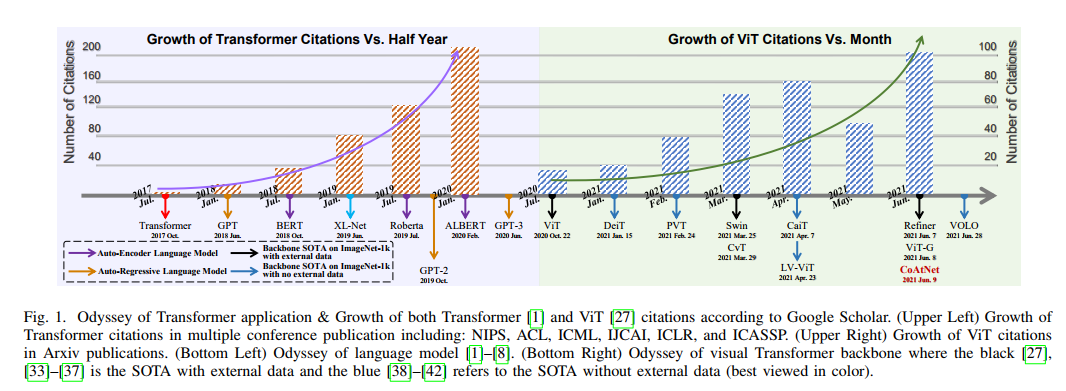
A survey of Transformer

Abstract

# 1 Introduction

Transformer [attention is all you need]是一个纯基于注意力机制的神经网络架构。Transformer 最早被提出于解决Sequence to sequence model [Sequence to Sequence Learning with Neural Networks]中的机器翻译任务，它强大的全局元素间的建模能力使它在“Machine Translation on WMT2014 English-German”的比赛中取得单模型State-of-the-art[attention is all you need]。如图1，Transformer逐渐成为NLP和CV领域主流的深度学习网络架构。之后的工作展示了Transformer先在足够的数据集上（如BERT的预训练数据集BooksCorpus (800M words) (Zhu et al., 2015) 和English Wikipedia (2,500M words).和GPT3的预训练数据集Common Crawl数据集）进行预训练，之后针对各种小数据集或各种特定的下游任务进行微调( fine-tune)，在Machine Translation， Sentiment Analysis， Question Answering等领域的不同数据集均能取得SOTA结果。

在计算机视觉领域，自Hinton et al.提出AlexNet [] 后，Yann LeCun，Wei Zhang，Alexander Waibel 等人在20世纪80年代末提出的卷积神经网络（CNN）直接成为计算机视觉领域最主流的模型。由于卷积神经网络具有稀疏连接、参数共享、平移不变等特性，使得卷积神经网络成为研究计算机视觉的主要范例。2017年Vaswani et al.提出基于自注意力机制的Transformer后，Transformer迅速在NLP领域取得优异的性能。Transformer在NLP领域的成功使得许多研究者开始试着将Transformer的架构迁移到计算机视觉领域的下游任务上。由于计算机视觉中像素间的信息密度远大于NLP任务中文本的信息密度，所以针对文本的Transformer模型架构并不直接适用于计算机视觉的任务。直到2020年，经过研究者们的设计呀牛，Transformer架构能够成功的迁移到计算机视觉领域的各项基本任务上，如图像分类、目标检测、图像生成、全景分割、视频处理。



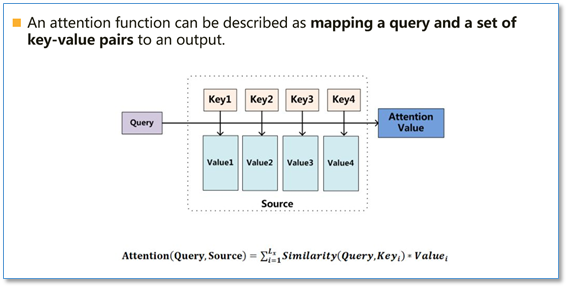
# 2 Transformer Architecture

Transform []是一种完全基于自注意力机制的模型架构，在最开始的时候Transformer是一个为了处理Sequence-to-Sequence Model中的机器翻译任务的新架构。Transformer兼具元素间长程依赖计算容易和并行处理的特性，相比于以往的模型在处理Seq2Seq任务中， CNN可以通过窗口滑动实现并行处理，由于CNN的局部性导致CNN难以计算远距离的元素之间的长程依赖，只能通过加深网络层数才能建模位置较远的元素；RNN在预测下一个元素时能包含之前输入元素的信息，即预测的元素根据输入的所有元素的信息综合产生，一定程度上可解决长程依赖问题，但每次只能产生一个输出无法并行输出。受限于RNN和CNN的架构问题， Vaswani et al.提出了Transformer架构。本模块使用七个小节来介绍Transformer架构：

1. **注意力机制**

2014-2018年，DeepMind团队Mnih, V[] et al.首先提出在RNN模型上使用Attention机制来进行图像分类。随后Bahdanau, D. et al.提出Attention Mechanism []，Minh-Thang Luong, Hieu Pham, Christopher D. Manning在[]中提出了两种 attention 的改进版本，即 global attention 和 local attention。Global Attention相较于[]能够更简单直接的在每一次生成目标词时，计算所有源语句隐藏状态的相似度。Local Attention只需要在计算时对源语句的某个子集计算相似度，之后基于子集生成文章向量(context vector)。Local attention可以视为Hard Attention和Soft Attention的混合体，因为他的计算复杂度要低于Global attention 和Soft attention，且local attention几乎处处可以微分，易于进行训练。等人提出的[]使用了大量的自注意力机制，凭借强大的建模能力，自注意力机制也成为了研究者们的研究热点，并使用自注意力机制在NLP任务上进行探索。

Attention函数的本质可以被描述成为一个查询(Query)到一系列(键Key-值Value)对的映射，如图：



计算Attention Score可以分为三步：

Step1：

Step2：

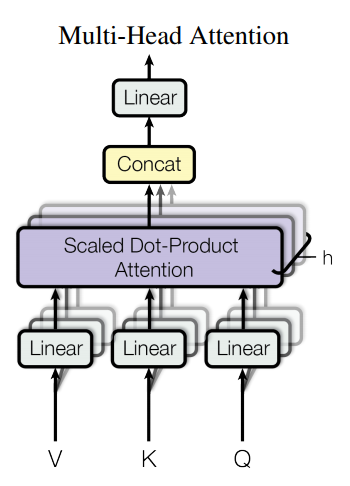
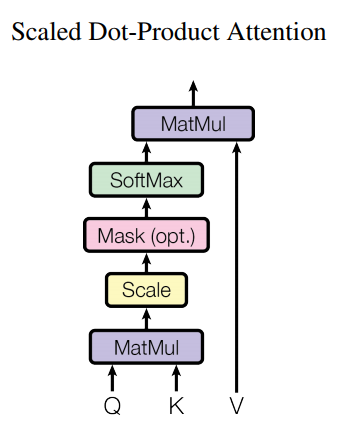
Step3：

其中Q是Query，K是Key，V是Value，通过函数计算注意分数，再经过softmax函数对注意力分数矩阵进行归一化处理，最后将加权到Value上。

1. **自注意力机制**

等人在Transformer中大量使用自注意力模块计算不同元素之间的相似度。作为Transformer中最重要的模块，该模块被分为部分，1)一个线性投影层将输入序列投影到三个不同的特征向量中(),其中n为序列长度，d是一个词向量的编码维度。其中向量的产生由下式进行计算：

其中,是线性投影矩阵， 是query和key向量的维度， 是value向量的维度。Query来自X的投影，键值对（Key-Value）来自Y的投影。这两个序列的输入方案被称为交叉注意力机制[]。一个注意力层如图：



左边单头自注意力机制；右边多头自注意力 From[Attention is all you need]

Transformer使用QKV(Query-Key-Value)来表示注意力机制，通过矩阵，Transformer自注意力的核心缩放点积注意力如下式：

其中N和M分别是Query和Key(或Value)的长度，分别是Query和Key(或Value)的维度，又被称作注意力权重矩阵，通过Softmax函数将注意力权重矩阵转化为标准分布，再分配给指定的Key元素，从而生成最终的输出向量，通过计算Q和K的点积后并除以来平滑梯度，防止梯度消失。

1. **多头自注意力机制**

因为单个注意力机制在受限的特征子空间进行建模，所以他建立的模型相对比较粗糙。为了使模型更加强大，V等人在[]中提出了多头自注意力(multi-head self-attention mechanism)(MHSA)，以提升自注意力模块的表现能力。单头注意力机制限制了模型只能集中注意力在几个元素上以致于忽略其他重要的元素，多头注意力使得模型具有更强大的表达能力，能从不同的角度(perspective)进行建模。通过不同的的Head关注不同的特征子空间，因为使用不同的Query，Key，Value矩阵，且矩阵随机初始化，这些矩阵可以将输入投影到不同的子空间当中。

给定输入向量和一定的头数h，输入向量首先被线性投影到三个矩阵中：Query，Key，Value。在每一个矩阵中使用头数h(例，当h=8时)划分向量，.形成矩阵组Q，K，V。

，.

多头自注意力的处理如下式：

其中，,,

1. **基于位置的前馈式网络**

除了Attention子层，Transformer中Encoder和Decoder每一层都含有一层基于位置的前馈式网络，其中包含一个GELU函数和一个全连接层，如下式：

MHSA的输出直接作为输入进入基于位置的前馈式网络，FFN在每一层Encoder或Decoder中的位置相同但是参数不同，所以能很好的线性拟合数据。

1. **位置编码**

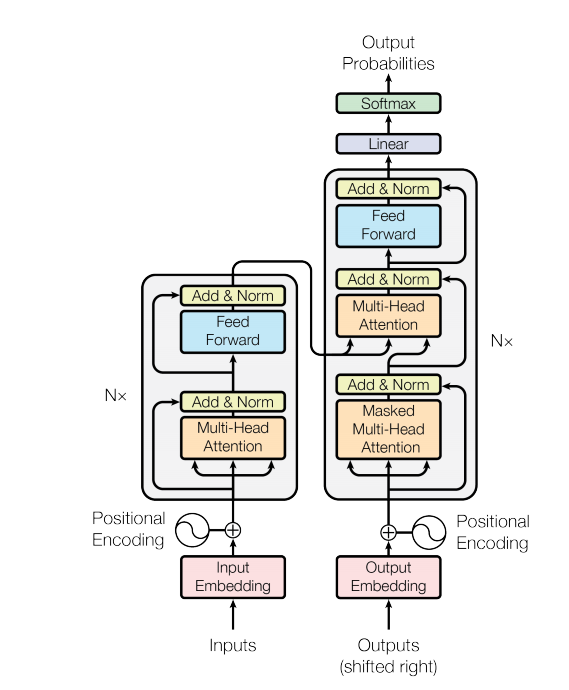
自注意力模块对输入的处理是不包含位置关系的，换而言之一次输入对于自注意力模块同等重要。为了使自注意力模块获得捕捉到词位置信息的能力，Transformer在输入加入了位置信息(Positional Encoding)：

其中为元素的位置，是维度，位置编码的每一个维度均有一个正弦波与之对应。

之后的研究有专注于改进位置编码的模型，1)绝对位置表示，由等人提出的BERT，由等人提出的FLOATER，他们均为每一个位置学习一组Position Encoding来表示绝对位置；2）相对位置表示，由等人提出的Music Transformer，由等人提出的T5，由等人提出的DeBERTa等，通过编码元素之间的相对距离学习元素之间成对的关系。还有一种位置编码的方式为可学习式的嵌入，这类方式得到的位置编码更加灵活，如[]，但这类方式存在问题，若后续输入序列大于训练时最长的序列则无法使用学习的位置编码进行嵌入。

1. **模型概览**

Fig是Transformer的整体架构图。 Transformer整体的架构由N× Encoder连接 N× Decoder组成。其中Encoder除第一层需要对输入加上位置编码外，所有Encoder都由一个自注意力模块和一个基于位置的前馈式网络组成，且两个模块的输出都会经过残差连接和层归一化进行处理；Decoder相比于Encoder多了一个Mask Multi-Head Attention，且Decoder中第二个自注意力模块的键值对(Key-Value pair)来自Encoder，其余结构相同。



1. **模型分析**

|  |  |  |  |
| --- | --- | --- | --- |
| Layer Type | Complexity per Layer | Sequential Operations | Maximum Path Length |
| Self-Attention |  |  |  |
| Recurrent |  |  |  |
| Convolutional |  |  |  |
| Self-attention(restricted) |  |  |  |

**Table 1：表中n是序列的长度，d是维度，k是卷积核大小，r限制自注意力的范围大小。表中比较了不同类型网络的每一层的复杂度、最小处理序列次数(并行处理)、最大处理长度。表来自[]**

从表1中可以得出，自注意力模块并行化的进行所有位置的连接，RNN则需要的顺序处理。当序列长度n小于维度d时，自注意力模块的计算比RNN的计算更快。为提升处理超长序列的能力，va等人设计了restricted Self-attention将注意力计算限制在一定的范围中，提高Transformer对长序列的建模能力。

# 3 Transformer’s Architecture in CV

## 3.1 Transformer Backbone for Computer vision

SENet

NLNet

GCNet

ViT/DeiT/Swin/iGPT/MoCoV3

BoTNet/CeiT/DeiT

AlexNet/ResNet/DenseNet

Backbone for Computer Vision

Convolution

Attention

CNN

CNN +Transformer

Transformer

图表示计算机视觉中的两大模型分类：卷积神经网络和注意力机制。图来自[]

在Transformer被用于机器翻译任务之前，计算机视觉领域使用主干的模型为卷积神经网络，如[AlexNet,ResNet,DenseNet]。Kaim等人提出的ResNet[]对计算机视觉领域产生了深远的影响。当研究者们发现Transformer在NLP领域大放异彩的时候，研究者们思考能否对于计算机视觉领域能否直接使用Transformer模型进行处理。由于文本和图片的信息密度不同，直接将Transformer模型用于计算机视觉领域的想法没有成功。到2020年，等人提出的[]重新掀起了Transformer在计算机视觉领域的应用。由于ViT根据图像进行独特的设计，规定一定大小的矩形块(如ViT里的Token大小14\*14)里的像素的集合称为Token，将所有Token送入一个线性投影层，之后将线性拟合后的Token输入Encoder中进行建模。受到[]的启发，在将Tokens输入到Encoder之前创建了一个类似BERT的[class] Token，用于储存所有元素之间交互的信息。经过Encoder后取[class] Token通过MLP进行预测。ViT的优异的表现迅速吸引了研究人员的目光，之后针对ViT的改进层出不穷[]。

为了建立文本与图像之间的模型，Liu et al.提出了滑动窗口的分层Vison Transformer[]。该模型同ViT一样是Transformer在计算机视觉领域里的主干网络。

## 3.2 CNN增强Transformer

由于Transformer模型是对所有元素之间进行建模，当输入的数据巨大(如文本冗长，图片尺寸大)的情况，模型计算需要很多的计算资源，并且某些元素之间的建模不是必要的。所以在Transformer提出后的NLP领域，研究者们针对注意力计算提出了稀疏注意力[]，线性化注意力[]，低秩自注意力[]等不同方向的改进。当Transformer引入计算机视觉后，针对大量的像素，直接进行像素级的自注意力计算成本巨大，所以研究者们借助CNN对图像进行特征提取之后输入Transformer进一步进行元素间的建模，这一思路减少了大量的冗余无效的计算，增加模型对图像建模的效率。这些工作总的可以概括为：软近似[]、直接局部处理[]、针对位置编码的表示[]、组合结构[]。

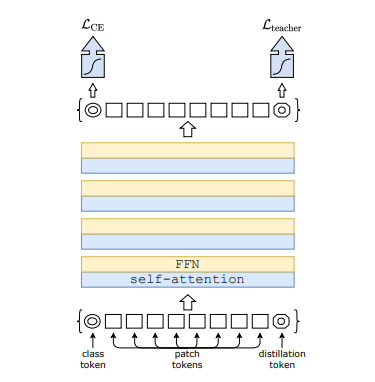
相关的工作：

**ConViT:**由于卷积架构的硬性归纳偏置能够有效的对图像进行学习，但是他的性能几乎已经达到上限。Vision Transformer依赖于大量数据的训练，在计算机视觉领域某些下游任务的表现超过了以CNN为基准的模型，但前者性能仍有提升的空间。等人[]提出了结合卷积架构和Vision Transformer的优势的模型ConViT。ConViT提出了一种门控位置的自注意力模型(GPSA)，该模型通过“软”卷积的归纳偏置，模仿卷积核来初始化GPSA，使每一个注意力头通过调节位置与信息的注意力门控参数来避开位置性，形成一种类似卷积的ViT架构。在ImageNet数据集中ConViT的表现优于DeiT。

GPSA的表达式如下：

其中类似卷积核一样可学的嵌入，是一个固定的相对位置嵌入，是可学习的门控参数。

**DeiT：**ViT这个纯基于注意力的网络架构被证实能适应计算机视觉的下游任务并取得好的表现。但训练ViT所需要的训练数据和时间都是很大的，从而限制了将模型的应用。等人提出了一个依赖于Token的蒸馏并完全针对Transformer设计的师生策略，确保学生网络根据注意力向教师网络学习。DeiT使用ImageNet进行训练无需引入额外数据能在参数量为(86M)的情况下取得Top-1准确率为83.1%的表现。



**Soft distillation:**

其中为教师模型的输出，为学生模型的输出，τ为蒸馏温度，KL为KL散度，为CE-loss，y为真值，ψ为softmax，λ为权重。

**Hard distillation：,**

其中

和 ViT 的 class token 类似，distillation token 是在图像块序列后面加上的token。然后它经过 Transformer的Encoder 之后的输出用于和 teacher model 的输出计算损失。作者发现class token和distillation token会收敛于不同的向量，cos距离为0.06，随着层数的加深，两者embedding的相似度逐渐上升，最终cos距离为0.93，说明他们希望得到相似但不相同的target。为了证明distillation token的有效是由于knowledge distillation的作用，而不是只由于多了一个token导致的，作者做了对比试验，不是在后面加distillation token，而是再加一个class token，发现他们最终收敛结果的相似性达到0.999，并且性能弱于前者，证明了加入distillation token的意义。

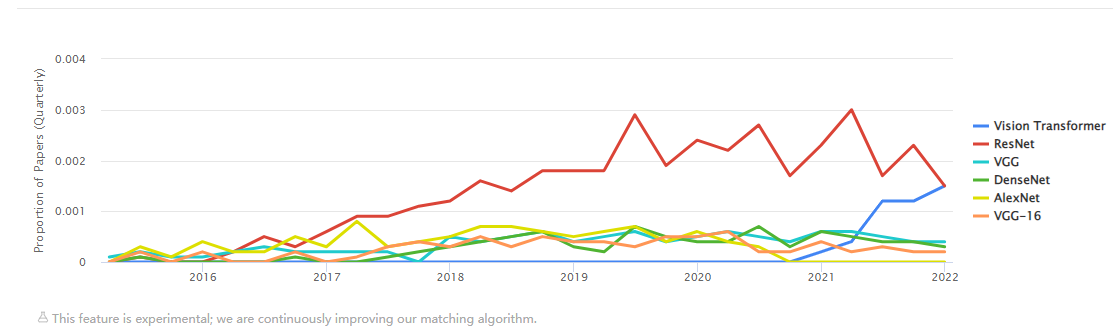
## 3.3 Transformer增强CNN

由于卷积架构对计算机视觉任务的归纳偏置，使得CNN对于处理计算机领域任务具有天然优势。但由于CNN感知图像是通过一定数量的卷积核扫过整张图片，受限于卷积核的大小，CNN的视野具有局限性，只能查看到当前元素一定范围内的元素，所以对于目标检测这一类需要计算长程依赖的任务进行特殊化的设定。由于Transformer可以远距离按建模两个元素之间的关系，使得模型对图像进行处理的时候模型的视野不局限于相邻元素，所以为了解决CNN视野局限的问题，研究人员使用借助Transformer远距离建模两个元素之间的特性来加强CNN[]。

**BoTNet：**BoTNet的工作是将ResNet的最后三个Stage中，使用全局的自注意力来代替空间卷积操作。为了解决计算机视觉任务中图像数据量过大的问题(如处理的一张图片的像素大小为1024\*1024)和自注意力模块的记忆和计算的复杂度随空间维度的增加呈现四次方的增长[]，使训练和推理的开销巨大。BoTNet针对上述问题做出如下的改进： (1) 通过卷积从大型图像中学习抽象和低分辨率的特征图谱。(2) 使用全局自注意力来处理并聚合卷积提取的特征图谱。此工作在不少计算机视觉领域的下游任务取得良好的成绩，比如取代了ResNet模型在COCO实例分割任务的最佳单一的模型。之后通过微调，在ImageNet数据集上达到了84.7%的Top-1准确率，并且在在TPU-v3的硬件训练效率是流行效率计算模型的1.64倍[]。

**VTs:**对于计算机视觉的处理主要是(1)将图像统一转化为排列的像素矩阵，(2)对已局部提取出的高纬度特征进行卷积。可是由于卷积操作不区分各个元素的重要性，且无论语义信息直接对各元素进行建模，导致元素间建立模型的视野不具有全局性。在[]中，等人(1)通过建模Tokens来表示图像中各部分的语义关系，(2)引入Transformer模型来计算不同Token之间的语义关系。结果现实，VTs能根据上下文有意识地关注不同的图像部分。这与直接将像素输入Transformer有了鲜明的对比，因为直接使用像素需要更大数量级的计算资源。通过比较VTs与CNN，在ImageNet 中使用更少的FLOPs和更少的参数将ResNet的准确性提高了4.6到7个百分点；基于VTs的特征金字塔网络能在减少FPN模块中FLOPs 6.5倍的情况下实现了0.35个百分点的mloU。

# 4 Transformer’s Application In Computer Vision



图表 “from paper with code”

Transformer[]在2017年提出，在那之后广泛的用于NLP领域，研究人员们提出许多NLP领域的重要论文[]，推动了NLP领域的发展。研究者们在早期也尝试将Transformer应用到计算机视觉领域[]，但表现都不是太好。从2020年ViT的提出，Transformer在CV领域中展现了革命性的提升，研究者们有可能可以将CV和NLP通过Transformer架构统一起来。这一结果有助于:(1)易于视觉和语言的联合建模[]。(2)两个领域的研究经验可以相互借鉴，从而加快各自领域的发展。

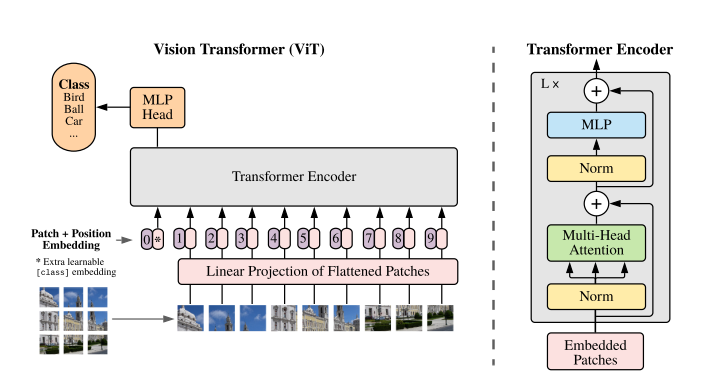
## 4.1 High/mid Level Vision

自从ViT提出后研究者们致力于使用或改进ViT研究Transformer在计算机视觉下游任务的应用。如图像分类[]，目标检测[]，分割任务[]。

### 4.1.1 Transformer for Classification

**ViT：ViT**由Dosovitskiy et al.提出的第一个基于Transformer的计算机视觉领域的主干结构[]。原始的Transformer的输入需要是序列化的Token，但图片是由一个一个的像素组成稍大一点的图片(如1024\*1024)像素有1048576个，这对于Transformer模型的计算要求是很大的。ViT提出将图片分为一定数量固定大小的Patch，通过对Patch建模，进而处理图片。

ViT首先划分一定数量固定大小的Patch将所有Patch输入一个线性投影层，再随机初始化一个同等大小的[class] Token用于储存Token间的信息，将线性拟合后的Token加上1维的位置编码输入Encoder，经过Encoder对所有元素进行建模过后取[class] Token用于预测。ViT的整体架构如下图：

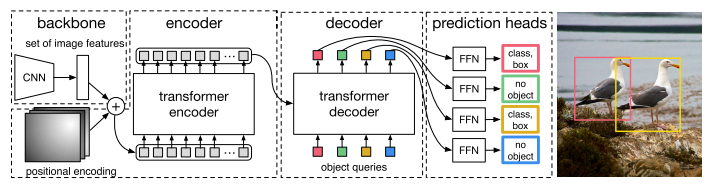


ViT在超大的数据集进行预训练过后针对各个计算机视觉领域的下游任务进行微调，能消耗更少的训练资源取得超过CNN模型性能的SOTA结果。同时ViT也启发了计算机视觉自监督预训练领域。

### 4.1.2 Transformer for Detection

在Transformer被用于目标检测任务之前，主流目标检测模型有[]。但基于Transformer的目标检测模型有不凡的表现[]。

**End-to-End Object Detection(DETR)：**DETR是最先提出的基于Transformer的端到端的直接进行集合预测的目标检测模型。



图表 1 DETR的架构图，来自[]

DETR将CNN和Transformer相结合，取代了以往的模型需要手工设计的工作，并且具有不错的性能。DETR采用的CNN为BackBone提取图片特征，Transformer对提取的特征进行进一步的建模，通过Decoder直接一次性产生图中所有类别的预测。

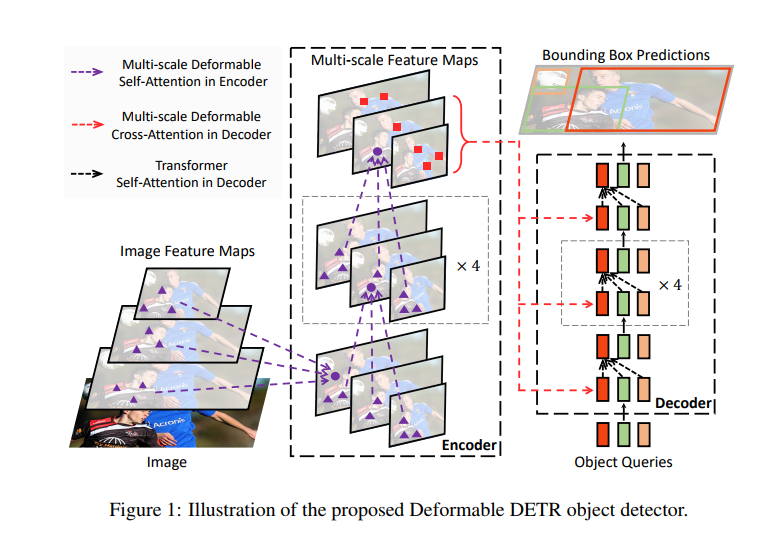
DETR的三大特点：(1)End to End任务模式(原始图片输入无需处理直接做出预测无需人工设计);(2) 设计了双边匹配损失(bipartite matching loss)，基于预测的box和ground truth boxes的二分图匹配计算loss的大小，从而使得预测的box的位置和类别更接近于ground truth。(3)集合预测，使用Transformer的encoder-decoder架构一次性生成 N 个box prediction。其中N是一个事先设定的、远大于image中object个数的一个整数。

DETR通过Decoder之后的分类分支(class)和回归分支(bounding box)进行预测，其中DETR的输出张量的维度为(N,class+1)和(N,4)，其中“class+1”表示已标注的类别数目和一个其他类别，4代表Bounding box的中心点坐标，高和宽。在训练中双边匹配算法表示预测类别与真实类别一一对应的损失值：

类似于学习锚点匹配的标签分类方法[]。使用匈牙利算法计算所有Bounding box和所有类别预测的两两广义距离，距离最近表示越可能是最优匹配关系：

综上所述，DETR提出了一种新的端到端目标检测的方式，在COCO数据集上取得与优化后的Faster R-CNN基线相当的结果，同时DETR这种灵活的结构全景分割的任务中也取得具有竞争力的结果。DETR的提出也带来了新的挑战，尤其是在训练和优化检测小型目标的方面。

**Deformable DETR：**Xizhou Zhu et al在[]的基础上提出了Deformable DETR[]。Deformable DETR为了解决DETR中:(1)训练收敛慢;(2)对小样本的检测效果差的问题。



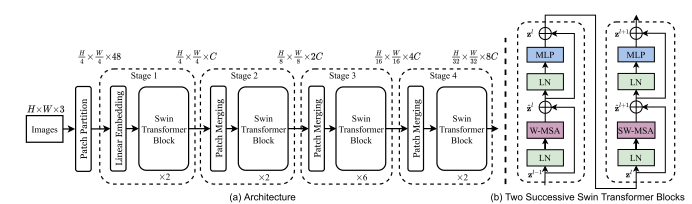
Deformable DETR相比于DETR在训练时少了10X的epochs但仍能取得较高的检测精度。收到特征网络金字塔的启发，将Transformer中的多头注意力改进为多尺度可变形的注意力机制(MSDA)：

MSDA模块将计算复杂度降低到了,将模型的推理速度提高了1.6倍。

### 4.1.3 Transformer for Segmentation

分割任务是计算机视觉领域内重要的任务之一，包含语义分割[]，实例分割[]，全景分割[]。

**Swin Transformer：**Liu et al.提出了一种使用滑动窗口沿着空间维度来建模全局和边界特征的模型[]。

****

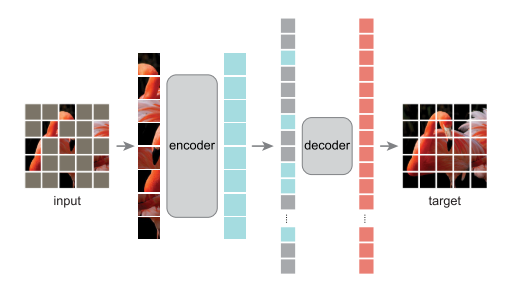
Swin Transformer的主要思想是将具有很强建模能力的Transformer结构和视觉信号的先验联合起来。通过浅层的小Patch到深层逐渐合并周围的Patch，同一Patch内的像素相互进行建模，不同Patch之间不计算相似度。类似于CNN的卷积核滑过图片的形式通过滑动窗口建立相邻Patch之间的关系，随着图像分辨率的提高计算复杂性从减少到了其中n是Patch的长度M是滑动窗口的大小，通过滑动窗口的方式能减少计算资源的消耗。

结果上，在目标检测的重要评测数据集 COCO 上，Swin Transformer 取得了单模型58.7的 box mAP 和51.1的 mask mAP，分别比此前最好的、没有扩充数据的单模型方法高出了+2.7个点和+2.6个点。此后，通过改进检测框架以及更好地利用数据，基于 Swin Transformer 网络的方法性能进一步取得了61.3的 box mAP 和53.0的 mask mAP，累计提升达+5.3 box mAP 和+5.5 mask mAP。在语义分割的重要评测数据集 ADE20K 上，Swin Transformer 也取得了显著的性能提升，达到了53.5 mIoU，比此前最好的方法高出+3.2 mIoU，此后随着分割框架和训练方法的进一步改进，目前已达到57.0 mIoU 的性能。

### 4.1.4 Transformer for Self-supervise

**MoCo V3：**

**MAE：**He et al. 提出了一种用于自监督任务的自编码的可扩展的视觉学习器[]。MAE是一个基于ViT的BERT化的一个模型，它把整个训练拓展到没有标号的数据集上，通过随机掩去图片上的Token再利用未被掩去的Token进行图像还原。MAE加速了Transformer在计算机视觉上的应用可能是未来影响最大的模型。



图表 MAE

MAE的思路很简单，通过Mask遮住一定数量的块之后重构缺失的像素，MAE的编码器只关注可见的块。在MAE中作者遮住了75%的块，若遮的块的数量少用传统的插值法即可重构缺失的像素值，遮住大部分块的话迫使模型能够学会更好地表征。

MAE的模型架构首先输入Patch，再对Patch做掩码操作，将没有掩码操作的Patch输入Encoder，再将掩码的Patch和Encoder的输出按位置进行排列输入进Decoder，通过Decoder重构Masked Patch的像素。MAE通过小数据集ImageNet 1K进行自监督训练，再做迁移学习，它在其他计算机视觉下游任务上表现好。

## 4.2 Low level vision

### 4.2.1 Image Generation

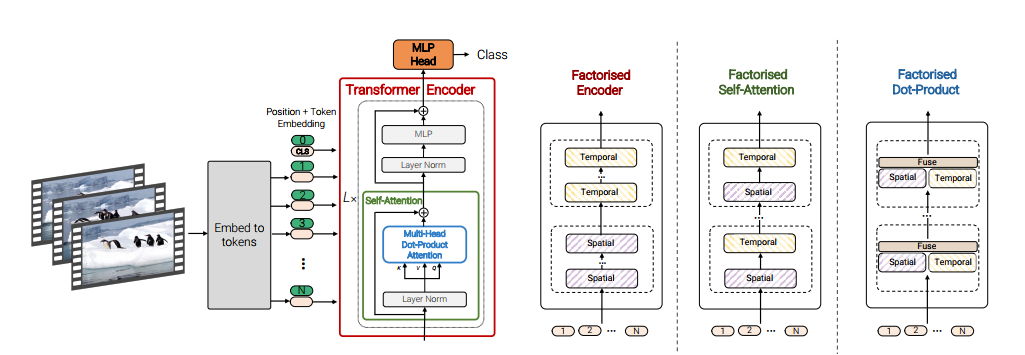
### 4.2.2 Image Enhencenment

## 4.3 Video Processing

**4.3.1 Video Classification**

视频分类任务是产生于视频相关的标签的任务，好的视频分类模型不仅能够精准的提供帧标签，还能在给定视频的各帧的特征和注释的情况下很好的描述视频[]。典型的任务分为(1)为视频分配一个或多个全局标签；(2)为每一帧分配一个或多个标签。

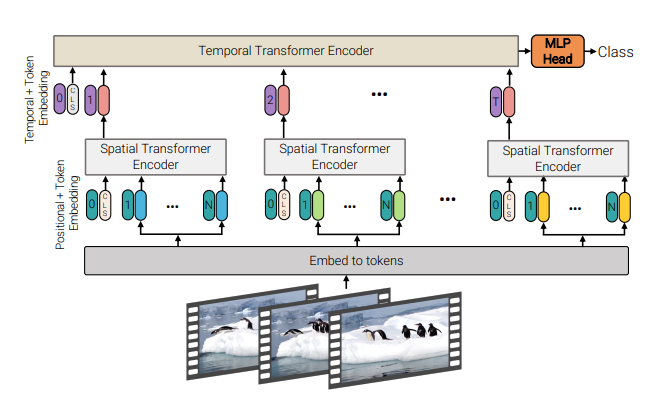
**ViViT：**Arnab et al.提出了一个纯基于Transformer的视频分类模型[]。该模型借助Transformer在图像分类上成功的经验，考虑时间和空间维度的标记，在Transformer中进行编码。虽然Transformer需要大量数据集进行训练，但该工作展示了如何在训练过程中有效得规范化模型，并利用预训练模型的图像在小数据集上进行训练。



图表 ViViT overview

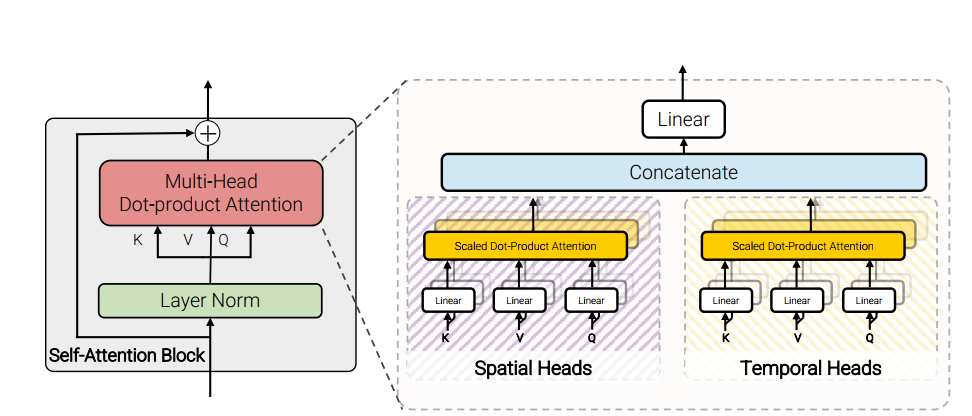
ViViT提出了四个注意力模型：

1. **时间空间注意力(Spatio-temporal attention):**与CNN架构的感受野随层数的加深感受野呈线性增长不同，每层Transformer都对所有的时空标记进行建模，因为他对所有对的关系进行建模，多头自注意力相对于Token的数量是具有二次复杂性的。
2. **因子化编码器(Factorised encoder):**

****

如上图，该模型由两个独立的Encoder组成，第一个为空间Encoder，目的为了从同一时间索引中提取Token之间的相互作用。经过的Transformer后能获得每一帧的表示。将所有帧的表示集合在一起由个Transformer层组成的时间编码器，来处理不同Token之间的相互作用。最后对该编码器输出的Token进行分类。

1. **因子化自注意力(Factorised self-attention)：**与模型1架构相同，但是先计算空间自注意力，在计算时间自注意力，因此架构中的每个自注意力块都有时空交互。为计算空间自注意力，首先将Token的大小从重置为t(通过表示)，同理，将时间的自注意力输入重置为。如下式：
2. **因子化点积注意力(Factorised dot-product attention)：**

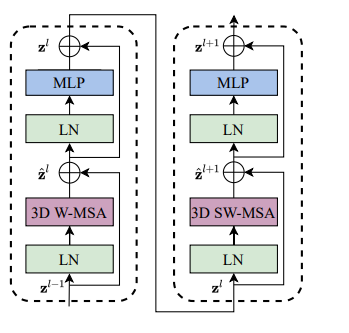
****

Model4如上图所示，计算复杂度与模型2和模型3相同的模型，使用不同head分别在时间和空间维度上计算每个Token的注意力权重。

该模型分别构建了**,**，该模型同时通过一半的注意力头计算另一半的注意力头计算分别关注时间维度和空间维度的Token，最终通过线投影来获得最终的输出。

ViViT在多个视频分类基准上取得了最先进的结果，包括Kinetics 400和600、Epic Kitchens、Something-Something v2和Moments in Time。在时间上取得了最先进的结果，超过了先前基于深度三维卷积网络的方法。

**Video Swin Transformer：**Cao et al.提出了在视频Transformer中引入局部性的归纳偏置[]，Video Swin Transformer的架构能够分解时间和空间的信息并进行全局的注意力计算，具体架构的实现是调整为图像设计的Swin Transformer[Swin Transformer]来实现的，同时也借助了预训练模型的力量。



Video Swin Transformer同Swin Transformer一样采用滑窗的方法，两个连续的Video Swin Transformer计算如下式：

其中和分别表示3D(S)W和FFN模块输出的特征，3DW-MSA和3DSW-MSA分别表示基于三维窗口的多头自注意力使用正则和移位的窗口分区配置。

Video Swin Transformer在视频识别的基线上实现了最先进的精准度。如在动作识别的验评集Kinetics-400取得了84.9%的top-1准确率，在验评集Kinetics-600上以小20倍的预训练数据和小3倍的模型大小取得了69.6%的Top-1准确率。

# 4.3.2 Object Tracking

物体追踪的任务是获取一组初始的物体检测，为每个初始检测创建一个唯一的ID，然后追踪每个物体在视频帧中的移动，保持ID分配。

**TransTrack:**简单在线和实时跟踪（SORT）是一种实用的多物体跟踪方法，重点是简单、有效的算法。在TransTrack中，Wojke et al.整合了外观信息以提高SORT的性能。由于这一扩展，TransTrack能够通过更长的闭塞期来追踪物体，有效地减少了身份转换的次数。本着原始框架的精神，TransTrack将大部分的计算复杂性放在离线预训练阶段，在这个阶段模型在大规模的人的重新识别数据集上学习深度关联度量。在在线应用中，TransTrack使用视觉外观空间中的近邻查询来建立测量与跟踪的关联。实验评估表明，TransTrack的扩展将身份切换的次数减少了45%，在高帧率下实现了具有竞争力的整体性能

## 4.4 Efficient Transformer

## 4.5 Other tasks

# 5.Disscusion&Conclusion

# 5.1 Conclusion

与卷积神经网络相比，Transformer现在已经成为计算机视觉领域最热门的研究方向。受惠于Transformer架构强大的建模能力和迁移学习能力，Transformer能在各项计算机视觉领域的任务种取得与CNN相当甚至更胜的表现。

# 5.2 Future Prospects

Transformer在计算机视觉领域的研究已经取得了很大的进展，并且已经显示出在多个基准上接近或超过CNN方法的SOTA结果。但Transformer仍不成熟，计算机视觉领域仍由CNN主导。

**冗余消除：**在NLP领域中冗余消除就是消除词与词之间的重复表达，但词的冗余度较低所以Transformer能在NLP领域成功应用；但在计算机视觉领域，以ViT为例，一张图片被分为一定数量固定大小的Token且每个Token包含16\*16个像素，由于图像的局部相关性，相邻的Patch之间有很大的相关性，所以图像像素存在较大的冗余。如何解决各类任务元素的冗余性问题，优化Transformer在计算机视觉领域的性能，会成为未来的一个研究点。

**效率优化：**在传统CNN中图片尺寸与计算复杂度呈线性关系，对于Transformer计算长程依赖关系的计算复杂度与图片尺寸（N）呈的关系。虽然Transformer的建模能力强大，但是需要更多的数据集和更长的训练时间和资源，所以进一步优化Transformer 的效率是未来的一个研究点。

**视频处理：**CNN在计算机视觉的视频处理领域还没有实现与人脸识别，目标检测所能达到的精度，Transformer易于计算长程依赖的特性，使得Transformer可能在时间和空间两个维度对视频信息建立强大的模型，成为计算机视觉视频处理领域的标准范例。Transformer对于视频信息的处理也将是未来关于Transformer研究的热点之一。

**多模态任务：**Transformer在NLP、CV和Audio等领域的成功应用，昭示着Transformer可能一统人工智能的各项领域。借助于Transformer的Encoder-Decoder架构，多模态的任务能够简单的通过Encoder进行信息特征提取，再通过Decoder进行解码，完成Encoder提取的信息再其他领域的表示，同时Transformer强大的建模能力也能够保证Encoder中信息的有效性和完整性。多模态任务也将是Transformer未来的研究热点之一。

**自监督学习：**在计算机视觉领域，虽然 Vision Transformers 可以取得比其他传统架构更好的结果，但它们的成功取决于对数据的相当大的需求。因此，以受监督的方式训练这些模型需要进行大量的标记工作，这并不总是可行或可持续的。因此，实现 Vision Transformers 的自监督方法可能是使这些模型不仅强大而且更容易应用于更广泛问题的一种可能方法。自监督学习也将是Vision Transformer未来的研究热点之一。