2022-1-6 ---- 2022-1-10

论文笔记

Transformers in computational visual media: A survey

- 链接：02-cvm0247.pdf (springer.com)

Abstract

- 使用self-attention机制，可以并行表示全局信息
- 通过**任务场景**将transformer分类并分析各任务场景的**key idea**
  - backbone design: review in details
  - high-level vision
  - **low-level vision**: focus on
  - **generation**: focus on
  - multi-modal learning
- 给予**量化比较**，为low-level vision以及generation任务展示图片
- 提供：**计算成本**以及**相关源码链接**

Introduction

transformer advantages:

- Transformers **learn with more inductive bias** and performs better when **trained on large datasets**
- Transformers provide a more **general architecture** suitable for most fields, including NLP, CV, and multimodal learning
- Transformers powerfully model long-range interactions in a computationally-efficient manner
- The learned representation of relationships is more **general and robust** than the local patterns from convolution modules

综述中提及的模型结构

Table 1 Recent visual transformers introduced in this survey

Area	Secondary area	Method	Contributions
Backbone network	Classification	T2T ViT [15]	An effective and efficient tokens-to-token module
		TNT [16]	The first to exploit the benefit of pixel-level relations
		CPVT [17]	An instance-level position embedding module
		ConViT [18]	Adaptive reception field in visual transformers
		DeepViT [19]	A Re-Attention module for deep-layer ViTs
		Swin Transformer [20]	A shifted-window based MSA & a deep-narrow module
		PiT [21]	The first to investigate the benefit of pooling in ViTs
		LocalViT [22]	A depth-wise convolution based module to exploit locality
	Visualization	Transformer-Explainability [23]	A better tool to visualize feature maps from ViT models
High-level vision	Detection	DETR [24]	First transformer-based detection SOTA model
		Deformable DETR [25]	An efficient attention module reducing time consumption
		UP-DETR [26]	An unsupervised pre-training method for DETR
		PVT [27]	A general transformer architecture for dense prediction
	Segmentation	VisTR [28]	First transformer-based segmentation model
		SegFormer [29]	A lightweight efficient segmentation transformer model
Low-level vision	Colorization	ColTran [30]	First transformer-based image colorization model
	Text-to-image	TIME [31]	Text-to-image generation
		DALL·E [32]	Zero-shot text-to-image generation framework
	Super resolution	IPT [11]	Image processing model
		TTSR [33]	Flexible application of transformer
	Image generation	TransGAN [34]	First pure transformer-based GAN for generation
		GANsformer [35]	A bipartite transformer
		VQGAN [36]	A transformer-based high-resolution image generator
	Image restoration	Uformer [37]	A transformer-based hierarchical encoder–decoder network
	Style transfer	StyTr^2 [38]	First transformer-based style transfer model
	Point cloud learning	PCT [39]	Among the first transformer-based point cloud models
Multi-modality learning	Two-stream model	ViLBERT [40]	The first proposed two-stream model for V+L tasks
	Single-stream model	UNITER [41]	A universal model for joint multi-modal embedding
	Mixed model	SemVLP [42]	First mixed single- and two-stream model

- 对于主干设计，当前的工作主要考虑两个方面
  - 在ViT中注入卷积的先验知识
  - 提高视觉特征的丰富性
- for high-level vision:
  - 基于DETR 的transformer检测模型
- for low-level vision and generation
  - colorization
  - text-to-image
  - super-resolution
  - image generation
- for multi-model learning
  - 回顾一些近期的具有代表性的工作 on vision-plus-language (V+L)

文章结构

- section 2 介绍visual transformer
- section 3 列出最近visual transformer的主干网络
- section 4 描述若干在物体检测领域的比较新的visual transformer设计
- section 5 介绍基于tranformer方法的多种low-level 视觉任务
- section 6 回顾最近在multi--modal learning的具有代表性的工作
- section 7 总结

Visual Transformer

- 若干MSA的变体: Reformer, Performer, and LinFormer

Backbone design

- 在ViT中注入卷积的先验知识
  - T2T-ViT , ConViT, PiT , and Swin Transformer
- 提高visual features的丰富性
  - TNT , CPVT, DeepViT and LocalViT

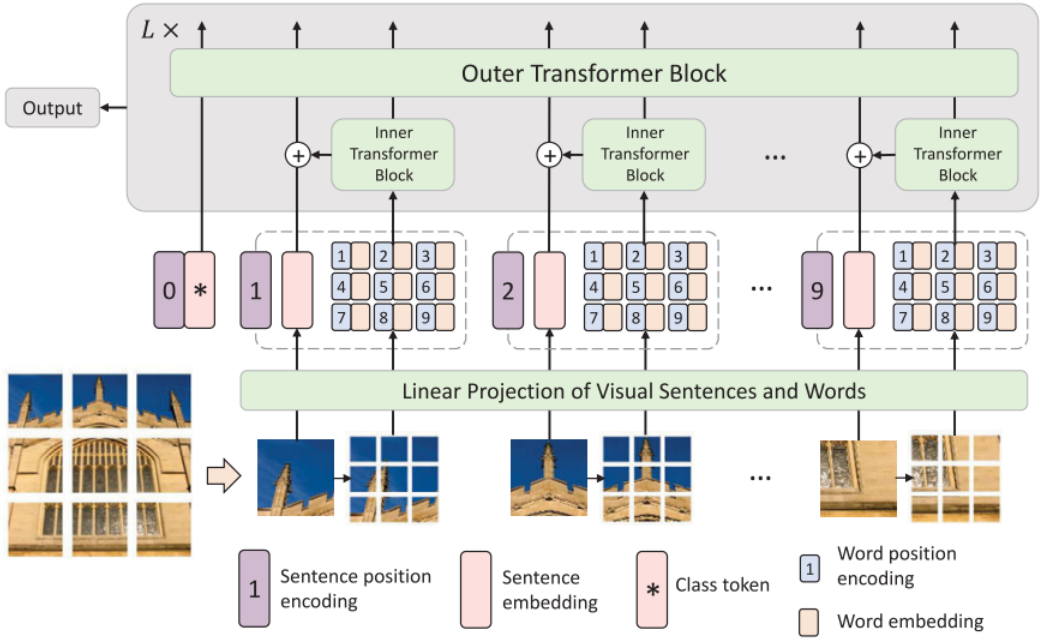
Latest-developments

T2T-ViT

- idea:
  - 关注原ViT模型将输入图像转换为tokens时，不能有效模拟**图像数据的空间结构**
  - 他们认为忽略图像的空间结构可能会导致训练效率变差、降低模型的性能
- solution：
  - 提出一个token-token模块，将空间信息注入到 tokenization of image patch
  - 逐渐降低token的长度，从而降低计算量，减少参数
- inspiration：
  - 根据**卷积**的架构，设计了一种**deep-narrow ViT 框架**
  - 这种架构可以减少参数的数量，增强训练的有效性

TNT

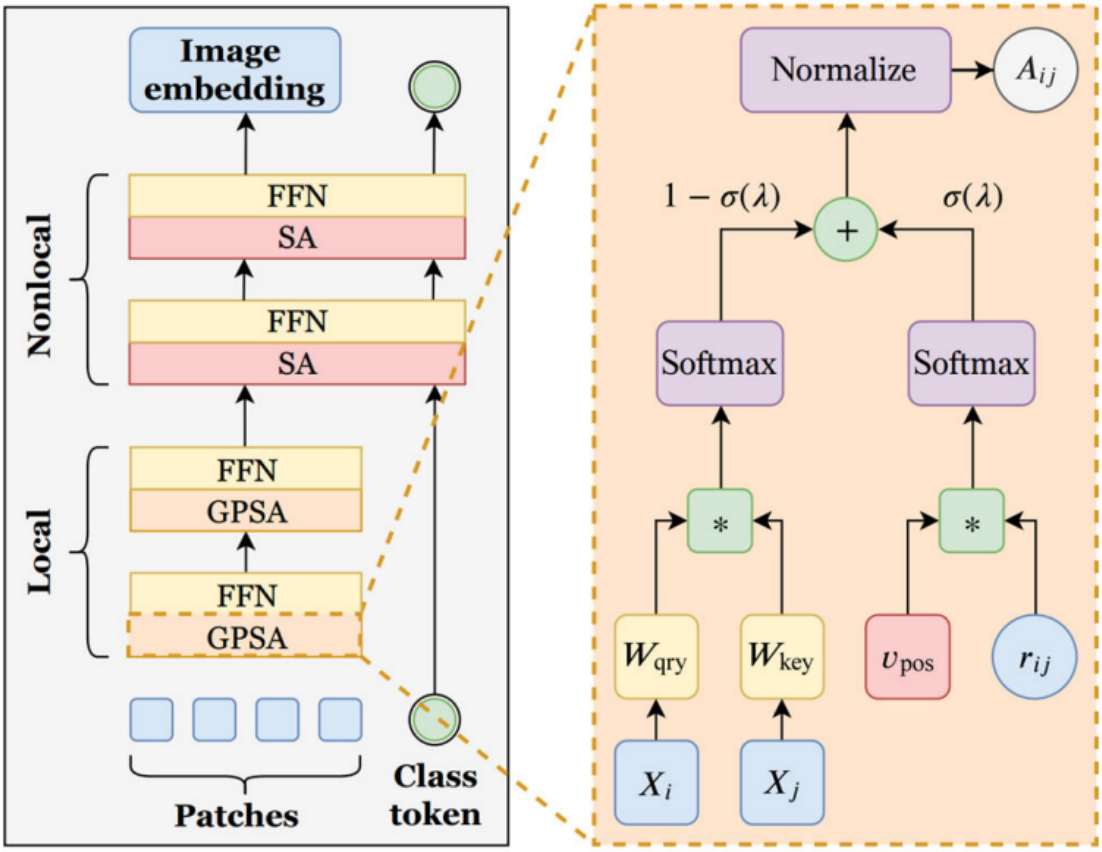
- a **Transformer-iN-Transformer (TNT)** framework
- idea
  - 进一步利用图像数据中内在空间结构信息
- solution：
  - 在学习有用的**视觉特征**时考虑patch and pixel level relations
  - 模型结构



    - 提出一个TNT块，有效以及高效利用像素级表示
    - 介绍一个额外的transformer ---- Inner T-Block
        - 在每一个patch中为像素级关系建模
        - 然后加强patch-level特征----通过计算像素级信息
- result：
    - 81.3% top-1 classification accuracy on ImageNet

ConViT

- a ViT model with soft convolutional inductive biases (ConViT)
- idea:
    - CCsay: 卷积更适合提取图像的特征信息
    - 卷积的inductive bias使得其训练时可以节约更多的样本，将卷积的inductive bias运用于transformer中，可以提高对样本的利用率【1】
    - 给transformer一个适应性的接受域
- core block: a gated positional self-attention(GPSA) module



- GPSA block有两个分支：
  - $\boldsymbol{W}_{\text{qry}}$ or $\boldsymbol{W}_{\text{key}}$ 用于建模全局 或 长范围关系
  - $\boldsymbol{v}_{\text{pos}}$ 用于建模局部区域的关系
- 同时，为了适应性权衡两个分支，引入一个可学习参数 λ
  - 在所有层以及MSA中所有heads里 初始化为1
- GPSA模块，可以在训练过程中适应性扩展self-attetion 接受域

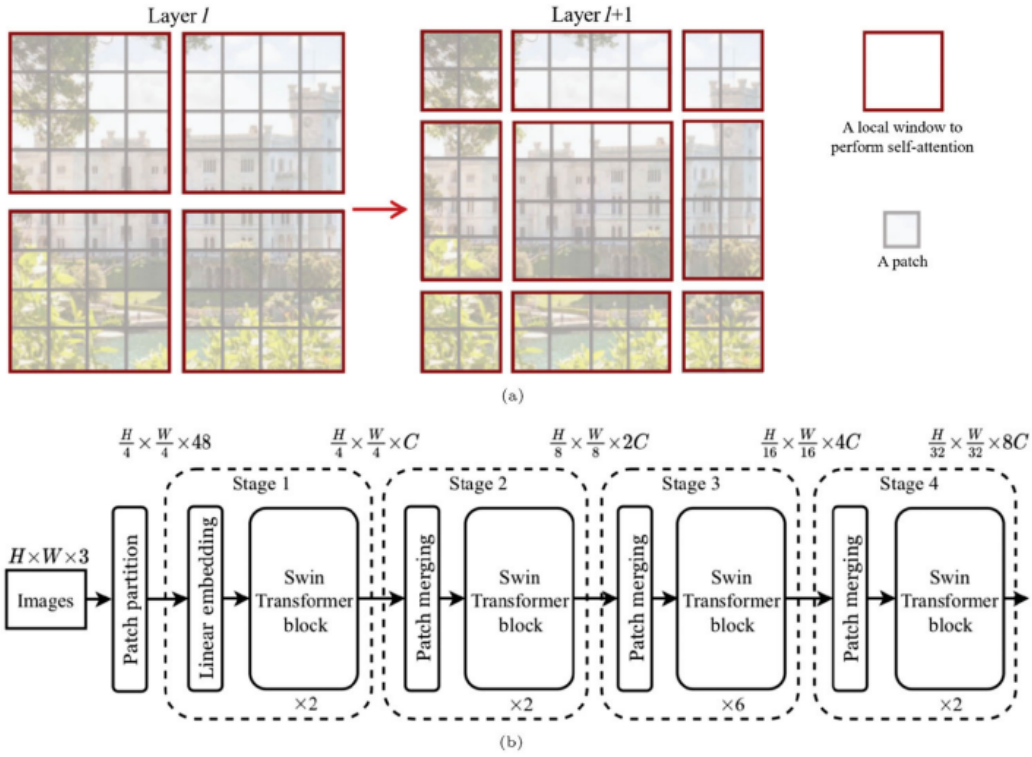
CPVT

- a novel design of position embedding module
- idea：
  - 更进一步加强从ViT中提取视觉特征的丰富性
- solution：
  - a conditional position embeddings for various inpu tokens 类似于动态神经网络设计
  - 在实现中
    - 还以空间方式重新排列输入标记，并应用卷积操作以一种可学习的方式提取位置嵌入。这样，它们在标记化过程中也保持了局部邻域信息，有利于分类性能
- Two further ViT models：
  - LeViT
  - CoaT
- 原论文abstract：
  - 我们提出了一种用于visual transformer的条件位置编码(CPE)方案[10,30]。
  - 不像以前的固定或可学习的位置编码是预定义的，独立于输入tokens，而CPE是动态生成的，并以输入tokens的局部邻域为条件。因此，CPE可以很容易地推广到比模型在训练过程中所见

过的更长输入序列。
    - 此外，CPE可以在图像分类任务中保持所需的平移不变性，从而提高分类精度。CPE可以通过一个简单的位置编码生成器(PEG)轻松实现，并且可以无缝地集成到当前的Transformer框架中。
    - 在PEG的基础上，提出了条件位置编码的视觉转换器(CPVT)。我们证明了CPVT与那些学习过的位置编码相比具有视觉上相似的注意地图。得益于条件位置编码方案，我们 ImageNet分类任务上获得了与vision transformer相比的最先进的结果 (state-of-the-art)。

Swin Transformer

- idea：
    - 图像数据包含大量空间信息，deep-narrow CNN 架构的成功
    - 视觉实体变化大，在不同场景下视觉Transformer性能未必好【2】
    - 图像分辨率高，像素点多，Transformer基于全局自注意力的计算导致计算量较大【2】
- 模型结构



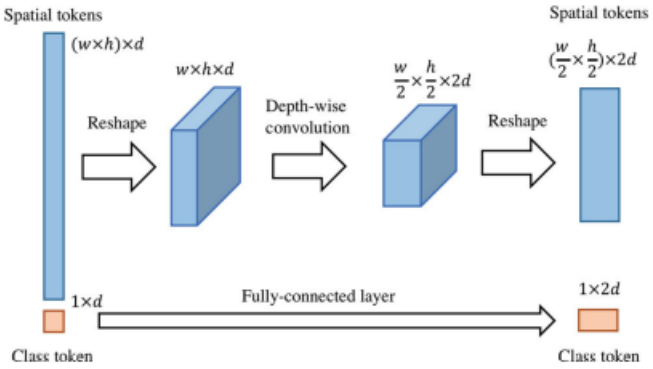
Deep ViT

- idea：
    - 分层以及层的缩放是CNN architecture的重要部分
    - 发现：在ViT上，transformer block堆叠20个时，性能会达到饱和
        - 注意力崩溃
            - 从一个MSA module提取出来的注意力特征图共享越来越相似的模式
            - 从而导致大量信息冗余 以及 低训练效率
    - 如果MSA头部之间的通信被提升，那么头与头之间的冗余将会被减少
- solution：
    - 基于上述驱动，提出一个简单有效的 Re-Attention module
        - 提供一个，可学习参数，以方便在一个MSA模块内的头与头之间通信
- abstract：
    - 我们考虑了参考**图像分割**的问题。给定一个输入图像和一个自然语言表达式，目标是**分割图像中由语言表达式引用的对象**。在这一领域的现有工作将语言表达和输入图像分开对待。他们没

有充分捕捉到这两种模式之间的长期相关性。在本文中，我们提出了一个跨模态的自我注意(CMSA)模块，该模块可以有效地捕捉语言和视觉特征之间的长期依赖关系。我们的模型可以自适应地聚焦于参考表达式中的信息词和输入图像中的重要区域。此外，我们提出了一种门控多级融合模块来选择性地整合图像中对应不同层次的自注意跨模态特征。该模块控制不同级别特征的信息流。我们在四个评估数据集上验证了所提出的方法。我们提出的方法一贯优于现有最先进的方法。

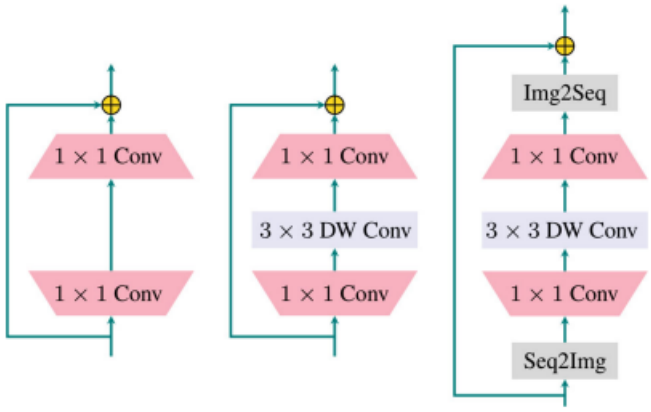
PiT

- idea:
  - 考虑到池化层的在卷积架构中对模型能力以及泛化性能的重要性
  - 提出在ViT中利用池化模块
- model



LocalViT

- idea
  - 学习ViT和CNN架构之间的不同，他们发现ViT擅长建模全局关系，但缺乏局部方案去学习局部区域的交互
  - 局部机制对于图像数据的空间结构建模是非常重要和有用的
- model：



Comparision on ImageNet

Table 2 Classification accuracy on ImageNet [64] for various visual transformers

Method	Image size	FLOGs (G)	#Param (M)	Acc (%)	Source (GitHub)
Convolution-based neural network					
ResNet [1]	224^2	4.1	25.6	76.2	—
RegNetY-4G [3]	224^2	4.0	21	80.0	facebookresearch/pycls
RegNetY-16G [3]	224^2	16.0	84	82.9	
EfficientNet-B0 [2]	224^2	0.4	5.3	77.1	rwightman/gen-efficientnet-pytorch
EfficientNet-B1 [2]	224^2	0.7	7.8	79.1	
EfficientNet-B3 [2]	300^2	1.8	12	81.6	
EfficientNet-B5 [2]	456^2	9.9	30	83.6	
EfficientNet-B7 [2]	600^2	37.0	66	84.3	
Visual transformer					
ViT [9]	384^2	55.4	86	77.9	google-research/vision_transformer
	384^2	190.7	307	76.5	
DeiT [65]	224^2	4.6	22	79.8	facebookresearch/deit
	384^2	55.4	86	83.1	
T2T ViT [15]	224^2	5.2	21.5	80.7	yitu-opensource/T2T-ViT
TNT [16]	224^2	5.2	23.8	81.3	huawei-noah/noah-research/tree/master/TNT
	224^2	14.1	65.6	82.8	
CPVT [17]	224^2	—	23	81.5	Meituan-AutoML/CPVT
	224^2	—	88	82.3	
ConViT [18]	224^2	5.4	27	81.3	—
	224^2	17	86	82.4	
DeepViT [19]	224^2	—	27	82.3	zhoudaquan/dvit_repo
	224^2	—	55	83.1	
Swin Transformer [20]	224^2	4.5	29	81.3	microsoft/Swin-Transformer
	384^2	47.0	88	84.2	
PiT [21]	224^2	4.6	22.1	81.9	naver-ai/pit
	224^2	12.5	73.8	84.0	
LocalViT [22]	224^2	4.6	22.4	80.8	ofsoundof/LocalViT

Visualization of ViT

参考文献

【1】 [论文阅读]ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases - 知乎 (zhihu.com)

【2】 Swin Transformer: Hierarchical Vision Transformer using Shifted Window