Dual attention network for Scene Segmentation

Jun Fu [1,3] Jing Liu* [1] Haijie Tian [1] Yong Li [2]
Yongjun Bao [2] Zhiwei Fang [1,3] Hanqing Lu [1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Business Growth BU, JD.com [3] University of Chinese Academy of Sciences
{jun.fu,jliu,zhiwei.fang,luhq}@nlpr.ia.ac.cn,hjtian_bit@163.com,{liyong5,baoyongjun}@jd.com

Abstract

文章建立在自注意力机制上，提出了Dual Attention Network去集成局部特征和全局依赖。分别为空间和通道模块，并且都是扩展在FCN的上。在Cityscapes, PASCAL Context and COCO Stuff 上达到新的sota，其中在Cityscapes上达到81.5%的IoU分数

Contribution

- We propose a novel Dual Attention Network (DANet) with self-attention mechanism to enhance the discriminant ability of feature representations for scene segmentation.

- A position attention module is proposed to learn the spatial interdependencies of features and a channel attention module is designed to model channel interdependencies. It significantly improves the segmentation results by modeling rich contextual dependencies over local features.

- We achieve new state-of-the-art results on three popular benchmarks including Cityscapes dataset , PASCAL Context dataset and COCO Stuff dataset .

Related Work

Semantic Segmentation

FCN Deeplabv2 Deeplabv3 PSPNet DAG-RNN PSANet EncNet

Self-attention Modules

Attention is all you need(nisp2017)

第一次提出了self-attention的注意力机制

non-local neural network(cvpr2018)

将self-attention运用到cv领域

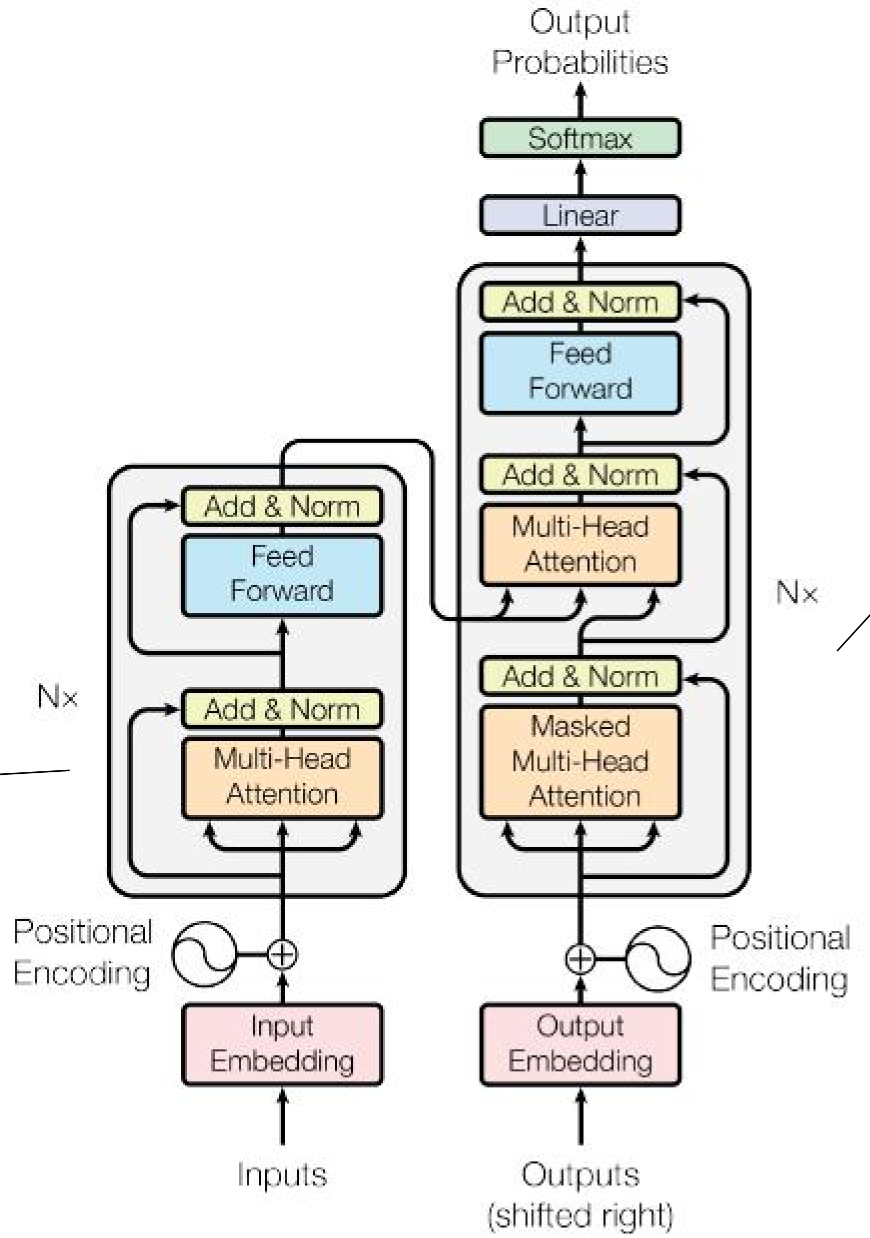
Self Attention

Model Architecture

提出了Transformer描绘输入和输出之间的全局依赖

Encoder: The encoder is composed of a stack of N = 6 identical layers. Each layer has two
sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, position-wise fully connected feed-forward network

Encoder

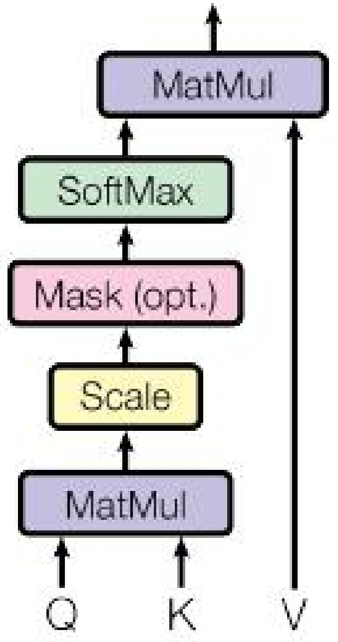


Decoder

Self Attention

Attention

Scaled Dot-Product Attention



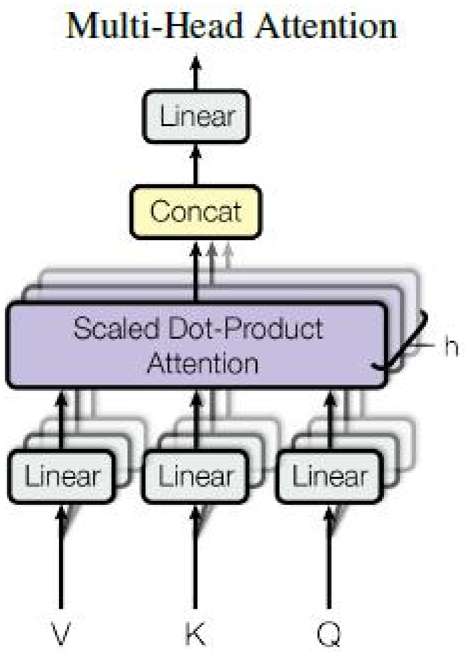
Input : Q,K是dk维， V是dv维

Output :

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Self Attention

Attention



在多个通道上使用Scaled Dot-Product Attention然后进行Concat

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head_1}, ..., \mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

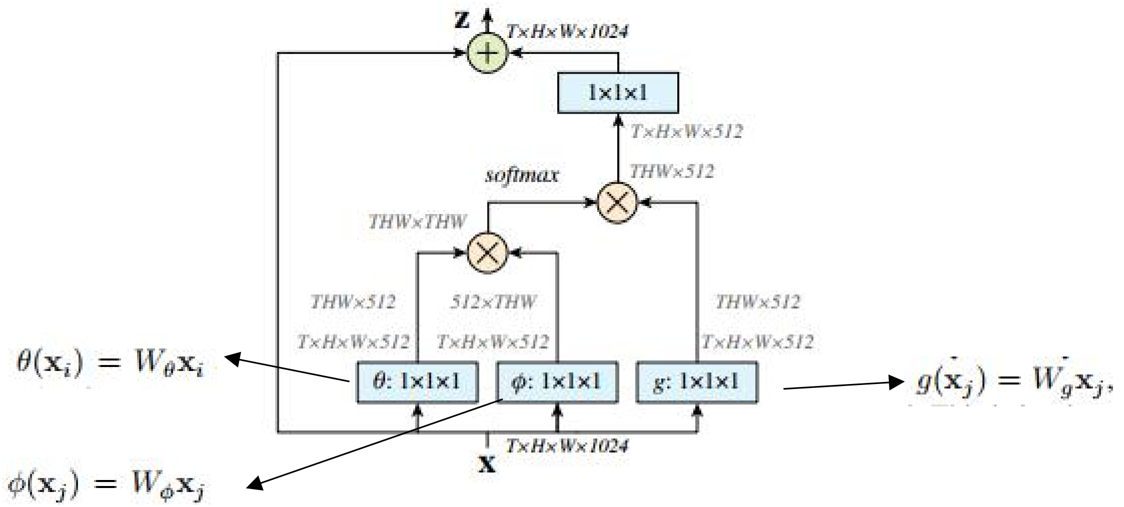
Self Attention

Non-local Neural Network将Self-Attention的机制运用到cv领域

文章提出了关于f选择的四个版本
Gaussian、Embedded Gaussian、Dot product、Concatenation

Dot product

$$f(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \phi(\mathbf{x}_j).$$

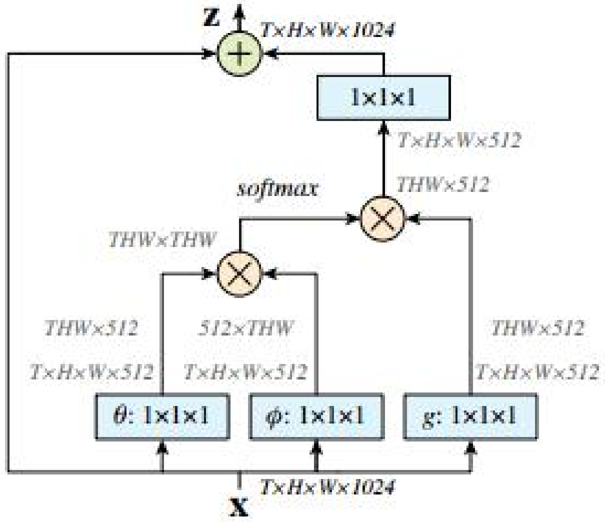


A spacetime non-local block

Self Attention

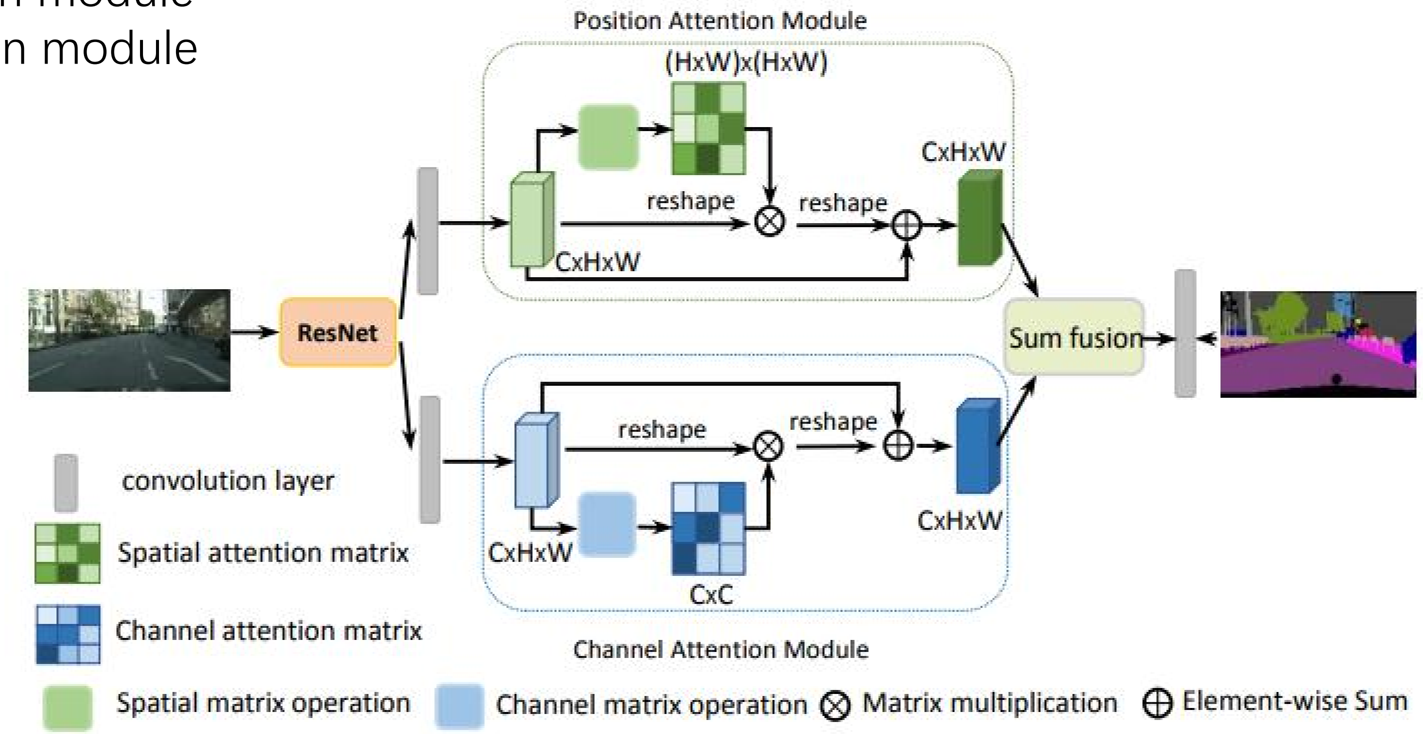
Non-local Block

$$\mathbf{z}_i = W_z \mathbf{y}_i + \mathbf{x}_i,$$



DANet

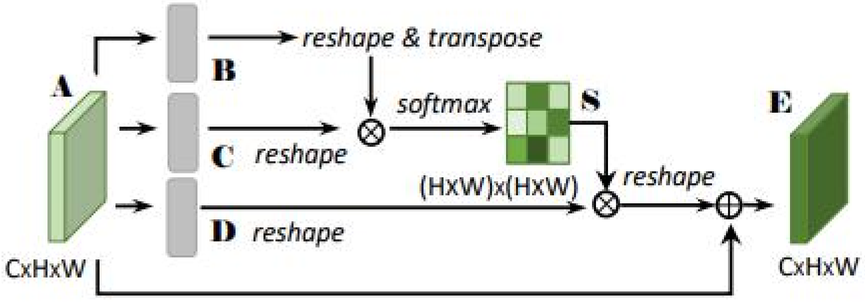
1.Position attention module
2.Channel attention module



DANet

Why?
Spatial dependcies between any two positions regardless of distance

- The first step is to generate a spatial atten- tion matrix which models the spatial relationship between any two pixels of the features
- Next, we perform a matrix multiplication between the attention matrix and the original features.
- Third, we perform an element-wise sum operation on the above multiplied resulting matrix and original features to obtain the final representations reflecting long- range contexts.

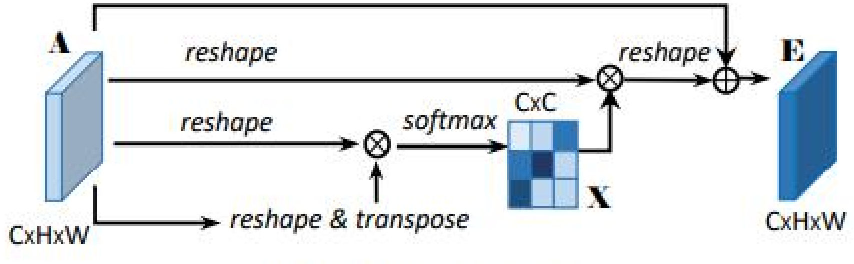


A. Position attention module

① 特征图A通过1x1卷积得到B、C、D，通道数变成原来的1/8
② B(CXHXW)通过转置reshape操作变成（NXC）然后和reshape后的C（CXN）点乘在通过softmax激活得到S（NXN），Dreshape（CXN）点乘S在reshape成（CXHXW）
③ 引入尺度系数α(初始值为0，通过训练学习乘得到的特征图，再加上A得到最后的输出E

DANet

Why?
Each channel map of high level features can be regarded as a class-specific response, and different semantic responses are associated with each other



B. Channel attention module

① 对A做reshape(CxN)和reshape与transpose(NxC)
② 将得到的两个特征图相乘，再通过softmax得到channel attention map X(C×C)
③ 接着把X的转置(CxC)与reshape的A(CxN)做矩阵乘法，再乘以尺度系数β，再reshape为原来形状，最后与A相加得到最后的输出E
④ 其中β初始化为0，并逐渐的学习得到更大的权重

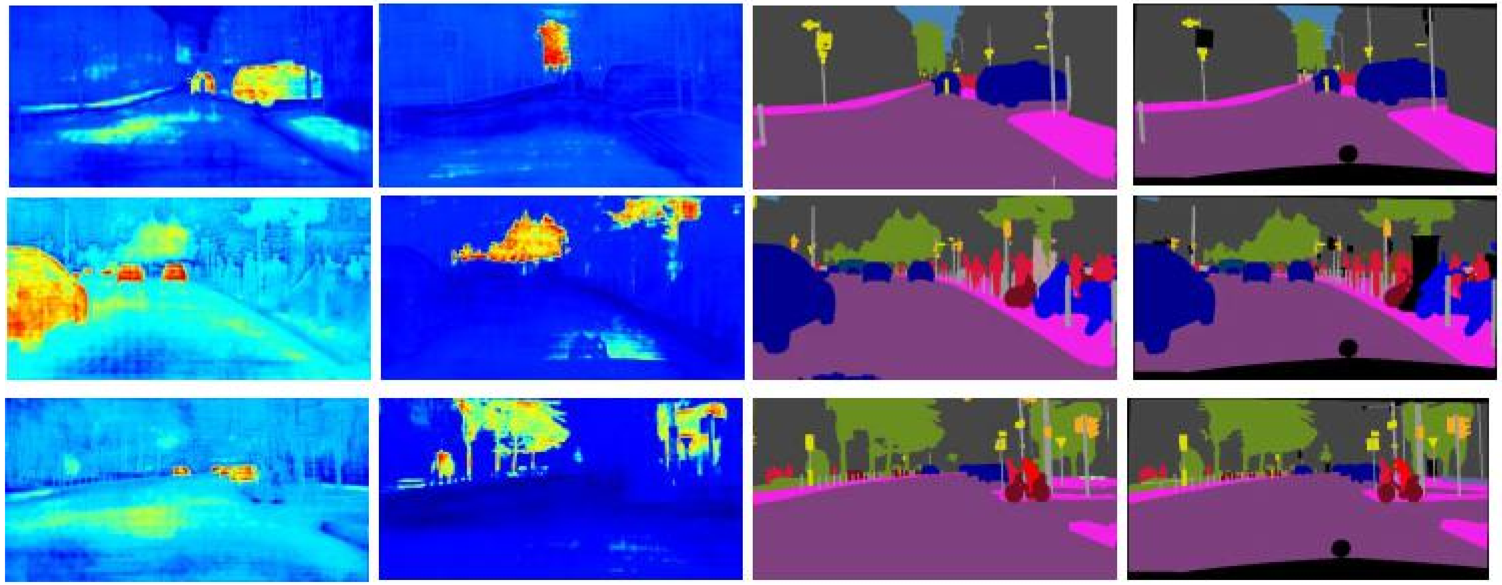
Results on Cityscapes Dataset

Method	BaseNet	PAM	CAM	Mean IoU%
Dilated FCN	Res50			70.03
DANet	Res50	✓		75.74
DANet	Res50		✓	74.28
DANet	Res50	✓	✓	76.34
Dilated FCN	Res101			72.54
DANet	Res101	✓		77.03
DANet	Res101		✓	76.55
DANet	Res101	✓	✓	77.57

Table 1: Ablation study on Cityscapes val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

Results on Cityscapes Dataset

Results on Cityscapes Dataset



Channel map #11 Channel map #4 Result Groundtruth

Results on PASCAL VOC 2012 Dataset

Method	**BaseNet**	**PAM**	**CAM**	**Mean IoU%**
Dilated FCN	Res50			75.7
DANet	Res50	✓	✓	79.0
DANet	Res101	✓	✓	80.4

Table 4: Ablation study on PASCAL VOC 2012 val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

Method	**Mean IoU%**
FCN [13]	62.2
DeepLab-v2(Res101-COCO) [3]	71.6
Piecewise [11]	75.3
ResNet38 [10]	82.5
PSPNet(Res101) [29]	82.6
EncNet (Res101) [27]	**82.9**
DANet(Res101)	82.6

Table 5: Segmentation results on PASCAL VOC 2012 testing set.

Results on Other Dataset

Method	Mean IoU%
FCN-8s [13]	37.8
Piecewise [11]	43.3
DeepLab-v2 (Res101-COCO) [3]	45.7
RefineNet (Res152) [10]	47.3
PSPNet (Res101) [29]	47.8
Ding et al.(Res101) [6]	51.6
EncNet (Res101) [27]	51.7
Dilated FCN(Res50)	44.3
DANet (Res50)	50.1
DANet (Res101)	**52.6**

Table 6: Segmentation results on PASCAL Context testing set.

Method	Mean IoU%
FCN-8s [13]	22.7
DeepLab-v2(Res101) [3]	26.9
DAG-RNN [18]	31.2
RefineNet (Res101) [10]	33.6
Ding et al.(Res101) [6]	35.7
Dilated FCN (Res50)	31.9
DANet (Res50)	37.2
DANet (Res101)	**39.7**

Table 7: Segmentation results on COCO Stuff testing set.

Thoughts

① 结合特征金字塔的思想，从不同尺度选取特征图进行position attention

② A采用更深的卷积得到B,C,D类似增加残差深度