



综述：图像处理中的注意力机制



极市平台

已认证的官方帐号

已关注

692 人赞同了该文章

本文对图像处理中的注意力机制进行了全面综述，介绍了注意力机制的基本概念和分类，并对多种方法进行了具体解读。本文作者@xys430381_1，仅作学术分享，著作权归作者所有，如有侵权，请联系后台作删文处理。

重磅好文：

- 微软亚研：对深度神经网络中空间注意力机制的经验性研究
- 论文：An Empirical Study of Spatial Attention Mechanisms in Deep Networks
- 论文阅读：图像分类中的注意力机制(attention)
介绍了Spatial transformer networks、Residual Attention Network、Two-level Attention、SENet、Deep Attention Selective Network
- 计算机视觉中的注意力机制 (Visual Attention)

▲ 赞同 692

● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

赞同 692



分享

- [Attention机制的文章总结](#)
[计算机视觉技术self-attention最新进展](#)
- [ECCV2018-注意力模型CBAM](#)
[【论文复现】CBAM: Convolutional Block Attention Module](#)
[BAM: Bottleneck Attention Module算法笔记](#)
- [基于注意力机制的细腻度图像分类](#)
[讲解RACNN](#)
此外还有MACNN等方法。

目录

- 概要
 - 为什么需要视觉注意力
 - 注意力分类与基本概念
- 软注意力
 - The application of two-level attention models in deep convolutional neural network for fine-grained image classification---CVPR2015
 - 1. Spatial Transformer Networks(空间域注意力)---2015 nips
 - 2. SENET (通道域) ---2017CPVR
 - 3. Residual Attention Network(混合域)---2017
 - Non-local Neural Networks, CVPR2018]
(#Nonlocal_Neural_Networks_CVPR2018_120)
 - Interaction-aware Attention, ECCV2018
 - CBAM: Convolutional Block Attention Module(通道域+空间域), ECCV2018
 - DANet: Dual Attention Network for Scene S
 - CCNet



赞同 692



分享

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

- 注意增强型卷积
 - PAN: Pyramid Attention Network for Semantic Segmentation(层域)---CVPR2018
 - Multi-Context Attention for Human Pose Estimation
 - Tell Me Where to Look: Guided Attention Inference Network
- 硬注意力
 - 一种通过引入硬注意力机制来引导学习视觉回答任务的研究
 - 1. Diversified visual attention networks for fine-grained object classification---2016
 - 2. Deep networks with internal selective attention through feedback connections (通道域)---NIPS 2014
 - 3. Fully Convolutional Attention Networks for Fine-Grained Recognition
 - 4. 时间域注意力(RNN)
- 自注意力
- RelatedWorks
- 自注意力的缺点和改进策略
- 自注意力小结



赞同 692



分享

概要

为什么需要视觉注意力

计算机视觉（computer vision）中的注意力机制（attention）的基本思想就是想让系统学会注意力——能够忽略无关信息而关注重点信息。为什么要忽略无关信息呢？

注意力分类与基本概念

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

该文分为：硬注意力、软注意力、此外，还有高斯注意力、空间变换

就注意力的可微性来分：

1. Hard-attention，就是0/1问题，哪些区域是被 attentioned，哪些区域不关注.硬注意力在图像中的应用已经被人们熟知多年：图像裁剪（image cropping）
硬注意力（强注意力）与软注意力不同点在于，首先强注意力是更加关注点，也就是图像中的每个点都有可能延伸出注意力，同时强注意力是一个随机的预测过程，更强调动态变化。当然，**最关键的是强注意力是一个不可微的注意力，训练过程往往是通过增强学习(reinforcement learning)来完成的。**（参考文章：Mnih, Volodymyr, Nicolas Heess, and Alex Graves. “Recurrent models of visual attention.” Advances in neural information processing systems. 2014.）

硬注意力可以用Python（或Tensorflow）实现为：

```
g = I[y:y+h, x:x+w]
```

上述存在的唯一的问题是它是不可微分的；你如果想要学习模型参数的话，就必须使用分数评估器（score-function estimator）关于这一点，我的前一篇文章中有对其的简要介绍。

1. Soft-attention，[0,1]间连续分布问题，每个区域被关注的程度高低，用0~1的score表示。



赞同 692



分享

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏



软注意力的关键点在于，这种注意力更关注区域或者通道，而且软注意力是确定性的注意力，学习完成后直接可以通过网络生成，**最关键的地方是软注意力是可微的**，这是一个非常重要的地方。可以微分的注意力就可以通过神经网络算出梯度并且前向传播和后向反馈来学习得到注意力的权重。然而，这种类型的软注意力在计算上是非常浪费的。输入的黑色部分对结果没有任何影响，但仍然需要处理。同时它也是过度参数化的：实现注意力的sigmoid 激活函数是彼此相互独立的。它可以一次选择多个目标，但实际操作中，我们经常希望具有选择性，并且只能关注场景中的一个单一元素。由DRAW和空间变换网络（Spatial Transformer Networks）引入的以下两种机制很好地解决了这个问题。它们也可以调整输入的大小，从而进一步提高性能。

就注意力关注的域来分：

1. 空间域(spatial domain)
2. 通道域(channel domain)
3. 层域(layer domain)
4. 混合域(mixed domain)
5. 时间域(time domain)：还有另一种比较特殊的强注意力实现的注意力域，时间域(time domain)，但是因为强注意力是使用reinforcement learning来实现的，训练起来有所不同

一个概念：Self-attention自注意力，就是 feature map 间的自主学习，分配权重（可以是 spatial，可以是 temporal，也可以是 channel间）



赞同 692



分享

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

The application of two-level attention models in deep convolutional neural network for fine-grained image classification—CVPR2015

1. Spatial Transformer Networks(空间域注意力)—2015 nips

Spatial Transformer Networks (STN) 模型[4]是15年NIPS上的文章，这篇文章通过注意力机制，将原始图片中的空间信息变换到另一个空间中并保留了关键信息。

这篇文章认为之前pooling的方法太过于暴力，直接将信息合并会导致关键信息无法识别出来，所以提出了一个叫空间转换器（spatial transformer）的模块，将图片中的空间域信息做对应的空间变换，从而能将关键的信息提取出来。

spatial transformer其实就是注意力机制的实现，因为训练出的spatial transformer能够找出图片信息中需要被关注的区域，同时这个transformer又能够具有旋转、缩放变换的功能，这样图片局部的重要信息能够通过变换而被框盒提取出来。



赞同 692



分享

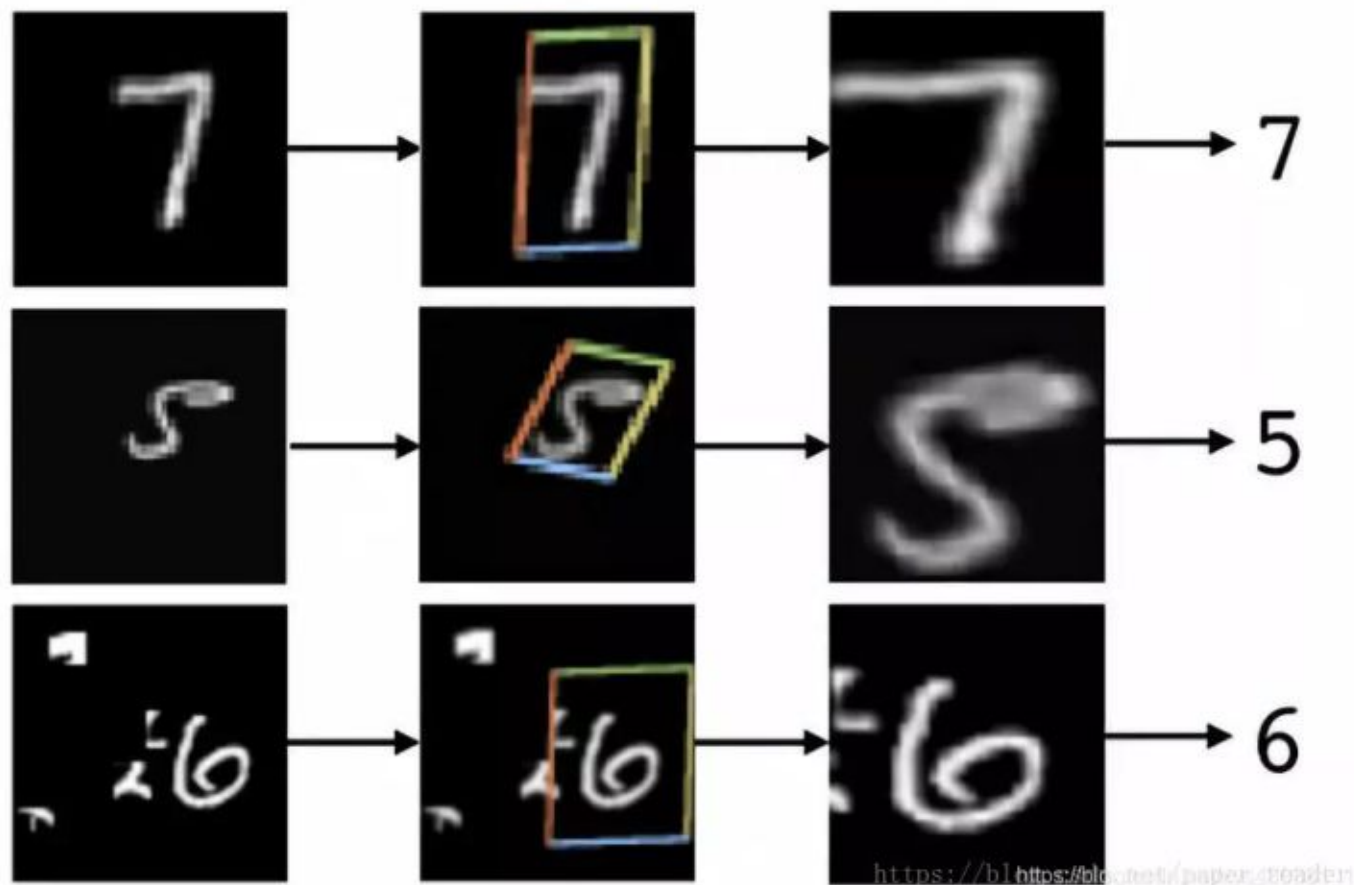
▲ 赞同 692 ▼

● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏



(a)列是原始的图片信息，其中第一个手写数字7没有做任何变换，第二个手写数字5，做了一定的旋转变换，而第三个手写数字6，加上了一些噪声信号；

赞同 692

分享

▲ 赞同 692

14 条评论

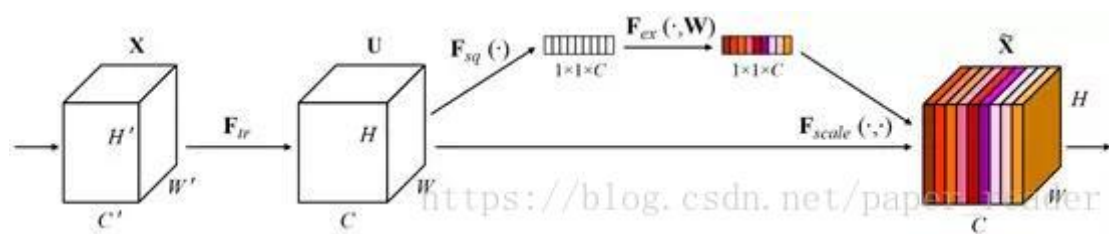
分享

♥ 喜欢

★ 收藏

©列中是通过spatial transformer转换之后的特征图，可以看出7的关键区域被选择出来，5被旋转成为了正向的图片，6的噪声信息没有被识别进入。

2. SENET（通道域）—2017CPVR



中间的模块就是SENet的创新部分，也就是注意力机制模块。这个注意力机制分成三个部分：挤压(squeeze)，激励(excitation)，以及scale(attention)。

流程：

1. 将输入特征进行 Global AVE pooling，得到 1_1_Channel
2. 然后bottleneck特征交互一下，先压缩 channel数，再重构回channel数
3. 最后接个 sigmoid，生成channel 间0~1的 attention weights，最后 scale 乘回原输入特征

详见 [《论文阅读笔记—SENET》](#)

3. Residual Attention Network(混合域)—2017



赞同 692



分享

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

来，这样就防止mask之后的信息量过少引起的网络层数不能堆叠很深的问题。

文提出的注意力mask，不仅仅只是对空间域或者通道域注意，这种mask可以看作是每一个特征元素（element）的权重。**通过给每个特征元素都找到其对应的注意力权重，就可以同时形成了空间域和通道域的注意力机制。**

很多人看到这里就会有疑问，这种做法应该是从空间域或者通道域非常自然的一个过渡，怎么做单一域注意力的人都没有想到呢？原因有：

- 如果你给每一个特征元素都赋予一个mask权重的话，mask之后的信息就会非常少，可能直接就破坏了网络深层的特征信息；
- 另外，如果你可以加上注意力机制之后，残差单元（Residual Unit）的恒等映射（identical mapping）特性会被破坏，从而很难训练。

该文章的注意力机制的创新点在于提出了**残差注意力学习(residual attention learning)**，不仅只把mask之后的特征张量作为下一层的输入，同时也将mask之前的特征张量作为下一层的输入，这时候可以得到的特征更为丰富，从而能够更好的注意关键特征。



赞同 692



分享

▲ 赞同 692 ▼

● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

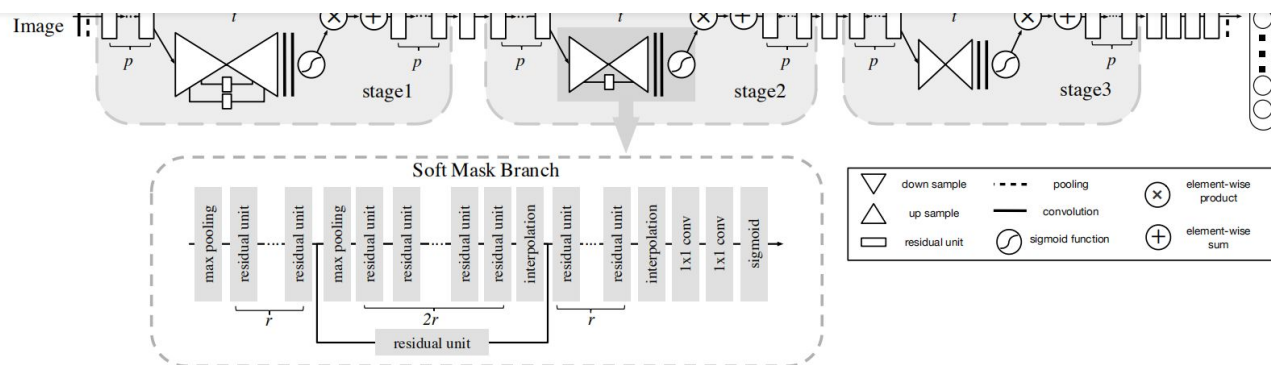


Figure 2: Example architecture of the proposed network for ImageNet. We use three hyper-parameters for the design of Attention Module: p , t and r . The hyper-parameter p denotes the number of pre-processing Residual Units before splitting into trunk branch and mask branch. t denotes the number of Residual Units in trunk branch. r denotes the number of Residual Units between adjacent pooling layer in the mask branch. In our experiments, we use the following hyper-parameters setting: $\{p = 1, t = 2, r = 1\}$. The number of channels in the soft mask Residual Unit and corresponding trunk branches is the same.

https://blog.csdn.net/xye430381_1

文章中模型结构是非常清晰的，整体结构上，是三阶注意力模块(3-stage attention module)。每一个注意力模块可以分成两个分支(看stage2)，上面的分支叫主分支(trunk branch)，是基本的残差网络(ResNet)的结构。而下面的分支是软掩码分支(soft mask branch)，而软掩码分支中包含的主要部分就是残差注意力学习机制。通过下采样(down sampling)和上采样(up sampling)，以及残差模块(residual unit)，组成了注意力的机制。

模型结构中比较创新的残差注意力机制是：

$$H_{i,d}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x)$$

H是注意力模块的输出，F是上一层的图片张量特征，N

$$f_1(x_{i,c}) = \frac{1}{1+\exp(-x_{i,c})}$$
$$f_2(x_{i,c}) = \frac{x_{i,c}}{\|x_i\|}$$
$$f_3(x_{i,c}) = \frac{1}{1+\exp(-(x_{i,c}-\text{mean}_c)/\text{std}_c)}$$

1. f_1 是对图片特征张量直接sigmoid激活函数，就是混合域的注意力；
2. f_2 是对图片特征张量直接做全局平均池化（global average pooling），所以得到的是通道域的注意力（类比SENet）；
3. f_3 是求图片特征张量在通道域上的平均值的激活函数，类似于忽略了通道域的信息，从而得到空间域的注意力。

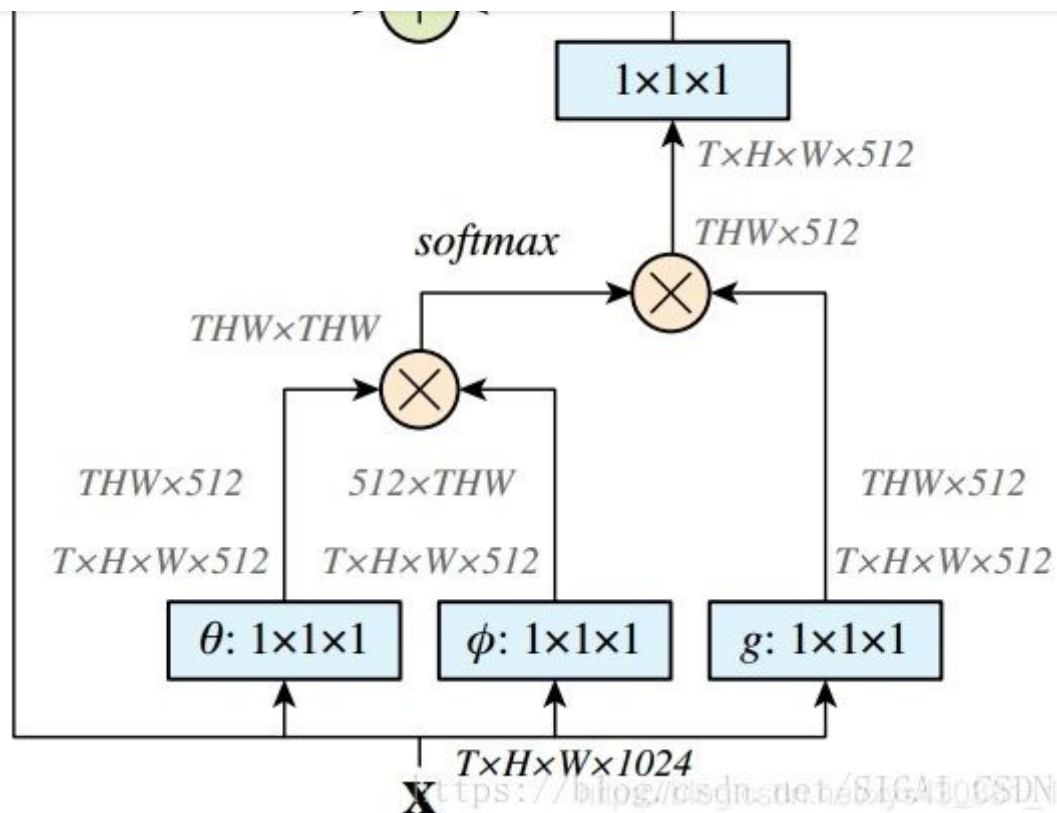
Non-local Neural Networks, CVPR2018

FAIR的杰作，主要 inspired by 传统方法用non-local similarity来做图像 denoise

主要思想也很简单，CNN中的 convolution单元每次只关注邻域 kernel size 的区域，就算后期感受野越来越大，终究还是局部区域的运算，这样就忽略了全局其他片区（比如很远的像素）对当前区域的贡献。

所以 **non-local blocks** 要做的是，捕获这种 **long-range** 关系：对于2D图像，就是图像中任何像素对当前像素的关系权值；对于3D视频，就是所有帧中的所有像素，对当前帧的像素的关系权值。

网络框架图也是简单粗暴：



文中有谈及多种实现方式，在这里简单说说在DL框架中最好实现的 Matmul 方式：

1. 首先对输入的 feature map X 进行线性映射（说白了就是 $1 \times 1 \times 1$ 卷积，来压缩通道数），然后得到 θ , ϕ , g 特征
2. 通过 reshape 操作，强行合并上述的三个特征除通道数外的维度，然后对 进行矩阵点乘操作，得到类似协方差矩阵的东西（这个过程很重要，计算出特征中的自相关性，即得到每帧中每个像素对其他所有帧所有像素的关系）
3. 然后对自相关特征 以列 or 以行（具体看矩阵 g 的形

map X 残差一下，完整的 bottleneck

嵌入在 action recognition 框架中的attention map 可视化效果：



图中的箭头表示，previous 若干帧中的某些像素 对最后图（当前帧）的脚关节像素的贡献关系。由于是soft-attention，其实每帧每个像素对对其有贡献关系，图中黄色箭头是把响应最大的关系描述出来。

总结

Pros: non-local blocks很通用的，容易嵌入在任何现有的 2D 和 3D 卷积网络里，来改善或者可视化理解相关的CV任务。比如前不久已有文章把 non-local 用在 Video ReID [2] 的任务里。

Cons: 文中的结果建议把non-local 尽量放在靠前的层里，但是实际上做 3D 任务，靠前的层由于 temporal T 相对较大，构造及点乘操作那步，超多的参数，需要耗费很大的GPU Memory

Interaction-aware Attention, ECCV2018

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

这篇文章扯了很多 Multi-scale 特征融合，讲了一堆 story，然并卵；直接说重点贡献，就是在 non-local block 的协方差矩阵基础上，设计了基于 PCA 的新loss，更好地进行特征交互。作者认为，这个过程，特征会在channel维度进行更好的 non-local interact，故称为 Interaction-aware attention

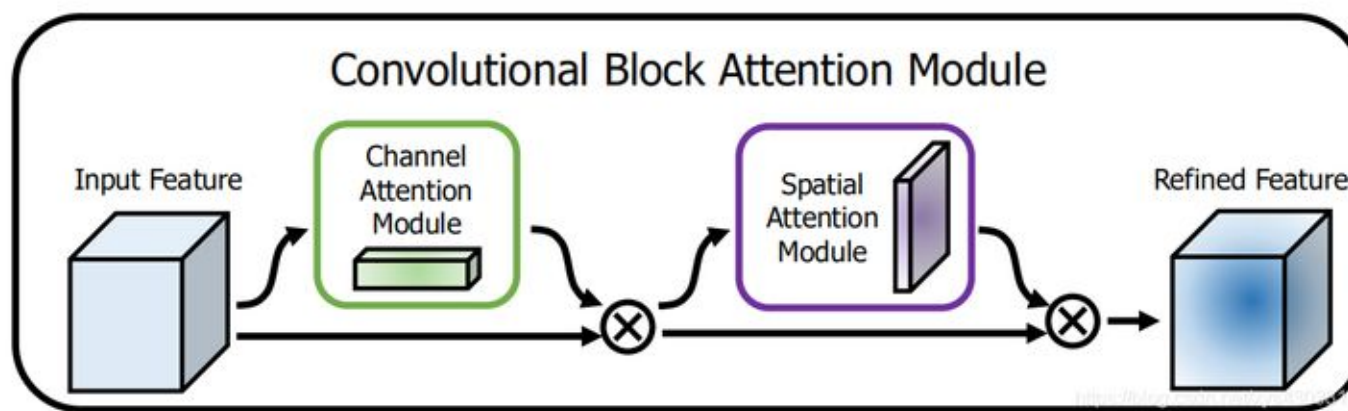
那么问题来了，怎么实现 通过PCA来获得 Attention weights呢？

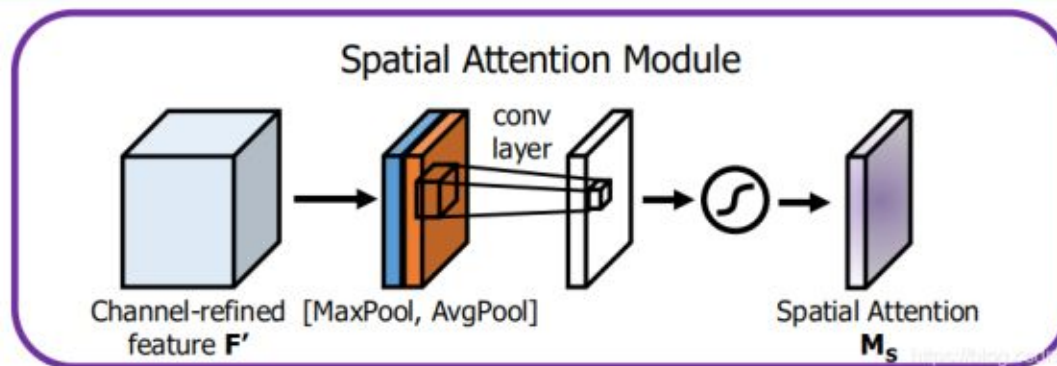
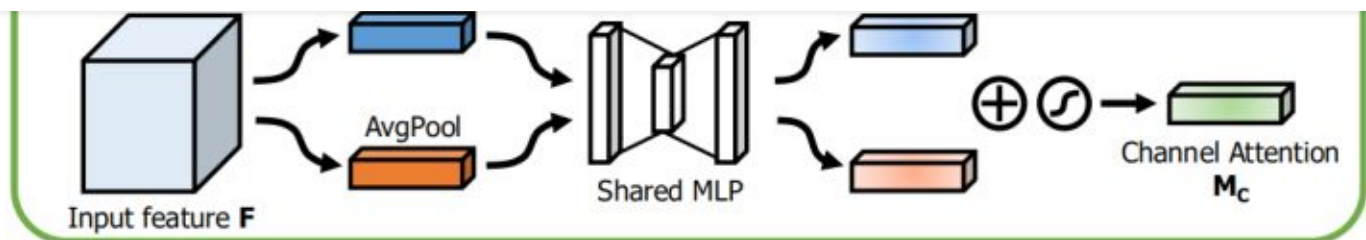
文中不直接使用 协方差矩阵的特征值分解 来实现，而是使用下述等价形式：

CBAM: Convolutional Block Attention Module(通道域+空间域), ECCV2018

这货就是基于 SE-Net [5]中的 Squeeze-and-Excitation module 来进行进一步拓展，

具体来说，文中把 channel-wise attention 看成是教网络 Look 'what' ；而spatial attention 看成是教网络 Look 'where' ，所以它比 SE Module 的主要优势就多了后者





通道注意力公式：

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F_{avg}^c) + W_1(W_{max}^c))) \end{aligned}$$

空间注意力公式：（空间域注意力是通过对通道axis进行AvgPool和MaxPool得来的）

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

CBAM 特别轻量级，也方便在端部署。

DANet: Dual Attention Network for Scene Segmentation (空间域+通道域)

CPVR2019

赞同 692

14 条评论

分享

喜欢

收藏

把Self-attention的思想用在图像分割，可通过long-range上下文关系更好地做到精准分割。

主要思想也是上述文章 **CBAM** 和 **non-local** 的融合变形：

把deep feature map进行spatial-wise self-attention，同时也进行channel-wise self-attetnion，最后将两个结果进行 element-wise sum 融合。

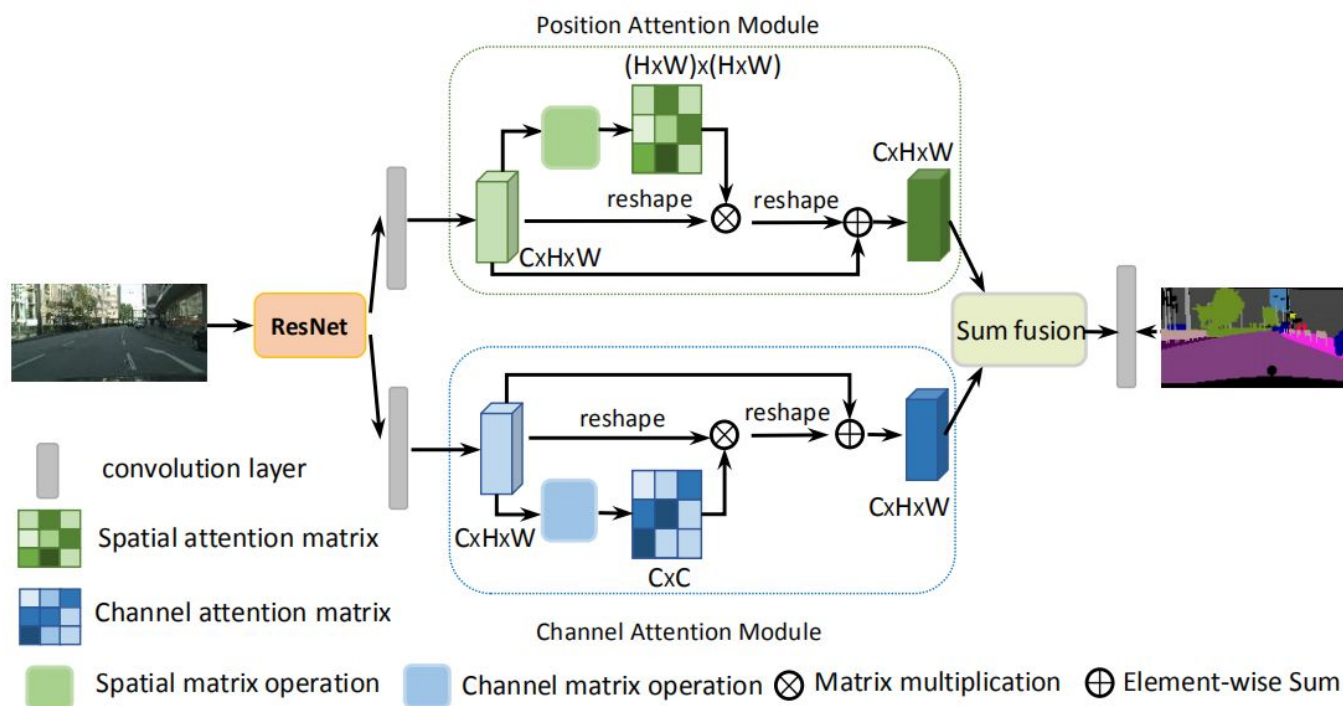
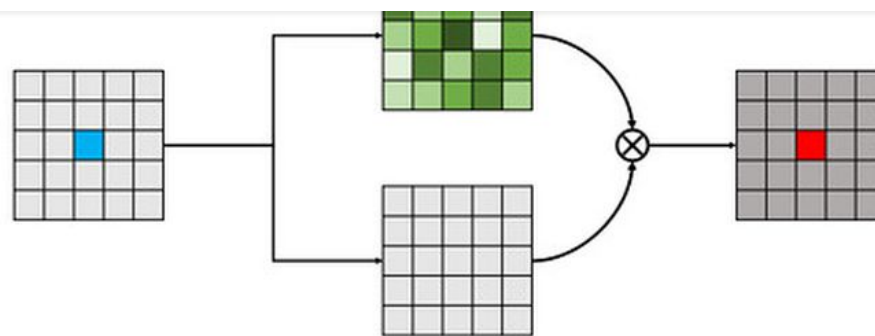


Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

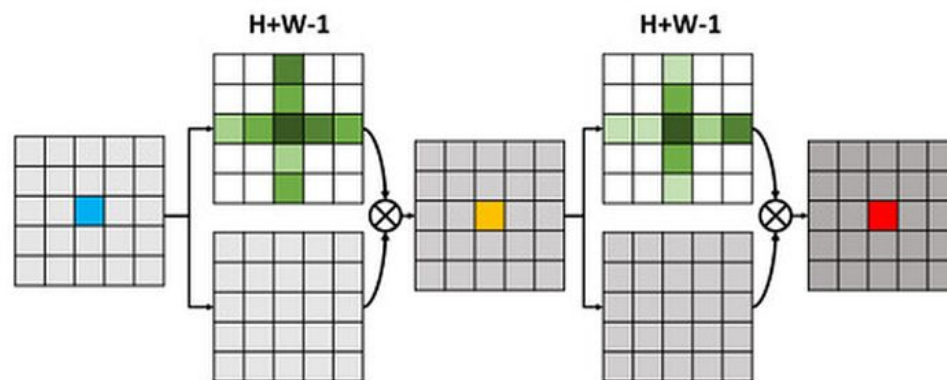
这样做的好处是：

CCNet

本篇文章的亮点在于用了巧妙的方法减少了参数量。在上面的DANet中，attention map计算的是所有像素与所有像素之间的相似性，空间复杂度为 $(H \times W) \times (H \times W)$ ，而本文采用了criss-cross思想，只计算每个像素与其同行同列即十字上的像素的相似性，通过进行循环(两次相同操作)，间接计算到每个像素与每个像素的相似性，将空间复杂度降为 $(H \times W) \times (H + W - 1)$ ，以图为例为下：

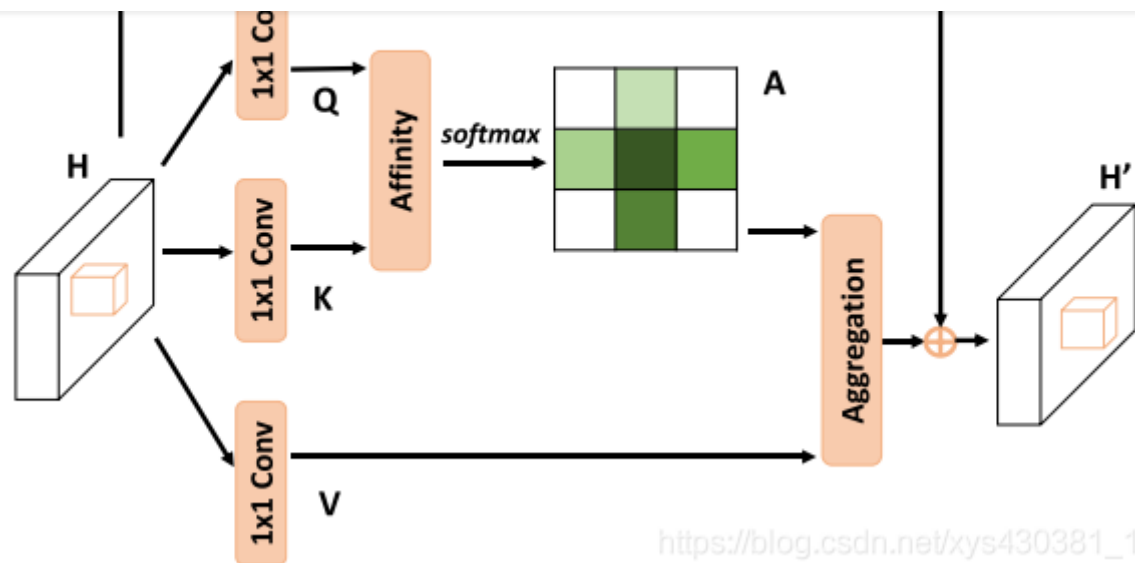


(a) Non-local block

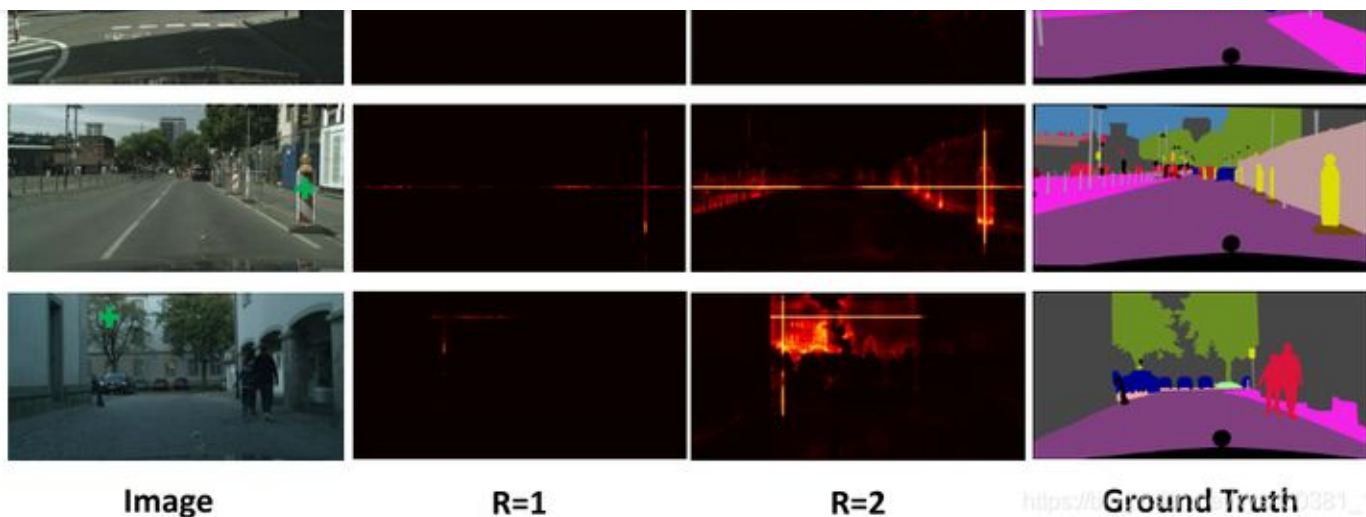


(b) Criss-Cross Attention block

整个网络的架构与DANet相同，只不过attention模块有所不同，如下图：在计算矩阵相乘时每个像素只抽取特征图中对应十字位置的像素进行点乘，计算相似度。



经过一轮此attention计算得到的attention map如下图R1所示，对于每个元素只有十字上的相似性，而通过两轮此计算，对于每个元素就会得到整张图的相似性，如R2。



得到此结果的原因如下图，经过一轮计算，每个像素可以得到在其十字上的相似性，对于不同列不同行(不在其十字上)的像素是没有相似性的，但是这个不同行不同列像素同样也进行了相似性计算，计算了在其十字上的相似性，那么两个十字必有相交，在第二次attention计算的时候，通过交点，相当于是间接计算了这两个不同列不同行像素之间的相似性。

实验结果达到了SOTA水平，但没有计算全部像素的attention方法准确率高。

OCNet

- [OCNet: Object Context Network for Scene Parsing \(Microsoft Research\)论文解析](#)
- [图像语义分割\(13\)-OCNet: 用于场景解析的目标语义网络](#)

摘要

论文侧重于语义分割中的语义聚集策略，即不再逐像素

▲ 赞同 692



● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

具体实现受到自注意力机制的影响包含两个步骤：1)计算单个像素和所有像素之间的相似性从而得到目标语义和每一个像素的映射；2)得到目标像素的标签。结果比现有的语义聚集策略例如PPM和ASPP这些不区别单一像素和目标语义之间是否存在属于关系的策略更加准确。

不得不说，这篇论文和DANet撞车了，而且撞的死死的，用的同样的核心内容

GCNet:Non-local Networks Meet Squeeze-Excitation Networks and Beyond

Non-local Networks Meet Squeeze-Excitation Networks and Beyond[论文解读](#)

GCNet 网络结构结构了non-local network和Squeeze-excitation networks.我们知道non-local network(NLNet) 可以捕获长距离依赖关系。可以发现NLnet的网络结构采用的是自注意力机制来建模像素对关系。*在这篇文章中non-local network的全局上下文在不同位置几乎是相同的，这表明学习到了无位置依赖的全局上下文，因此这样导致了大量的计算量的浪费。作者在这里提出了一种简化版的模型去获得全局上下文信息。*使用的是query-independent(可以理解为无query依赖)的建模方式。同时更可以共享这个简化的结构和SENet网络结构。因此作者在这里联合了这三种方法产生了一个global context(GC) block

注意增强型卷积

用自注意力增强卷积：这是新老两代神经网络的对话（附实现）

PyTorch 实现地址

PS：启示就是在卷积算子的基础上，加入了多个注意力（每个注意力头都是一个非局部卷积块，然后将这些head concat到一起。

▲ 赞同 692



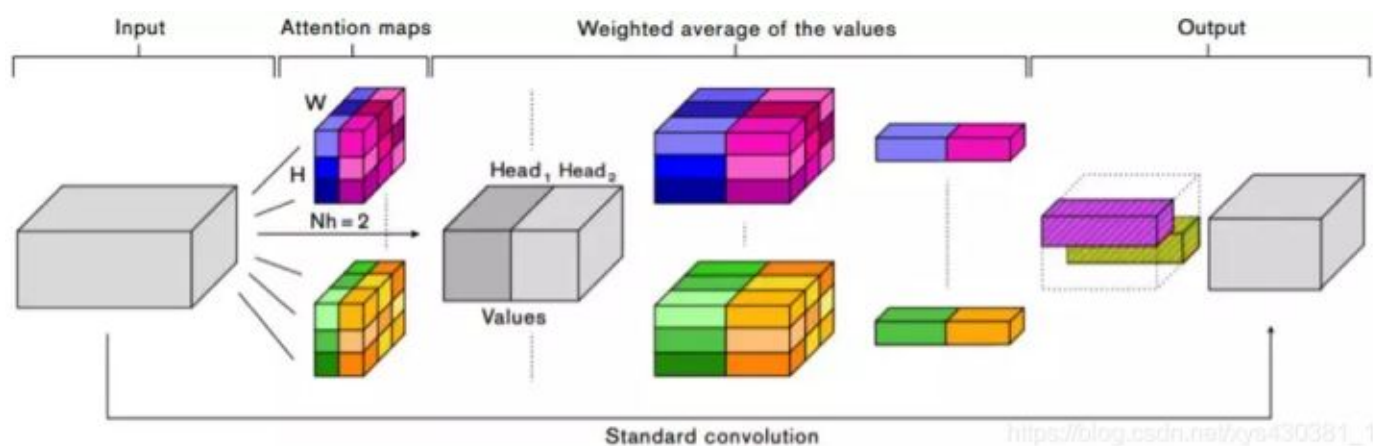
● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

卷积运算有一个显著缺陷，即仅在局部近邻上工作，也由此会错失全局信息。另一方面，自注意则是获取长程交互性方面的一项近期进展，但还主要应用于序列建模和生成建模任务上。在这篇论文中，我们研究了将自注意（作为卷积的替代）用于判别式视觉任务的问题。我们提出了一种全新的二维相对自注意机制，研究表明这足以在图像分类任务上替代卷积作为一种单独的原语。我们在对照实验中发现，当结合使用卷积与自注意时所得到的结果最好。**因此我们提出使用这种自注意机制来增强卷积算子，具体做法是将卷积特征图与通过自注意产生的一组特征图连接起来。**



```

# X has shape [B, H, W, F_in]
conv_out = tf.layers.conv2d(X, Fout - dv, k)
# [B, Nh, HW, dvh or dkh]
flat_q, flat_k, flat_v = compute_flat_qkv(X, dk, dv)
# [B, Nh, HW, HW]
logits = tf.matmul(flat_q, flat_k, transpose_b=True)
if relative:
    h_rel_logits, w_rel_logits = relative_logits(q)
    logits += h_rel_logits
    logits += w_rel_logits
weights = tf.nn.softmax(logits)
# [B, Nh, HW, dvh]
attn_out = tf.matmul(weights, flat_v)
attn_out = tf.reshape(v, [B, Nh, H, W, dv // Nh])
attn_out = combine_heads_2d(v) # [B, H, W, dv]
attn_out = tf.layers.conv2d(attn_out, dv, 1)
return tf.concat([conv_out, attn_out], axis=3)

```

<https://blog.csdn.net/zys430381>

单个注意力head

给定一个形状为 (H, W, F_{in}) 的输入张量，我们将其展开为一个矩阵 $X \in \mathbb{R}^{HW \times F_{in}}$ ，并如 Transformer 架构提出的那样执行多头注意。则单个头 h 的自注意机制输出为

$$O_h = \text{Softmax} \left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_k^h}} \right) (XW_v)$$

所有头的输出可以连接起来：

$$MHA(X) = Concat [O_1, \dots, O_{Nh}] W^O$$

我们使用不同规模不同类型的模型（其中包括 ResNet 和一种当前最佳的可移动式受限网络）进行了广泛的实验，结果表明注意增强能在 ImageNet 图像分类与 COCO 目标检测任务上实现稳定的提升，同时还能保证参数数量大体接近。尤其值得提及的是，我们的方法在 ImageNet 上实现的 top-1 准确度优于 ResNet50 基准 1.3% 我们的方法还在 COCO 目标检测上超过 RetinaNet 基准 1.4 mAP。

注意增强仅需极少量的计算负担就能实现系统性的改善，并且**在所有实验中都明显优于流行的 Squeeze-and-Excitation 通道式注意方法。**

实验还有个让人惊讶的结果：在 ImageNet 上全自注意模型（注意增强的一种特例）的表现仅略逊于对应的全卷积模型，这说明自注意本身就是一种强大的图像分类基本方法。

PAN: Pyramid Attention Network for Semantic Segmentation(层域)—CVPR2018

亮点1：论文是将Attention机制与金字塔结构结合作为本文的亮点，这样可以在高层语义指导的基础上来提取相对与较低层的精确的密集特征，取代了其他方法里面的复杂的空洞卷积dilated和多个编码解码器的操作，跳出了以往常常用到的U-Net结构；

亮点2：:采用了一个全局pooling进行底层特征的权值加权，对特征的map起到的选取的作用。

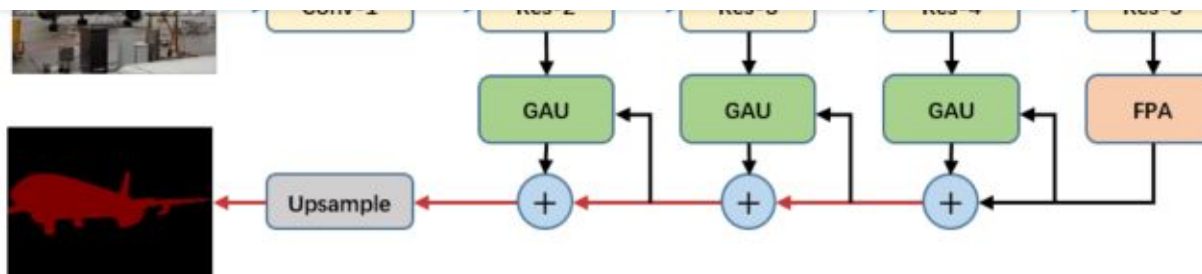


Figure 2: Overview of the Pyramid Attention Network. We use ResNet-101 to extract dense features. Then we perform FPA and GAU to extract precise pixel prediction and localization details. The blue and red lines represent the downsample and upsample operators respectively.

模块一：FPA（FeaturePyramid Attention）特征金字塔Attention

- 解决的问题：不同的scale大小的图片以及不同大小的物体给物体分割带来了困难
- 现有的方法：类似于PSPNet、DeepLab采用空间金字塔pooling实现不同的尺度以及多孔金字塔池化ASPP结构，问题一：pooling容易丢失掉局部信息，问题二：ASPP因为是一种稀疏的操作会造成棋盘伪影效应，问题三：只是简单地多个scale concat缺乏上下文的信息，没有关注上下文信息情况下效果不佳（下图作图为现有的方法），该部分处理主要是用在处理高层特征上的操作。
- 提出的方案：如右图所示，在提取到高层特征之后不再进行pooling的操作，而是通过三个连续的卷积实现更高层的语义，我们知道更高层的语义会更加接近ground truth的情况，会关注一些物体信息，所以用更高层的语义来作为一种Attention的指导，与高层特征做完 1×1 卷积不变化大小的情况下进行相乘，也就是加强了具有物体信息的部位带有的权值，得到了带有Attention的输出，同时因为金字塔卷积的结构采用下采样和上采样的操作，所以也解决不同物体不同scale的问题。

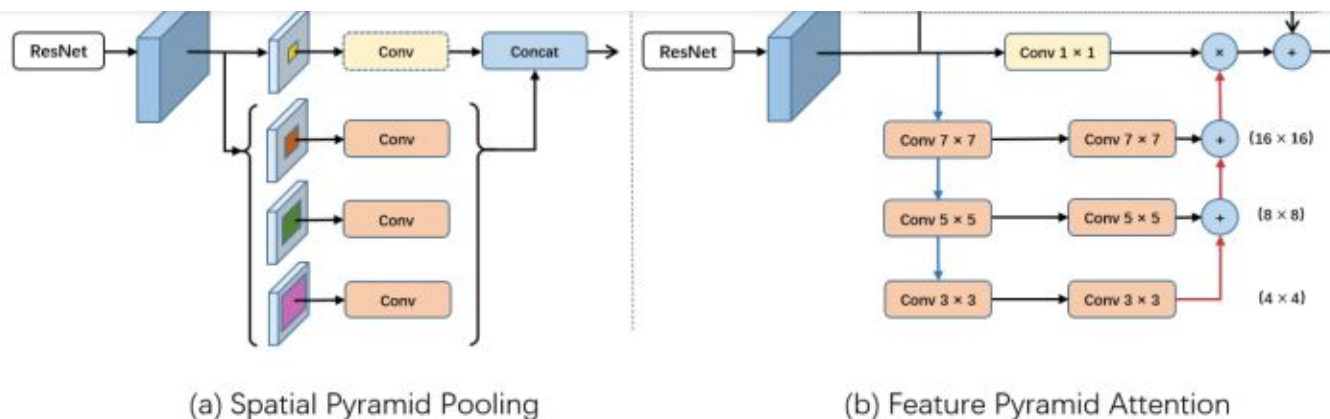


Figure 3: Feature Pyramid Attention module structure. (a) Spatial Pyramid Pooling structure. (b) Feature Pyramid Attention module. '4 × 4, 8 × 8, 16 × 16, 32 × 32' means the resolution of feature map. The dotted box means the global pooling branch. The blue and red lines represent the downsample and upsample operators respectively.

模块二：

- 解决的问题：对于高层的特征常常可以实现有效的分类，但是重构原始图像的解析度或者说 predict 上无法精细地实现。
- 现有的方法：类似于 SegNet、Refinenet、提拉米苏结构等等都是采用了 U-Net 的结构，采用了解码器 decoder 也就是反卷积之类再加上底层的特征，一层层地往上累加以便恢复图像细节，论文中讲到了这种虽然是可以实现底层和高层的结合以及图像重构，但是 computation burden
- 提出的方案：如下图所示，抛弃了 decoder 的结构，原始形式是直接底层特征加 FPA 得到的高层特征，但在 skip 底层特征的时候论文采用了高层特征作为指导设置了相应的权重，使得底层与高层的权重保持一致性，**高层特征采用了 Global**

- 我们对低层次特征执行 3×3 的卷积操作，以减少 CNN 特征图的通道数。
- 从高层次特征生成的全局上下文信息依次经过 1×1 卷积、批量归一化 (batch normalization) 和非线性变换操作 (nonlinearity)，然后再与低层次特征相乘。
- 最后，高层次特征与加权后的低层次特征相加并进行逐步的上采样过程。
- 我们的 GAU 模块不仅能够更有效地适应不同尺度下的特征映射，还能以简单的方式为低层次的特征映射提供指导信息。

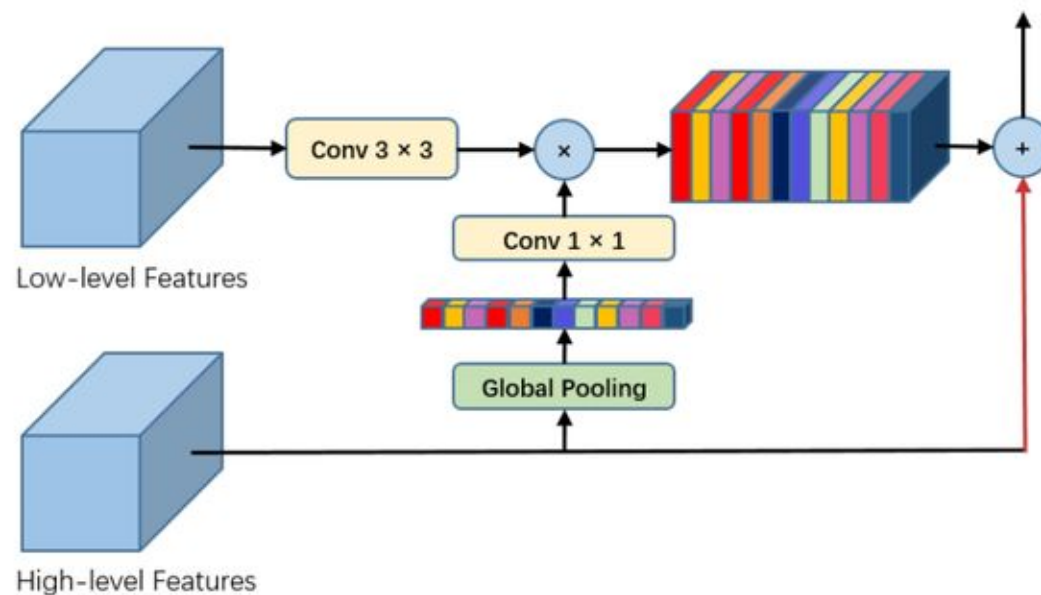


Figure 4: Global Attention Upsample module structure

硬注意力

一种通过引入硬注意力机制来引导学习视觉回答任务的研究

软注意力机制已在计算机视觉领域取得了广泛的应用和成功。但是我们发现硬注意力机制在计算机视觉任务中的研究还相对空白。而硬注意力机制能够从输入信息中选择重要的特征，因此它被视为是一种比较软注意力机制更高效、直接的方法。本次，将为大家介绍一种通过引入硬注意力机制来引导学习视觉回答任务的研究。此外结合 L2 正则化筛选特征向量，可以高效地促进筛选的过程并取得更好的整体表现，而无需专门的学习过程。

1. Diversified visual attention networks for fine-grained object classification—2016

2. Deep networks with internal selective attention through feedback connections (通道域)—NIPS 2014

提出了一种Deep Attention Selective Network (dasNet)。在训练完成后，通过强化学习 (Separable Natural Evolution Strategies) 来动态改变attention。具体来说，attention调整的是每个conv filter的权重（和SENet一样有木有，都是channel维度）。policy是一个neural network，RL部分的算法如下：

```

1: while True do
2:    $images \leftarrow \text{NEXTBATCH}(n)$ 
3:   for  $i = 0 \rightarrow p$  do
4:      $\theta_i \sim \mathcal{N}(\mu, \Sigma)$ 
5:     for  $j = 0 \rightarrow n$  do
6:        $a_0 \leftarrow \mathbb{1}$  {Initialise gates  $a$  with identity activation}
7:       for  $t = 0 \rightarrow T$  do
8:          $v_t = \mathbf{M}_t(\theta_i, x_i)$ 
9:          $o_t \leftarrow h(\mathbf{M}_t)$ 
10:         $a_{t+1} \leftarrow \pi_{\theta_i}(o_t)$ 
11:      end for
12:       $L_i = -\lambda_{\text{boost}} d \log(v_T)$ 
13:    end for
14:     $\mathcal{F}[i] \leftarrow f(\theta_i)$ 
15:     $\Theta[i] \leftarrow \theta_i$ 
16:  end for
17:   $\text{UPDATESNES}(\mathcal{F}, \Theta)$  {Details in supplementary material.}
18: end while

```

<http://logos.dccnet.org/way02019/>

其中每次while循环代表一次SNES迭代，M表示训练好的CNN， μ 和Sigma是policy参数对应的分布超参，p是采样p个policy参数，n是随机抽取n张图片。

3. Fully Convolutional Attention Networks for Fine-Grained Recognition

本文利用基于强化学习的视觉 attention model 来模拟物体分类。这个框架模拟人类视觉系统的识别过程，通

▲ 赞同 692

▼

● 14 条评论

➤ 分享

♥ 喜欢

★ 收藏

次物体part。每一个 glimpse的位置作为一个 action，图像和之前glimpse的位置作为 state，奖励衡量分类的准确性。本文方法可以同时定位多个part，之前的方法只能一次定位一个part

4. 时间域注意力(RNN)

这个概念其实比较大，因为计算机视觉只是单一识别图片的话，并没有时间域这个概念，但是[7]这篇文章中，提出了一种基于递归神经网络（Recurrent Neural Network，RNN）的注意力机制识别模型。

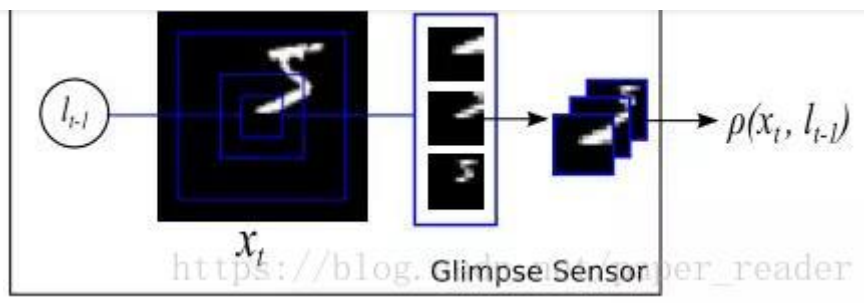
RNN模型比较适合的场景是数据具有时序特征，比如使用RNN产生注意力机制做的比较好的是在自然语言处理的问题上。因为自然语言处理的是文本分析，而文本产生的背后其实是有一个时序上的关联性，比如一个词之后还会跟着另外一个词，这就是一个时序上的依赖关联性。

而图片数据本身，并不具有天然的时序特征，一张图片往往是一个时间点下的采样。但是在视频数据中，RNN就是一个比较好的数据模型，从而能够使用RNN来产生识别注意力。

特意将RNN的模型称之为时间域的注意力，是因为这种模型在前面介绍的空间域，通道域，以及混合域之上，又新增加了一个时间的维度。这个维度的产生，其实是基于采样点的时序特征。

Recurrent Attention Model [7]中将注意力机制看成对一张图片上的一个区域点的采样，这个采样点就是需要注意的点。而**这个模型中的注意力因为不再是一个可以微分的注意力信息，因此这也是一个强注意力（hard attention）模型**。这个模型的训练是需要使用**增强学习（reinforcement learning）**来训练的，训练的时间更长。

这个模型更需要了解的并不是RNN注意力模型，因为这个模型其实在自然语言处理中介绍的更详细，更需要了解的是这个模型的如何将图片信息转换成



这个是模型中的关键点，叫Glimpse Sensor，我翻译为扫视器，这个sensor的关键点在于先确定好图片中需要关注的点（像素），这时候这个sensor开始采集三种信息，信息量是相同的，一个是非常细节（最内层框）的信息，一个是中等的局部信息，一个是粗略的略缩图信息。

这三个采样的信息是在 l_{t-1} 位置中产生的图片信息，而下一个时刻，随着 t 的增加，采样的位置又开始变化，至于 l 随着 t 该怎么变化，这就是需要使用增强学习来训练的东西了。

自注意力

- [自注意力机制在计算机视觉中的应用](#)
- [用Attention玩转CV，一文总览自注意力语义分割进展](#)

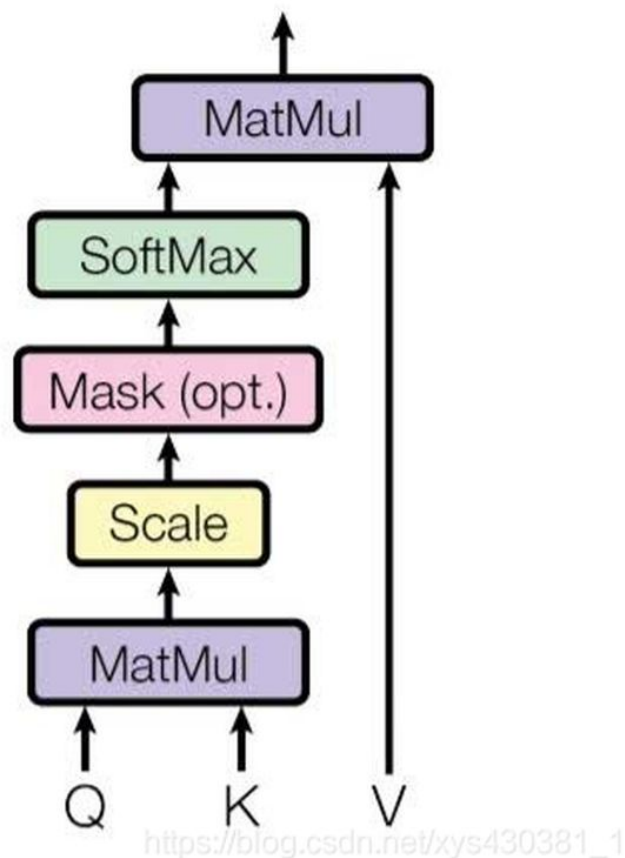
自注意力机制是注意力机制的改进，其减少了对外部信息的依赖，更擅长捕捉数据或特征的内部相关性。

在神经网络中，我们知道卷积层通过卷积核和原始特征的线性结合得到输出特征，由于卷积核通常是局部的，为了增加感受野，往往采取堆叠卷积层的方

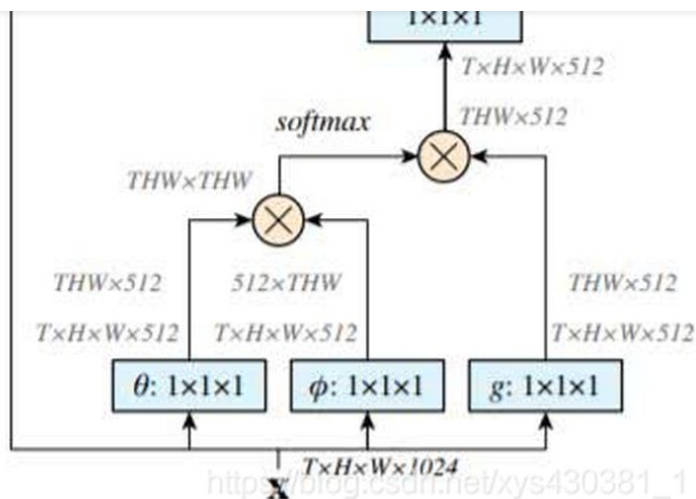
自注意力机制 (self-attention)[1] 在序列模型中取得了很大的进步；另外一方面，上下文信息 (context information) 对于很多视觉任务都很关键，如语义分割，目标检测。**自注意力机制通过 (key, query, value) 的三元组提供了一种有效的捕捉全局上下文信息的建模方式。**接下来首先介绍几篇相应的工作，然后分析相应的优缺点以及改进方向。

Related Works

Attention is all you need [1] 是第一篇提出在序列模型中利用自注意力机制取代循环神经网络的工作，取得了很大的成功。其中一个重要的模块是缩放点积注意力模块 (scaled dot-product attention)。文中提出 (key, query, value) 三元组捕捉长距离依赖的建模方式，如下图所示，key和query通过点乘的方式获得相应的注意力权重，最后把得到的权重和value做点乘得到最终的输出。



Non-local neural network [2] 继承了(key, query, value) 三元组的建模方式, 提出了一个高效的 non-local 模块, 如下图所示。在Resnet网络中加入non-local模块后无论是目标检测还是实例分割, 性能都有一个点以上的提升 (mAP), 这说明了上下文信息建模的重要性。



Danet [3]是来自中科院自动化的工作，其核心思想就是通过上下文信息来监督语义分割任务。作者采用两种方式的注意力形式，如下图所示，分别是spatial和 channel上，之后进行特征融合，最后接语义分割的head 网络。思路上来说很简单，也取得了很好的效果。

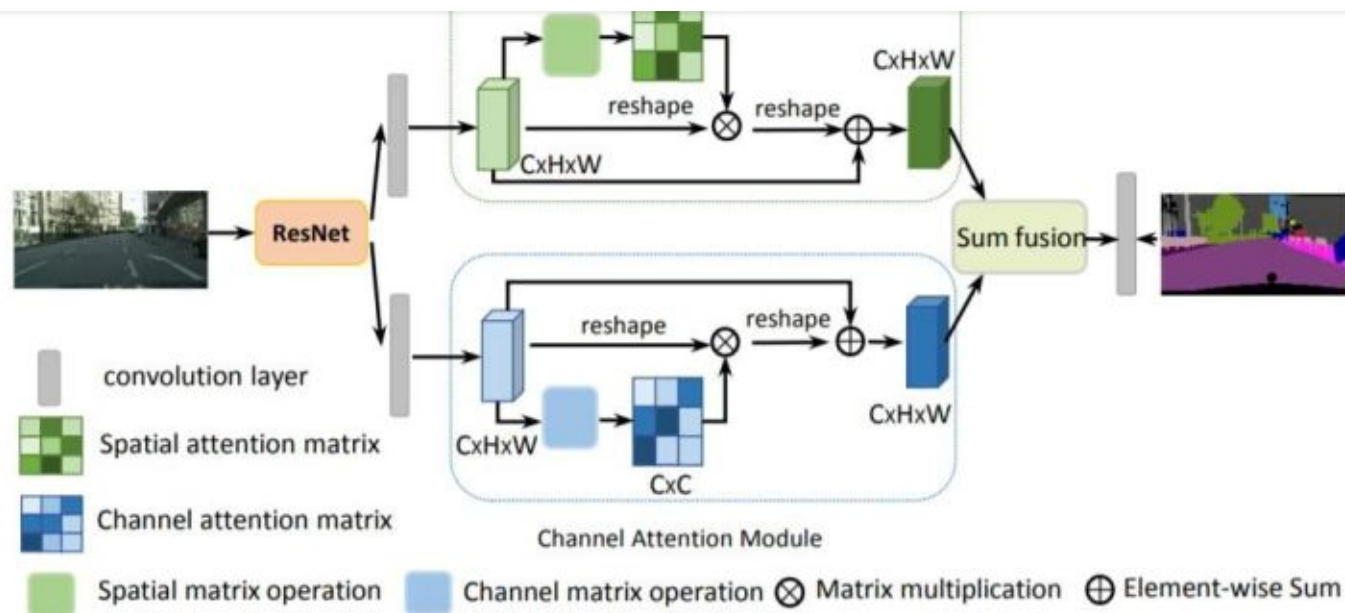
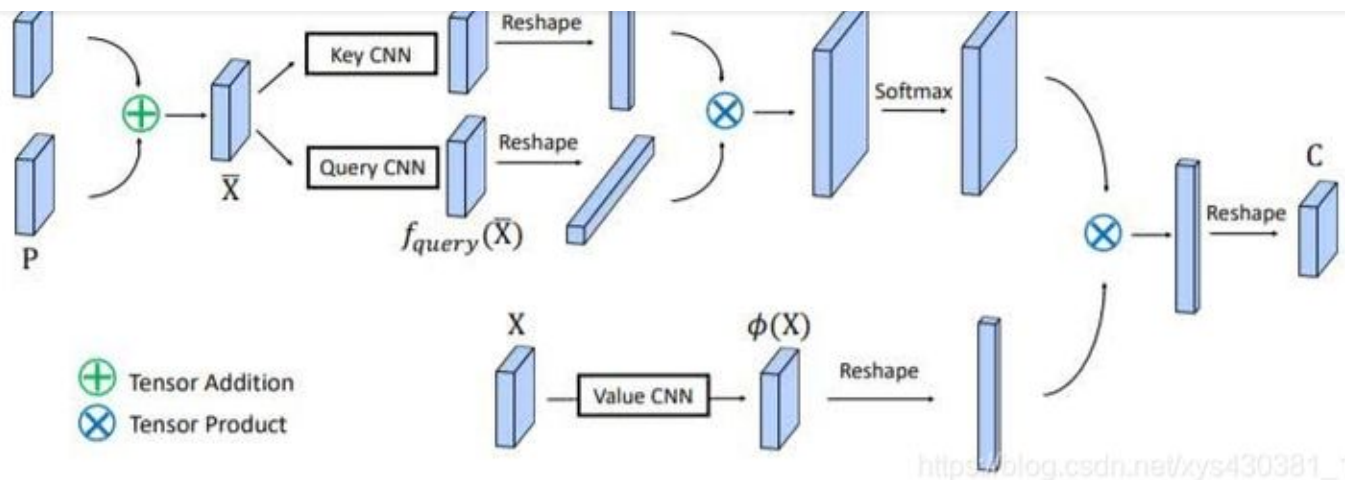
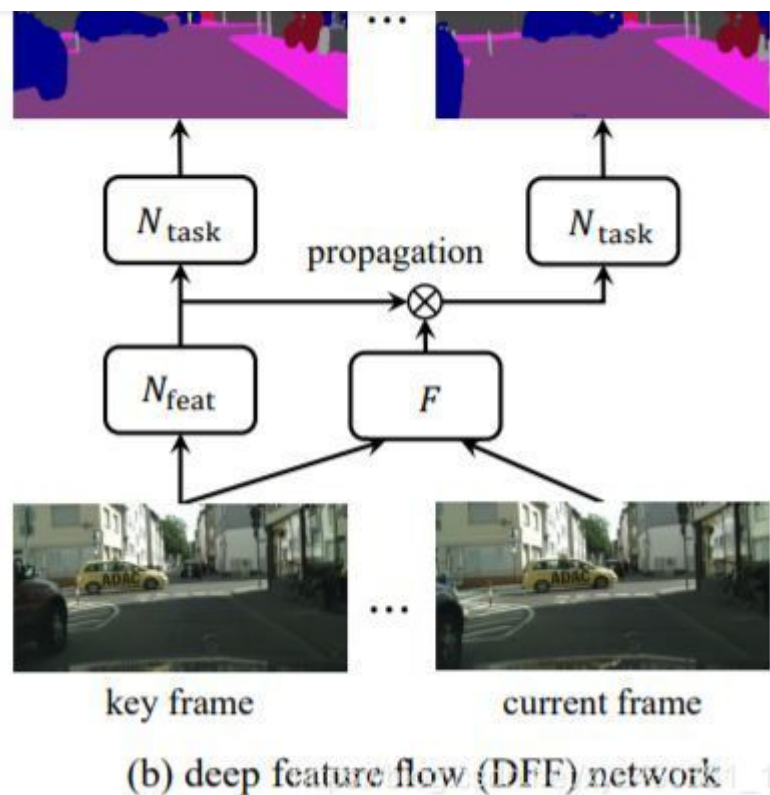


Figure 2: An overview of the Dual Attention Network. (Best viewed in color) blog.csdn.net/xys430381_1

Ocnet[4]是来自微软亚洲研究所的工作。同样它采用 (key, query, value) 的三元组，通过捕捉全局的上下文信息来更好的监督语义分割任务。与Danet [3]不同的是它仅仅采用spatial上的信息。最后也取得了不错的结果。



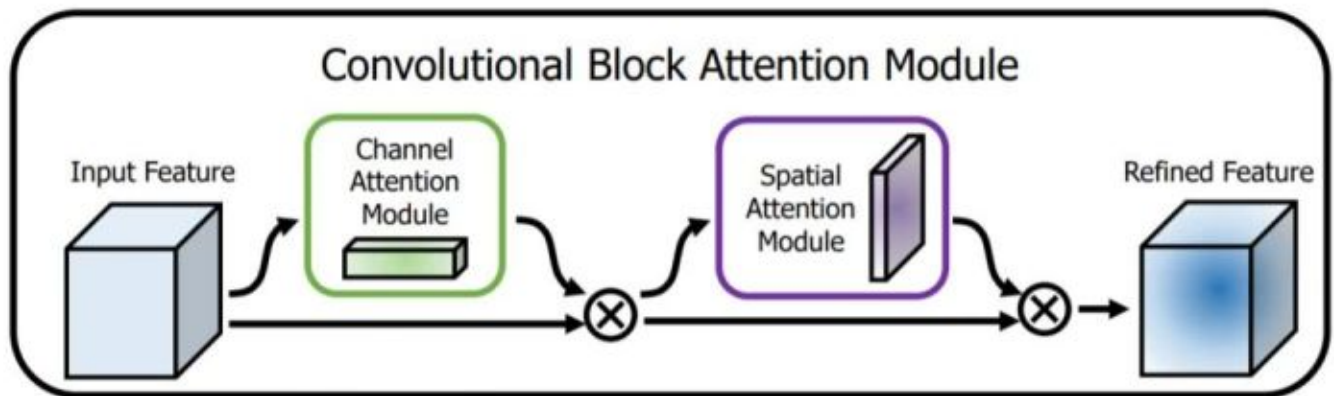
DFF [5] 是来自微软亚洲研究所视觉计算组的工作。如下图所示，它通过光流来对视频不同帧之间的运动信息进行建模, 从而提出了一个十分优雅的视频检测框架DFF。其中一个很重要的操作是 warp, 它实现了点到点之间的对齐。在此以后出现了很多关于视频检测的工作, 如, FGFA[6], Towards High Performance [7]等, 他们大部分都是基于warp这个特征对其操作。由于光流网络的不准确性以及需要和检测网络进行联合训练, 这说明现在视频检测中的光流计算其实不准确的。如何进行更好的建模来代替warp操作, 并且起到同样的特征对其的作用是很关键的。通常而言我们假设flow运动的信息不会太远, 这容易启发我们想到通过每个点的邻域去找相应的运动后的特征点, 具体做法先不介绍了, 欢迎大家思考 (相关操作和自注意力机制)。



自注意力的缺点和改进策略

前面主要是简单的介绍了自注意力机制的用途，接下来分析它的缺点和相应的改进策略，**由于每一个点都要捕捉全局的上下文信息，这就导致了自注意力机制模块会有很大的计算复杂度和显存容量。**如果我们能知道一些先验信息，比如上述的特征对其通常是一定的邻域内，我们可以通过限制在一定的邻域内来做。另外还有如何进行高效的稀疏化，以及和图卷积的联系，这些都是很开放的问题，欢迎大家积极思考。

CBAW [10] 提出了结合spatial和channel的模块，如下图所示，在各项任务上也取得很好的效果。



最后介绍一篇来自百度IDL的结合channel as spatial的建模方式的工作 [11]。本质上是直接在 (key, query, value) 三元组进行reshape的时候把channel的信息加进去，但是这带来一个很重要的问题就是计算复杂度大大增加。我们知道分组卷积是一种有效的降低参数量的方案，这里也采用分组的方式。但是即使采用分组任然不能从根本上解决计算复杂度和参数量大的问题，作者很巧妙的利用泰勒级数展开后调整计算key, query, value的顺序，有效的降低了相应的计算复杂度。下表是优化后的计算量和复杂度分析，下图是CGNL模块的整体框架。

自注意力小结

自注意力机制作为一个有效的对上下文进行建模的方式，在很多视觉任务上都取得了不错的效果。同时，这种建模方式的缺点也是显而易见的，一是没有考虑channel上信息，二是计算复杂度仍然很大。相应的改进策，一方面是如何进行spatial和cha进行捕捉信息的稀疏化，关于稀疏的好处是可以更加鲁

编辑于 2020-09-21

深度神经网络

计算机视觉

注意力机制

文章被以下专栏收录



极市平台

原创计算机视觉技术干货分享，微信公众号：极市平台



计算机视觉

计算机视觉相关知识介绍



统计与机器学习

推荐阅读

▲ 赞同 692

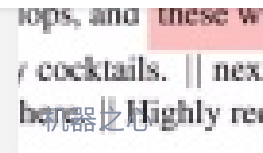


💬 14 条评论

➦ 分享

♥ 喜欢

★ 收藏



14 条评论

⇌ 切换为时间排序

写下你的评论...



Wet sand

2020-09-21

推荐一篇 CVPR 2020 的 Exploring Self-attention for Image Recognition，我自己读觉得满有趣的

5



极市平台 (作者) 回复 Wet sand

2020-09-21



赞



随风杀 回复 Wet sand

03-15

你好，Exploring Self-attention for Image Recognition，这篇的代码你泡通了吗

赞



代號C

你好，时间域注意力引用的文献7是哪一篇呀？

▲ 赞同 692 ▼

💬 14 条评论

➦ 分享

❤️ 喜欢

★ 收藏



非是藉秋风

2020-09-22

Attention is nothing but interaction

1



kk1201

03-09

想请教下什么是长距离依赖关系(long-range dependencies)? 新手不是很懂

赞



邝鹏飞

2020-12-18

感觉就是罗列，没有系统性的分析

赞



Bailuo chris 回复 邝鹏飞

01-22

罗列出来，别人一看就知道从哪儿入手了呀。毕竟大家要解决的问题都是不太相同的。

2



吾生而有涯

2020-10-27

我感觉卷积不就是一种注意力吗？注意图片上的一些特征。注意力机制是“更注意了”吗？

赞



u 呼吁 回复 吾生而有涯

03-13

感觉卷积偏局部，注意力机制偏全局吧

3



战术卷卷卷 回复 吾生而有涯

赞同 692



14 条评论


分享

喜欢

收藏

我觉得卷积是local networks. 没法关注到glo



姑娘眉眼弯弯 

2020-09-26

非常详细了，感谢~




赞



Crimson Sky

2020-09-21

收藏了 



赞



数据de遐想

2020-09-21

棒



赞

▲ 赞同 692 ▼



💬 14 条评论

🔗 分享

❤️ 喜欢

★ 收藏