Research Investigation for Transformer for Vision

（Correlation（relationship） is the core of the core，相关性才是核心中的核心）

**Summary**

（Correlation is the core of the core，相关性才是核心中的核心）。Correlation is even more important than local features.

**Introduction**

1. Transformer for Vision是从NLP领域借鉴过来的用到CV领域的优秀算法，Transformer（Attention is all you need）借助Attention机制在NLP领域部分取代了传统RNN结构，对长序列（长期依赖）的语言处理有着很好的作用。它有着Encoder和Decoder的特殊结构，突破了旧有RNN和LSTM的序列时间长度的瓶颈，达到了极好的水平。[1]2020年借助CNN对短时序列更好的特征提取能力，Facebook的conformer（Transformers with convolutional context for ASR）的提出使模型性能有更进一步提升。从NLP领域发展来看，CNN和Attention的结合取得了很好效果。

2. ViT（vision transformer）[2]是Google在2020年提出的直接将transformer应用在图像分类的模型，后面很多的Vision Transformer的工作都是基于ViT进行改进的。ViT的思路很简单：直接把图像分成固定大小的patchs，然后通过线性变换得到patch embedding，这就类比NLP的words和word embedding，然后将patch embedding输入到Transformer encoder里面就可以进行相应的图像分类任务。

3.Transformer的具体介绍可以查看[1]，它给我的最大感受就是（Correlation is the core of the core，相关性才是核心中的核心）。但和论文中的理解殊途同归。最重要的不是某一核心特征，而是整体之间的关系。

4．为什么重要？CNN是特殊的Self-Attention，Self-Attention是CNN的拓展，这非常重要。也就是说我们有了更加普遍更加高级的手段。这有严格数学证明（https://arxiv.org/abs/1911.03584）On the Relationship between Self-Attention and Convolutional Layers。local multi-head dot-product self attention blocks can completely replace convolutions。注意力机制具有全局视野，能够充分考虑到不同特征部分的相关性，已经有证明得出在训练样本比较大时，Self-Attention的上限更高，更具未来性。同时在Transformer for Vision领域使用transformer研究CV开始相对还没有多久，手段还比较初级，还有很大的提升空间。它让我看到了至少在图片合成领域，如何解决目前合成图片的五官相对位置不正确的问题，有了新的思路和新的解决方案。在CV领域，Transformer for Vision训练过程和训练样本数量都要求很高，需要新的手段来解决这些问题。最后，Transformer提供了一种多模态融合能力，CNN不擅长融合其他模态的信息，例如文字、标签、语音、时间等等各种相关空间离散且抽象的信息，Transformer可以。

5.总结最近工作见附表，我也注意到 [Bottleneck Transformers for Visual Recognition](https://arxiv.org/abs/2101.11605)这种最新进展，将ResNet中的第4个block中的bottleneck替换为MHSA（Multi-Head Self-Attention）模块，形成新的模块，取名叫做Bottleneck Transformer (BoT) 。最终由BoT这样的block组合成的网络结构就叫做BotNet。最终基于BotNet的Mask R-CNN的实例分割在coco数据集上取得了44.4% Mask AP ，49.7% Box AP。在分类任务中，在 ImageNet上取得了84.7%的top-1准确性。并且比 EfficientNet快2.33倍。依旧使用CNN的主体结构，性能大幅提升，计算量相对不是很大。

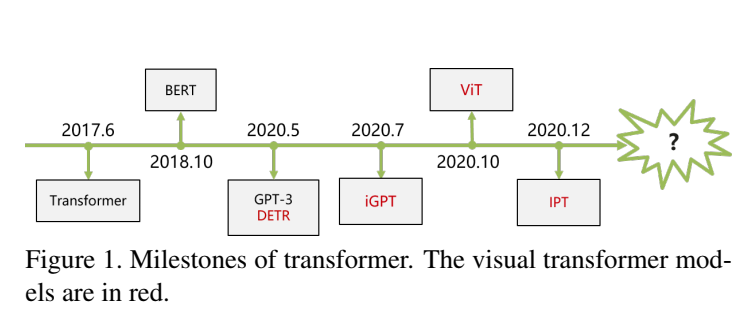
GitHub：<https://github.com/leaderj1001/BottleneckTransformers>

<https://github.com/lucidrains/bottleneck-transformer-pytorch>

CNN与self-attention相结合的：

1. by augmenting feature maps for image classification
2. by further processing the output of a CNN using self-attention

**Methodological Approachs**



1. **CNN**

没有全局视野，深度学习的核心算法之一，取得了很大的成就。

合理性：CNN滤除无效信息，表征有效信息。具有很强特征提取能力，但视野太窄，犹如孔中窥豹。

1. **VIT：AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE**

其实核心问题就是考虑如何把图像数据H\*W\*C,序列化成一个一个词那种结构，自然就想到将图片crop成一个一个patch，假设有N个patch,维度为p\*p\*C,reshape加concate一下就变成个N\*p^2C,也就类似词向量。

1. **.图像转序列②.Position embeddings** **③.learnable embedding④.输入transformer encoder**

实验结果:

在mid-sized的数据上训练，精度比同等规模的ResNet网络低几个百分点；

在大规模的数据集上训练，迁移到较小规模的数据集上结果>= state of the art。（特别是，最佳模型在ImageNet上达到88.55%，在ImageNet-ReaL上达到90.72%，在CIFAR-100上达到94.55%，在19个任务的VTAB套件上达到77.63%。）

直接放个工程：https://github.com/lukemelas/PyTorch-Pretrained-ViT

合理性：把NLP大发神威的纯transformer应用到了CV领域。**ViT中不同的query是share key set的，这会使得内存访问非常友好而大幅度提速。**ViT的出现改变了很多固有认知**1. locality（局部性）；2. translation invariance（平移不变性）。**

问题：复杂，计算效率低,收敛困难。

**3.Look Closer to See Better：Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition**

在这篇文章里面，作者们提出了一个基于 CNN 的注意力机制，叫做 recurrent attention convolutional neural network（RA-CNN），该模型递归地分析局部信息，从局部的信息中提取必要的特征。同时，在 RA-CNN 中的子网络（sub-network）中存在分类结构，也就是说从不同区域的图片里面，都能够得到一个对鸟类种类划分的概率。除此之外，还引入了 attention 机制，让整个网络结构不仅关注整体信息，还关注局部信息，也就是所谓的 Attention Proposal Sub-Network（APN）。这个 APN 结构是从整个图片（full-image）出发，迭代式地生成子区域，并且对这些子区域进行必要的预测，并将子区域所得到的预测结果进行必要的整合，从而得到整张图片的分类预测概率。

合理性：想法很有趣，重要的部分走近点放大看。缺点，要是有两个及多个重要目标呢？

**4 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

论文原文:  
[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](https://arxiv.org/abs/1810.04805)

官方代码和预训练模型  
Github: <https://github.com/google-research/bert>

**BERT优点**

Transformer Encoder因为有Self-attention机制，因此BERT自带双向功能

因为双向功能以及多层Self-attention机制的影响，使得BERT必须使用Cloze版的语言模型Masked-LM来完成token级别的预训练

为了获取比词更高级别的句子级别的语义表征，BERT加入了Next Sentence Prediction来和Masked-LM一起做联合训练

为了适配多任务下的迁移学习，BERT设计了更通用的输入层和输出层

微调成本小

**BERT缺点**

随机遮挡策略略显粗犷，推荐阅读《Data Nosing As Smoothing In Neural Network Language Models》

[MASK]标记在实际预测中不会出现，训练时用过多[MASK]影响模型表现;

每个batch只有15%的token被预测，所以BERT收敛得比left-to-right模型要慢（它们会预测每个token）

BERT对硬件资源的消耗巨大（大模型需要16个tpu，历时四天；更大的模型需要64个tpu，历时四天。

**5.** [**GPT模型（Generative Pre-Training）**](https://www.cnblogs.com/yifanrensheng/p/13167796.html)

均是采用的transformer的结构，而GPT是单向的。

GPT 使用 Transformer 的 Decoder 结构，并对 Transformer Decoder 进行了一些改动，原本的 Decoder 包含了两个 Multi-Head Attention 结构，GPT 只保留了 Mask Multi-Head Attention，如下图所示。

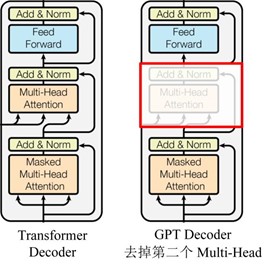


图2 GPT模型

Image GPT

Image GPT的要点：

使用与NLP中的GPT-2相同的transformer架构

无监督学习，无需人工标记

需要更多的计算来生成有竞争力的表示

学习到的特征在低分辨率数据集的分类基准上实现了SOTA性能

合理性：图像版GPT3打破语言与视觉界线



图3 GPT3用一句话生成图片

https://www.youtube.com/watch?v=-6Xn4nKm-Qw

**6.End-to-end object detection with transformers（DERT）**

即同时得到N个推理结果，然后计算真值和预测这两个集合的差异，并用二分图loss以匹配预测和真相对象。不过由于DERT的速度比较慢，之后有ACT等方案的解决，甚至引入蒸馏网络的MTKD方法等。缺点：训练周期长，对小物体检测效果差。

7. **Pre-Trained Image Processing Transformer（IPT）**

paper：<https://arxiv.org/pdf/2012.00364.pdf>

提出了一种处理低水平视觉任务（超分辨率、图像去雾、图像去噪）的预训练模型。

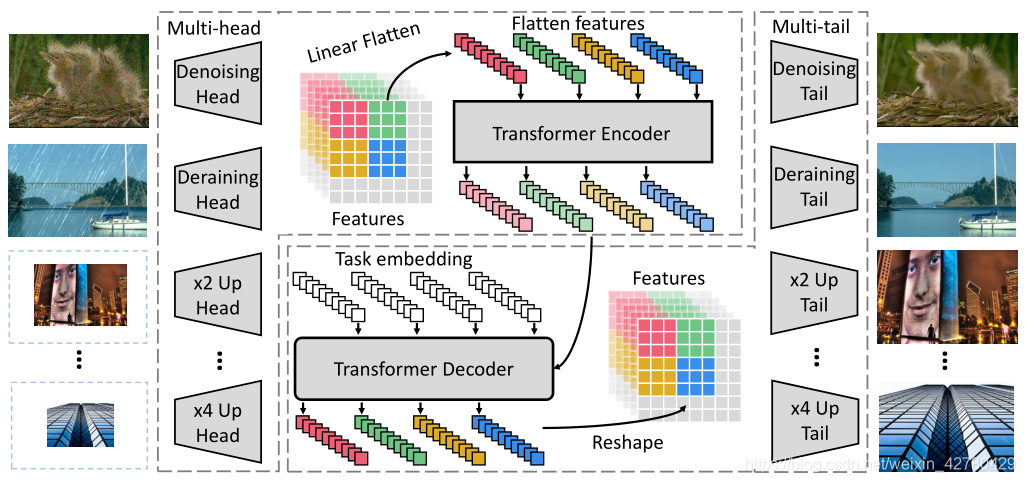


图4 IPT结构示意

1.头部（Heads）

2.编码器（Transformer encoder）

3.解码器（Transformer decoder）

4.尾部（Tails）

**在超分辨率、去噪、去雨任务上表现SOTA！**性能优于IGNN、RDN和RCDNet等网络

**Findings：**

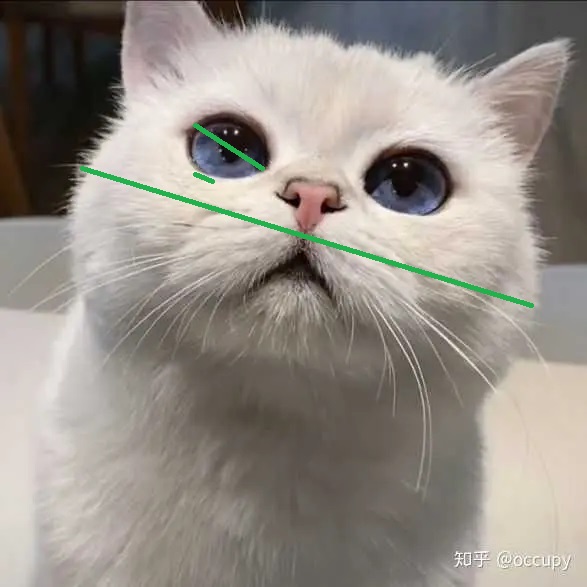


图5 一张猫脸示意图

短距离举例 中距离举例

当我们展现一只猫的图片出来，构成这只猫的点阵之间存在相关性，否则点阵组成的图片不能叫做猫。构成猫的眼睛的点之间存在很强相关性，同理猫嘴和猫鼻子也是。从随机理论拓展，构成猫的点阵图中的任何点都和自身以及其它的任何一个非本身点存在大小不同的相关性。像素点之间的距离是不同的，最小是相邻像素点，最大的可以横跨整张图片，猫的眼睛和嘴间的距离不大不小。像素点间的距离小一般相关距离就会小，但也有特例，比如猫脸边缘的点距离小但相关距离却要很大。相关距离短的点相关性更强，猫脸中的点和非猫脸中的点相关性就很弱，说明它们之间的相关距离较长。而猫的眼睛之间相关性很强，相关距离相对较短。相关性强的点通过非线性表征会在高维空间中收敛在一起，形成各自的相关空间。不同相关空间所占的体积和维度、特性各不相同但各自收敛并有着一定互相包含。

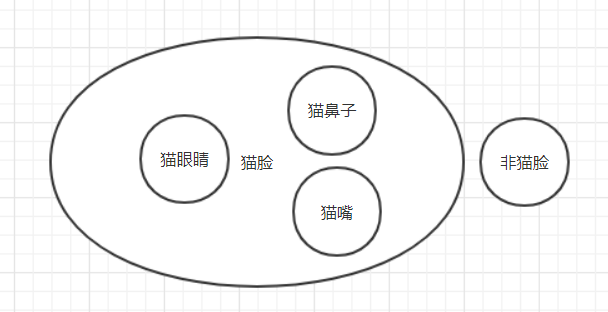


图6 相关收敛空间示意

从下图可知Transformer的self-attention确实就是在计算序列之间的相关性，我们通过对transformer网络的学习，抑制掉不重要的非猫脸部分，突出猫脸部分。这实际上就是在划分收敛的相关空间，同时对部分相关空间进行抑制（非猫脸部分也会收敛到与猫脸不同的相关空间）。这时，我们就“attention“到了猫脸。

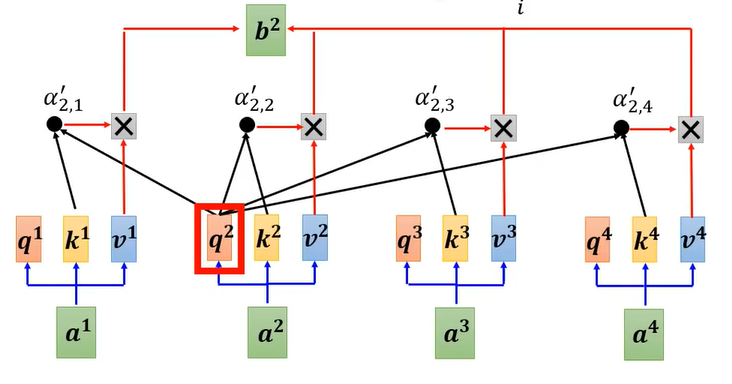


图7 transformer部分计算方法



图8 原图片



图9 注意后的图片

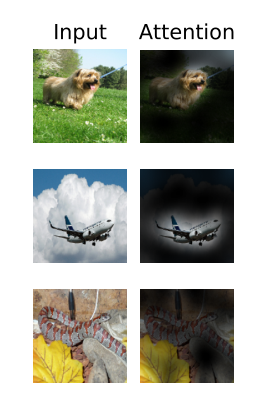


图10 再举一例

从CV角度来讲，全局感受野特别重要。用Attention机制来替换旧有各种在CV领域利用人为设定的大小不同感受野达到特定目的的手段是未来几年可以预见的发展方向。有理论严格证明，CNN是特殊的Self-Attention，这非常重要。CNN的conv算子一个问题就是感受野比较局限，为了扩大网络的关注区域，需要卷积-池化-堆叠多层这种结构，带来的问题其实是“有效/真实”感受野是以某个中心为原点向外高斯衰减的，因此CNN通常的有效attention就是图中的某一两个比较重要的parts。

我们需要注意到，文本和语音尤其是文本本身已经是对信息进行了特征提取，是相对抽象的，我们要想表征语音信息所需要的非线性空间相对较小，需要的网络其实可以不特别深。而图片只是点阵图，并没有进行特征提取，它包含了各种丰富的无用信息，我们需要的只是其中的一部分有效信息。在CNN中，我们利用非线性特性构建深层网络对图片的空间进行非线性表征，利用卷积网络对图片进行特征提取（滤波）来滤除无用信息。这里面包含两个重点，一个是滤除无用信息，另一个是对信息进行非线性表征。今天部分论文提出完全使用transformer来处理图像，不仅参数量和计算量特别大，而且很难训练使其收敛。其实Transformer的结构中存在非线性层，是具有一定的特征提取与表征能力的，对于语音和文字的抽象特性来说，特征本身就被提取的，所以transformer的非线性能力够用。但对于图片来说却很难做到收敛。同时直观说transformer看到的局部特征只是大概的，不是特别清晰。局部特征还看不清呢，所以训练中并不容易过拟合。

图片的相关性重不重要呢？从识别角度讲不那么重要，因为我只要看到部分特征，比如猫的眼睛、猫的嘴和鼻子我基本就可以判定这是只猫了，而五官的排布不重要，这是基于部分特征的联想与补全。但是在图片合成和进一步提高识别率的角度，猫的眼睛、嘴和鼻子的排布的位置是重要的，它们之间具有相关性，只有特定的排布才能认为这是只猫，把嘴或者鼻子弄反了，这是在CNN算法中常发现的问题。当我们能看到猫脸的相对排布时（构成猫脸的点阵每个点之间的相关性时），即便我看不太清楚猫的眼睛的具体局部结构，我的判定能力其实也是要比只看眼睛的局部特征也要强的。我们可不可以通过增加层数来增加transformor的能力呢，我本人认为不可以，因为它提取的是点阵之间的相关性，相关性是不具备传递性的，两个点之间相关性强和其它两个点之间相关性强其实是没啥关系的。在相关角度，甚至我们可以用transformer来看两个图片是不是来自同一张图片的部分截取。

局部特征同样重要，而CNN网络特别适合做这个，通过CNN来实现局部特征的提取和降低参数量，提高非线性能力是重要的，把CNN和transformer结构结合在一起，哪怕是利用已有的conformer来替换Transformer for Vision中的网络结构，都可以尝试一下，应该会有所提升。我也坚信，这种结合的方法必将会在近几年取得更大突破。尤其是用transformer来替代旧有方式实现全局感受，一定可以取代传统深度方法取得突破的。

有个思考，对于视频来说是否可以结合CNN和Transformer来实现更好的特性呢。毕竟transformer是在端到端的网络是在时域中展开的，是否有机会增加视频每一帧的核心语义的相关性呢？

**References：**

[1]（Li Hongyi和Wang Shusen的视频课程，

https://www.bilibili.com/video/BV1Xp4y1b7ih?from=search&seid=10483150581334169827，https://www.bilibili.com/video/BV1BQ4y1R7V7?from=search&seid=4328378749261091290）

或者（Original Paper[Attention Is All You Need]

https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf). Ashish Vaswani\*, Noam Shazeer\*, Niki Parmar\*, Jakob Uszkoreit\*, Llion Jones\*, Aidan N. Gomez\*, Łukasz Kaiser\*, Illia Polosukhin\*. NIPs 2017.）

附表

Classification

[2][AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE](https://arxiv.org/pdf/2010.11929.pdf). Alexey Dosovitskiy∗, Lucas Beyer∗, Alexander Kolesnikov∗, Dirk Weissenborn∗, Xiaohua Zhai∗, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. Arxiv 2020.

An alternative way to scale attentionis to apply it in Blocks of varying sizes(Weissenbornetal.,2019), in the extreme case only along individual axes(Hoet al., 2019; Wang et al., 2020a).

Detection

[Fast Convergence of DETR with Spatially Modulated Co-Attention](https://arxiv.org/pdf/2101.07448.pdf). Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, Hongsheng Li. Arxiv 2021.

[End-to-End Object Detection with Adaptive Clustering Transformer](https://arxiv.org/pdf/2011.09315.pdf). Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, Hao Dong. Arxiv 2020.

[Toward Transformer-Based Object Detection](https://arxiv.org/pdf/2012.09958.pdf). Josh Beal\*, Eric Kim\*, Eric Tzeng, Dong Huk Park, Andrew Zhai, Dmitry Kislyuk. Arxiv 2020.

[Rethinking Transformer-based Set Prediction for Object Detection](https://arxiv.org/pdf/2011.10881.pdf). Zhiqing Sun\*, Shengcao Cao\*, Yiming Yang, Kris Kitani. Arxiv 2020.

[UP-DETR: Unsupervised Pre-training for Object Detection with Transformers](https://arxiv.org/pdf/2011.09094.pdf). Zhigang Dai1, Bolun Cai, Yugeng Lin, Junying Chen. Arxiv 2020.

[DEFORMABLE DETR: DEFORMABLE TRANSFORMERS FOR END-TO-END OBJECT DETECTION](https://arxiv.org/pdf/2010.04159.pdf). Xizhou Zhu∗, Weijie Su∗, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai. Arxiv 2020.

[End-to-End Object Detection with Transformers](https://arxiv.org/pdf/2005.12872.pdf). Nicolas Carion\*, Francisco Massa\*, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. ECCV 2020.

**Segmentation**

[Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers](https://arxiv.org/pdf/2012.15840.pdf). Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, Li Zhang. Arxiv 2020.

[End-to-End Video Instance Segmentation with Transformers](https://arxiv.org/pdf/2011.14503.pdf). Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, Huaxia Xia. Arxiv 2020.

**Tracking**

[TransTrack: Multiple-Object Tracking with Transformer](https://arxiv.org/pdf/2012.15460.pdf). Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, Ping Luo. Arxiv 2020.

**Image Synthesis**

[Taming Transformers for High-Resolution Image Synthesis](https://arxiv.org/pdf/2012.09841.pdf). Patrick Esser*, Robin Rombach*, Bjorn Ommer. Arxiv 2020.

**Action Understanding**

[Video Action Transformer Network](https://arxiv.org/pdf/1812.02707.pdf). Rohit Girdhar, Joao Carreira, Carl Doersch, Andrew Zisserman. CVPR 2019.

**3D Vision Tasks**

**Point Cloud Processing**

[PCT: Point Cloud Transformer](https://arxiv.org/pdf/2012.09688.pdf). Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, Shi-Min Hu. Arxiv 2020.

[Point Transformer](https://arxiv.org/pdf/2012.09164.pdf). Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, Vladlen Koltun. Arxiv 2020.