	57.6	69.4	64.7
conv-add	07	70.0	**75.3**	**80.6**	69.2	55.1	51.4	75.8	81.6	84.3	50.3	76.0	67.6	**81.2**	81.6	74.1	76.9	36.9	69.7	65.1	75.5	70.7
GA-SCC	07	**71.0**	74.6	79.9	68.7	56.5	**52.4**	77.3	**82.8**	83.0	52.5	**76.2**	**69.3**	79.2	**83.3**	**76.8**	**77.1**	**44.1**	70.3	65.9	78.7	**72.0**

Table 2. Detection result on PASCAL VOC 2007 test set (trained on PASCAL VOC 2007 trainval set only). The Faster RCNN[7] is use the VGG16 as backbone.

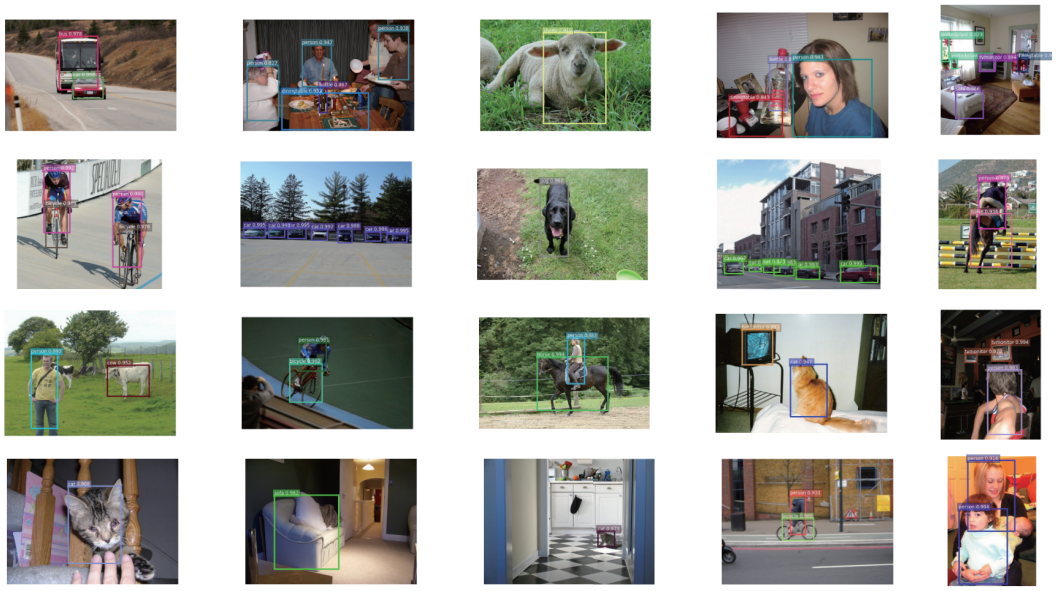

Fig. 4. Some example in PASCAL VOC 2007 test set, the model is trained on PASCAL VOC 2007 training set only. The annotation of box is category label and softmax score during $[0, 1]$. Here, we use threshold 0.8 to display the result.

In the future, we will test and improve our model based on other datasets like VOC2012 and MS COCO. Due to the size of VOC 2007, we just import two convolutional layers in the current framework. As a quick improvement, we will stack more convolutional layers for larger datasets. Furthermore, we will try to integrate the global and local context together in a unified framework. Here, we also raise an insightful thought based on this work: can we connect multiple SCCs to further refine the feature representation, since the input and output of the SCC possess the same data structure and similar semantic contents.

5. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China under Project No. 61472103.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings*

of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[6] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[10] J Long, E Shelhamer, and T Darrell, "Fully convolutional networks for semantic segmentation," in *Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[11] Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, et al., "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2403–2412.

[12] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[13] Alexander Toshev and Christian Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1653–1660.

[14] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, "Structured feature learning for pose estimation," in *CVPR*, 2016.

[15] Xingyu Zeng, Wanli Ouyang, Junjie Yan, Hongsheng Li, Tong Xiao, Kun Wang, Yu Liu, Yucong Zhou, Bin Yang, Zhe Wang, et al., "Crafting gbd-net for object detection," *arXiv preprint arXiv:1610.02579*.

[16] Spyros Gidaris and Nikos Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1134–1142.

[17] Pedro O. Pinheiro, Tsung Yi Lin, Ronan Collobert, and Piotr Dollr, *Learning to Refine Object Segments*, Springer International Publishing, 2016.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," *arXiv preprint arXiv:1612.03144*, 2016.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, pp. 541–551, 1989.

[21] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *arXiv preprint arXiv:1512.01274*, 2015.