**Abstract:**

In this survey, we provide a comprehensive review of various attention mechanisms in computer vision and categorize them according to approach, such as channel attention, spatial attention, temporal attention and branch attention

作者在本文全面审视了CV中的注意力机制并将他们分类成通道、空间、时间和分支注意力

**Intro:**

In a vision system, an attention mechanism can be treated as a dynamic selection process that is realized by adaptively weighting features according to the importance of the input.

在视觉系统中，注意力机制可以看成根据输入的重要性自适应理解权重特征的一个动态选择过程

Progress can be coarsely divided into four phases. The first phase begins from RAM, pioneering work that combined deep neural networks with attention mechanisms. In this phase, recurrent neural networks(RNNs) were necessary tools for an attention mechanism.

注意力机制的应用进程可以分成四个时期，第一个时期是RAM，开创性的结合了深度神经网络和注意力机制。RNN也是注意力机制的重要工具。

At the start of the second phase, Jaderberg et al proposed the STN which introduces a sub-network to predict an affine transformation used to to select important regions in the input. Explicitly predicting discriminatory input features is the major characteristic of the second phase; DCNs are representative works.

STN引入子网来预测用于选择输入中重要区域的仿射变换，DCN也是一个有代表性的工作。明确歧视性输入特征是第二阶段的主要特点

The third phase began with SENet that presented a novel channel-attention network which implicitly and adaptively predicts the potential key features. CBAM and ECANet are representative works of this phase.

第三阶段开始与SENet，提出了全新的通道注意力机制网络，它隐式且自适应的提出潜在关键特征。CBAM和ECANet是这个阶段代表性工作。

The last phase is the self-attention era. Self-attention was firstly proposed in and rapidly provided great advances in the field of natural language processing, Wang et al. took the lead in introducing self-attention to computer vision and presented a novel non-local network with great success in video understanding and object detection.

最后一阶段是自注意力机制，自注意力机制最先提出于自然语言处理领域，wang等人把它介绍到CV，并提出了non-local网络，在视频理解和目标检测表现很好，其他具体提出网络见Figure3

The main contributions of this paper are:

• a systematic review of visual attention methods, covering the unified description of attention mechanisms, the development of visual attention mechanisms as well as current research,

• a categorisation grouping attention methods according to their data domain, allowing us to link visual attention methods independently of their particular application, and

• suggestions for future research in visual attention.

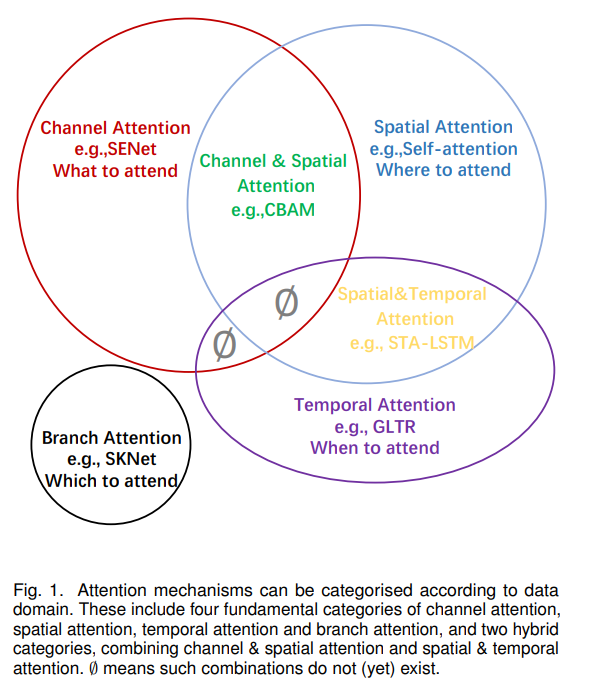
本文主要的贡献是：对注意力机制的统一描述，注意力机制的发展；根据数据域对注意力机制进行分类；对未来视觉注意力的研究方向的建议

**Conclusion:**

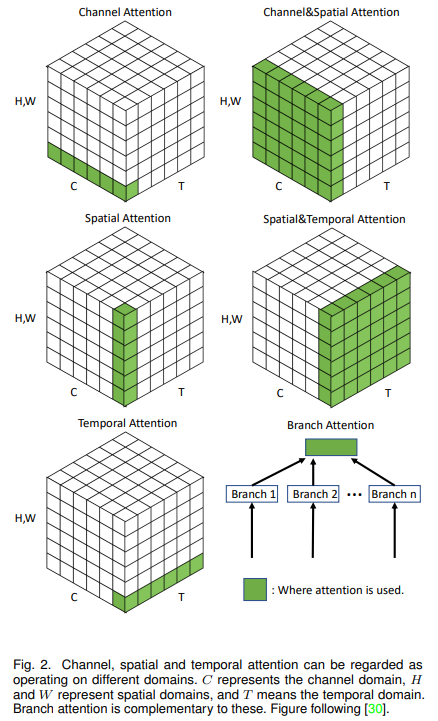
Attention mechanisms have become an indispensable technique in the field of computer vision in the era of deep learning. This survey has systematically reviewed and summarized attention mechanisms for deep neural networks in computer vision. We have grouped different attention methods according to their domain of operation, rather than by application task, and show that attention models can be regarded as an independent topic in their own right. We have concluded with some potential directions for future research.

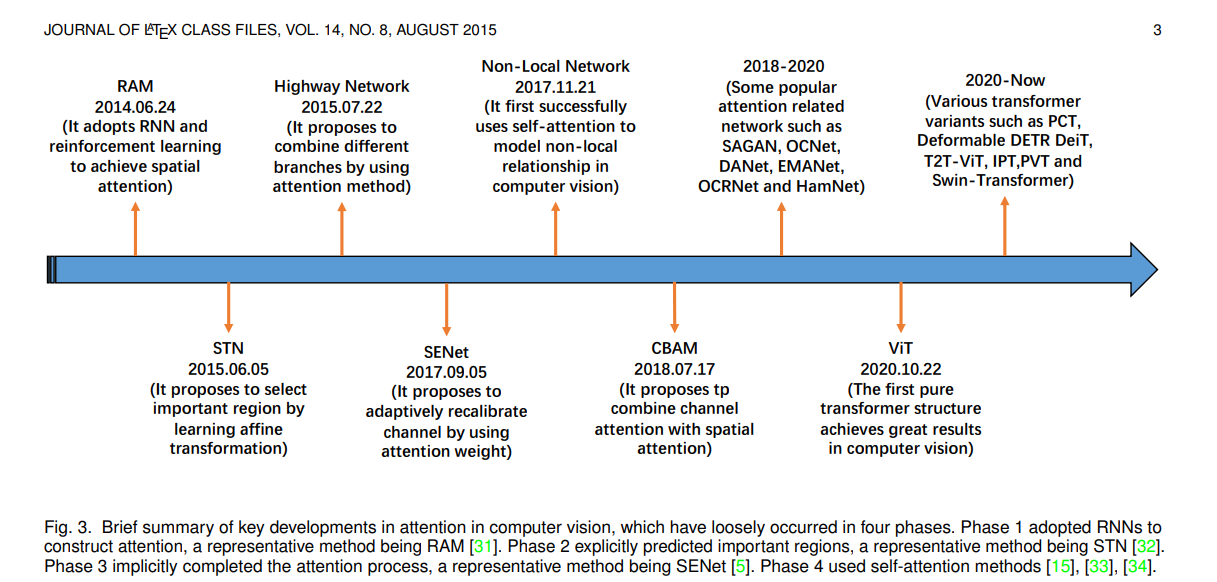
注意力机制变成了CV领域中深度神经网络的不可缺少的技术，本文系统的回顾了深度神经网络中的注意力机制，并根据他们的操作领域对他们进行分类，表明了它们可以作为独立的话题。本文还对未来有潜力的方向进行了总结。

**Figure：**

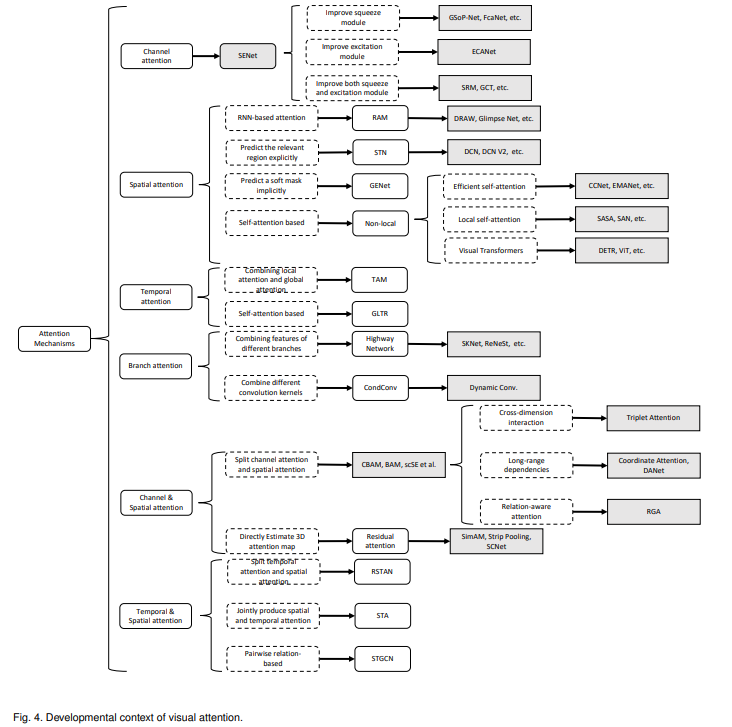
注意力机制由数据领域可以被分类成4种基本类型和两个复合类型，通道、空间、时间和分支注意力

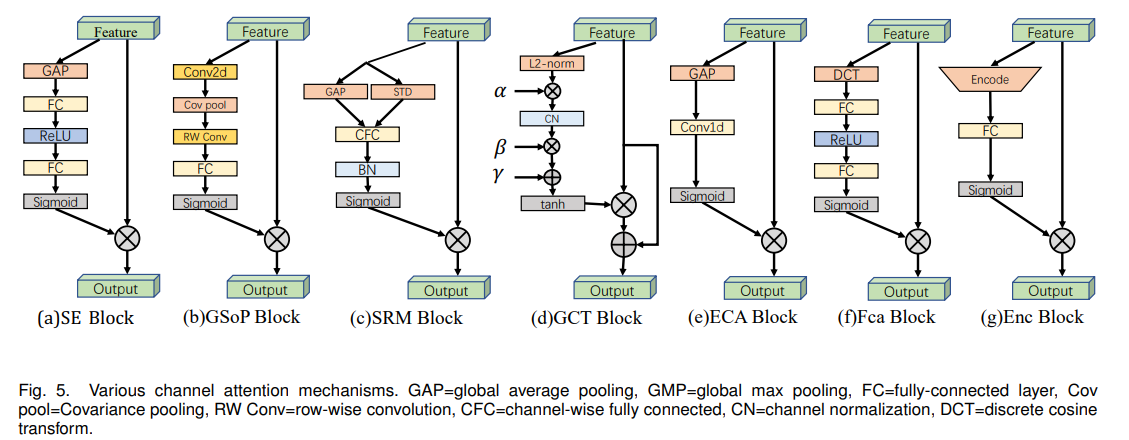
和空间&时间注意力，通道&空间注意力

通道，时间，空间注意力机制可以认为在不同的领域内工作，C代表通道领域，H、W代表空间领域，T代表时间领域，分支注意力和他们是互补的



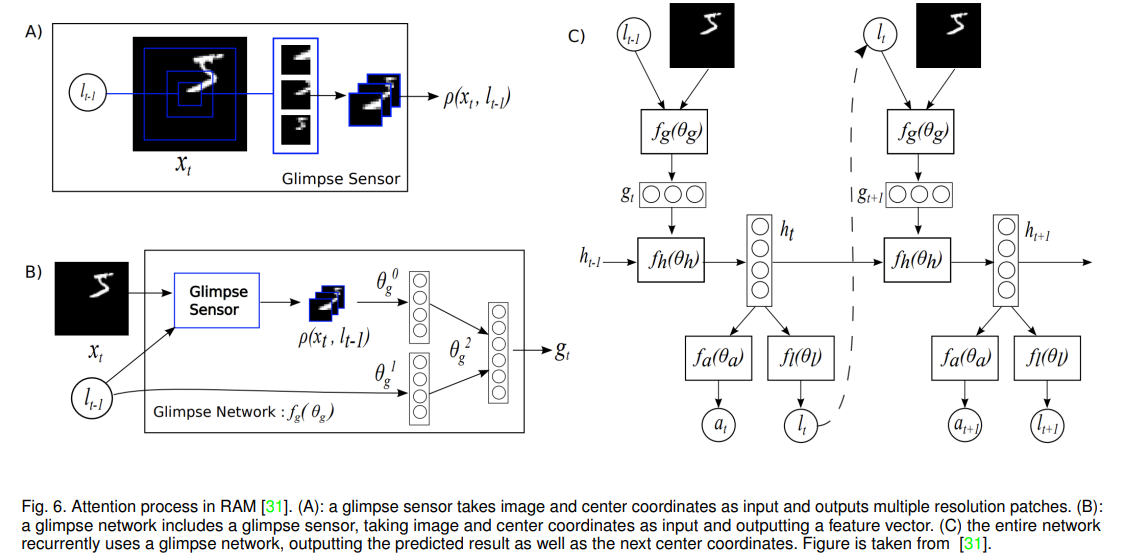
简要总结了注意力机制在CV中的关键发展，有四个阶段，第一个阶段是用RNN去构建注意力，第二阶段明确预测了重要区域，STN是一种代表性方法，阶段3是隐含地完成注意力过程，第四阶段使用了自注意力方法。

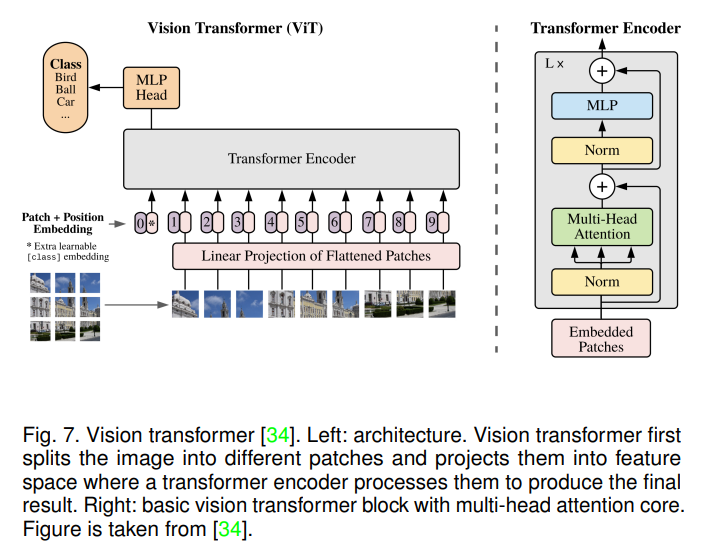
这张图主要讲述了视觉注意力的发展和分类内包含的网络



Then we discuss various channel attention methods along with their development process respectively

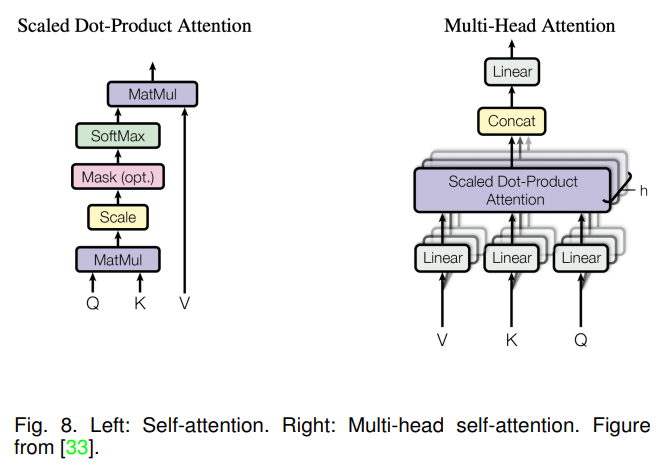
不同的通道注意力

RAM中的注意力过程，glimpse传感器将拍照和中心坐标作为输入然后输出多种分辨率的patches。glimpse网络循环输入照片和中心坐标和输出的特征向量，再次输出

（左）ViT结构

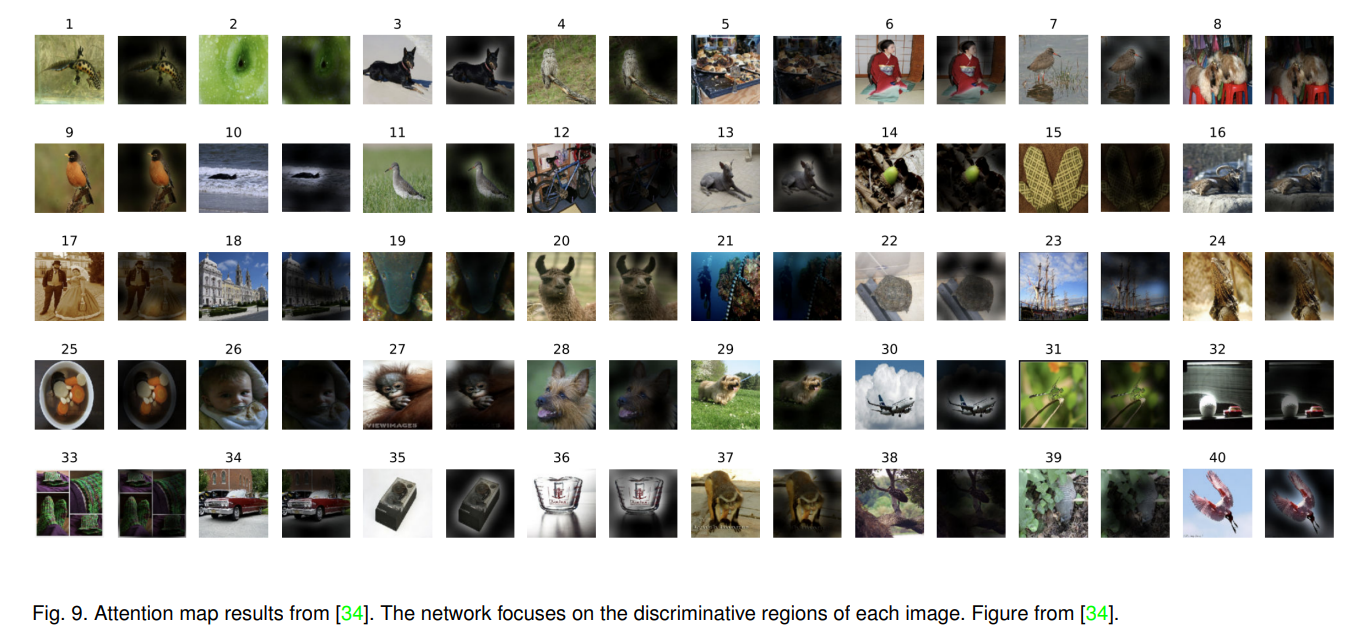
（右）基本VIT块，含多头注意力块

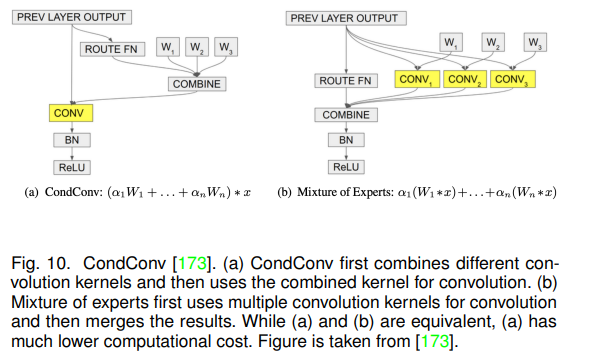
vision transformer首先先把图片分成不同的图像块，然后把他们投影到特征空间中，transformer编码器处理他们生成最终结果



（左）自注意力

（右）多头自注意力

transformer的注意力图，网络集中于有判别力的每个图像的区域



CondConv具有结合不同卷积核的能力，通过一系列可学习的权重，最后对特征图进行卷积

a和b等价，但是a计算开销小得多

**Other Surveys:**

unlike previous surveys, we provide a classification which groups various attention methods according to their data domain, rather than according to their field of application.

不同于其他论文，本文提供了一种分类方法，基于数据域而不是应用域

**Attention methods in computer vision:**

1.General form

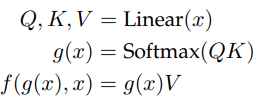


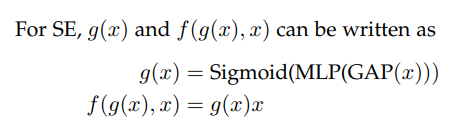
Here g(x) can represent to generate attention which corresponds to the process of attending to the discriminative regions. f(g(x), x) means processing input x based on the attention g(x) which is consistent with processing critical regions and getting information.

上式输入为x，g（x）是处理关键区域和获取信息的注意力，f（g（x），x）是处理的过程

Here we take self-attention and squeeze-and-excitation(SE) attention as examples.

self-attention:





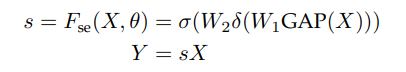
2.Channel Attention:

In deep neural networks, different channels in different feature maps usually represent different objects. Channel attention adaptively recalibrates the weight of each channel, and can be viewed as an object selection process.

在深度神经网络中，不同特征图中的不同通道通常表示不同的物体。通道注意力可以自适应地重新校准每个通道的权重，可以看作是一个物体选择过程。

1)SENet：

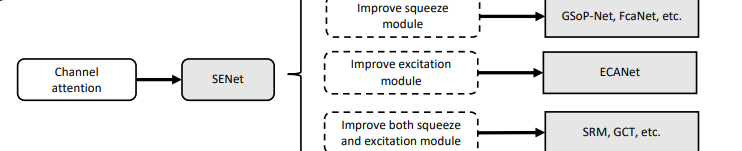
SE blocks are divided into two parts, a squeeze module and an excitation module

GAP global average polling

激励模块通过全连接层和非线性层捕捉通道关系并输出为注意力向量，然后再乘以输入特征来缩放它

SE块强调重要通道、抑制噪声

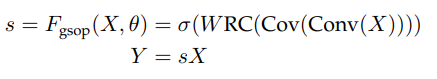
然而，SE模块的squeeze块通过全局平均池化层不能很好的抓住复杂全局信息，excitation模块的全连接层提高了模型复杂度和计算难度。后面的网络改进了这两点，如



2)GSoP-Net

To improve the squeeze module by using a global second-order pooling (GSoP) block to model high-order statistics while gathering global information.

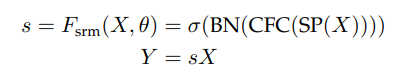
GSoP block can be formulated as:



Conv(·) reduces the number of channels, Cov(·) computes the covariance matrix and RC(·) means row-wise convolution.

3)SRM

Lee et al. proposed the lightweight style-based recalibration module (SRM).Style pooling (SP(·)) which combines global average pooling and global standard deviation pooling

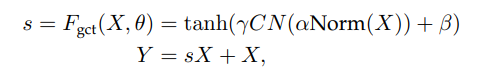
CFC轻量级全连接层

SP是结合了全局平均池化和全局标准偏差池化

It also adopts a lightweight channel-wise fully-connected (CFC) layer, in place of the original fully-connected layer, to reduce the computational requirements.

4)GCT

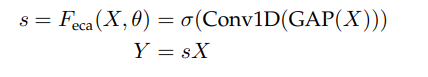
Yang et al. propose the gated channel transformation (GCT) to efficiently collect information while explicitly modeling channel-wise relationships



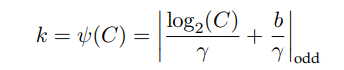
X进入先经过计算l2-Norm of each channel，然后经过可学习的α和β参数，再进行通道归一化操作（CN），γ为可学习的参数，最后经过非线性层tanh。输出Y增加了残差

由于它的轻量级参数，可以被加入到每个CNN卷积层后面

5)ECANet



它通过1\*1卷积来控制模型复杂度而不是直接降维，它只考虑每个通道附近的通道之间的相互关系

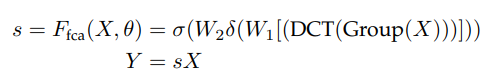
where γ and b are hyperparameters. |x|odd indicates the

nearest odd function of x.

k决定了卷积层的卷积核大小

它提升了excitation模块，并可用于各种CNNs中

6)FcaNet

DCT-discrete cosine transform离散余弦变换

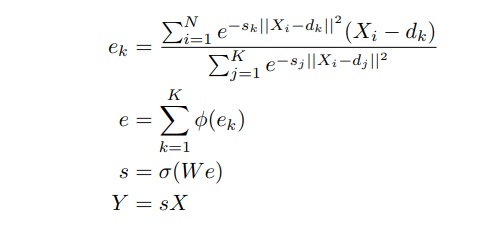
全局平均池化只是离散余弦变换的一个特例

他们将X不同通道分离，高宽不变，然后通过2D DCT（可以使用预训练结果降低复杂度），然后通过ReLU和Sigmoid得到s

这个网络在分类任务中取得极佳成果

7)EncNet

context encoding module (CEM)：

dk和sk是可学习的参数

φ是BN和ReLU

CEM不但增强了类相关的特征图，还通过SE-loss使网络平等的对待大的和小的物体，由于其轻量化的结构，CEM可以仅以较低开销成为各种类型的主干

8)Bilinear Attention

bi-attention block：



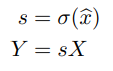
φ表示用来降维的嵌入函数，

UTri（）表示提取矩阵的上三角函数

Vec（）表示向量化计算

ω and ϕ是内置函数

输出的x用于计算外部逐点注意力机制的空间通道注意力权重



双注意力块使用双线性池化来模拟每个通道上的局部特征对之间的交互，同时保留空间信息。与基于注意力的其他模型相比，使用所提出的AIA（Attention-in-Attention）可以更多地关注高阶统计信息。双注意力可以融入任何CNN主干中，以提高其表示能力并抑制噪声。

3.Spatial Attention

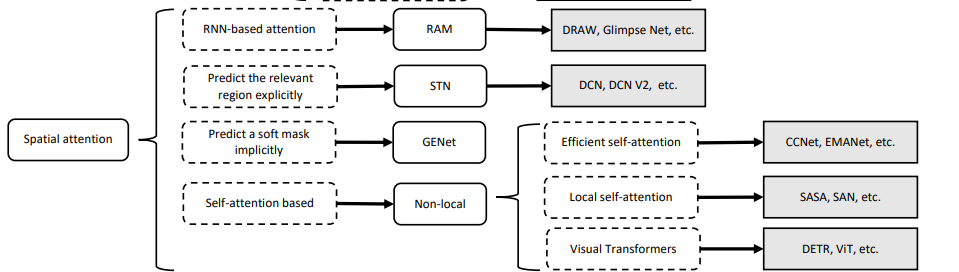


Figure4：为空间注意力的分类

1)RAM

Mnih et al. proposed the recurrent attention model (RAM) that adopts RNNs and reinforcement learning (RL) to make the network learn where to pay attention.

RAM的结构在Figure6下有说明

the RAM has three key elements: (A) a glimpse sensor, (B) a glimpse network and (C) an RNN model.

RAM提供了一种简单但有效的方法来将网络集中在关键区域上，从而减少了网络执行的计算次数，特别是对于大型输入，同时提高了图像分类结果。

2) Glimpse Network

The proposed deep recurrent visual attention model consists of a context network, glimpse network, recurrent network, emission network, and classification network.

首先，上下文网络将下采样图片作为输入，输出循环网络的初始状态和第一瞥的位置。

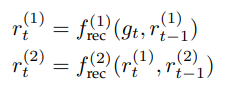
然后，时间步长为t，根据一瞥xt和其位置lt，glimpse网络提取有用信息，表示为：



fimage(X) and floc(lt) are non-linear functions which both output vectors having the same dimension, and

used for fusing information from two branches（融合两个分支的信息）

然后循环网络根据已有信息，输出如下

rt为当前循环网络的隐藏状态

emission 网络预测了下一个glimpse，表示为：



最后，分类网络基于循环网络的状态输出预测

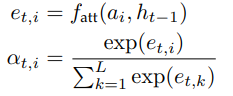


与在整个图像上运行的CNN相比，它所提出的模型的计算成本要低得多，并且因为它在每一步只处理一个glimpse，所以可以自然地处理不同大小的图像。循环注意力机制还提高了鲁棒性，同时缓解了过拟合问题。该传递线可以融入任何最先进的CNN主干或RNN单元中。

3) Hard and soft attention

To visualize where and what an image caption generation model should focus on, Xu et al. introduced an attention-based model as well as two variant attention mechanisms, hard attention and soft attention.

这个模型目的是通过每个时间step产生字幕，所以采用了LSTM网络作为解码器

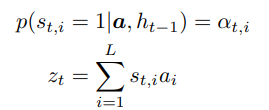
ai为特征向量，ht-1为先前的隐藏状态

fatt为MLP

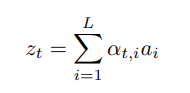
可以被解释成位置i为需要关注的位置的可能性（hard-attention）

或者位置i对下一个单词的相对重要性（soft-attention）

hard-attention上下文向量由此得到：

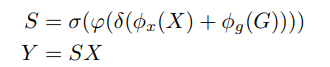


soft-attention的上下文向量由此得到：



4)Attention Gate

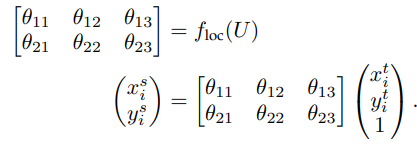
Oktay et al. proposed a simple and yet effective mechanism, the attention gate (AG), to focus on targeted regions while suppressing feature activations in irrelevant regions



为了实现1\*1的卷积

5)STN

Jaderberg等人提出了空间变换网络（STN），该网络使用一个显式的过程来学习对平移、缩放、旋转和其他更一般的扭曲的不变性，使网络关注最相关的区域。STN是第一个明确预测重要区域并提供具有变换等变性的深度神经网络的注意力机制。

θ是可学习的矩阵，floc是任何可微函数，比如轻权重全连接网络

U是输入的特征图

6）Deformable Convolutional Networks

与STN类似，Dai等人提出了可变形卷积网络（deformable ConvNets），使其对几何变换具有不变性，但它们以不同的方式关注重要区域。具体来说，可变形ConvNets不学习仿射变换。它们将卷积分为两个步骤：首先从输入特征图上在一个规则网格R上采样特征，然后通过加权求和使用卷积核聚集采样的特征。

可变形ConvNets自适应地选择重要区域并扩大卷积神经网络的接受域，这在目标检测和语义分割任务中非常重要

7）Self-attention and variants

自注意力机制在NLP领域大放异彩，最近也成为了CV领域的工具，但是自注意力方法对图像有二次复杂度，所以大多数变种都是为了减少其计算复杂度，比如

CCNet：把密连通图变成稀疏图，减少计算复杂度



EMANet：采用期望最大化（EM）算法，

ANN：采用spatial pyramid pooling，减少了计算复杂度

GCNet，GloRe，OCRNet，A2Net，HamNet，EANet

这些网络要么改进了一些性能，要么降低计算复杂度

下面这些网络使用自注意力机制代替CNNs

SASA，LR-Net，SAN

8) Vision Transformers

Dosovitskiy et al proposed the vision transformer (ViT) which is the first pure transformer architecture for image processing.

Figure7 、8说明了它的结构（HMA）

9)GENet

Inspired by SENet, Hu et al. designed GENet to capture long-range spatial contextual information by providing a recalibration function（重新校准功能） in the spatial domain.

The gather-excite module is lightweight and can be inserted into each residual unit like an SE block. It emphasizes important features while suppressing noise

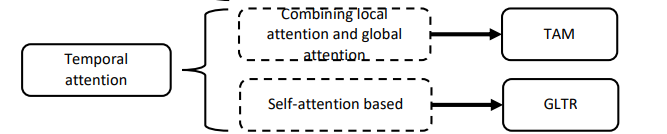
10)PASNet

Zhao et al. presented the novel PSANet framework to aggregate global information. It models information aggregation as an information flow（信息聚合为信息流） and proposes a bidirectional information propagation mechanism（双向信息传播机制） to make information flow globally.

It can be added to the end of a convolutional neural network as an effective complement to greatly improve semantic segmentation.

4.Temporal Attention

Temporal attention can be seen as a dynamic time selection mechanism determining when to pay attention, and is thus usually used for video processing.



1) Self-attention and variants

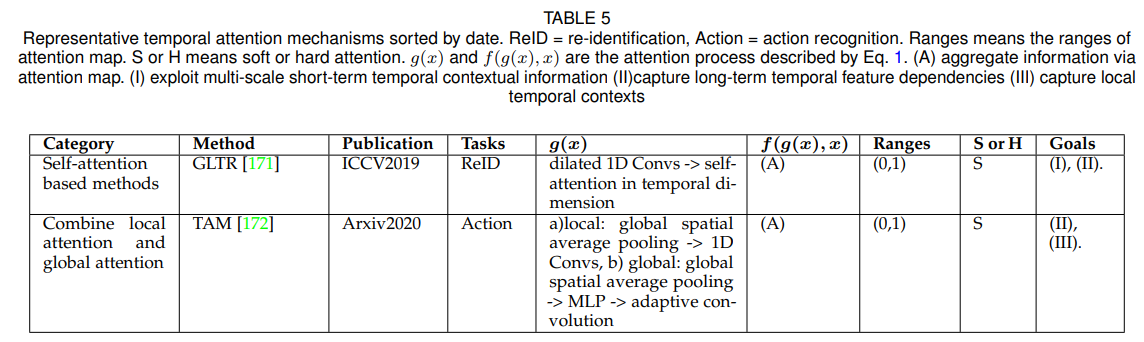
Li et al. proposed a global-local temporal representation (GLTR) to exploit multi-scale temporal cues（利用多尺度时间线索） in a video sequence.

GLTR由用于本地时间上下文学习的膨胀时间金字塔（DTP）和用于捕获全局时间交互的时间自注意力模块组成。

它可以被纳入任意CNN主干网，然而其二次时间复杂度限制了其应用

2)TAM

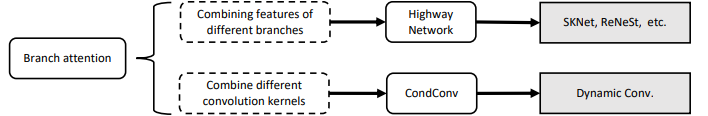
为了高效且灵活地捕捉复杂的时序关系，Liu等人提出了一个时序自适应模块(TAM)。该模块采用自适应核来捕获全局上下文信息，相比于GLTR中的自注意力机制，具有更低的时间复杂度。

借助本地分支和全局分支，TAM可以捕获视频中的复杂时序结构，并以较低的计算成本增强每个帧的特征。由于其灵活性和轻量级设计，TAM可以添加到任何现有的2D CNNs中。

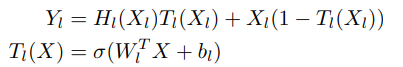
通过数据区分的代表性的时间注意力机制

5. Branch Attention

Branch attention can be seen as a dynamic branch selection mechanism: which to pay attention to, used with a multi-branch structure.



1) Highway networks

Inspired by the long short term memory(LSTM) network, Srivastava et al. proposed highway networks that employ adaptive gating mechanisms to enable information flows across layers to address the problem of training very deep networks.信息跨层流动，Xl、Yl为输出

让它能够使用简单的梯度下降算法训练非常深的网络

2)SKNet

神经科学领域的研究表明，视觉皮层神经元会根据输入刺激自适应地调整其感受野（receptive field，RF）的大小。这启发Li等人提出了一种名为选择性核（selective kernel，SK）卷积的自动选择操作。SK卷积通过三个操作实现：拆分、融合和选择。

在拆分操作中，对特征图应用具有不同核大小的转换以获得不同大小的RF。然后通过逐元素求和将所有分支的信息融合在一起以计算门控向量。该门控向量用于控制来自多个分支的信息流。最后，在门控向量的引导下，通过聚合所有分支的特征图获得输出特征图。

ResNeSt also adopts this attention mechanism to improve the CNN backbone in a more general way, giving excellent results on ResNet and ResNeXt.

3) CondConv

In order to more efficiently increase the capacity of convolutional neural networks, Yang et al. proposed a

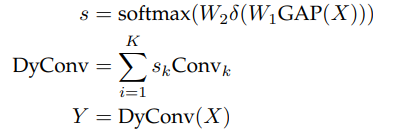
novel multi-branch operator called CondConv



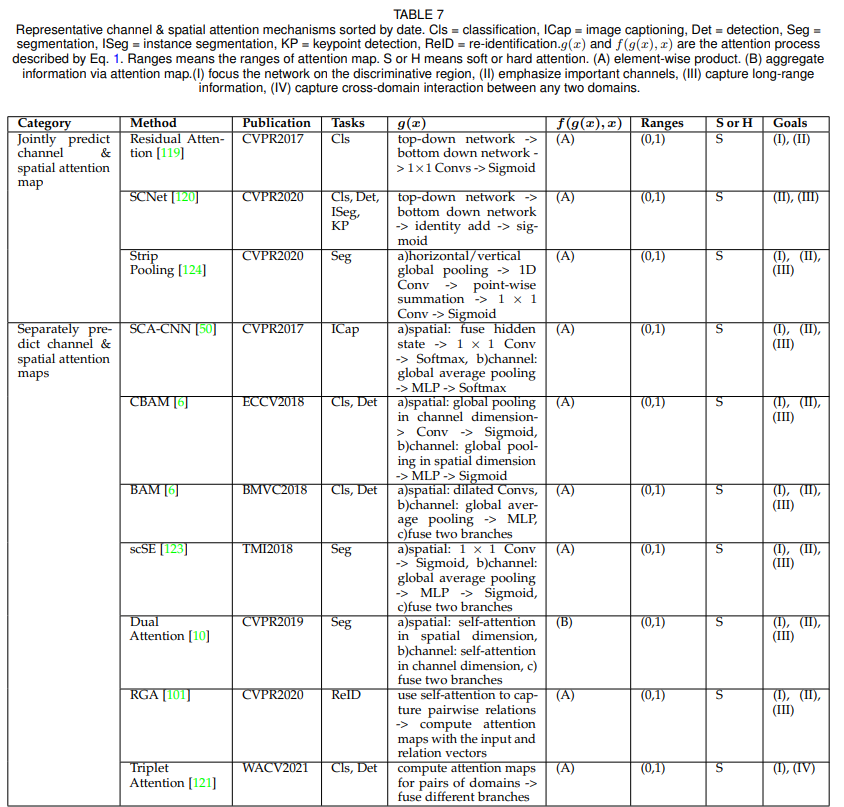
结构图在Figure10

4)Dynamic Convolution

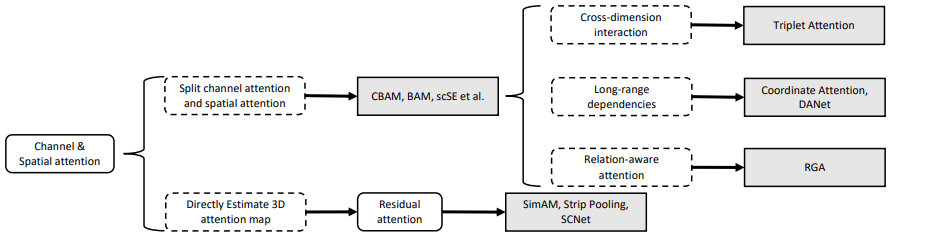
轻量级CNN极低的计算成本限制了网络的深度和宽度，进而降低了其表示能力。为了解决上述问题，Chen等人提出了动态卷积这一新颖操作符设计，该设计在增加表示能力的同时，几乎不增加额外的计算成本，并且不会改变网络的宽度或深度，这一点与CondConv并行。



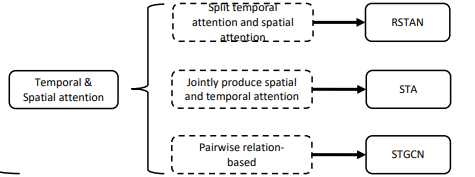
6. Channel & Spatial Attention

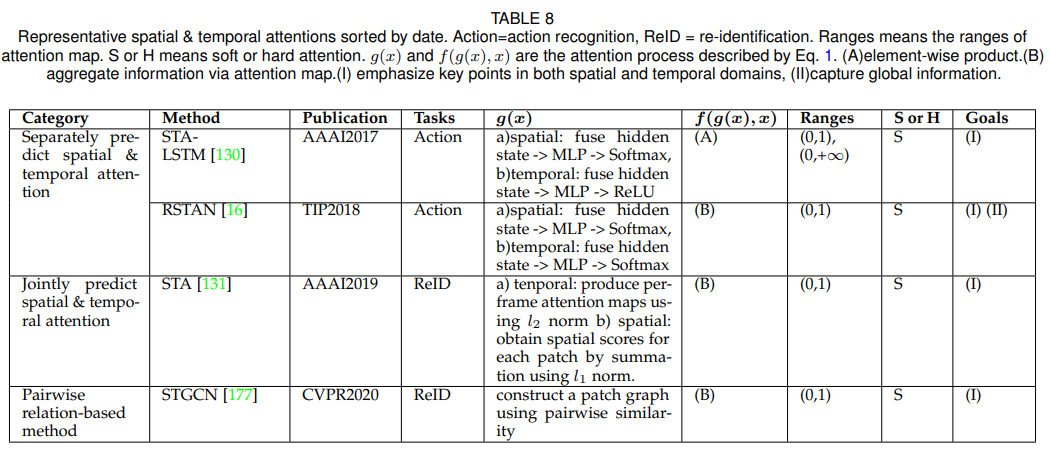


通道&空间注意力结合了通道注意力和空间注意力的优点，自适应地选择重要的物体和区域。残差注意力网络在通道&空间注意力领域中具有开创性，强调空间和通道维度上信息特征的重要性。它采用由多个卷积组成的自下而上的结构来产生一个3D（高度、宽度和通道）注意力图。然而，它的计算成本较高且接收场有限。

为了利用全局空间信息，后来的工作通过引入全局平均池化以及解耦通道注意力与空间通道注意力来提高特征的辨别能力，以提升计算效率。其他工作将自注意力机制应用于通道&空间注意力以探索成对交互。还有进一步的工作采用空间-通道注意力机制来扩大接收场。

7) Spatial & Temporal Attention



空间与时间注意力结合了空间注意力和时间注意力的优点，自适应地选择重要的区域和关键帧。一些作品RSTAN , STA-LSTM分别计算时间注意力和空间注意力，而其他作品如spatiotemporal attention (STA)产生联合时空注意力图。还有专注于捕捉成对关系的进一步作品spatiotemporal graph convolutional network (STGCN)。

**FUTURE DIRECTIONS**

1. 注意力机制的充分必要条件

很难找到所有注意力机制的充分必要条件，对于注意力机制的充分必要条件仍然值得探索

2. 通用的注意力块

是否存在一个通用的注意力块，可以利用各种类型的注意力机制。例如，一个软选择机制（分支注意力）可以根据所执行的具体任务在通道注意力、空间注意力和时间注意力之间进行选择。

3. 特性和可解释性

通常基于注意力的模型只能给出直观的感觉，而不是精确的理解。但在如医疗诊断和自动驾驶系统，有更严格的要求，需要更好地描绘方法的运作方式，包括失败模式。开发可描绘和可解释的注意力模型可以使它们更广泛地适用。

4. 稀疏激活

注意力机制可以产生稀疏激活。一个启示为：即稀疏激活可以在深度神经网络中实现出色的性能

而这激励我们探索哪种架构可以模拟人类视觉系统。

5．基于注意力的预训练模型

在NLP取得了巨大成功，在CV领域，预训练和注意力模型相结合应该进一步探索：训练方法、模型结构、预训练任务和数据规模都值得研究。

6.优化

基于注意力的网络和卷积神经网络是不同的模型；不同的优化方法可能对不同的模型效果更好。为注意力模型探索新的优化方法可能是值得的。

7.部署

卷积神经网络具有简单、统一的结构，这使得它们易于部署到各种硬件设备上。

值得尝试寻找可以广泛部署的简单、高效和有效的基于注意力的模型。