Vision Transformer

1.ViT:An image is worth 16x16 words: Transformers for image recognition at scale(2020)

论文链接

本文提出了Vision Transformer，将transformer架构应用到图片分类问题，除了预处理不同，其余就是一个用于分类的transformer编码器

- 想感慨的是，这篇论文是除了Yolo v3那个技术报告外读的最顺畅的一个。一方面，ViT尽量不改变transformer结构（为了方便的直接使用nlp领域已经在硬件上高效实现的transformer结构），另一方面attention is all you need是我读的第一篇论文，读的很仔细，还印象深刻了属于是。
- 预处理：为了得到序列输入，将一张图片分割为多个patch，维度为**patch数量*(patch长*宽*通道数**)，将一个patch的特征作为一个token，且通过可训练的线性映射得到D维patch embedding；为了保留位置信息，ViT也使用了1维position embedding（2维效果没啥提升）；为了实现分类任务，在序列开始加入了一个可训练的[class]token，其最终状态作为分类的特征
- inductive bias:文中认为，CNN具有translation equivariance和locality等inductive bias（这是模型自身的一种先验），这是优点但也会受限（不如模型自己学习到）。transformer的优势在于inductive bias更少（只有MLP和position embedding），空间关系必须从头开始学，因此在大数据集上训练时优于CNN（更好的先验）。
- 微调：在微调时，remove预训练的分类头然后重新初始化进行训练。当训练的图像分辨率高于预训练时，为了保证预训练的position embedding有效，在保持patch-size不变的同时，根据patch的相对位置对embedding进行二维插值

- 论文中提到，当在中等数据集上训练时，transformer的表现不如CNN，但优势体现在数据集更大的时候。ViT通过在大型数据集上预训练，后微调得到了sota表现。
- 本文还提到一种混合模型，先用CNN提取patch的特征，再对其patch & position embedding作为输入

2.Swin Transformer: Hierarchical Vision Transformer using Shifted Windows(2021)

论文链接

本文提出了一种新的vision transformer结构Swin transformer，利用shifted window降低计算复杂度，并通过patch merge获得多尺度特征图，后续可以类似于FPN或U-Net方法进行dense prediction任务

- 背景：本文认为，将transformer应用于Vision时需要考虑两个域之间的两个差别，一为视觉实体具有不同的尺寸，二为视觉任务多需要高分辨率输入，而transformer对输入为平方复杂度。为了解决这两个问题，Swin transformer分别使用了层次特征图和计算局部self-attention的方法
- 结构：预处理与ViT类似，将图片分为patch后计算embedding（在这里无需加入position embedding），输入两个级联的Swin transformer block后进行patch merge，即令相邻的patch(2*2=4个)concatenation成4d张量后经线性层降为2d，从而使特征图长宽变为一半，相当于步长为2的下采样，将结果再输入两个级联的Swin transformer block，重复这个过程
- Swin transformer block包括级联的两部分，他们的多头自注意力层(MSA)不同。首先将输入第一个transformer block，其MSA为w-MSA，对每个无重叠的window(每个window包含M*M个patch)分别计算自注意力；将第一个block的结果输入第二个，其MSA为SW-MSA，对特征图进行shifted window分割，后对新的window（许多window尺寸小于M*M，具体看论文）计算自注意力
- w-MSA使自注意力的计算转为线性复杂度，SW-MSA建立w-MSA的不同window之间的关系，丰富了全局特征。
- 论文中提出了一种高效mask方法计算shifted window的自注意力，具体看论文
- 本文使用了相对位置偏差，在计算自注意力时加入。因为M^2与M^2的patch之间有(2M-1)*(2M-1)种相对位置关系（每个维度2M-1），所以训练一个(2M-1)*(2M-1)维度的bias矩阵，计算时从中取值即可

Semantic Segmentation

1.FCN:Fully Convolutional Networks for Semantic Segmentation(2015)

论文链接

本文使用全卷积网络实现了pixel-pixel,端到端的语义分割模型，还提出了一种利用多尺度特征的方法

- 背景：分类网络CNN取得了很好的效果，想迁移到语义分割任务——去掉分类层、将FC换为conv、加入上采样实现dense predict。非线性层使CNN只能接受固定尺寸的输入，而每个FC可以等效为一个卷积层，因此使用FCN可以接受任意尺寸的输入。
- 分类任务中，卷积网络在提取特征的过程中会不断地下采样，使特征图尺寸不断下降，这使top层的特征分辨率较低不适应于pixel-wise的语义分割任务，需要让分类网络适应dense predict。本文检验了overFeat中提出的shift-and-stitch方法（没使用），最终使用了上采样方法——反卷积/双线性插值（最后一次上采样将反卷积初始化为双线性插值，再学习），和pixel loss实现了dense predict
- 结合高分辨率浅层和低分辨率高层的语义特征，FPN应该是对此有所借鉴。在对top层（第五层）上采样32倍时，FCN-8s将第五层先2倍上采样再与经过1*卷积的第四层相连接，将结果2倍上采样，再与经过1*1卷积的第三层相加，将结果8倍上采样得到与原图尺寸一致的输出，从而结合了多个尺度的特征图。（如果融合更low的层收益递减）

- top特征图的通道数为C（类别数），因此相当于特征图的每个点为C维张量（每个类的得分），信息太少了！不利于后面大尺度的上采样，这在U-net中进行了改进，在上采样部分仍保留了丰富的特征通道

2.U-Net: Convolutional Networks for Biomedical Image Segmentation(2015)

论文链接

这是一篇用于医学图像的语义分割论文，但提出的U-net是一个广泛取得优秀结果的模型

- U-net也采用了全卷积网络，与FCN相似。先前向传递一个CNN获得下采样的一系列特征图，将top层特征图经过两个3*3卷积层后，进行一系列上采样(*2)，每次上采样后，将结果与**下采样过程中对应的特征图**裁剪后拼在一起(concatenation)，经过两个3*3卷积层后进行下一次上采样，最后一次上采样后使用1*1卷积层后的每个像素的分类。上采样和下采样过程比较对称，形成一个U型结构（论文中的图片很清晰）
- 一个比较重要的点。U-net中绝大部分使用的是3*3卷积层，没有pad！所以每经过一次卷积层，特征图尺寸都会-2，因为这个原因上采样和下采样对应的特征图尺寸有所区别，需要将下采样的特征图裁剪后concatenation。文中认为在边缘pad会使边缘像素的特征随深度增加而越来越模糊，特征图尺寸下降也与下述overlap-tile策略有关
- 也许是医学图像的问题（分辨率太大），也可能是当时的设备限制（内存小），也可能是因为数据量小（切片增加数据量），U-net使用了overlap-tile策略，将图片切片成m*m的patch，并对patch进行padding（即取patch周围的上下文像素），使padding后的patch经过U-net后（尺寸会降低）尺寸恰为m*m。对于图片边缘的patch，可能有些方向没有上下文来padding，这时使用镜像padding，用patch作镜面对称。通过这种方式，可以实现对任意大图像进行无缝切割后进行预测，每个patch也获得了上下文信息。
- 与FCN在上采样有一个不同，FCN上采样时直接对分类分数上采样，显然很不准；U-net在上采样时保留丰富的特征，在最后才用1*1卷积层分类

> FCN在结合下采样特征图时将其1*1卷积后直接相加，U-net先concatenation再经过3*3卷积融合，FPN将其经过1*1卷积后相加再经过3*3卷积融合

- 为了提高对“接触的目标”的区分，本文使用了加权交叉熵损失，使用了一个公式（见论文），在训练前对每个GT图计算权重图，这种方法会使目标间的小背景具有较高的权重
- 医学图像分割任务的一个挑战为有标注数据很少，本文使用了数据增强，其中随即弹性形变的效果最好

3.Segnet: A deep convolutional encoder-decoder architecture for image segmentation(2016)

论文链接

网络结构与U-net类似，先下采样再上采样最后分类，提出了一种新的上采样方法，减小内存。虽然文章很长，但创新点有限

- Segnet的动机是实现高效的场景理解结构，更注重于优化时间和内存消耗，同时在各项指标上具有竞争力。
- SegNet应用了encoder-decoder结构（下采样和上采样阶段），encoder为FCN，卷积+BN+ReLU+最大池化得到该尺寸的特征图，decoder先上采样再接卷积层再BN再ReLU，最终实现像素级分类
- 关键点：在encoder中，只记录特征图max pooling时最大值的索引，从而使需要记录的特征信息大大降维。上采样时用了max pooling indices，根据encoder中对应特征图池化时的最大索引，实现上采样（对

应索引取值，其余置零）。上采样后的特征图是稀疏的，后接三个（卷积层再BN再ReLU）得到稠密的特征图用于下一阶段的上采样。上采样不需要学习也提高了效率。
- 实验表明，使用全部encoder时的特征图可以得到最好的效果，但在内存受限时SegNet可以提高表现

4.PSPNet:Pyramid scene parsing network(2017)

论文链接

本文提出了应用了Pyramid pooling module的PSPNet，可以聚合不同区域的上下文特征，并加入了一个辅助loss来训练深度ResNet

- 背景：全局信息和上下文关系对场景分析（语义分割）是重要的，简单的使用全局池化会损失空间关系而导致歧义，因此提供了一种金字塔池化，从而建立全局场景的先验。
- 将图片输入主干网络得到top特征图，将其按照不同尺寸池化，池化后有N*N个bin(N=1,2,3,6)，N=1时便为最一般的全局池化，这样可以得到不同尺度子区域的representation，不同水平的上下文信息。对每个池化后的context representation，用一个1*1卷积层将N*N尺寸的维度降为1\N，从而保持个水平全局特征之间的权重。之后分别进行上采样（双线性插值），使尺寸恢复为原特征图大小，再将这四个与原特征图concatenation，进行卷积以得到最后预测
- 对于主干网络，使用了ResNet和扩张卷积，在训练时，除了对最后一层的特征图进行预测，还加入了一个辅助损失，在res4b22残差块进行预测，共同反向传播更新网络，帮助优化学习过程。（前者的权重更大）

> 感觉PSP和目标检测中的SPP思想基本一样

5.Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs(2017)

论文链接

本文提出了Deeplab v2，在v1的基础上改进，因为v1的论文没看，所以读的有些粗糙，一些细节没弄清楚，之后若用到再细研究

- Deeplab的主要特点为：应用了空洞卷积(atrous concolution)；使用ASPP模块(atrous spatial pyramid pooling)；使用了CRF(Conditional Random Field)，这个方法在后续版本被抛弃
- 背景：应用于分类任务的CNN建构对空间变换具有一定的鲁棒性，这对分割问题不利——降低了分辨率、处理不同尺度物体、定位精度下降，第一条的三个特点分别解决这三个挑战
- 空洞卷积：空洞卷积可以在保持特征图视野大小的同时矿大感受野。为了扩大感受野，过去会增加步长或池化，会降低特征图视野大小，本Deeplab应用了空洞卷积，将Resnet第五个池化层及之后的池化换为步长为2的空洞卷积，从而由原来的下采样32倍变为下采样8倍。之后再用双线性插值上采样8倍，恢复原图像尺寸进行预测

> 空洞卷积可能导致grid problem，即感受野扩大，但某些最邻近的像素被忽略，可以通过连续使用不同尺寸的空洞卷积来时感受野铺满

- ASPP：在预测时，为了获得多尺度特征，对特征图进行了4个尺度下的空洞卷积，后分别又接了卷积层，将得到的4个输出和一个全局池化值（先全局池化再插值，细节不清楚）五部分concatenation起来，进行最后的预测

> 因为没时间读v1的论文了（大概也不太重要吧，而且现在是transformer时代了），可能一些细节没搞懂，以后再说

6.RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation(2017)

论文链接

本文也是为了解决下采样过程中导致的分辨率下降问题，提出RefineNet利用下采样过程中的所有信息来细化富含语义信息的top层特征图，且帮助其上采样，还提出一种链式残差池化

- 背景：分辨率下降是语义分割任务常见的挑战，一种方法为下采样后通过反卷积等方式上采样，但其实没有利用细粒度特征；另一种方法为Deeplab提出的空洞卷积，在保持视野大小的同时扩大感受野，但一方面会增加许多高维卷积占用很大内存，另一方面空洞卷积也是一种下采样，潜在的丢失一些信息。本文提出了RefineNet，使用下采样过程中的多尺度的、高分辨率特征图，细化帮助上采样时语义信息丰富、分辨率低的特征图，思想和FPN比较类似
- RefineNet下采样时使用的ResNet主体结构，利用了第二个池化层开始的特征图(1/4--1/32)。将1/32的特征图输入RefineNet4(这是一个block)，输出1/32的新特征图，再和1/16特征图一起输入RefineNet3，输出1/16的新特征图，依次下去，直到得到融合了细粒度特征的1/4特征图，做softmax再双线性插值
- RefineNet块里做了什么：先将1/2个特征图（对应Resnet块的特征图和上一个RefineNet块的输出）分别输入两个级联的RCU(残差卷积单元),每个RCU包括两个3*3卷积和ReLU和残差链接，其中除了RefineNet4的输出维度为512其余为256（RCU的目的是将预训练的适用于分类的特征图适应于分割任务，一种解释罢了）；将输出进行multi-resolution fusion，分别输入3*3卷积（将维度统一为最低的）和上采样（将尺寸统一为最大的），再相加；将输出进行Chained Residual Pooling，将输入进行级联的带残差链接的池化+卷积块，也就是每进行一次池化+卷积，都与这次的输入相加再输入到下一个池化+卷积（这样可以得到丰富的不同尺度的池化特征，并通过卷积权重加起来，认为这样可以有效捕捉背景上下文特征）；将输出通过一个RCU得到最终输出。
- 在整个网络中，应用了丰富的残差思想，既有短程（块内）的残差连接，又在上采样时与下采样时的特征图连接，是梯度更容易的传到靠前的参数中，有利于端对端训练

7.SERT:Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers(2021)

论文链接

将纯transformer结构应用到语义分割任务，使用encoder-decoder架构，提出SETR，创新点不大

- 预处理：和ViT一样，先分成patch，再映射到patch embedding，加上position embedding作为输入
- encoder:24个tranformer encoder块，相应的有24张特征图
- decoder:语义分割的难点在于，将特征图的尺寸恢复到原图分辨率，本文提出了三种decoder方式
    - Naive:将encoder最后一层特征图reshape成3D后，先用卷积层将维度转为类别数，在双线性插值到原尺寸
    - PUP:将encoder最后一层特征图reshape成3D后，交替上采样*2和卷积层
    - MLA(multi-Level feature Aggregation):和FPN类似，取M个encoder的特征图，先reshape成3D，再分别经过卷积层和4倍上采样，再加入一个横向连接，分别经过卷积层，再按维度concatenation,最后经过卷积层和4倍上采样得到原尺寸

1.Deeplab v3:Rethinking atrous convolution for semantic image segmentation(2017)

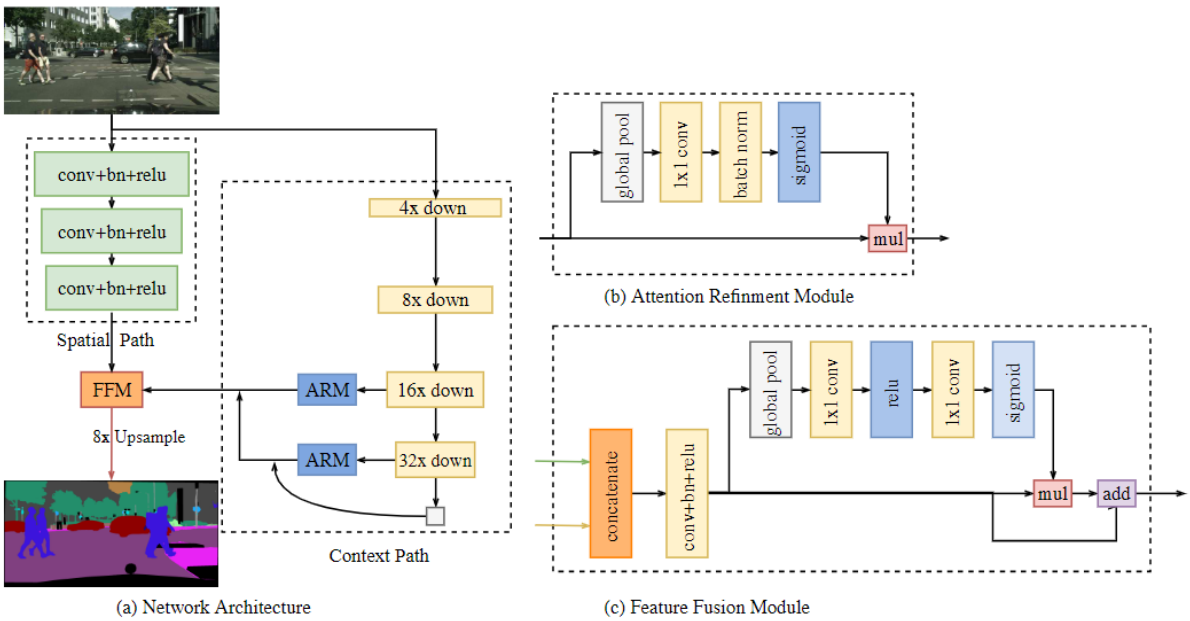
论文链接 在deeplab v2的基础上进行了改进，提出了级联的或并联的应用了空洞卷积的模块，均优于v2

- 级联：将Resnet的后几个block改成空洞卷积，输入与输出分辨率不变，每个block之间、和内部的卷积层之间空洞卷积的膨胀系数均有区别，一方面防止grid problem，另一方面扩大感受野
- 并联（ASPP）：改进了两点，加入了BN；空洞卷积的膨胀系数太大的话，无效点（padding）的数量大大增加，达不到扩大感受野的目的，因此加入了Image-level特征（全局池化层），后接1*1卷积和上采样，与ASPP输出拼接在一起

2.Bisenet: Bilateral segmentation network for real-time semantic segmentation(2018)

论文链接

本文提出了一种双边分割模型Bisenet，实现效果和效率的均衡

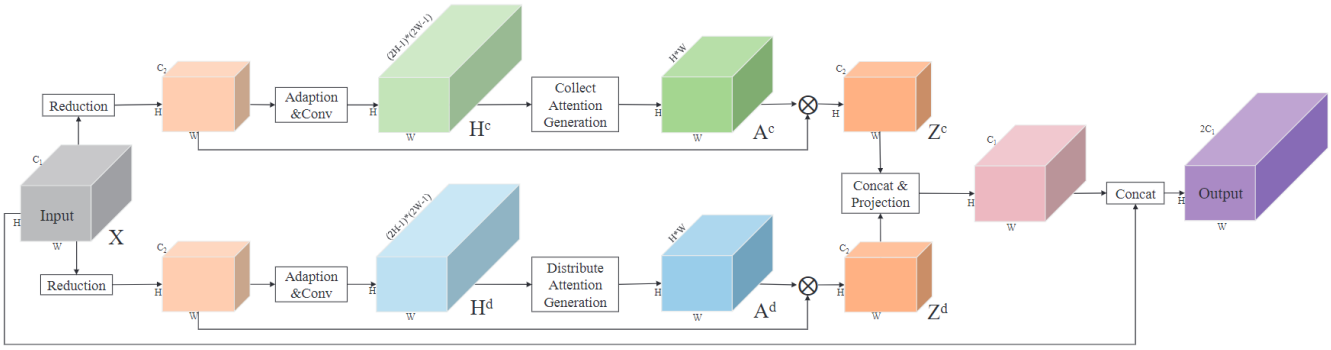


- 主要特点是，网络包括两条路径，context path和spatial path。前者通过快速的下采样pretrained主干网络Xception，扩大感受野。获得较低分辨率的含丰富语义特征的特征图，后接ARM(Attention refinement module)，其中包含全局池化；后者仅有三个卷积层，下采样8倍（因此尽管尺寸大但计算量不大），保留了原图像丰富的空间特征。
- 因为两条路径的信息的level不同，因此用FFM结合这两部分的特征。

3.Psanet: Point-wise spatial attention network for scene parsing(2018)

论文链接

引入point-wise注意力，考虑相对位置的同时考虑全局信息，每个点都自适应的通过一个可学习的注意力映射与其他所有点链接

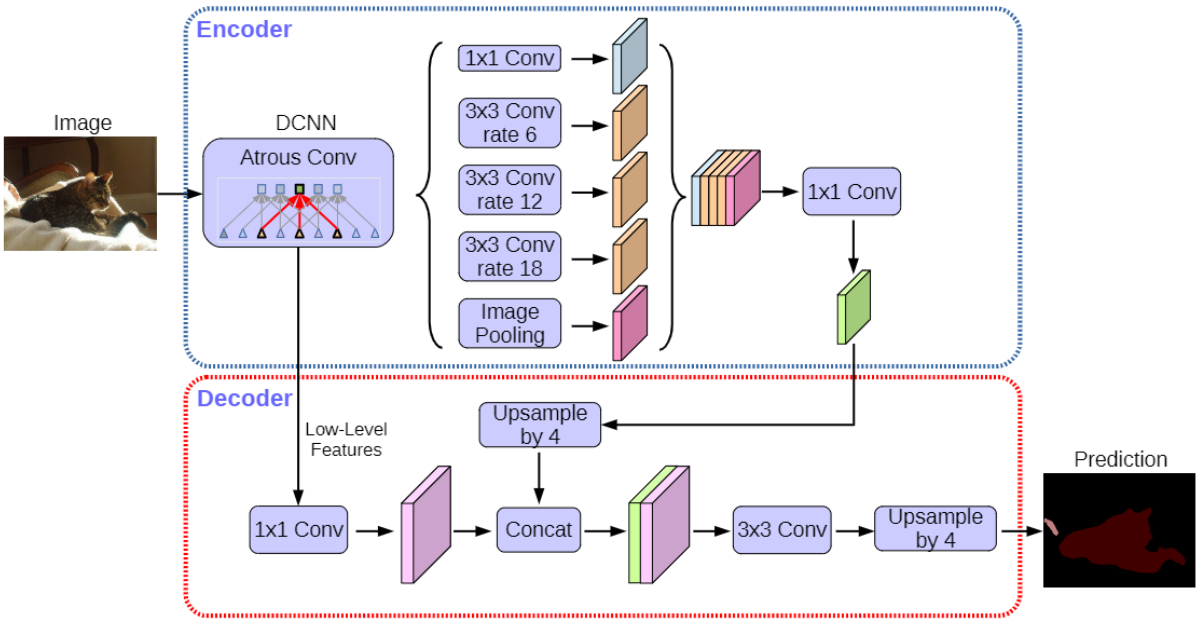


- 建立了一个双向信息传播路径，逐点注意力包括两部分，第一部分为其他点j对目标点i预测的重要性，第二部分为目标点i对其他点j的重要性，这两部分对特征图上每个点都是H*W维。文中先生成了一个2H-1*2W-1维的特征图，通过聚焦于它的不同位置，获得每个点H*W维注意力，得到attention map，将注意力图按公式可得每个点的特征。
- (输入特征图为H*W/C_2) collect中H*W*(H*W)维的attention map，每个点的H*W维向量表示H*W每个点对该点的注意力分数，对应加权求和每个点的C_2维向量，可得该点的输出特征；distribution部分的attetion map，每个点的H*W维向量表示该点对H*W个点的重要性，所以求输出特征时，取全局每个点的H*W维特征中的第i维作为该全局点对目标点i的注意力加权，累加可得输出特征

4.Encoder-decoder with atrous separable convolution for semantic image segmentation(2018)

论文链接

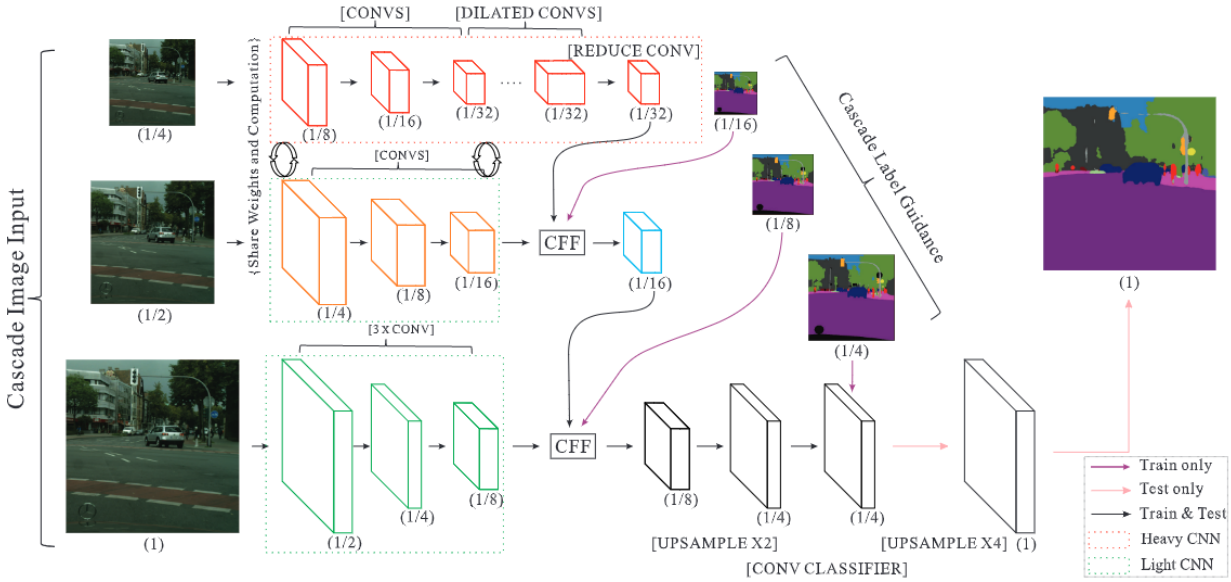
为了在保证分辨率的同时加入语义信息，deeplab v3使用空洞卷积替代池化，从而保证尺寸的同时扩大了感受野，但是这种方法不如encoder-decoder对边界信息更细节。因此，deeplav v3+结合了encoder-decoder结构，将v3作为一个强大的encoder，之后加了一个简单的decoder，还探索了深度可分离空洞卷积（应用在ASPP和decoder）



6.Icnet for real-time semantic segmentation on high-resolution images(2018)

论文链接

提出了一个实时语义分割框架ICNet，利用级联图片输入，融合不同尺寸的特征图，实现coarse-to-fine预测，在低分辨率特征图使用完整网络，在高分辨率部分使用轻量级网络，从而显著减小计算量。



- 在CFF(cascade feature fusion)模块，使用双线性插值和空洞卷积实现不同尺寸特征图的融合

使用辅助损失，每个尺寸的特征图都会被用来预测并计算损失，最终损失会加权

7.Non-local neural networks(2018)

论文链接

本文提出一种 non-local 操作，和一个通用的non-local block，将self-attention统一到non-local的范式中，并提出了一些其他可能的选择。

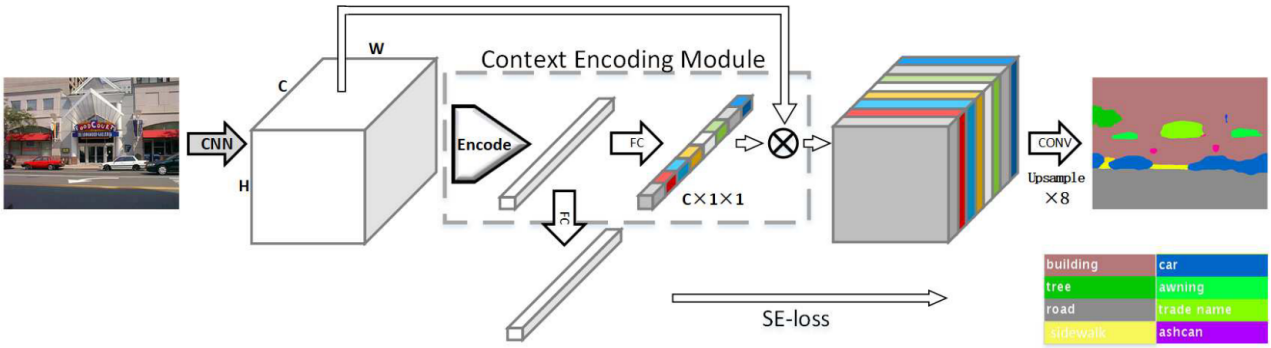
$$\mathbf{y}_i = \frac{1}{\mathcal{C}(\mathbf{x})} \sum_{\forall j} f(\mathbf{x}_i, \mathbf{x}_j) g(\mathbf{x}_j). \qquad (1)$$

Here i is the index of an output position (in space, time, or spacetime) whose response is to be computed and j is the index that enumerates all possible positions. $\mathbf{x}$ is the input signal (image, sequence, video; often their features) and $\mathbf{y}$ is the output signal of the same size as $\mathbf{x}$. A pairwise func-

8.EncNet:Context encoding for semantic segmentation(2018)

论文链接

提出了 Context Encoding Module,编码上下文信息，类似于SENet，对特征图的每个通道加权



- ENC模块中的encoder layer，通过传统方法得到K个语义词，利用softmax加权得到每个像素对每个语义词的残差特征，累加得整张图对每个语义词的残差特征
- 将encoder layer的输出input全连接层，得到每个通道的权重
- 引入了辅助任务，SE-loss，GT可以从分割GT中获得，每个类别的二元交叉熵

10.DANet:Dual attention network for scene segmentation(2019)

论文链接

为了更好的捕捉上下文信息（全局信息）和通道间的联系，本文提出了一种双注意力网络DANet，使用两个注意力模块来得到更好的特征表示

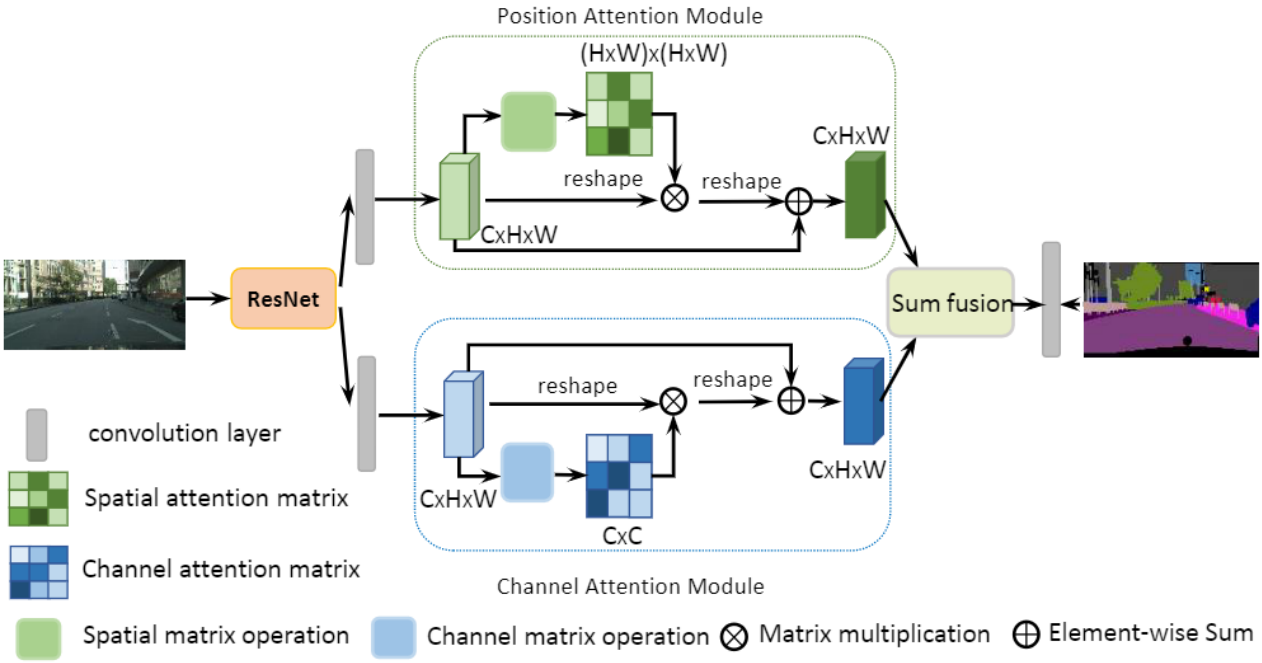


Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

position attention module，计算特征图H*W维度的自注意力，得到(H*W)*(H*W)的注意力分数矩阵，计算加权值；channel attention module，计算特征图通道维度的自注意力，得到C*C的注意力分数矩阵，再计算加权

值。最后将二者融合。