https://www.zhihu.com/question/36591394/answer/69124544

一、背景

视觉注意力机制是人脑特有的一种对信号处理的机制，人类视觉通过观察全局图像，选取一些局部重点关注区域，然后对这些区域投入更多注意力来获取更多的细节信息，抑制其他无用信息。

二、基本思想

Attention mechanism的本质是模仿人类视觉注意力机制，**学习出一个对图像特征的权重分布，再把这个权重分布施加在原来的特征上，为后面任务如图像分类、图像识别等提供不同的特征影响，使得任务主要关注一些重点特征，忽略不重要特征**，提高任务效率。

任务聚焦/解耦：通过将任务分解，设计不同的网络结构（或分支）专注于不同的子任务，重新分配网络的学习能力，从而降低原始任务的难度，使网络更加容易训练。

例：Mask R-CNN，将分类和分割解耦，当box branch已经分好类时，segmentation branch只需关注分割，不关注类别，使得网络更加易训练。当训练样本是狗时，生成狗的mask的网络连接只需聚焦于狗的样本，只有这个类别的loss才会被反传，其他类别不会对连接权重更新。（https://blog.csdn.net/yideqianfenzhiyi/article/details/79422857）

三、attention mechanism分类

3.1 soft Attention 和Hard Attention

Kelvin Xu等人与2015年发表论文《Show, Attend and Tell: Neural Image Caption Generation with Visual Attention》，在Image Caption中引入了Attention，当生成第i个关于图片内容描述的词时，用Attention来关联与i个词相关的图片的区域。Kelvin Xu等人在论文中使用了两种Attention Mechanism，即Soft Attention和Hard Attention。我们之前所描述的传统的Attention Mechanism就是Soft Attention。Soft Attention是参数化的（Parameterization），因此可导，可以被嵌入到模型中去，直接训练。梯度可以经过Attention Mechanism模块，反向传播到模型其他部分。

相反，Hard Attention是一个随机的过程。Hard Attention不会选择整个encoder的输出做为其输入，Hard Attention会依概率Si来采样输入端的隐状态一部分来进行计算，而不是整个encoder的隐状态。为了实现梯度的反向传播，需要采用蒙特卡洛采样的方法来估计模块的梯度。

两种Attention Mechanism都有各自的优势，但目前更多的研究和应用还是更倾向于使用Soft Attention，因为其可以直接求导，进行梯度反向传播。

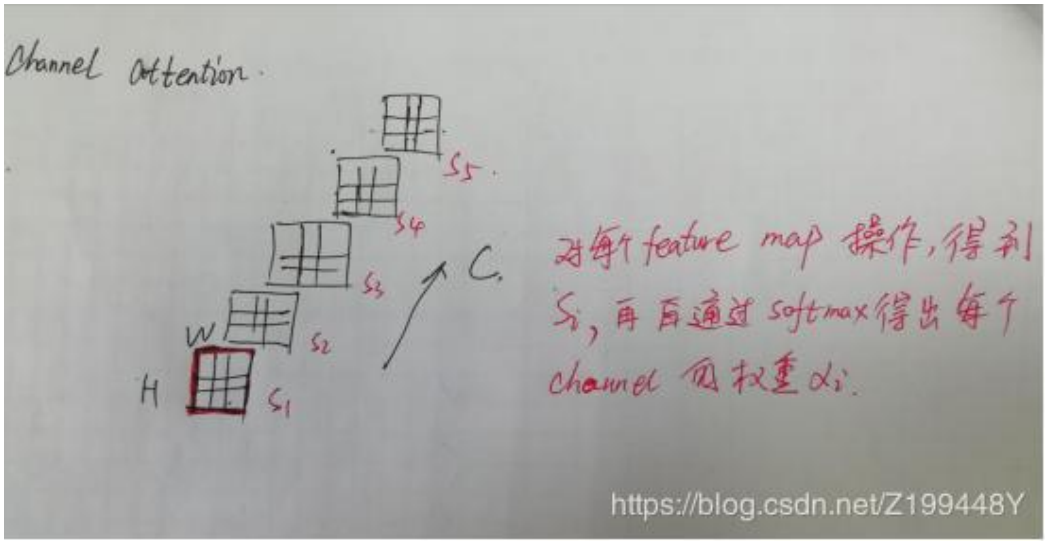
Attention mechanism可分为Soft attention和Hard attention，soft是保留所有分量进行加权，hard是以某种策略选取部分分量。

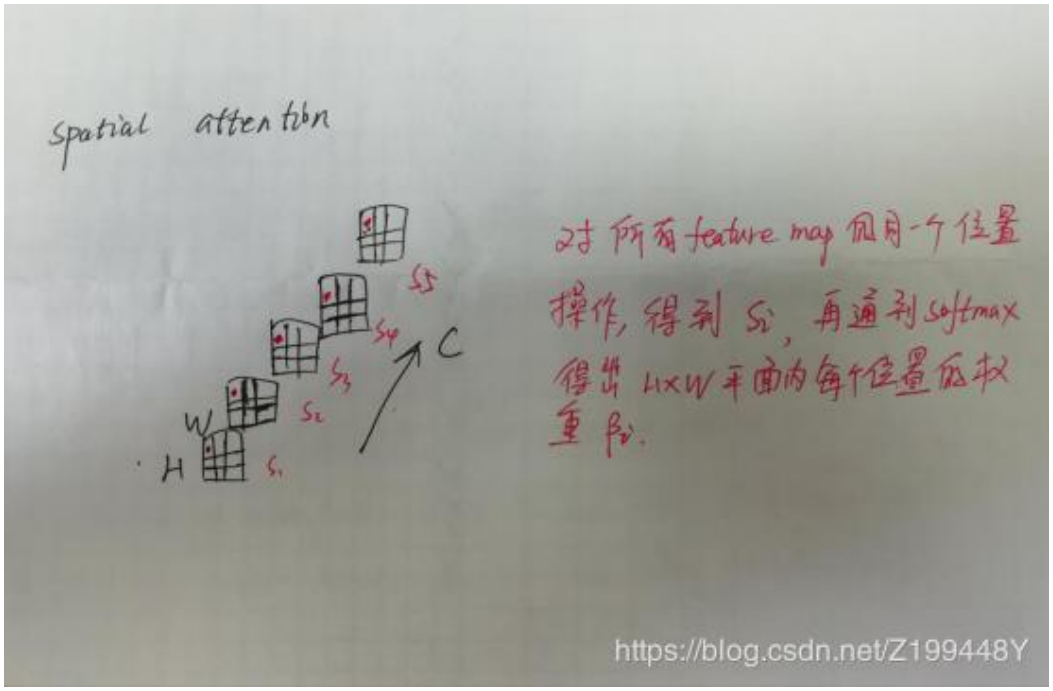
Attention mechanism可以加权在原图上，例如《Recurrent model of visual attention》（Google DeepMind 2014）和《Multiple object recognition with visual attention》（ICLR 2015）。

Attention mechanism**可以加权作用在空间尺度**（Spatial attention）上，给不同空间区域加权，也**可以作用在channel尺度**（channel attention）上，给不同通道特征加权，甚至可以给特征图上每个元素加权。

Attention mechanism可以加权作用在不同时刻历史特征上，例如machine translation。

通常对于C*H*W的feature map，spatial attention的H*W平面权重不同，C权重相同；channel attention的C权重不同，H*W平面权重相同。Channel attention关注“是什么”，spatial attention关注“在哪儿”。





四、attention设计

如何计算权重一般分为两个步骤：

1. 设计一个打分函数，针对每个attention向量，计算出一个score，打分依据就是和attention所关注的对象（实质是一个向量）的相关程度，越相关，所得值越大；
2. 对所得到的K个score Si，通过softmax函数，得到最后的权重，即：

关键点：如何结合具体问题，设计出所要关心的attention，然后将attention向量加入到model中，作为计算score的依据。

五、文章中attention mechanism的设计

1.《Squeeze-and-Excitation Networks》（CVPR 2018 oral / ImageNet2017图像分类任务冠军）

1》任务：图像分类

2》思想：考虑特征通道之间的关系提高网络性能。

3》具体操作

Squeeze操作，沿channel维度压缩特征，将二维的特征变为一个实数，且output维度和input的特征通道数相同，它表示特征通道上响应的全局分布，且使得靠近input的层可以获得全局的感受野。

Excitation操作，通过参数w为每个特征通道生成权重，其中参数w被学习用来显示的建模特征通道间的相关性。

Reweight操作，将excitation输出的权重当作特征选择后的每个特征的重要程度，然后通过乘法加权到每个通道上，完成在通道上对原始特征的重新标定。

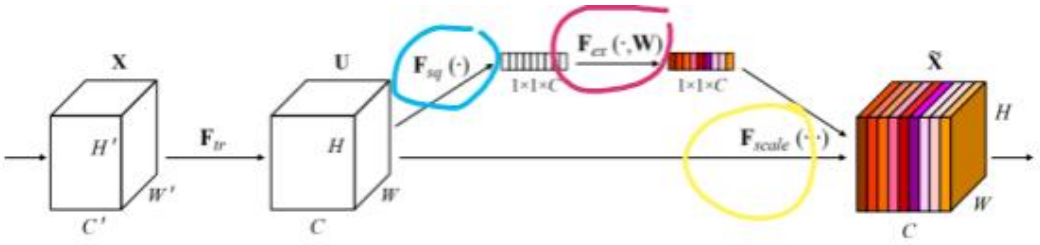


Figure 1: A Squeeze-and-Excitation block.

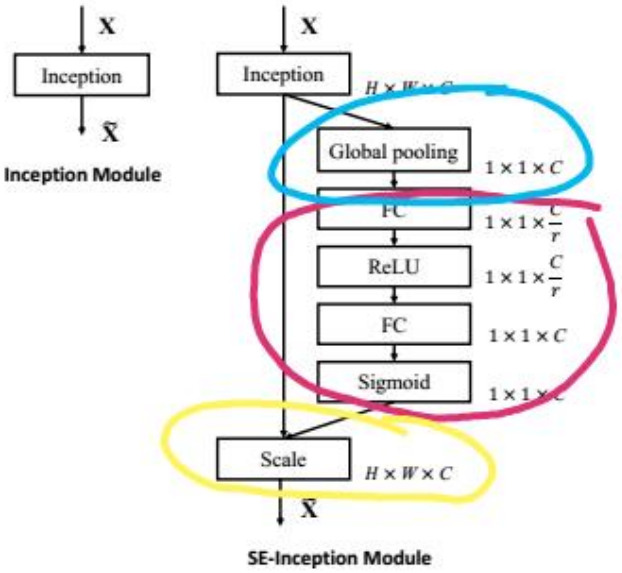


Figure 2: The schema of the original Inception module (left) and the SE-Inception module (right).

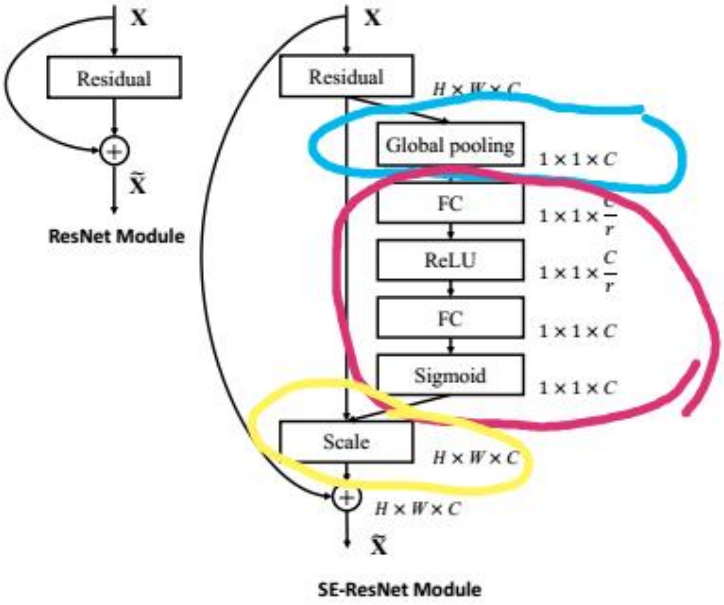


Figure 3: The schema of the original Residual module (left) and the SE-ResNet module (right).

备注：1、蓝色-Squeeze操作，Global average pooling用来计算channel-wise的统计量；粉色-Excitation操作；黄色-Reweight操作；

2、用两个FC层比用一个FC层的好处：具有更多非线性，可以更好拟合通道间复杂的相关性；极大的减少了参数量和计算量。缺点：不能保持spatial information

2.其他attention设计的文章

http://www.pianshen.com/article/9778303586/