Journal of Information Technology Research
Volume 15 • Issue 1

Research of Self-Attention in Image Segmentation

Fude Cao, Shandong Institute of Commerce and Technology, China*
Chunguang Zheng, Shandong Institute of Commerce and Technology, China
Limin Huang, Shandong Institute of Commerce and Technology, China
Aihua Wang, Shandong Institute of Commerce and Technology, China
Jiong Zhang, Shandong Institute of Commerce and Technology, China
Feng Zhou, Shandong Institute of Commerce and Technology, China
Haoxue Ju, Shandong Institute of Commerce and Technology, China
Haitao Guo, Shandong Institute of Commerce and Technology, China
Yuxia Du, Shandong Institute of Commerce and Technology, China

ABSTRACT

Although the traditional convolutional neural network is applied to image segmentation successfully, it has some limitations. That's the context information of the long-range on the image is not well captured. With the success of the introduction of self-attentional mechanisms in the field of natural language processing (NLP), people have tried to introduce the attention mechanism in the field of computer vision. It turns out that self-attention can really solve this long-range dependency problem. This paper is a summary on the application of self-attention to image segmentation in the past two years.

KEYWORDS

Convolutional Neural Networks, Image Segmentation, Self-Attention

I. INTRODUCTION

In the field of computer vision, image segmentation is a very important basic research direction. In general, image segmentation is to divide the pixels in the image into different parts (add different labels) according to certain rules. There are super pixels segmentation, semantic segmentation, instance segmentation, and panoramic segmentation (Ren and Malik, 2003). This paper mainly refers to semantic segmentation and instance segmentation. The former is to assign a category label to each pixel in the image (for example, cars are blue, buildings are brown, etc.). So even different people are represented by the same color, without distinguishing individuals in the same class. But the latter instance segmentation method is similar to object detection, but the output of instance segmentation is a mask instead of a bounding box. Instance segmentation does not need to label each pixel. So it only needs to find the edge contour of the object of interest and distinguish individuals.

We know that the beginning of image segmentation using deep learning is FCN (Shelhamer et al., 2017). The principle is to modify the classification convolution neural network (such as ResNet or VGG network, etc.) into a fully convolution network. FCN first enlarges the resolution of the picture, then through a series of convolution operations, and does an average pooling to the n × n feature map

DOI: 10.4018/JITR.298619

*Corresponding Author

and finally upsamples to obtain the final prediction image. Because this network consists entirely of convolutional layers, we call it a fully convolutional network. However, a network consisting entirely of convolutional layers has a big problem which is even a large convolution kernel will only have a small perceptual domain in its implementation. However, the segmentation tasks require a very large perceptual domain. In order to effectively increase the perceptual domain, there are many convolutional networks with dilation (Deeplabs v1 (Chen et al., 2017), v2 (Chen et al., 2018), v3 (Chen et al., 2017), and v3+ (Chen et al., 2018)) and multiple pooling networks PSPNet (Zhao et al., 2017). However, neither of these two methods really establish the connection between every pixel in the image, especially the connection between long-distance pixels.

At this time, the attention mechanism has achieved very good results in NLP. So people thought of introducing attention mechanism to computer vision. Attention mechanism is the first to imitate the internal process of human observation behavior which is a mechanism to align internal experience with external sensation, thereby increasing the observation precision of some areas. The basic idea is to let the system learn to pay attention focusing only on important information and ignoring irrelevant information. So it can quickly extract important features of sparse information (Hu, 2020). It is therefore widely used in NLP tasks, especially in machine translation. The self-attention mechanism is an improvement of the attention mechanism, which reduces the dependence on external information for capturing the internal correlation of features effectively. Although the self-attention mechanism is first proposed elsewhere, it is widely used in machine translation from the paper "Attention Is All You Need". The self-Attention here is also called Transformer (Vaswani et al., 2017). In computer vision, self-attention is to perform autonomous learning between feature maps and automatically assign weights (Mnih et al., 2014). Because in image segmentation context information is very critical, self-attention can provide useful and effective solutions to context modeling, especially the context of remote pixels.

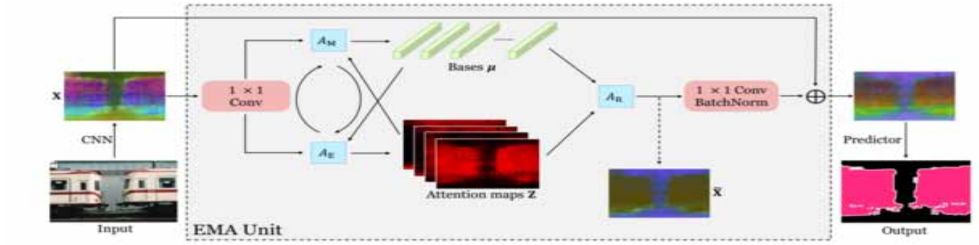
The following will introduce a series of papers on self-attention mechanism application in the field of image segmentation for solving the problem of long-range dependence. Based on this, we consider whether the self-attention module may replace the convolution module. Finally this survey is to demonstrate that the self-attention module can replace the convolution module, thus pointing out the direction for future research. Attention in this paper refers to self-attention.

II. APPROACH

1. *Attention is All You Need* (Vaswani et al., 2017)

(Vaswani et al., 2017) is the first work to propose the use of self-attention mechanism to replace recurrent neural networks in sequence models, and has achieved great success. One important module is the scaled dot-product attention module. The paper proposes that triples (Key, Query, Value) are a modeling method for capturing long-distance dependencies. As shown in the following Figure 1, Key and Query obtain the corresponding attention weights by dot product, and finally the obtained weights dot product with Value to get the final output.

Figure 1. (left) Scaled Dot-Product Attention. (right) Multi-Head Attention. (Vaswani et al., 2017)

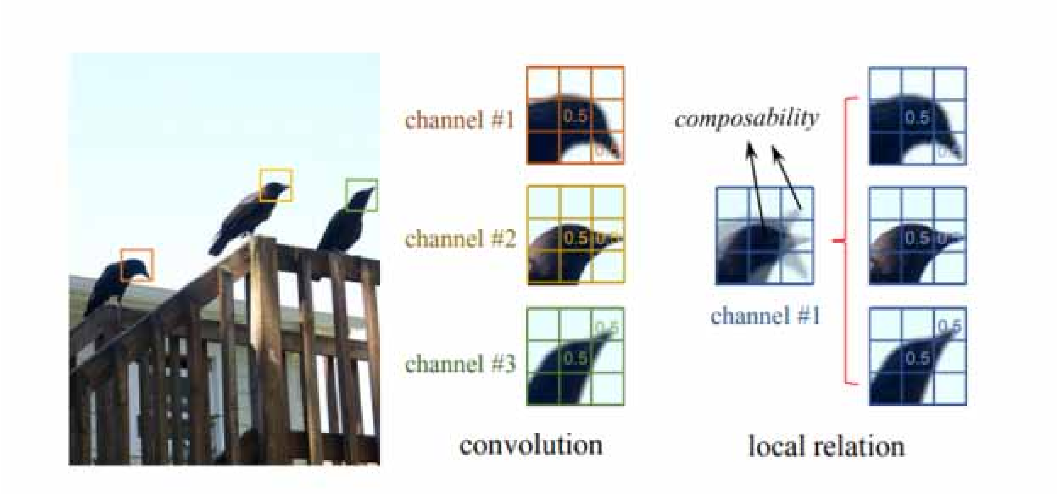


The above is for NLP domain. In the field of image segmentation, self-attention refers to the three matrices Q (Query), K (Key), and V (Value) from the same input (image). Query is a concept to be retrieved, and Key is equivalent to the Key in the web library or dictionary. Use the set where Key is to express Query. And Value is a feature transformation of all Keys. According to Query, a weighted summation of Value is performed using the similarity between Query and Key to obtain the feature after Query. We can find the attention is doing the non-local weighted average operation actually.

2. Import Attention: Non-Local (Wang et al., 2017)

In the field of computer vision, the most successful introduction of Attention is Non-local neural networks (Wang et al., 2017), which simply uses (Vaswani et al., 2017) for video classification tasks, thus capturing the long-range dependence that cannot be captured by convolution. For 2D images, long-range dependence refers to the relationship weight of any pixel in the image to the current pixel, while for 3D video, it refers to the relationship weight of all pixels in all frames to the pixels of the current frame. As shown in Figure 2, in order to guess what object the ball is in, we need to refer to the information of each position. Therefore, we hope that the feature of x_i can be updated to the relationship with all pixels in all frames.

Figure 2 A spacetime non-local operation (Wang et al., 2017)



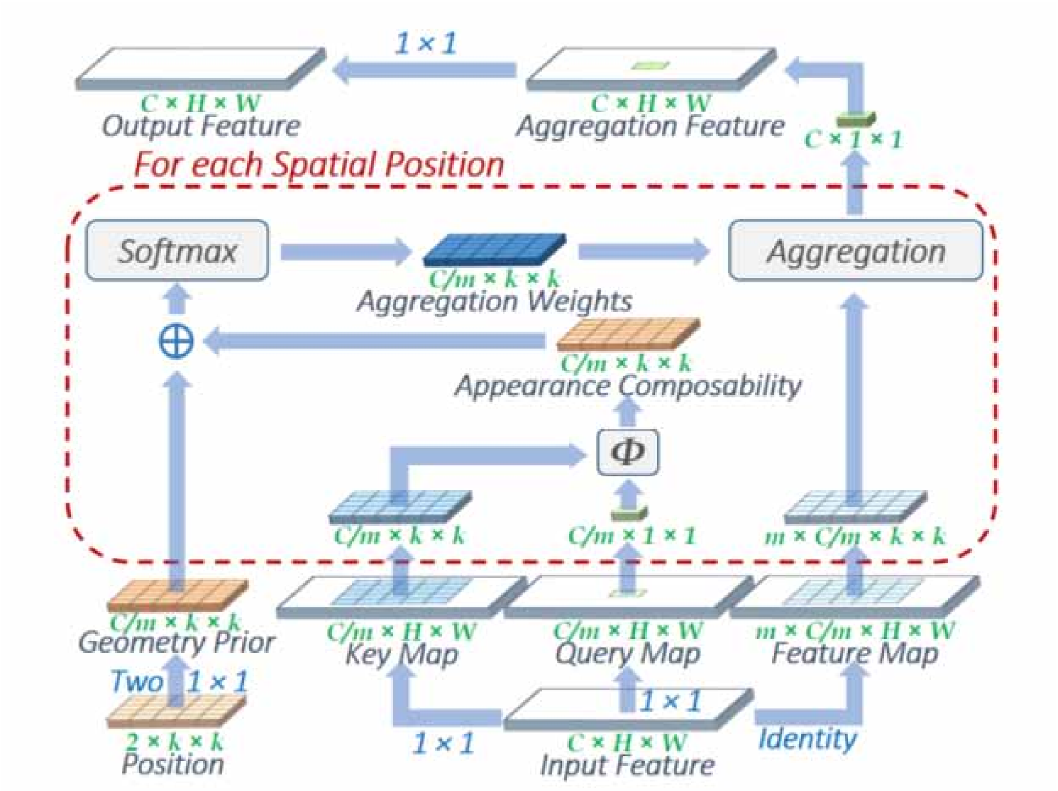
The author uses the following equation (1) to calculate the relationship between x_i and all other pixels x_j. The f function models the relationship between x_i and x_j. C(x) is a normalization of f, and $g(x_j)$ is a transformation of the pixel x_j which we want to refer to. Since we are updating x_i pixel, refer to x_j of all pixels. Here f uses the Embedded Gaussian method (formula 2) as the similarity function (there are three other choices in the paper). The specific implementation process is shown in Figure 3. θ is called query, which is a 1×1×1 convolution of the original image x, φ is called key, is also a 1×1×1 convolution, and they both do dot products, and then through softmax function, and then through g(called value) transformation of x dot product. This is a residual connection (Equation 3). From the above, we can see that non-local is a de-noising method for image (Buades et al., 2005). The computational complexity here is O(NNC), N is the number of pixels in the image (W×H). Floating-point operations (FLOPs) will be very large if there are a lot of pixels in the image. Experimental results show that the accuracy of video classification and image segmentation is greatly improved.

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{1}$$

$$f(x_i, x_j) = e^{\theta(x_i)^T \varphi(x_j)} \tag{2}$$

$$z_i = W_z y_i + x_i \tag{3}$$

Figure 3 A spacetime Non-local block (Wang et al., 2017)



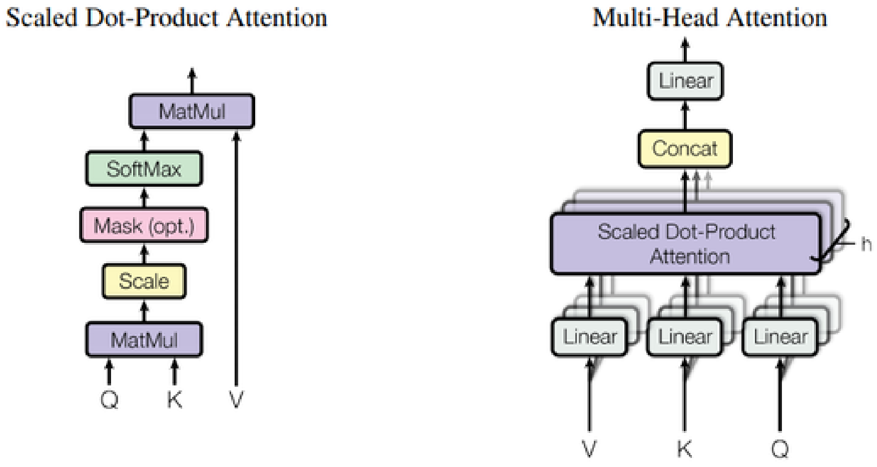
Other neural networks, such as PSANet (Zhao et al., 2018), DANet (Fu et al., 2020), OCNet (Yuan, et al., 2018) and CCNet (Huang et al., 2020), are neural networks which import attention mechanisms or are related to non-local networks ideas. Here PSANet (Zhao et al., 2018) is related to non-local networks, and introduces the self-attention mechanism in a similar way. First, a correlation between x_i and x_j is modeled (Equation 4). The features of xi and x_j are considered, as well as the correlation between their positions. After the similarity is normalized, x_j is averaged as an update to x_i. The process is divided into two parts: the first part is called Collect, and the second part is called Distribute (Equation 5). Function F of Collect is simply calculated as the distance between the pixels related to x_i only. Distribute is only calculated the distance associated with x_j and the distance between it and x_i. i refers to the place where it is collected, and j refers to the place from which it started. The final correlation is expressed as the sum of the two.

$$z_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F(x_i, x_j, \Delta_{ij}) x_j \tag{4}$$

$$z_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(x_i) x_j + \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(x_j) x_j \tag{5}$$

PSANet contains two attentions, equivalent to two heads in transformer (Vaswani et al., 2017). The two channels play the roles of collect and distribute respectively. As shown in the Figure 4, the channel above is Collected, and the channel below is Distribute. First, the input image of a very high-dimensional image (such as 2048-dimensional image) should be reduced in lower dimension (for example 512-dimensional), and then transformed into a (2H-1)×(2W-1) dimensional matrix through 1×1×1 convolution, and then the operation Collect (or Distribute) attention generation should be performed. Testing in different data sets (such as PASCAL VOC, Cityscapes dataset etc.), the segmentation accuracy is better.

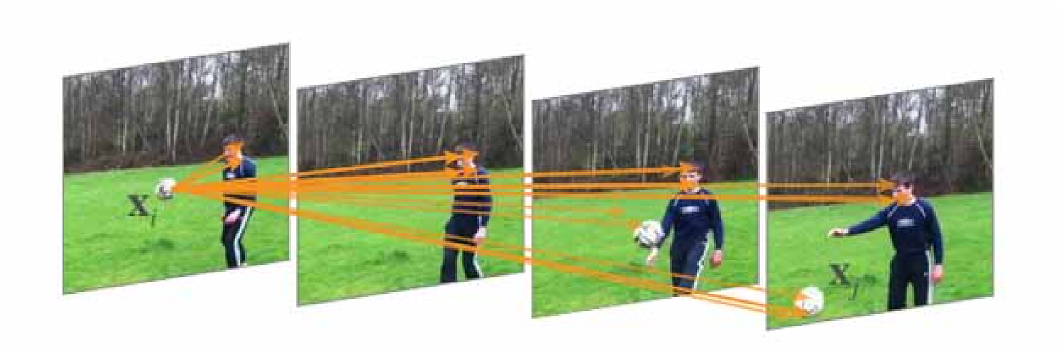
Figure 4 The structure of PSA network module (Zhao et al., 2018)



3. Optimize Attentional: A^2-Nets *(Chen et al., 2018)*

As mentioned above, the disadvantage of non-local, which first introduces self-attention mechanism, is that the computational complexity is too large. Although several papers mentioned above can be said to optimize non-local to varying degrees, none of them is ideal. The most important contribution to the optimization of attention in mathematical form is A^2-nets (Chen et al., 2018), and the "Low-rank Reconstruction" thinking of mapping and reflection in this paper has inspired a lot of exploration. The idea of A^2-nets is that a graph is made up of several key elements, and the authors want to first discover these combination factors and then map them back. If the number of image pixels is denoted as N and the number of combination factors ("Global Descriptors" in Figure 5) is denoted as K, then K here is significantly less than N, so the image reconstructed with K will have the effect of sparse low-rank. As shown in figure 5, we first learn an Attention map and Attention vectors (double Attention) from the original feature map, and then map them back for refactoring.

Figure 5 An A^2-Nets block (Chen et al., 2018)



We can intuitively see from Figure 6 that the figure on the left shows the computational complexity O(NNC) of non-local, while the figure on the right shows the computational complexity O(NCC) of A^2-nets. Since the number of C is much smaller than the number of N, there is a significant difference in their computational complexity. The whole process can be divided into two steps: a process of compression from HW to K ("Global Descriptors" in Figure 5), and the process of anti-compression from K to HW. The K here is different for different images, so it is an adaptive process. So A^2-nets are very widely used.

Because A^2-net (Chen et al., 2018) proposed the double attention block and inspired the thinking on how to do the mapping and reflection. Relevant research neural networks include SGR (Liang et al., 2018), beyond Grids (Li et al., 2018), GloRe (Chen et al., 2020), LatentGNN (Zhang et al., 2019), APCNet (He et al., 2020) and EMANet (Li et al., 2019). EMANet (Li et al., 2019) proposed EM algorithm to solve mapping and reflection. EM algorithm is the maximum likelihood solution used to solve the hidden variable model. In Figure 7, the hidden variables are regarded as the mapping matrix, and the model parameters are K descriptors. E updates mapping matrix of attention maps Z, M updates descriptor μ. After T iterations, the feature graph is reconstructed by using the transpose of the mapping matrix (and normalization) as the reflection matrix, as shown in Figure 7. EM algorithm itself guarantees convergence to the local maximum value of the likelihood function, so the descriptors and mapping relations iterated by EM algorithm are more guaranteed to meet representativeness than those simply learned through network. It can be said that EMANet has done the best in the field of image semantic segmentation currently.

Figure 6 (left) Non-local block computational process. (right) A²-nets block computational process (Wang et al., 2018) (Chen et al., 2018)

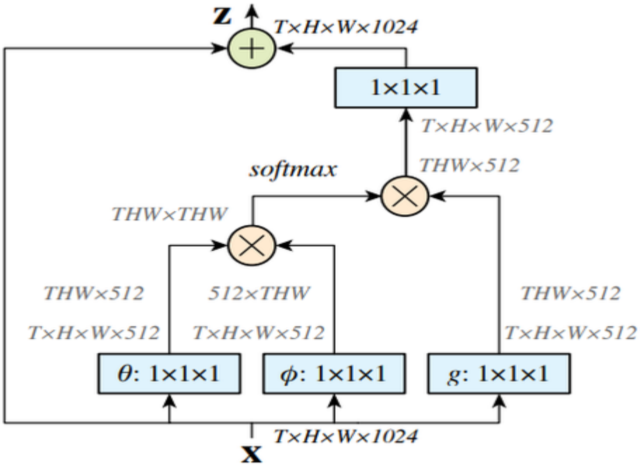
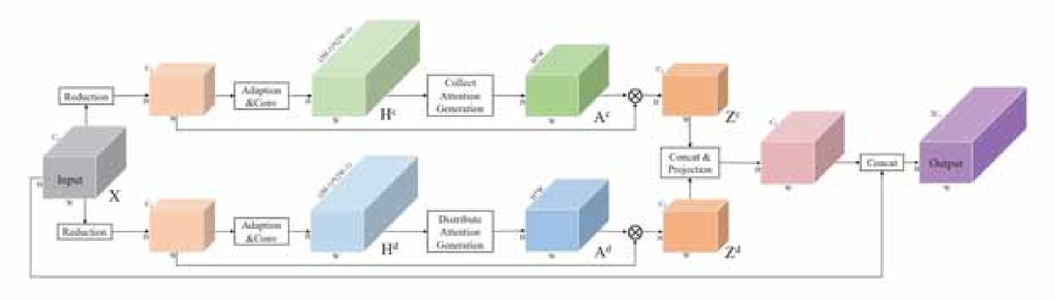


Figure 7 The structure of EMA Unit (Li et al., 2019)

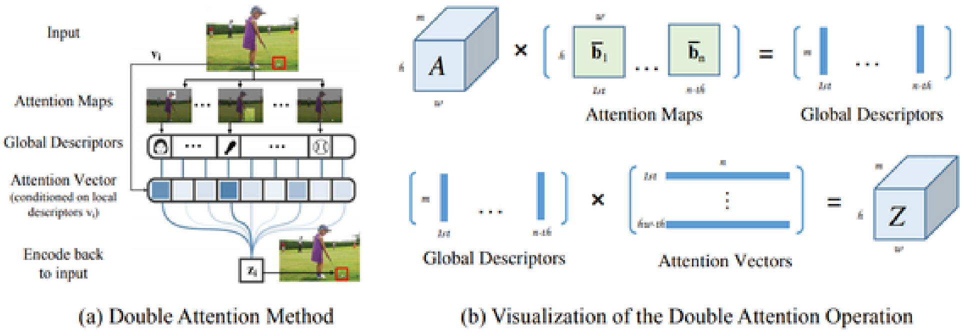


4. Attention Replaces Convolution: LR-Net (Hu et al., 2019)

The goal of Non-Local is to supplement the convolutional networks, hoping to complement the deficiencies of global information in convolutional neural network modeling. But the approach of Local Relation Network (LR-Net) (Hu et al., 2019)is to replace convolution, which requires that LR-net be a relational network (Battaglia et al., 2018). Under the premise of relational network, the learning scope is required to be local, because only local relationship can build the information bottleneck, and then learn the patterns in the picture well through limited data because only the bottleneck can force it to learn some patterns. The second is to introduce geometric priors, because the geometric position relationships in vision are very important, and that's the most important mode of operation of convolution. The third is that the self-attention module calculates the similarity of the key/ query through the vector dot product, but the author further finds that the value of the key/ query is also a scalar not need to be a vector which can save many parameters and calculations. In this way, we can build multi groups of relationship modules within a limited budget. The conceptual difference between the Local Relation Layer and the Convolution Layer is that the former has computational composability, instead of using a global convolution template, as shown in Figure 8.

The local relation layer replaces all spatial convolution layers, thereby generating a LR-Net without a spatial convolution layer. The 1 × 1 convolution here is actually a transformation of features and cannot be called a convolution. As shown in Figure 9, the local relation layer uses the standard Query, Key, and Value triples for feature correlation calculations. Geometry Prior is designed for spatial correlations, and this operator boldly replaces 3 × 3 convolutions. While saving parameters, it also has denaturation such as rotation. Its performance is not inferior to convolutional neural networks (such as ResNet (He et al., 2016)), and it is slightly better.

Figure 8 Illustration of convolution layer and local relation layer (Hu et al., 2019)

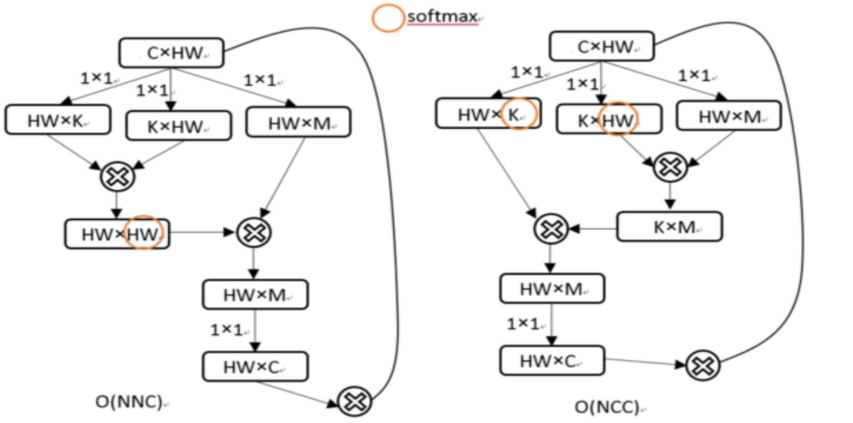


(a) Double Attention Method

(b) Visualization of the Double Attention Operation

In the AACnet (Bello et al., 2020) paper, the authors investigated the problem of using self-attention (as an alternative to convolution) for discriminative visual tasks. The author proposes a new two-dimensional relative self-attention mechanism, and research shows that this is sufficient to replace convolution with self-attention as a separate module on image classification tasks. However, the authors found in controlled experiments that the best results are obtained when convolution is combined with self-attention.

Figure 9 The local relation layer (using self-attention module, no spatial convolutional layer) (Hu et al. 2019)



III. DATA SETS

Image segmentation most commonly used data sets mainly introduce three: PASCAL VOC, MS COCO and Cityscapes. The first two are the most vital datasets for semantic segmentation (Chen et al. 2018), (Chilamkurthy et al. 2017).

The PASCAL VOC provides standardized image data sets for object class recognition. It also provides a common set of tools for accessing the data sets and annotations, enables evaluation and comparison of different methods. So it can be evaluated the performance of various image segmentation methods by running a challenge on this dataset (Everingham et al.,2015). The PASCAL VOC is a relatively old data set, it provides 20 categories, including people, cars and so on. There are 6,929 labeled pictures, which provide class-level labeling and distance-level labeling. Most images in this dataset have a foreground or two surrounded by highly diverse backgrounds. It implicitly leads to bias towards algorithms containing detection techniques.

The MS COCO is a large-scale object detection, segmentation, and captioning dataset (Cordts et al., 2016). This is by far the largest semantic segmentation dataset. It provides 80 categories, and more than 330,000 pictures, of which 200,000 pictures are marked. The number of object instances in the entire data set exceeds 1.5 million. The latest papers are all experiments on this dataset, because this is the most difficult and the most challenging data set.

The Cityscapes Dataset focuses on semantic understanding of urban street scenes (Cordts et al., 2016). It has 30 detailed categories. Five thousand of the images are finely annotated to the pixel level. There are also 20,000 images with rough markings. It can also provide class-level segmentation and distance-level segmentation.

IV. DISCUSSION

Here we use the experimental results to deeply understand and think about the self-attention mechanism. Table 1 shows that the segmentation accuracy on the PASCAL VOC dataset has been greatly improved after using the self-attention module. But EMANet's thinking based on mapping and reflection is deeper, so the effect of low-rank reconstruction is better. As for the comparison on the COCO and Cityscapes data set, please refer to the papers (Wang et al., 2018) (Zhao et al., 2018) in detail, because in the papers introduced in this paper, there is no horizontal comparison between them, so they are not compared here.

Table 1 Comparison on the PASCAL VOC test set

Approach	Backbone	mIoU	Source
Wide ResNet (He et al., 2016)	WideResNet-38	84.9	Reported in the paper (Li et al., 2019)
PSANet (Zhao et al., 2018)	ResNet-101	85.7	Reported in the paper (Zhao et al., 2018) (Li et al., 2019)
EMANet (Li et al., 2019)	ResNet-101	87.7	Reported in the paper (Li et al., 2019)

(mIoU: Mean Intersection over Union. First, IoU = Area of Overlap /Area of Union Second, averaged over each category)

From Table 2, we find that ResNet, Non-local and A^2-net parameter quantities have little difference, but the computational complexity is vastly different. Especially when ResNet-26+NL has 4@Conv3&4 blocks, it can be said that FLOPs will double (from 9.3G to 21.3G), but ResNet-26+A^2 only increases less than 1.5G. Their image segmentation accuracy is not much different.

Next from Table 3 we can see that it is possible to replace the convolution with a Non-Local block in a relational network, and the accuracy can exceed the convolutional neural network from Table 3. But if it is not in a relational network, simply using Non-local (NL-26 in Table 3) will not work well. This shows that using self-attention to replace convolution as a separate module is currently feasible, but it is not good, and it needs to be operated under a relational network or in conjunction with convolution (Bello et al., 2020).

Table 2 Comparisons between performance from multiple nonlocal blocks and multiple double attention blocks on Kinetics dataset (Kay et al., 2017). In order to be suitable for image segmentation, we only focus on top-1 clips accuracy. (Chen et al., 2018)

Approach	+N Blocks	#params	FLOPs	Clip@1(%)
ResNet-26 (He et al., 2016)	None	7.043M	8.3G	50.4
ResNet-26+NL (Wang et al., 2018)	1 @Conv4	7.312M	9.3G	51.7
	2 @Conv4	7.581M	10.4G	52.0
	4 @Conv3&4	7.719M	21.3G	52.4
ResNet-26+A^2 (Chen et al., 2018)	1 @Conv4	7.312M	8.7G	52.3
	2 @Conv4	7.581M	9.2G	52.5
	4 @ Conv3&4	7.719M	10.1G	53.0

Table 3 Comparison Local relation networks with non-local neural networks (Hu et al., 2019)

Approach	top-1	top-5	# params	FLOPs	
ResNet-26 (He et al., 2016)	72.8	91.0	16.0M	2.6G	
NL-26 (Wang et al., 2018)	47.7	72.1	17.3M	2.6G	
LR-Net-26 (Hu et al., 2019)	75.7	92.6	14.7M	2.6G	
LR-Net-26-NL(Hu et al., 2019)	76.0	92.8	37.1M	5.6G	

V. CONCLUSION

When convolutional neural networks and recurrent neural networks became popular, attention mechanisms appeared. Not only can the attention module be embedded in any layer, but also the spatial convolution layer can be replaced with the attention-based component block, so as to build a large-scale pre-trained model. In the field of image segmentation, according to the papers in recent years, we can see that from non-local neural networks, self-attention non-local modules have been embedded, and A^2-Nets has optimized self-attention of non-local networks until replace the convolutional layer with non-local blocks in the relational network. We believe that the self-attention module shines in the field of computer vision including image segmentation, so we suggest that the future research may be focus on the self-attention module.

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Battaglia, P. W., Hamrick, J. B., & Bapst, V. (2018). *Relational inductive biases, deep learning, and graph networks*. arXiv:1806.01261

Bello, I., Zoph, B., Le, Q., Vaswani, A., & Shlens, J. (2020). Attention Augmented Convolutional Networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Buades, A., Coll, B., & Morel, J. M. (2005). A non-local algorithm for image denoising. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on.* IEEE. doi:10.1109/CVPR.2005.38

Chen, Y., Rohrbach, M., & Yan, Z. (2020). Graph-Based Global Reasoning Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Chen, Y., Kalantidis, Y., Li, J., Yan, S., & Feng, J. (2018). A2-Nets:Double attention networks. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).*

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. doi:10.1109/TPAMI.2017.2699184 PMID:28463186

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation*. arXiv:1706.05587

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *European Conference on Computer Vision: Computer Vision – ECCV 2018, 11211*, 833-851. doi:10.1007/978-3-030-01234-2_49

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. *Computer Science*, (4), 357–361.

Chilamkurthy, S. (2017). *A 2017 Guide to Semantic Segmentation with Deep Learning*. Academic Press.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/CVPR.2016.350

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2015, January). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*(1), 98–136. doi:10.1007/s11263-014-0733-5

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., & Fang, Z. (2020). Dual Attention Network for Scene Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

He, J., Deng, Z., Zhou, L., Wang, Y., & Qiao, Y. (2020). Adaptive Pyramid Context Network for Semantic Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/CVPR.2016.90

Hu, D. (2020). *An Introductory Survey on Attention Mechanisms in NLP Problems*. Intelligent Systems and Applications. doi:10.1007/978-3-030-29513-4_31

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2020). CCNet: Criss-Cross Attention for Semantic Segmentation. In *2019 International Conference on Computer Vision (ICCV).* arXiv:1811.11721v2

Kay, W., Carreira, J., Simonyan, K., Zhang, B., & Zisserman, A. (2017). *The kinetics human action video dataset*. arXiv:1705.069502

Xia, L., Zhong, Z., Wu, J., Yang, Y., Lin, Z., & Hong, L. (2019). *Expectation-Maximization Attention Networks for Semantic Segmentation*. arXiv:1907.13426

Yin & Gupta. (2018). Beyond Grids: Learning Graph Representations for Visual Recognition. Advances in Neural Information Processing Systems, 31.

Liang, X., Hu, Z., Hao, Z., Liang, L., & Xing, E. P. (2018). Symbolic Graph Reasoning Meets Convolutions. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018).*

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, *3*. arXiv1406.6247

Ren, X., & Malik, J. (2003). Learning a Classification Model for Segmentation. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. doi:10.1109/ICCV.2003.1238308

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651. doi:10.1109/TPAMI.2016.2572683 PMID:27244717

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. arXiv:1706.03762.

Wang, G. Gupta, & Kaiming. (2018). Non-local neural networks. In *Computer Vision and Pattern Recognition (CVPR)*. arXiv:1711.07971

Yuan, Y., & Wang, J. (2018). *OCNet: Object Context Network for Scene Parsing*. arXiv:1809.00916.

Zhang, S., Yan, S., & He, X. (2019). *LatentGNN: Learning Efficient Non-local Relations for Visual Recognition*. arXiv:1905.11634

Zhao, H., Yi, Z., Shu, L., & Shi, J. (2018). PSANet: Point-wise spatial attention network for scene parsing. *European Conference on Computer Vision: Computer Vision – ECCV 2018, 11213*, 270-286.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid Scene Parsing Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. doi:10.1109/CVPR.2017.660