ViT:2020

AN IMAGE IS WORTH 16×16 WORDS: TRANSFORMER FOR IMAGE RECOGNITION AT SCALE

论文背景

transformer结构在NLP中使用颇为广泛，注意力机制在视觉任务中主要与卷积结合，或者替换卷积网络中的某些组成部分，同时保持卷积神经网络的整体结构不变。

之前的尝试大都受限于自注意力SA模块的计算复杂度与输入成二次方，单像素输入会使得计算消耗很大。

切分Patch的方法之前也有人提出过，但是由于切分的patch过小，只能运用在小分辨率图像上

论文创新

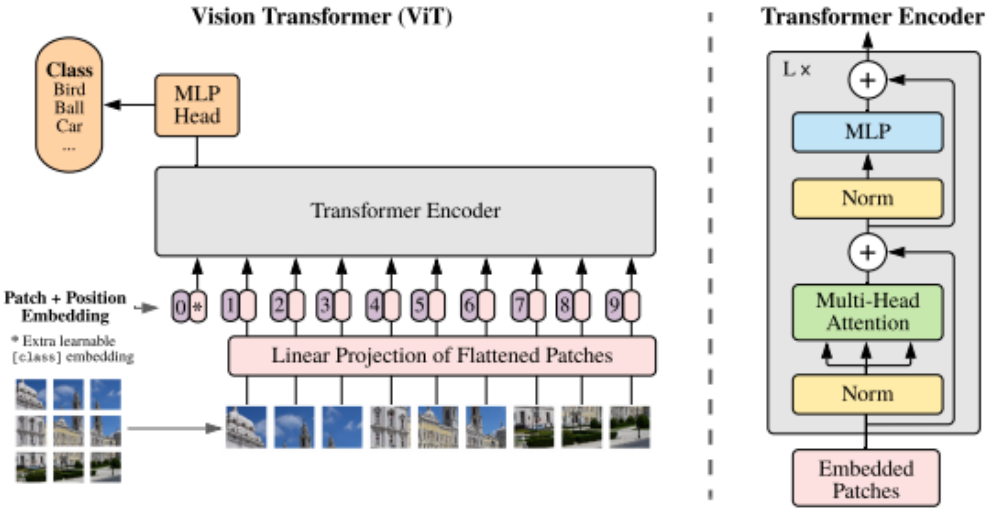
在本文中我们证明了在视觉任务中卷积并不是必须的，并实现了一个纯注意力的网络可以很好执行图像分类任务，并在大量数据集下训练之后迁移到中型和小型数据集下。

切分patch：

> An overview of the model is depicted in Figure 1. The standard Transformer receives as input a 1D sequence of token embeddings. To handle 2D images, we reshape the image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer uses constant latent vector size D through all of its layers, so we flatten the patches and map to D dimensions with a trainable linear projection (Eq. 1). We refer to the output of this projection as the patch embeddings.

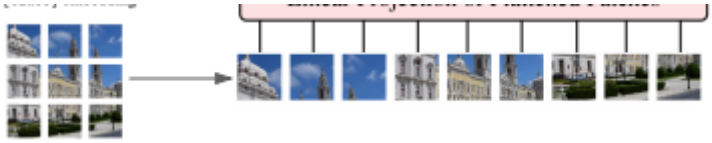
论文方法

对transformer原结构做了最小的改动

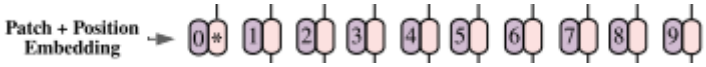


切分patch：

$H \times W \times C \longrightarrow N \times (P^2 \cdot C)$



借助BERT的class token的方法，我们设置了一个可学习的embedding嵌入编码后的patches 序列中



位置编码：

standard learnable 1Dposition embeddings

模型通过位置嵌入的相似性来学习对图像内的距离进行编码

没有观察到使用二维编码效果更好

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}}; \mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \cdots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{pos}, \qquad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D} \tag{1}$$

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, \qquad \ell = 1 \ldots L \tag{2}$$

$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\mathbf{y} = \text{LN}(\mathbf{z}_L^0) \tag{4}$$

输入序列可以由CNN的特征图形成

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

自注意力机制能够整合整个图像的信息，我们通过根据注意力权重计算信息在整合过的图像空间中的平均距离，这种距离类似CNN中的接受野大小

实际效果

在小型数据集和中型数据集上的表现略差于卷积网络，认为是transformer缺少CNN中的归纳偏置

CNN中具有很强的先验假设，比如局部性、平移不变性等，但是在ViT中只有MLP层具有这些特性。而且由于切分Patch之后，Patch之间的所有空间关系必须从头开始学习。

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

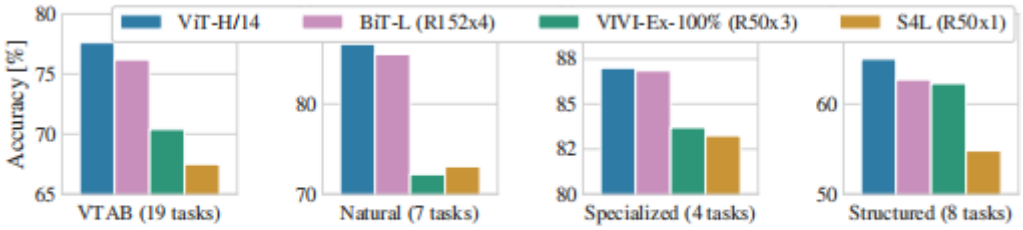
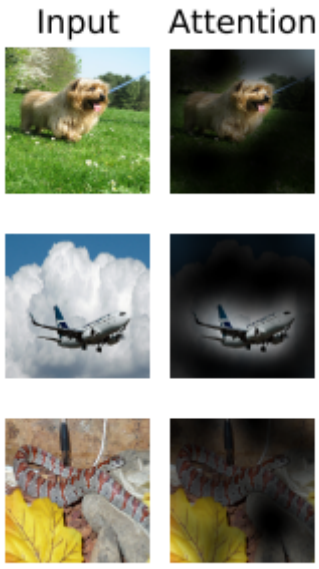


Figure 2: Breakdown of VTAB performance in *Natural*, *Specialized*, and *Structured* task groups.



个人见解

1.transformer在视觉领域中第一次取得比肩CNN的效果

2.transformer归纳偏置小，主要是由于transformer缺少CNN中类似的先验知识，所以需要吃更多的数据去学习这些先验知识。

3.SA的计算二次复杂度依旧是一个待解决的问题

4.ViT的直筒型结构不能很好应用于下游任务当中