Dual-stream-network-for-visual-recognition

1、Motivation

transformer具有很强的全局建模能力，但是缺乏捕获局部模式的能力。为了解决这一问题作者提出了可以计算细粒度特征并高效融合的dual stream network。

过往的工作有许多将CNN引入Transformer来提升局部建模能力的方法。（CvT中将linear projection换成CNN，ContNet在token maps上做卷积）但这存在一些问题：1、卷积-注意力交替或将linear换成卷积的操作可能不够好；2、CNN和attention功能上的冲突可能会影响训练效果；3、attention不一定能在高分辨率的feature map上很好地捕获信息；4、纯attention的计算复杂度很高，降采样降低复杂度的操作也丢失了一部分局部信息。

2、Approach

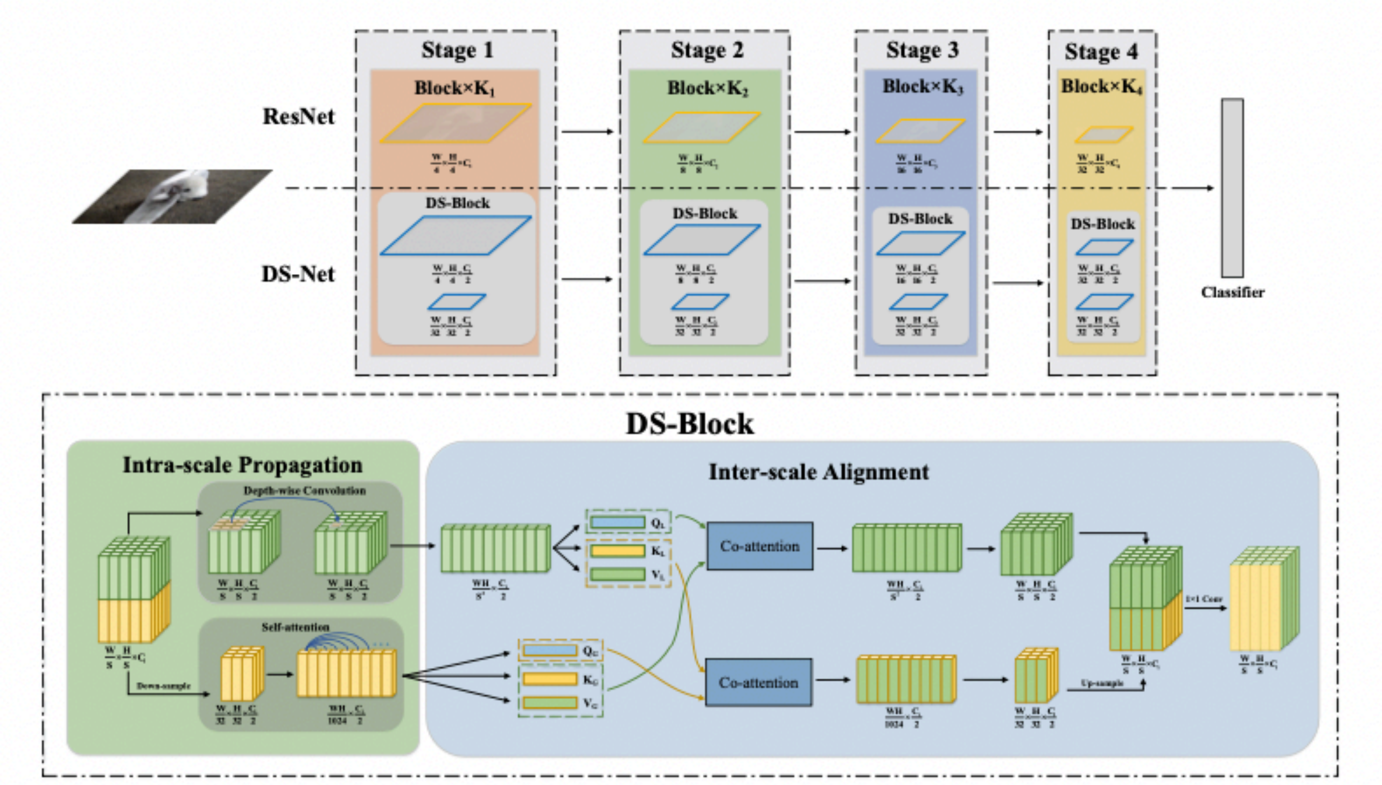


Figure 2: Illustration of the proposed DS-Net, including Intra-scale Propagation module and Inter-scale Alignment module. Compared to ResNet, in which only single resolution is processed, our DS-Net, instead, generates dual-stream representations via DS-Blocks.

2.1 overview

看完整体框架以后感觉其思路是沿用downsample的方式，将self-attention作用的分辨率控制在1/32 of the image降低计算复杂度并捕获global信息，而用dw卷积作用于更高的分辨率，捕获local信息。然后接一个co-attention模块实现跨尺度信息的融合。

比较好奇的是down sample是怎么做的，以及每个stage 的channel size如何确定。

2.2 Intra-scale propagation

将当前stage的输入feature map分成f_l和f_g两部分，该模块具体实现如下：

local representation

对于$f_l \in \mathbb{R}^{W_i \times H_i \times C_l}$，用3 x 3 DW卷积提取局部特征

$$f_L(i,j) = \sum_{m,n}^{M,N} W(m,n) \odot f_l(i+m, j+n),$$

这个地方原文对公式的表述可以学一下的。

global representation

首先将f_g flatten 到 $l_g = \frac{W}{32} \times \frac{H}{32}$长度，然后对之采用self-attention。

$$f_Q = f_g W_Q, \quad f_K = f_g W_K, \quad f_V = f_g W_V,$$

$$f_G = \mathrm{softmax}(\frac{f_Q f_K^T}{\sqrt{d}}) f_V,$$

其中$d = \frac{\frac{C_i}{2}}{N}$，N是self-attention的头数

2.3 Inter-scale Alignment

之所以需要有这样的一个模块，是因为作者发现global feature和local feature所关注的内容并不匹配，简单地将其融合（concatenation、element-wise addition和production）可能无法捕获两者的深层关系。

首先将获取的f_L拉平成时序特征，然后采用下面的方式实现特征融合：

$$\begin{aligned} Q_L &= f_L W_Q^l, \quad K_L = f_L W_K^l, \quad V_L = f_L W_V^l, \\ Q_G &= f_G W_Q^g, \quad K_G = f_G W_K^g, \quad V_G = f_G W_V^g, \end{aligned}$$

$$W_{G \to L} = \mathrm{softmax}(\frac{Q_L K_G^T}{\sqrt{d}}), \quad W_{L \to G} = \mathrm{softmax}(\frac{Q_G K_L^T}{\sqrt{d}}).$$

$$h_L = W_{G \to L} V_G, \quad h_G = W_{L \to G} V_L,$$

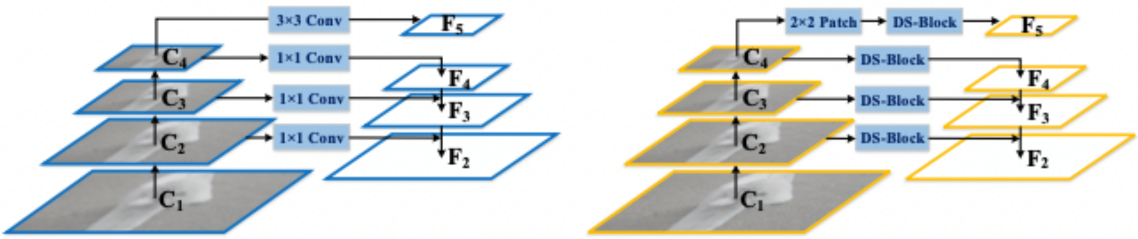
这个地方比较容易理解，值得注意的是两个co-attention模块的参数并不共享。

2.4 dual-stream feature pyramid network

Previous methods often cause large extra memory and computation costs, due to their complicated architectures and utilized high resolution feature maps

同时，全局的nonlocal又会丢失一些局部信息导致小物体检测效果不好。

因此，作者将dual-stream的设计引入FPN



(a) FPN (b) DS-FPN

Figure 3: The architecture of DS-FPN. C_i denotes the feature maps in stages from backbone, and F_i denotes the reconstructed features for detection and segmentation.

3、Experiment

三种结构：

Table 1: Detailed settings of DS-Net. Dconv denotes 3 × 3 depth-wise convolution, and MHSA denotes multi-head self-attention. C_i denotes the number of channels in ith stage. The feature dimension expansion ratio of each block is set to 4.

Stage	Input size	DS-Net-T	DS-Net-S	DS-Net-B
Stage 0	224×224	4×4, 64, stride=4, padding=0		
Stage 1	56×56	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_1=64\end{bmatrix}\times 2$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_1=64\end{bmatrix}\times 3$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_1=64\end{bmatrix}\times 3$
Stage 2	28×28	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_2=128\end{bmatrix}\times 2$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_2=128\end{bmatrix}\times 4$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_2=128\end{bmatrix}\times 4$
Stage 3	14×14	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_3=320\end{bmatrix}\times 4$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_3=320\end{bmatrix}\times 8$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_3=320\end{bmatrix}\times 28$
Stage 4	7×7	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_4=512\end{bmatrix}\times 1$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_4=512\end{bmatrix}\times 3$	$\begin{bmatrix}\text{Dconv}\\ \text{MHSA}-8\\ C_4=512\end{bmatrix}\times 3$
	7×7	global average pooling, 1000-d fc, softmax		

3.1 Ablation study

3.1.1 ratio of local to global feature

在前面的描述中，input feature被均匀分成两部分送入两个stream，如果不平衡地送入不同的stream会怎么样呢?

When α equals 0, only depth-wise convolution is performed, and when α equals 1, only self-attention is performed.

Table 2: DS-Net-T performance on ImageNet-1k validation set with different α.

α	0	0.25	0.5	0.75	1
Top-1(%)	77.1	78.0	**78.1**	77.9	77.6
Top-5(%)	93.3	**94.1**	**94.1**	94.0	93.9
Params (M)	8.6	8.7	9.1	9.8	10.7
FLOPs (G)	1.573	1.578	1.592	1.615	1.647
Throughput (Images/s)	3240	1733	1199	912	740

3.1.2 none is dispensable in DS-Block

Table 3: Ablations of removing components of DS-Net-T* on ImageNet-1k validation set.

Versions	DS-Net-T*	$w/o\ f_L$	$w/o\ f_G$	w/o $G \rightarrow L$	w/o $L \rightarrow G$	w/o $L \leftrightarrow G$
Top-1(%)	**79.0**	76.7	76.6	76.5	76.4	78.1
Top-5(%)	**94.8**	93.6	93.7	93.5	93.4	94.1

3.2 Image classification

Table 4: Comparison with the accuracy of other state-of-art methods on ImageNet-1k validation set. The input images are reshape to 224 × 224 resolution. DS-Net* represents the corresponding DS-Net version with Inter-scale Alignment module.

Method	Params (M)	FLOPs (G)	Throughput (Images/s)	Top-1 (%)
		ConvNet		
ResNet-18 [18]	11.8	2	-	69.9
ResNet-50 [18]	25.6	4.1	-	74.2
ResNet-101 [18]	44.5	7.8	-	77.4
RegNetY-8GF [35]	39.2	8	-	79.9
RegNetY-16GF [35]	83.6	15.9	-	80.4
		Transformer / Hybrid		
DeiT-T [42]	6	-	2536	72.2
CPVT-Ti [7]	6	-	-	72.4
T2T-ViT-12 [48]	6.9	-	-	76.5
ConTNet-S [47]	10.1	1.5	-	76.5
DS-Net-T (**ours**)	9.1	1.6	1199	**78.1**
DS-Net-T* (**ours**)	10.5	1.8	1034	**79.0** (+6.8)
DeiT-S [42]	22.1	4.6	940	79.9
CrossViT-15 [4]	27.4	5.8	640	81.5
T2T-ViT-14 [48]	22	5.2	-	81.5
ConTNet-M [47]	19.2	3.1	-	80.2
TNT-S [16]	23.8	5.2	-	81.3
CvT-13 [46]	20	4.5	-	81.6
PVT-Small [45]	24.5	3.8	820	79.8
CPVT-Small-GAP [7]	23	4.6	817	81.5
Swin-T [27]	29	4.5	766	81.3
DS-Net-S (**ours**)	19.7	3	582	**81.9**
DS-Net-S* (**ours**)	23	3.5	510	**82.3** (+2.4)
DeiT-B [42]	86	17.5	292	81.8
CrossViT-18 [4]	43.3	9	430	82.5
ConTNet-B [47]	39.6	6.4	-	81.8
PVT-L [45]	61.4	9.8	-	81.7
Swin-S [27]	50	8.7	437	83.0
DS-Net-B (**ours**)	48.8	7.6	387	**82.8**
DS-Net-B* (**ours**)	49.3	8.4	335	**83.1** (+1.3)

这里的DS-Net-B是没有Inter-scale -Alignment module的网络，带*的则是有的