



Attention is all you Need

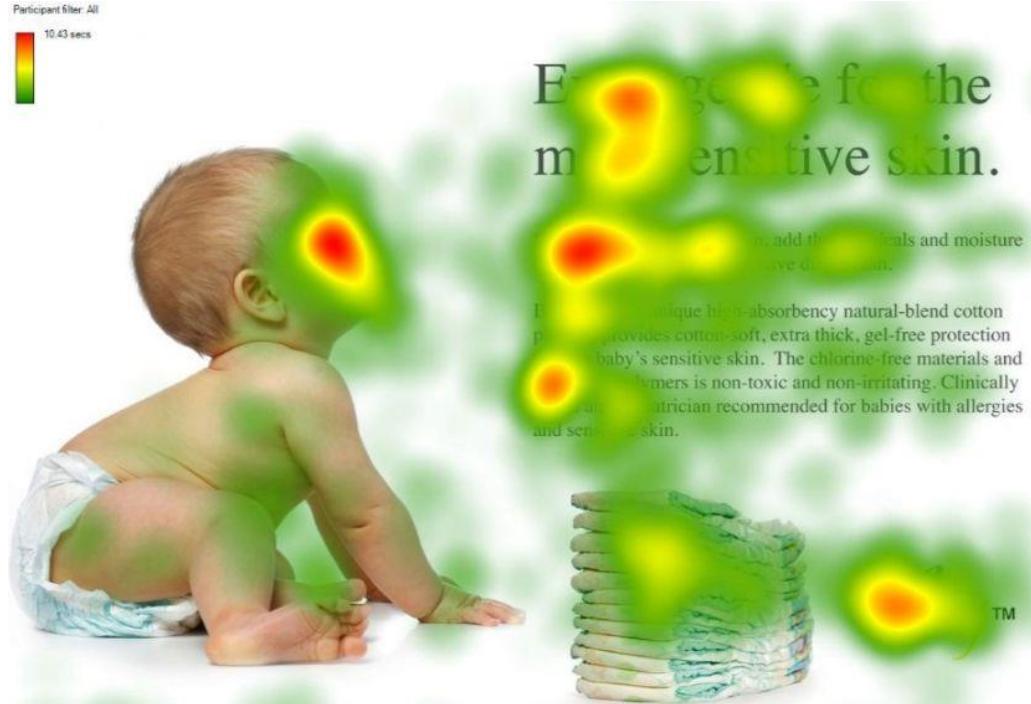
文豪

2021/4/30





Attention模型的介绍



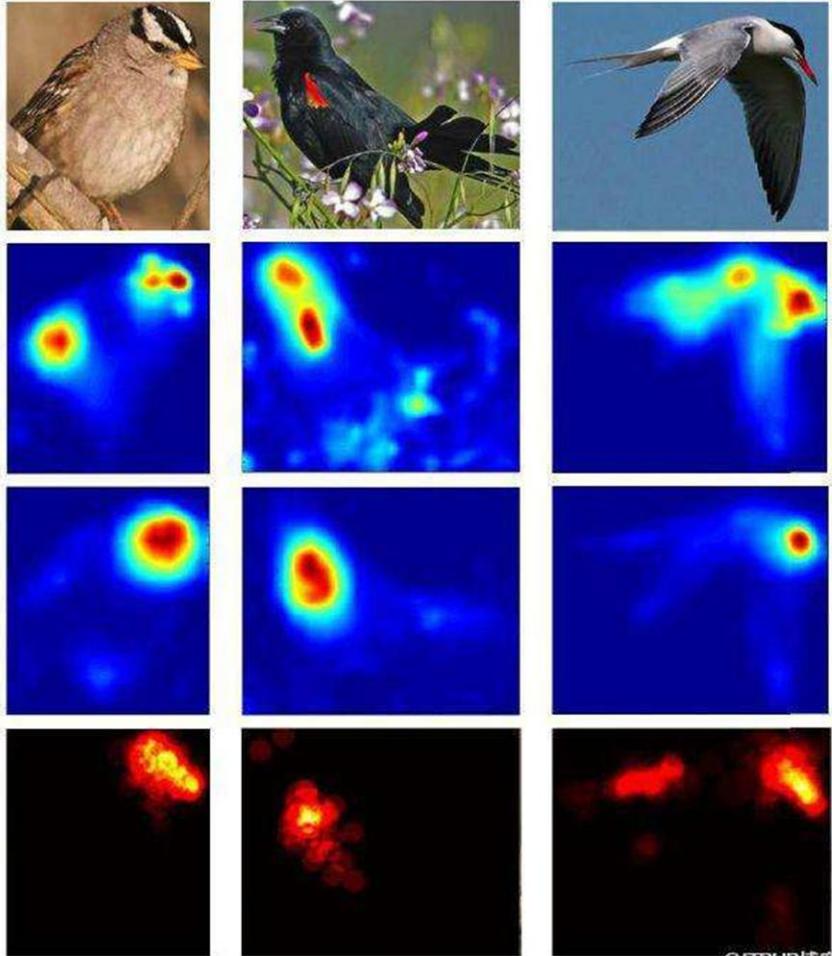
什么是Attention ?

注意力机制来源于对人类认知科学的研究，在认知过程中，由于信息处理的瓶颈，人类会选择性地关注所有信息的一部分，同时忽略其他可见的信息，即人类视觉的注意力机制。

Attention机制的两个主要任务：

- 1、决定需要重点关注的区域
- 2、分配有限的信息处理资源给重要的部分

Image



Attention机制的应用



Attention模型最初应用于图像识别，模仿人看图像时，目光的焦点在不同的物体上移动。当神经网络对图像或语言进行识别时，每次集中于部分特征上，识别更加准确。如何衡量特征的重要性呢？最直观的方法就是权重，因此，Attention模型的结果就是在每次识别时，首先计算每个特征的权值，然后对特征进行加权求和，权值越大，该特征对当前识别的贡献就大。



Attention模型的发展历程

计算机视觉方向Attention模型的发展简介：



计算机视觉方向Attention模型的分类：



★ Attention模型代表作

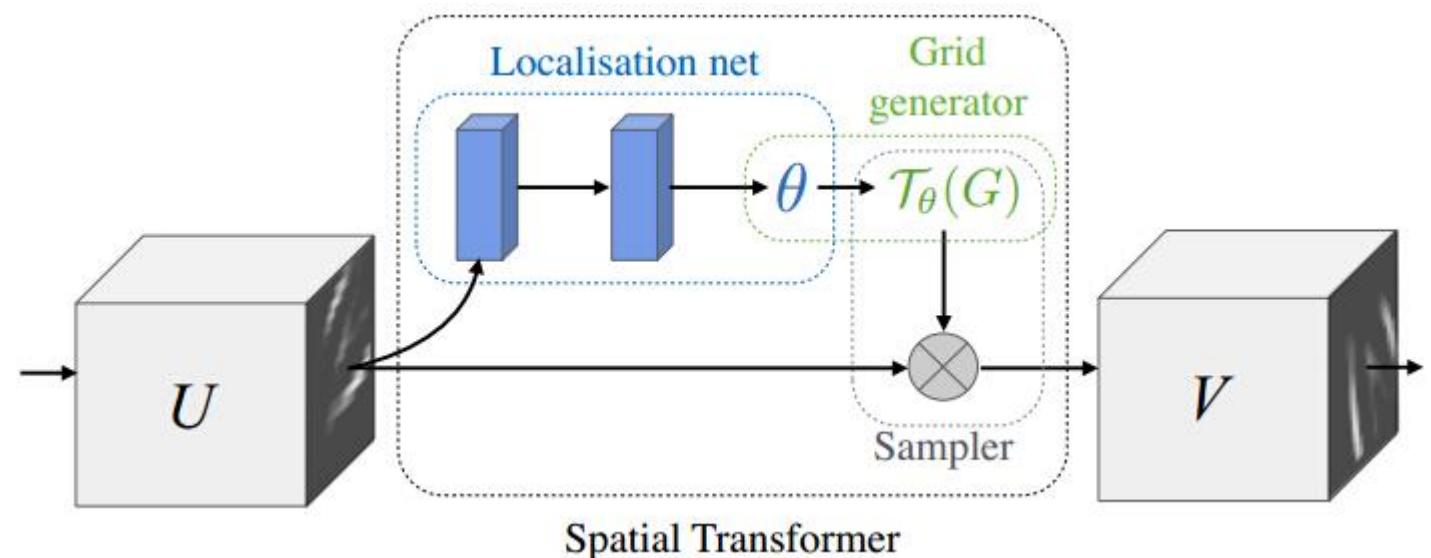
论文：Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. 2015: 2017-2025.

论文概述：卷积神经网络（CNN）已经被证明能够训练一个能力强大的分类模型，但与传统的模式识别方法类似，它也会受到数据在空间上多样性的影响。这篇Paper提出了一种叫做空间变换网络（Spatial Transform Networks， STN），该网络不需要关键点的标定，能够根据分类或者其它任务自适应地将数据进行空间变换和对齐（包括平移、缩放、旋转以及其它几何变换等）。在输入数据在空间差异较大的情况下，这个网络可以加在现有的卷积网络中，提高分类的准确性。

STN网络结构：

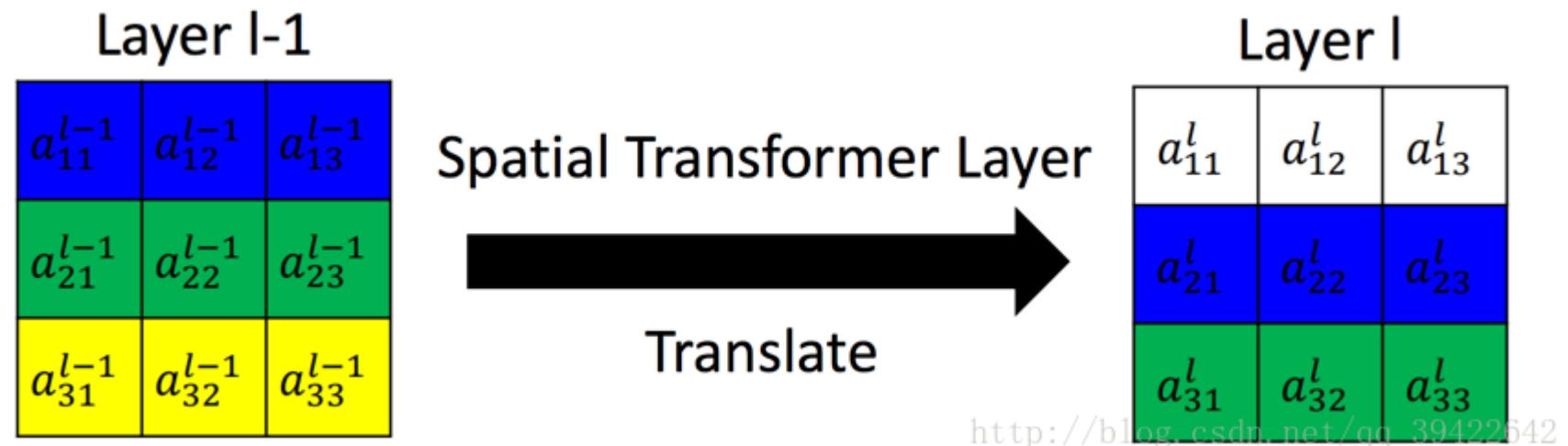
组成：

- Localisation net: 参数预测
- Grid generator: 坐标映射
- Sampler: 像素采样



Spatial Transformer的实现：

- Localisation Net: 作用就是通过一个子网络（全连接或者卷积网，再加一个回归层），生成空间变换的参数。 $a_{11}^l = w_{1111}^l a_{11}^{l-1} + w_{1112}^l a_{12}^{l-1} + w_{1113}^l a_{13}^{l-1} + \dots + w_{1133}^l a_{33}^{l-1}$

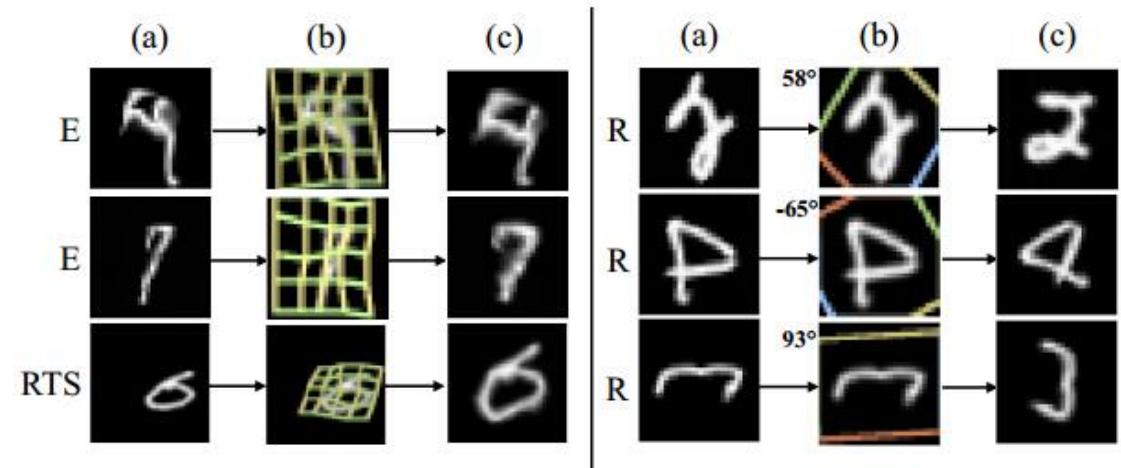


General Layer: $a_{nm}^l = \sum_{i=1}^3 \sum_{j=1}^3 w_{nm,ij}^l a_{ij}^{l-1}$ 通过调整这些权值，达到缩放、平移的目的。

- Grid：作用是依据预测的变换参数构建一个采样网格，它是一组输入图像中的点经过采样变换后得到的输出，其实就是一种映射关系 T_θ .

假设 U （不局限于输入图片，也可以是其它层输出的feature map）每个像素的坐标为 (x_i^s, y_i^s) ， V 的每个像素坐标为 (x_i^t, y_i^t) ，空间变换函数 T_θ 为仿射变换函数，那么 (x_i^s, y_i^s) 和 (x_i^t, y_i^t) 的对应关系可以写为：

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix}$$

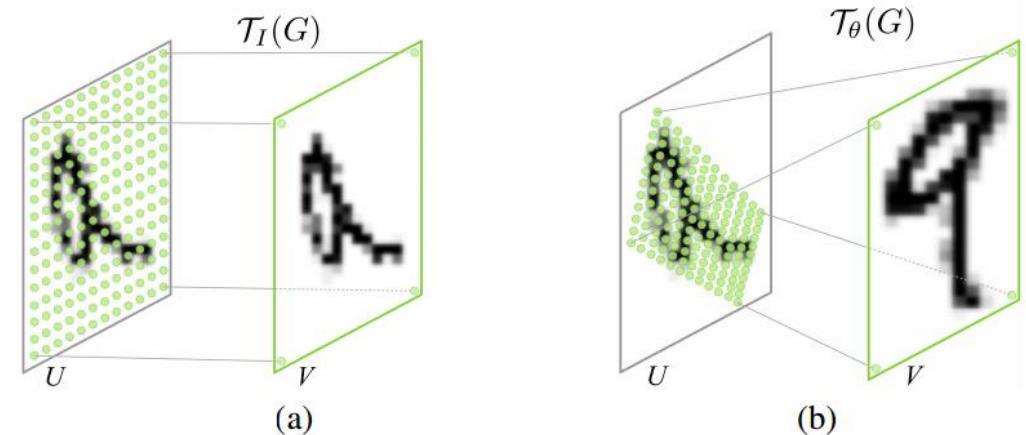


当然系数矩阵 A_θ 可以有其他的形式。

➤ Sampler: 作用是利用采样网格和输入的特征图同时作为输入产生输出，得到特征图经过变换之后的结果。

采样器利用采样网格和输入的特征图同时作为输入产生输出，得到了特征图经过变换之后的结果，如下：

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$



至此，STN网络的整个前向传播的过程完成，其与普通网络的不同之处就在于采样的过程中，需要有一个特定的网格作为引导。此外，作为一种网络，其学习能力是需要通过训练和反向传播的加持，才能不断提高网络的学习能力，改善效果。

➤ Backward propagation: 作用是计算参数梯度，优化代价函数。

输出对采样器的求导公式：

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \begin{cases} 0, & \text{if } |m - x_i^s| \geq 1 \\ 1, & \text{if } m \geq x_i^s \\ -1 & \text{if } m < x_i^s \end{cases}$$

$$\frac{\partial V_i^c}{\partial y_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0, & \text{if } |n - y_i^s| \geq 1 \\ 1, & \text{if } n \geq y_i^s \\ -1 & \text{if } n < y_i^s \end{cases}$$

输出对grid generator的求导公式：

$$\frac{\partial V_i^c}{\partial \theta} = \begin{pmatrix} \frac{\partial V_i^c}{\partial x_i^s} \cdot \frac{\partial x_i^s}{\partial \theta} \\ \frac{\partial V_i^c}{\partial y_i^s} \cdot \frac{\partial y_i^s}{\partial \theta} \end{pmatrix}$$

至此将以上几个部分合并起来，即为完整的STN网络。

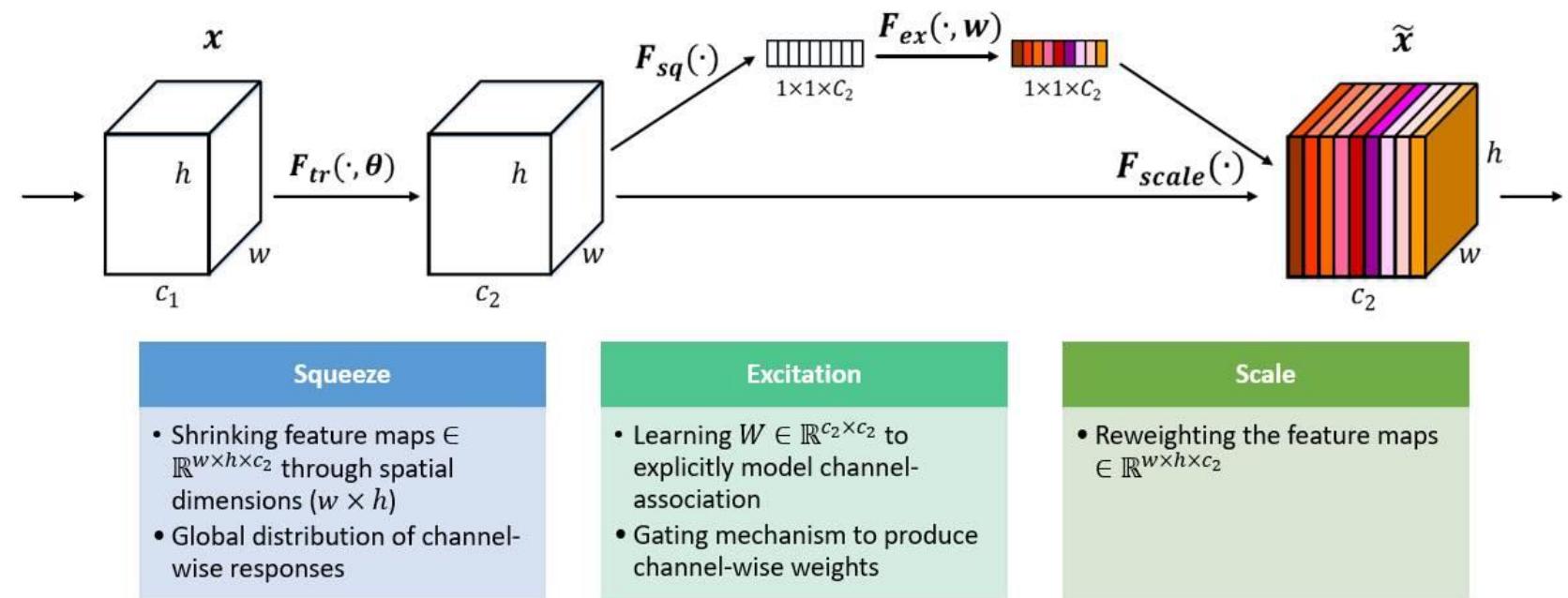
论文：Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

论文概述：卷积神经网络（CNN）的核心构件是卷积运算，它使网络能够通过在每一层的局部接受域内融合空间和通道信息来构造信息特征。在这项工作中，我们将重点放在渠道关系上，并提出一个新颖的体系结构单元，我们称之为“挤压和激励”（SE）块，该模块通过显式建模渠道之间的相互依赖性来自适应地重新校准渠道方式的特征响应。SE-Net赢得了最后一届ImageNet 2017竞赛分类任务的冠军。

SE-Net网络结构：

从本质上讲，卷积是对一个局部区域进行特征融合，这包括空间上（H和W维度）以及通道间（C维度）的特征融合。这个注意力机制分成三个部分：挤压(squeeze)，激励(excitation)，以及注意(attention)。

Squeeze-and-Excitation Module



➤ Squeeze:

Squeeze操作是顺着空间维度进行特征压缩，将每个二维的特征通道变成一个实数，这种操作有以下两个优点：

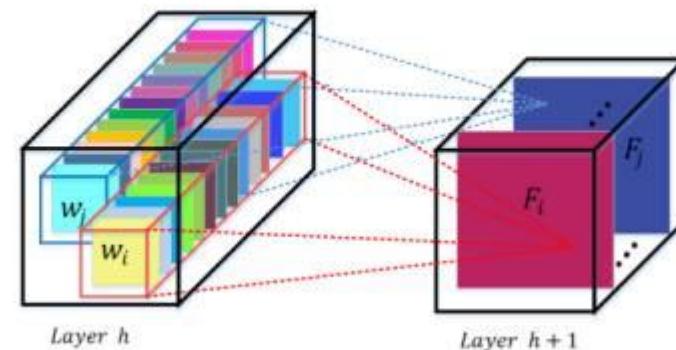
- 这个实数某种程度上具有全局的感受野，并且输出的维度和输入的特征通道数相匹配
- 它表征着在特征通道上响应的全局分布，而且使得靠近输入的层也可以获得全局的感受野

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_c(i, j), z \in R^C$$

Convolution

A convolutional filter is expected to be an informative combination

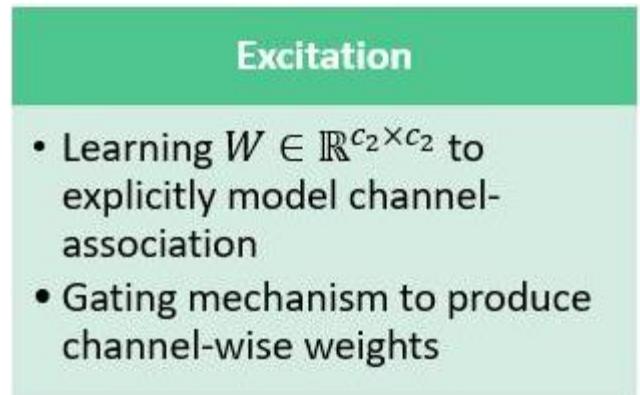
- Fusing **channel-wise** and **spatial** information
- Within local receptive fields



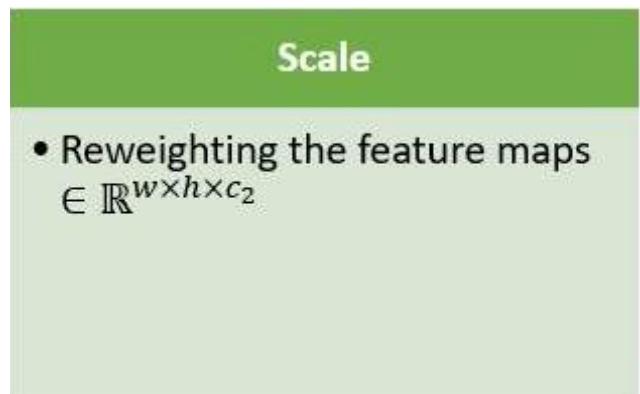
➤ Excitation & Reweight

Excitation: 它是一个类似于循环神经网络中门的机制。通过参数 来为每个特征通道生成权重，其中参数 被学习用来显式地建模特征通道间的相关性。

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \text{ReLU}(W_1 z))$$



Reweight: 我们将Excitation的输出的权重看做是进过特征选择后的每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定。



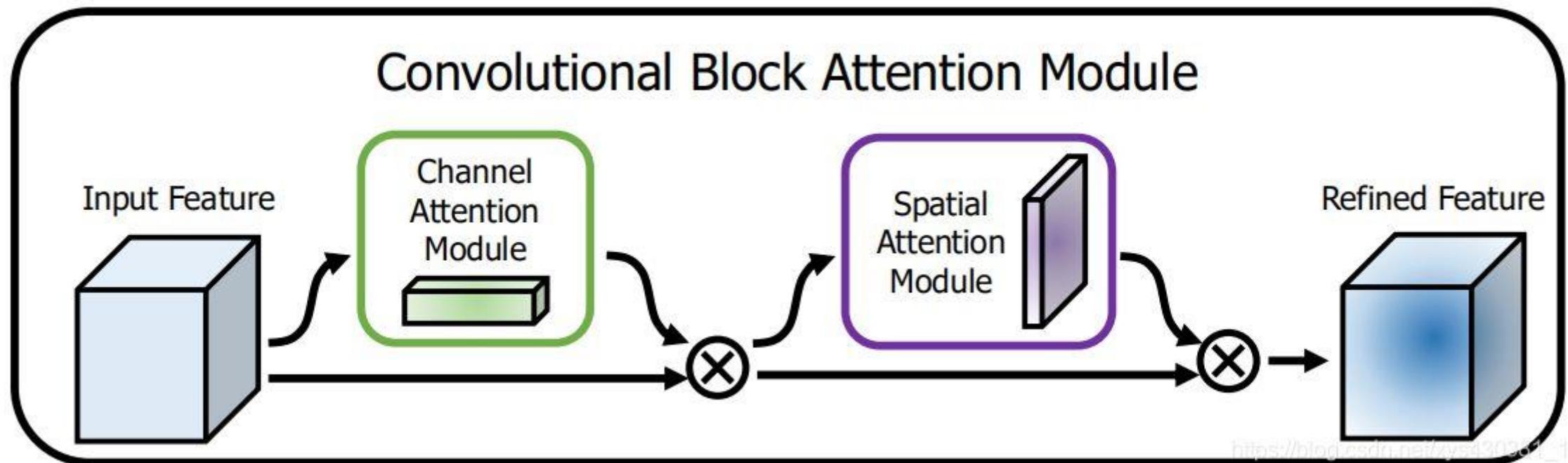
$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c$$

论文：Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

论文概述：提出了卷积块注意力模块（CBAM），这是一种简单有效的注意力模块，可以与任何前馈卷积神经网络集成。给定一个中间特征图，我们的模块会沿着两个独立的维度（通道和空间）依次推断注意力图，然后将注意力图与输入特征图相乘以进行自适应特征细化。我们的模块可与基础CNN一起进行端到端培训。我们通过在ImageNet-1K, MS COCO检测和VOC 2007检测数据集上进行的广泛实验来验证CBAM。我们的实验表明，使用各种模型对分类和检测性能的改进是一致的，证明了CBAM具有泛适用性。

CBAM网络结构：

- Channel Attention Module + Spatial Attention Module

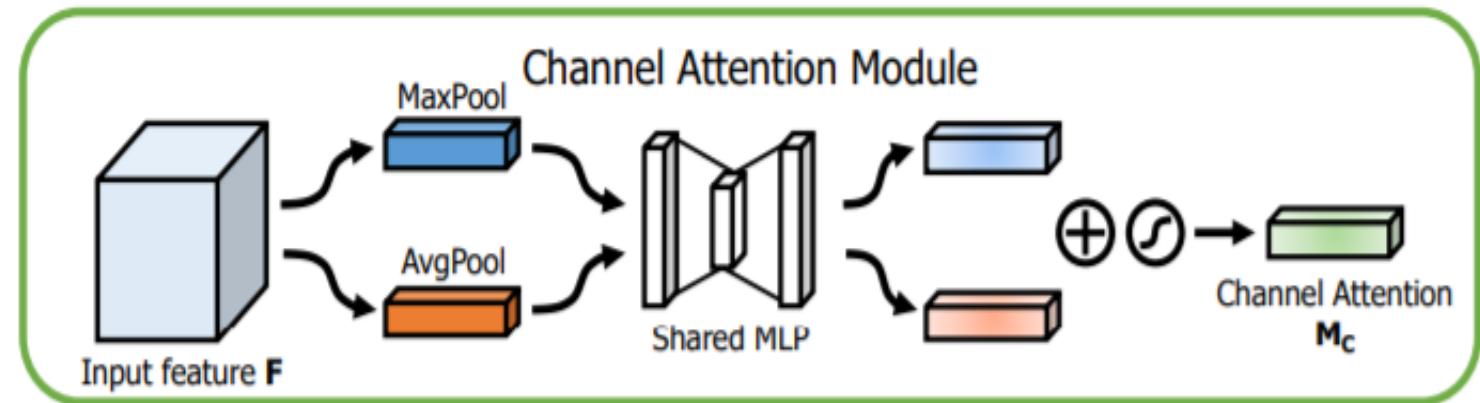


<https://blog.csdn.net/yeyi30301/>

Part A: Channel Attention Module

- 通道注意力机制，对于每一个输入的 feature \mathbf{F} ，在空间维度上都对其进行两种池化操作：
 - 全局最大池化
 - 全局平均池化

然后将池化后的特征输入到MLP层中进行处理，并将最大池化和平均池化经过处理之后的特征求和，输入到sigmoid函数中，得到Channel Attention权重 M_c 。

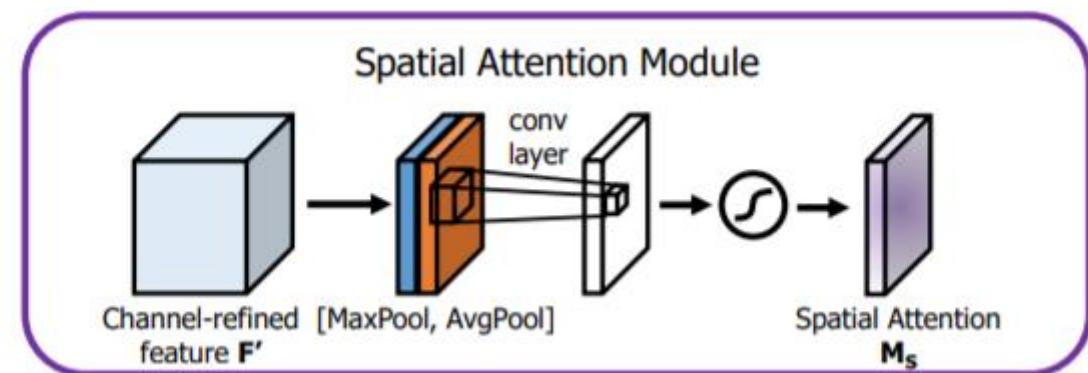


$$\begin{aligned} M_c(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{avg}^c)) + \mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{max}^c))) \end{aligned}$$

Part B: Spatial Attention Module

- 空间注意力机制，对于每一个输入的feature \mathbf{F} ，在通道维度上对其进行两种池化操作：
 - 全局最大池化
 - 全局平均池化

在通道维度上进行池化提取特征，提取的次数是高乘以宽，依次将特征提取完之后将其合并，得到一个2通道的特征图。



$$\begin{aligned}\mathbf{M}_s(\mathbf{F}) &= \sigma(f^{7 \times 7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])) \\ &= \sigma(f^{7 \times 7}([\mathbf{F}_{avg}^s; \mathbf{F}_{max}^s]))\end{aligned}$$

从上述模型看，注意力模型具备以下优点：

- 侧重点分明，会选择性关注输入图像的某些区域，符合视觉系统的特点
- 扩展了神经网络的能力，且易于集成到现有的深度学习网络结构中
- 有助于更好的解释并可视化神经网络模型



Presented by WH