Multi-kernel non-local neural network for semantic segmentation

Shengmin Yang, Huichao Sun, Mingzhu Zhang, Zhonggui Sun*
School of Mathematical Sciences, Liaocheng University, Liaocheng 252059, Shandong, China

ABSTRACT

As a milestone in semantic segmentation, non-local block (NLB) efficiently enhances the ability of regular convolutional neural networks in capturing long-range dependencies. From the view of mathematical modeling, NLB is based on a single Gaussian kernel. Existing works suggest that multi-kernel methods generally get more powerful performance in edge detection, which is crucial to image segmentation. Motivated by this consideration, we design a Multi-kernel Non-local Block (MKNLB). As expected, the proposed MKNLB exhibits excellent behaviors when being used in semantic segmentation. Additionally, with the distributive law of matrix multiplication, the complexity of its implementation is comparable to that of the standard NLB. Theoretical analyses and preliminary experiments on benchmark datasets both support the same conclusions.

Keywords: Image segmentation, deep neural network, self-attention, non-local block, multi-kernel non-local block

1. INTRODUCTION

As an important and challenging topic in computer vision, semantic segmentation marks semantic labels for every pixel in the image. Like many other applications, semantic segmentation has achieved impressive progress in recent years benefitting from the success of Deep Neural Networks (DNNs)[1,2]. Shelhamer et al.[3] proposed Fully Convolutional Network (FCN), which is a pioneering work in semantic segmentation using DNNs. Since then, the FCN-based approach[2,4] has been used in various segmentation scenarios. Relying on learnable convolutions, this kind of method can capture rich semantic information. However, the results are still not satisfactory. An important reason is that the localization of the convolution operation cannot utilize the global information in the image under study.

To address this problem, inspired by the technique in Natural Language Processing (NLP), Wang et al.[5] designed a simple and efficient Non-local Block (NLB) that combines non-local means[6] and CNN successfully. The work first introduced self-attention into computer vision and thus becomes a milestone in semantic segmentation. Meanwhile, the building block, i.e., NLB, can be plugged into many existing DNNs to improve their performance in applications. Thus, researchers began to devote more and more attention to it. The subsequent works mainly focus on reducing the complexity of the block[7,8].

In this paper, motivated by an earlier work[9] about the General Non-Local denoising model based on Multi-kernel-induced Measures (GNLMKIM), we design a novel non-local block called Multi-kernel Non-local Block (MKNLB). With multi-kernel strategy, MKNLB detects edges more efficiently and thus gets more powerful performance in segmentation. Additionally, with the distributive law of matrix multiplication, the computational burden of MKNLB is comparable to that of the standard NLB.

The effectiveness of the method is investigated using two benchmark semantic segmentation datasets (Cityscapes[10] and ADE20K[11]). With the indicator of mean intersection over union (mIoU), our approach significantly outperforms the methods using the standard NLB.

2. RELATED WORKS

2.1 Multi-kernel model for non-local means

Non-local Means (NLM)[6] is a classical filter that utilizes the dissimilarity measure between patches to operate in a non-local area (even the entire image). Actually, the mathematical model for non-local means is not unique. For example, in Reference[9], GNLMKIM employs multi Gaussian kernels to define the measure and applies Shannon regularizer to balance the linear relationship between various kernels. The specific model can be defined by:

* altlp@hotmail.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125066P
 · 0277-786X · doi: 10.1117/12.2662018

$$\begin{aligned} &\min_{x,\lambda} \quad \sum_{i\in\Omega}\sum_{j\in\Omega}\left(1-\sum_{t=1}^{k}\lambda_t G_t\left(x_i,x_j\right)\right)+p\sum_{t=1}^{k}\left(\lambda_t \ln\left(\lambda_t\right)\right) \\ &s.t. \quad \lambda_t \geq 0, \quad \sum_{t=1}^{k}\lambda_t = 1. \end{aligned} \tag{1}$$

where, i and j denote the pixel positions from the definition domain Ω of image x; x_i and x_j are the two corresponding image patches; $G_t(t=1,\ldots,k)$ is the Gaussian kernel used to measure the similarity between the two patches. Naturally, $1-\sum_{t=1}^{k}\lambda_t G_t(x_i,x_j)$ becomes the dissimilarity between patches. λ_t can be viewed as the importance of the single kernel G_t. p represents the regularization parameter that trades off the two terms of the model. As declared in Reference[9], the outputs of NLM can be derived from the optimization model under the single-kernel case. Meanwhile, with multi-kernel methods, the filters derived from the above model usually get more powerful ability in edge detecting. It is the fact that motivates us to modify NLB with multi-kernel strategy.

2.2 Non-local block

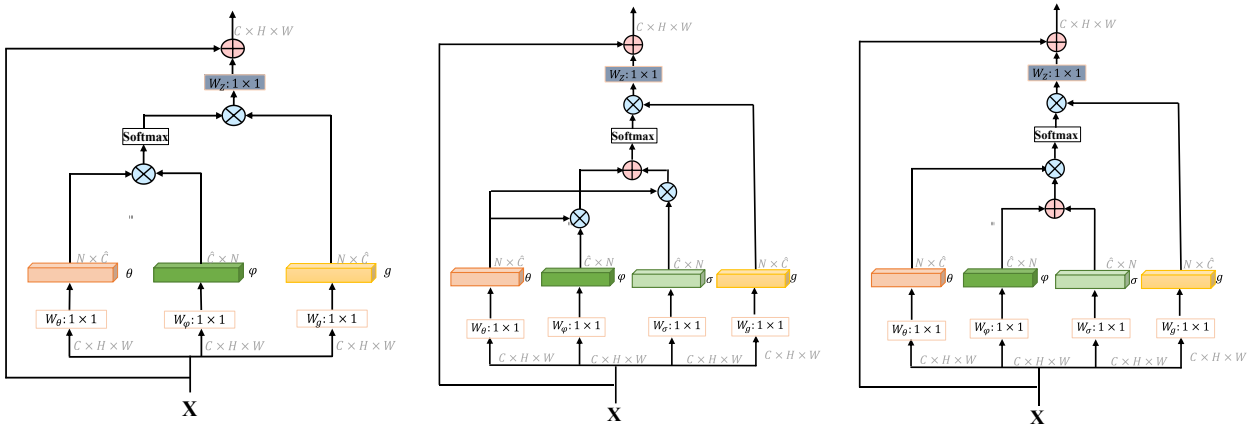
Non-local block, i.e., NLB[5], captures the long-range dependencies of pixels and thus becomes key to semantic segmentation. Specifically, the block is defined as:

$$z_i = x_i + W_z \sum_{j\in\Omega}\frac{f(x_i,x_j)}{S(x_i)}(W_g x_j) \tag{2}$$

where $f(x_i,x_j)$ denotes the similarity between position i and j in the input and $S(x)$ is the normalization factor. W_z and W_g are two 1 × 1 convolutions respectively.

For the similarity f, we take embedded Gaussian as an example to describe the definition in detail. Under this condition, $f(x_i,x_j)=e^{\theta(x_i)^T\varphi(x_j)}$, $\theta(x_i)=W_\theta x_i$, $\varphi(x_j)=W_\varphi x_j$, $\theta\in\Re^{\hat{C}\times H\times W}$, $\varphi\in\Re^{\hat{C}\times H\times W}$. W_φ and W_θ are two 1 × 1 convolutions. $\hat{C}$, H, W respectively indicate their channel number, input width and input height.

For an input $X\in\Re^{C\times H\times W}$ (C indicates the input channel number), the standard NLB with embedded Gaussian is shown in Figure 1a. Here, we intend to design a multi-kernel version of NLB, compared with the original one, which can get more powerful performance when being used in semantic segmentation.



(a) (b) (c)

Figure 1. (a): Architecture of a standard NLB; (b): Initial definition of MKNLB; (c): Architecture of MKNLB. $N=H\times W$.
Note: ⊗: Matrix multiplication; ⊕: Element-wise sum.

3. MULTI-KERNEL NON-LOCAL BLOCK

Here, we first give the definition of our multi-kernel non-local block (i.e., MKNLB) in Section 3.1. Then, by analyzing the complexity, its efficient implementation is designed in Section 3.2.

3.1 Initial definition

As aforementioned, the standard NLB is based on the single Gaussian kernel and the existing work indicated that a multi-Gaussian kernel generally gets better behaviors in edges[9]. The fact motives us to extend NLB to its multi-kernel version (named MKNLB) to improve the ability in segmentation.

For simplicity, we take two kernels in our MKNLB. The definition is

$$z_i = x_i + W_z \sum_{j \in \Omega} \frac{f_1(x_i, x_j) + f_2(x_i, x_j)}{S(x_i)} (W_g x_j) \tag{3}$$

where $f_1(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)}$, $f_2(x_i, x_j) = e^{\theta(x_i)^T \sigma(x_j)}$ are the two Gaussian kernels, $\sigma(x_j) = W_\sigma x_j$, W_σ is a 1 × 1 convolution, other symbols all have the same meanings as those in NLB formulated in equation (2).

For clarity, the definition of MKNLB is illustrated in figure 1b. Its effectiveness in semantic segmentation will be validated in the experiments part. Considering the multi-kernel strategy may increase the computational burden at first glance, we design an efficient implementation of MKNLB next, whose complexity is comparable to the standard NLB.

3.2 Efficient implementation

As shown in figure 1a, the similarity calculation with matrix multiplication, i.e., $\theta(x_i) \times \varphi(x_j)$, is the main computational burden in non-local block. Similarly in Reference[7], the operation can be simplified and expressed as:

$$\Re^{N \times \hat{C}} \times \Re^{\hat{C} \times N} \rightarrow \Re^{N \times N} \tag{4}$$

where $N = H \times W$. Therefore, this multiplication of matrices has a complexity of $O(\hat{C}N^2)$.

In our MKNLB defined in Figure 1b, due to the participation of the two kernels, the similarity part becomes $\theta(x_i) \times \varphi(x_j) + \theta(x_i) \times \sigma(x_j)$. At first glance, the computational burden may increase. Fortunately, it can be reduced with the distributive law of matrix multiplication. The specific expression is changed as follows:

$$\Re^{N \times \hat{C}} \times \Re^{\hat{C} \times N} + \Re^{N \times \hat{C}} \times \Re^{\hat{C} \times N} \rightarrow \Re^{N \times \hat{C}} \times (\Re^{\hat{C} \times N} + \Re^{\hat{C} \times N}) \tag{5}$$

Equation (5) indicates that, MKNLB can be implemented by performing the two-matrix addition first and then followed by one time matrix multiplication. Considering the former complexity is distinctly lower than the latter. Therefore, this implementation of MKNLB can be approximated as $O(\hat{C}N^2)$. That is, the complexity of our MKNLB is comparable to that of the standard NLB. For clarity, the implementation is illustrated in Figure 1c.

4. EXPERIMENTS

To evaluate the MKNLB, we conduct experiments for semantic segmentation on two benchmark datasets: Cityscapes[10] and ADE20K[11].

4.1 Datasets and evaluation metrics

Cityscapes: It consists of 5000 images from 50 different cities belonging 19 categories. In order to facilitate training, validation, and testing, the images have been divided into 2975, 500, and 1525 segments, respectively

ADE20K: The dataset contains 20210 images in the training dataset with 150 semantic classes, 2000 images make up the validation set, while 3352 make up the test set. As well known, the dataset is particularly challenging in semantic segmentation datasets due to complex scenarios.

Metrics: The mean intersection over union (mIoU) is used to evaluate all datasets.

4.2 Training details

During training, our code follows a standard frame from the semantic segmentation open resource library MMSegmentation[12]. Two Quadro Rtx 6000 GPUs are used for all experiments. We apply stochastic gradient descent (SGD) with the weight decay is 0.0005. The initial learning rate $\gamma_0 = 0.01$ is decayed following the poly learning rate policy, where γ_0 is multiplied by $1-\left(\frac{iter}{iter_{max}}\right)^{0.9}$. For Cityscapes, we set the batch size is 4 and randomly crop the input images to 512×512. For ADE20K, we set the batch size is 8 and randomly crop the input images to 512×1024. For the two datasets, we choose random flip and scale these images within [0.5, 2]. For all experiments, we select the pre-trained ResNet-101 as backbone framework.

4.3 Comparisons with other methods

In this section, we will analyze the results of the two datasets (Cityscapes[10] and ADE20K[11]) for semantic segmentation.

On one hand, we compare the proposed multi-kernel non-local network with five other methods on the Cityscapes validation set. Table 1 is shown the experiment outcomes of mIoU numbers. We trained all methods for 8K iterations. Based on the same backbone, the multi-kernel non-local network attains 77.59% mIoU. It can be observed that 2.68% mIoU better than the original non-local network. We also find that our method performs better than the previous methods by more than 0.71% mIoU.

On the other hand, we compare the performance of our method on the ADE20K validation set. The outcomes of mIoU numbers are shown in Table 2. We trained 8K iterations in this dataset to compare the performance with other methods. As we all know, the dataset is challenging to train due to a variety of image sizes, complex semantic information, and the difference between training and validation sets. Despite under this condition, our method also achieves 41.35% mIoU. It can still be 1.03% better than the original non-local network and also defeat other listed methods.

Table 1. Comparisons with the state-of-the-arts on the Cityscapes validation set.

Method	Backbone	mIoU (%)
CCNet[14]	ResNet-101	76.31
GCNet[8]	ResNet-101	76.52
DNL[13]	ResNet-101	76.8
ANN[7]	ResNet-101	76.88
NLB[5]	ResNet-101	74.91
Ours	ResNet-101	**77.59**

Table 2. Comparisons with the state-of-the-arts on the ADE20K validation set.

Method	Backbone	mIoU (%)
GCNet[8]	ResNet-101	39.7
DNL[13]	ResNet-101	39.81
ANN[7]	ResNet-101	41.09
NLB[5]	ResNet-101	40.32
Ours	ResNet-101	**41.35**

5. CONCLUSION

In this paper, we design an efficient block called Multi-kernel Non-local Block (MKNLB) for semantic segmentation. In

contrast to the standard Non-local block (NLB), the proposed MKNLB detects edges more efficiently and thus gets better performance when being used in image segmentation. Meanwhile, with the distributive law of matrix multiplication, we design an efficient implementation of MKNLB, whose complexity is comparable to the standard NLB. The segmentation experiments conducted on benchmark datasets (Cityscapes and ADE20K) validated its effectiveness. For future work, we would like to expand the applications of MKNLB to further vision tasks

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 11801249; the Nature Science Foundation of Shandong Province China under Grant ZR2020MF040, and in part by the Open Project of Liaoch eng University under Grant 319312101-01.

REFERENCES

[1] Badrinarayanan, V., Kendall, A. and Cipolla, R., "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481-95 (2017).
[2] Chen, L. C., Papandreou, G., Kokkinos, I., et al., "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Transactions on Pattern Analysis and Machine 40, 834-48 (2018).
[3] Shelhamer, E., Long, J. and Darrell, T., "Fully convolutional networks for semantic segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 3431-3440 (2017).
[4] Lin, G., Milan, A., et al., "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 1925-34 (2017).
[5] Wang, X., Girshick, R., He, K., et al., "Non-local neural networks," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 7794-803 (2018).
[6] Buades, A., Coll, B. and Morel, J. M., "A non-local algorithm for image denoising," Proc. IEEE Conf. on Computer Vision and Pattern Recognition vol 2, 60-5 (2005).
[7] Zhu, Z., Xu, M., Bai, X., et al., "Asymmetric non-local neural networks for semantic segmentation," IEEE/CVF Int. Conf. on Computer Vision, 593-602 (2019).
[8] Cao, Y., Xu, J., Lin, S., et al., "GCNet: Non-local networks meet squeeze-excitation networks and beyond," Proc. IEEE/CVF Int Conf. on Computer Vision Workshops, (2019).
[9] Sun, Z., Chen, S. and Qiao, L., "A general non-local denoising model using multi-kernel-induced measures," Pattern Recognit. 47, 1751-63 (2014).
[10] Cordts, M., Omran, M., Rehfeld, T., et al., "The cityscapes dataset for semantic urban scene understanding," Pro. IEEE Conf. on Computer Vision and Pattern Recognition, 3213-23 (2016).
[11] Zhou, B., Zhao, H., Puig, X., et al., "Scene parsing through ADE20K dataset," Pro. IEEE Conf. on Computer Vision and Pattern Recognition, 5122-30 (2017).
[12] MMSegmentation Contributors. MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark. (2020). https://github.com/open-mmlab/mmsegmentation
[13] Yin, M., Yao, Z., Cao, Y., et al., "Disentangled non-local neural networks," European Conference on Computer Vision, 191-207 (2020).
[14] Huang, Z., Wang, X., Huang, L., et al., "Ccnet: Criss-cross attention for semantic segmentation," Proc. of the IEEE/CVF Inter. Conf. on Computer Vision, 603-12 (2019).