Transformer in CV

一、Transformer的发展简述

Transformer是一个已经广泛应用在NLP领域的模型，例如机器翻译、问答系统、文本摘要、语音识别等等方向，它出自Google的论文《Attention is all you need》[2]，一经提出，就引起了很大的反响。

谷歌的Transformer模型最早是应用与机器翻译任务，它改进了RNN最被人诟病的训练慢的缺点，利用self-attention机制实现快速并行，并且Transformer会增加到非常深的深度，提升模型准确率。

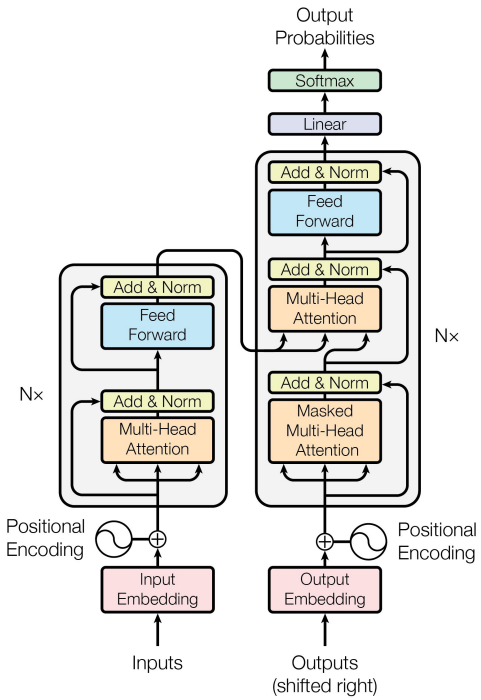
Transformer结构已经在许多自然与阳处理任务中取得了目前最好的成果，Transformer模型的一个主要突破是在20年年中发布的GPT-3，被授予NeurIIPS2020最佳论文。每一个NLPer都应该搞明白Transformer，因为它非常重要。

这里给出进阶参考资料：哈佛大学NLP研究组介绍Transformer原理和代码的文章：http://nlp.seas.harvard.edu/2018/04/03/attention.html

二、Transformer在CV中的结构

2.1 单纯型Transformer结构

首先来看看纯Transformer模型的主要结构，下图是Transformer的整体框架：



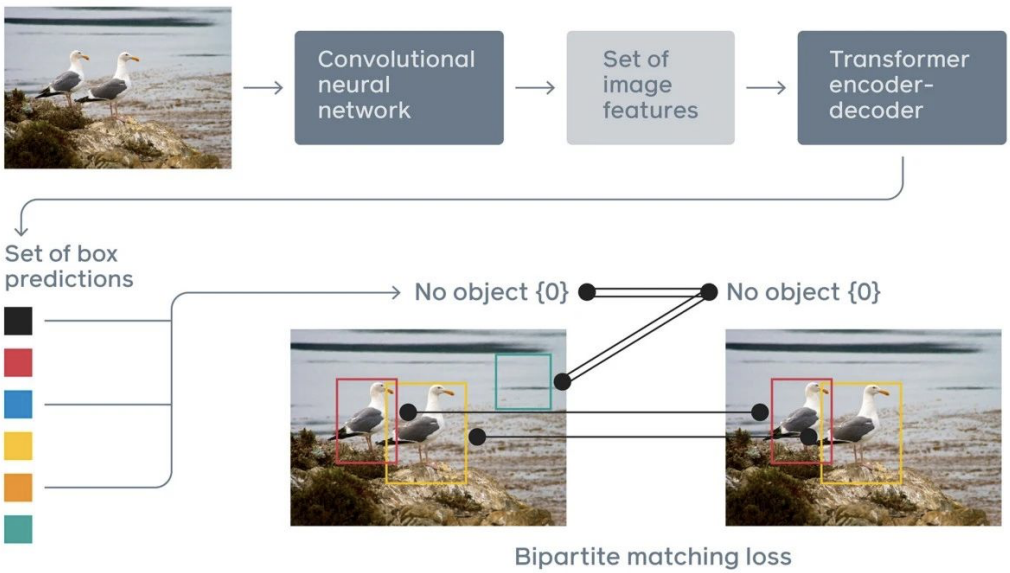
Transformer采用了编码器-解码器（encoder-decoder）结构。上图的左半边用NX框出，代表一层编码器，上图右半边用NX框出的部分，代表一层解码器，论文里的编码器和解码器分别有6层。

定义输入序列首先经过初始词嵌入（word embedding），再和positional encoding相加后，输入到编码器中。

输入编码器的句子首先会经过一个自注意力（multi-head self-attention）层，这层帮助编码器在对每个单词编码时，关注输入句子的其他单词。自注意力层的输出会传递到全连接前馈(feed-forward)神经网络中。在每一个层后，都有一个残差连接和归一化层。为了方便残差连接，模型中的所有子层，包括初始词嵌入层的输出向量维度均为512。六个编码器在结构上是相同的，但是他们之间并没有共享参数。
编码器输出序列经过的处理和输入序列一样，然后输入到解码器。解码器中包括三个层：自注意力（masked multi-head self-attention）层、注意力层(multi-head self-attention)和前馈（feed-forward）层。自注意力（masked multi-head self-attention）层的输入仅包含**当前位置之前的词语信息**，注意力层(multi-head self-attention)的输入包含编码器的输出信息。每个层后面同样都有残差连接和归一化层。
解码器的输出经过一个线性层，再接Softmax。线性层会将输入转化为一个超长向量，softmax层会将向量转化为概率，再通过适当的算法选择输出的词语。
注意力机制最早是应用在图像特征提取中，人观察图像时，并不会观察到每一个细节，而是把注意力放在最重要的部分。现在注意力机制被应用到了NLP任务中，并取得了很好的效果。多头注意力机制（Multi-head self-attention）是使用多个注意力机制进行单独计算，以获取更多层面的语义信息，然后将各个注意力机制获取的结果进行拼接组合，得到最终的结果。
比方说，这个句子是我们要翻译的输入：“The animal didn’t cross the street because it was too tired.”这句话中的“it”指的是什么？是指街道还是动物？对人类来说，这是一个简单的问题，但对算法而言却不那么简单。当模型处理“ it”一词时，自注意力机制使其能够将“it”与“animal”相关联。在模型处理每个单词（输入序列中的每个位置）时，自注意力使其能够查看输入序列中的其他位置，以寻找思路来更好地对该单词进行编码。
原理细节过多，难度较大，并没有深入和展开。针对原理细节可以阅读这篇文章：https://kexue.fm/archives/4765

2.2 混合型Transformer结构

DETR是将CNNs和Transformer相结合的混合结构，它是第一个成功地将Transformer作为pipeline中的主要构建块的目标检测框架。它与以前的SOTA方法(高度优化的Faster R-CNN)的性能匹配，具有更简单和更灵活的pipeline。



DETR的流程如下：

1. CNN学习图像的二维表示并提取特征
2. 将CNN输出加上位置编码，输入Transformer编码器
3. Transformer解码器通过输出嵌入到前馈网络预测类别和包围框

编码器和解码器内部结构与单纯的Transformer结构基本相同。对比传统的目标检测方法，例如Faster R-CNN有多个步骤进行锚的生成和NMS,而DETR放弃了这些手工设计的组件，显著地简化了物体检测pipeline。

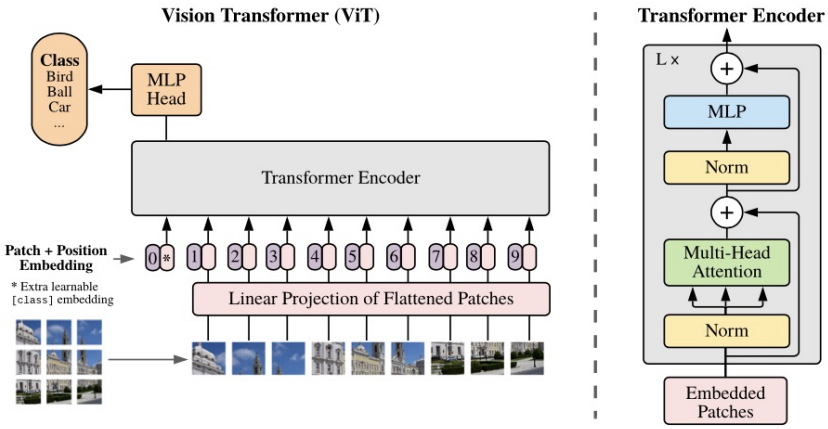
三、 Transformer在CV领域的应用

3.1 简介

Transformer模型在分类、检测、分割上的应用已经比较多，思路都是类似的，因此我将针对每个任务，我举出一篇论文作为例子。关于更多的Transformer在CV上的应用及相关工作，可以查看两篇综述文章[1][2]。

3.2 Transformer在图像分类任务上的应用

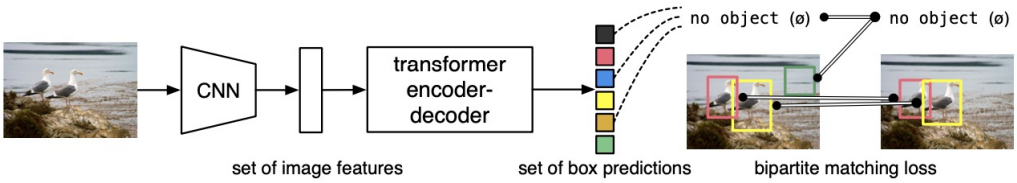
分类Vision Transformer,简称 ViT出自论文： An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[3]。



论文使用Transformer模型进行图像分类。把图像分成固定大小的patches，把patches看成words送入transformer的encoder，中间没有任何卷积操作，增加一个class token来预测分类类别。他在许多图像分类任务上也优于最先进的卷积网络，同时所需的预训练计算资源大大减少(至少减少了4倍)。

3.3 Transformer在图像检测任务上的应用

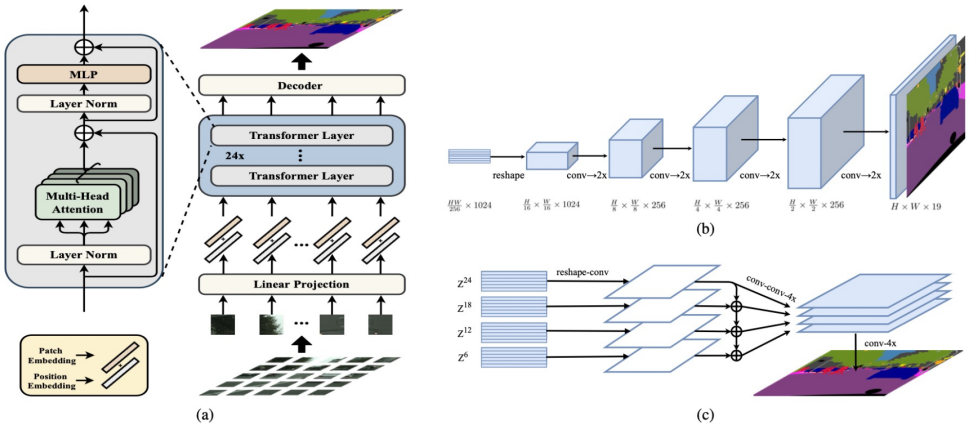
检测 DETR：End-to-End Object Detection with Transformers



论文使用Transformer进行物体检测和分割，先用CNN提取特征，然后把最后特征图的每个点看成word，这样特征图就变成了a sequence words，而检测的输出恰好是a set objects，Transformer正好适合这个任务。

3.4 Transformer在图像分割任务上的应用

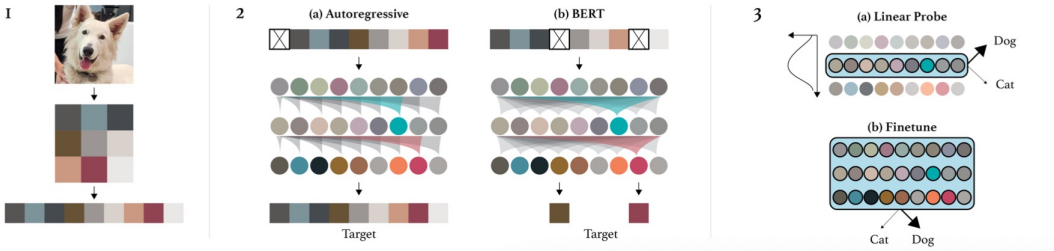
分割 SETR：Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers



用ViT作为的图像的encoder，然后加一个CNN的decoder来完成语义图的预测。

3.5 像素级图像补全

像素级补全：Generative Pretraining from Pixels



类似GPT文本补全，使用Transformer进行像素级图像补全。Image GPT是一个在像素序列上用图像补全训练的GPT-2 transformer 模型。首先，对原始图像进行预处理，将其调整为低分辨率，并将其重塑为一维序列。然后，选择两个训练目标之一，自回归下一个像素预测或掩蔽像素预测。最后，评估这些目标学习到的表示。
来自预训练的图像GPT的特征在一些分类基准上取得了最先进的性能，并在ImageNet上接近最先进的无监督精度。

四、Transformer的优缺点

4.1 Transformer的优点

虽然Transformer没有逃脱传统学习的套路，它也只是一个全连接（或者是一维卷积）加Attention的结合体。但是其设计不同于RNN或者CNN，并且取得了非常不错的效果。
Transformer是一个简单的、可扩展的结构，它可以不局限于NLP领域，也可以用于计算机视觉任务，是非常有科研潜力的一个方向，值得每个深度学习的相关人员仔细研究。
Transformer算法与传统算法相比，训练效率有显著的优势，符合目前的硬件（主要指GPU）环境。

4.2 Transformer的缺点

粗暴的抛弃RNN和CNN虽然非常炫技，但是它也使模型丧失了捕捉局部特征的能力，RNN + CNN + Transformer的结合可能会带来更好的效果。
Transformer失去的位置信息其实在NLP中非常重要，而论文中在特征向量中加入Position Embedding也只是一个权宜之计，并没有改变Transformer结构上的固有缺陷。
在DERT中检测小目标的性能较低，在ViT中当预训练数据集较小时，性能也不是很好。

五、参考文献及更多

[1]A Survey on Visual Transformer
[2]Transformers in Vision: A Survey
[3]Dosovitskiy, Alexey, et al. ”An image is worth 16x16 words: Transformers for image recognition at scale.” arXiv preprint arXiv:2010.11929 (2020).