# 背景

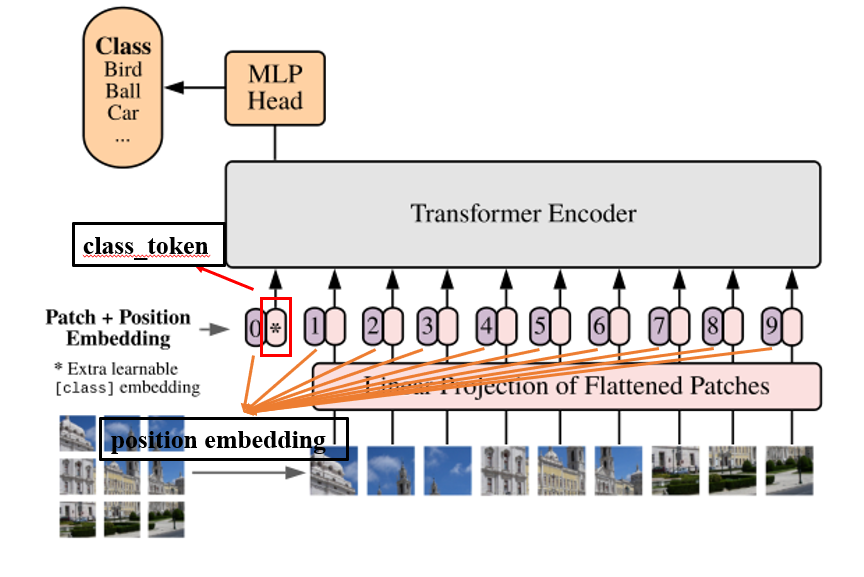
Transformer模型在NLP领域大火以后，受此影响，在计算机视觉领域中出现了许多关于将Transformer和卷积相结合的模型，是文中提到的Hybrid模型，有研究发现将CNN和Transformer混合起来可以提高精度，然而，又有研究证明在图像分类任务中纯Transformer模型在计算量和模型尺寸相同的前提下，也可以达到和纯卷积相同的准确率。

# 方法

## 整体结构

网络的整体结构如下图所示：

图 1

首先对输入图像进行卷积操作，使用的卷积核尺寸为16×16，步长为16，也就是将原图16×16个像素打包，通过一个映射层得到一个个的patch，然后和可学习的类别向量(class\_token)进行concat拼接，再和位置映射向量相加，接着输入到Transformer Encoder中，最后接一个MLP Head结构用于最终的分类任务。

## Embedding

由于标准的Transformer结构的输入是向量序列，即一个二维数据，但是图像的尺寸都是三维的，[C, H, W]，因此首先要考虑的就是输入数据的形状。

Embedding层首先通过卷积操作将一张图像划分为N个patches，每个patch的大小为[P, P]，，此时，patch的尺寸为[P2, C]，再通过一个线性映射将每个patch映射到一维向量中，即将[P2, C]→[D]。

## Class embedding & Position embedding

将刚才得到的二维token拼接一个class token，这个类别向量是可以学习的。并且在后续的内容中，通过Transformer Encoder后只需要进行一个切片操作就可以将类别信息单独提取出来，进行最后的分类任务。

Position embedding就是Transformer里提到的位置映射。

关于position embedding起到的作用，文中也做了实验。

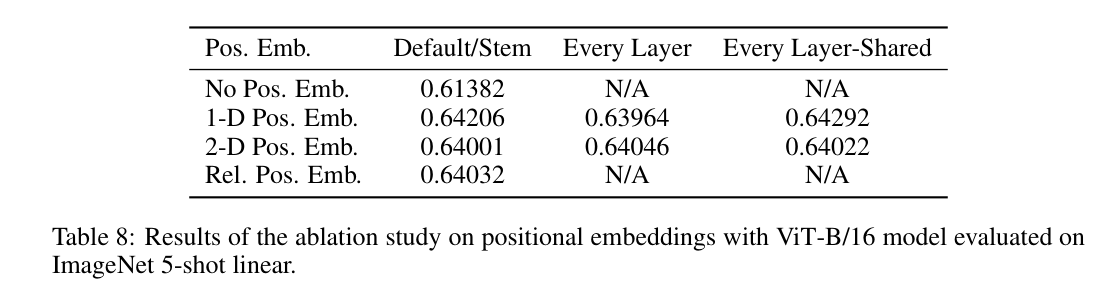


表 1

从表1中可以看出，不使用位置编码和使用位置编码，在最后结果准确率上差别还是很大的。不过，使用一维位置编码、二维位置编码、相对位置编码的差别不大，这是因为Transformer Encoder是对patch进行操作的，在patch中所含的空间信息很少。文中使用的是最简单的一维位置编码。

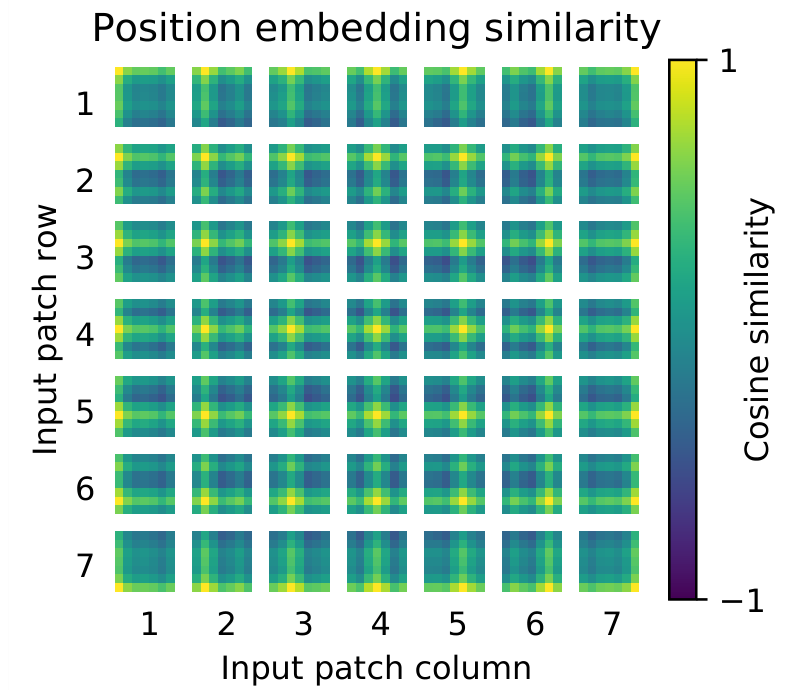


图 2

图2是将ViT-L/32的每一个patch的位置编码与其他的patch的位置编码进行了余弦相似度计算，比较亮的区域就代表着两者位置编码相似度高。从图2中可以看出，第一行第一列patch的位置编码与自己的位置编码相似度最大，位置相近的patch，其位置编码相似度也比较高。

## Transformer Encoder

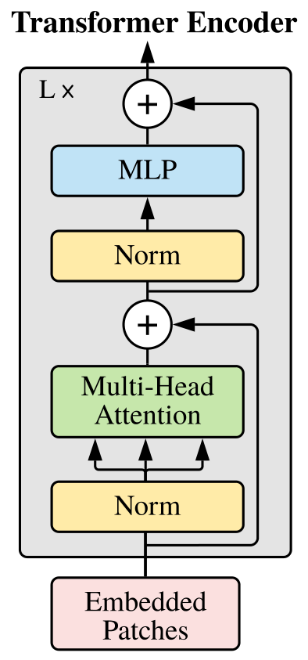


图 3

从图3中可以看出，Transformer Encoder模块可以分为两个部分，第一个部分是由一个Layer Normal层，一个多头注意力模块，最后添加了跳跃连接；第二个部分是由一个Layer Normal层，一个MLP模块，也进行了跳跃连接。然后将这整体结构重复L次。其中，MLP模块是由两个全连接层，以及一个GELU激活函数构成的。

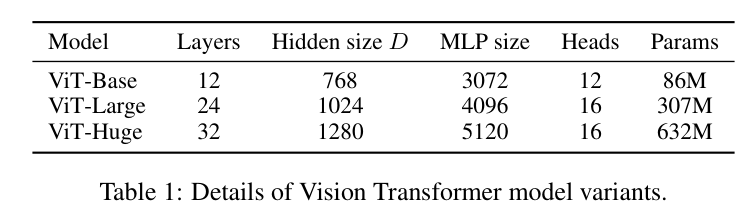


表 2

从表2中Layer参数指的是Transformer Encoder重复堆叠的次数。Hidden size D即在Embedding层将输入图像映射为token后，每个token的维度。MLP Size指的是Transformer Encoder中MLP模块第一个全连接层输出的维度，可以看出，MLP模块第一个全连接层将输入的向量的维度扩展了4倍。

## MLP Head

将通过Transformer Encoder后的向量的第一个维度信息以切片的方式提取出来，得到的即使class token，将class token输入到MLP Head中实现最后的分类任务，也就是下式实现的功能。



关于[class]token的提取，这里尝试过使用全局平均池化操作，其对比分析图如下所示：

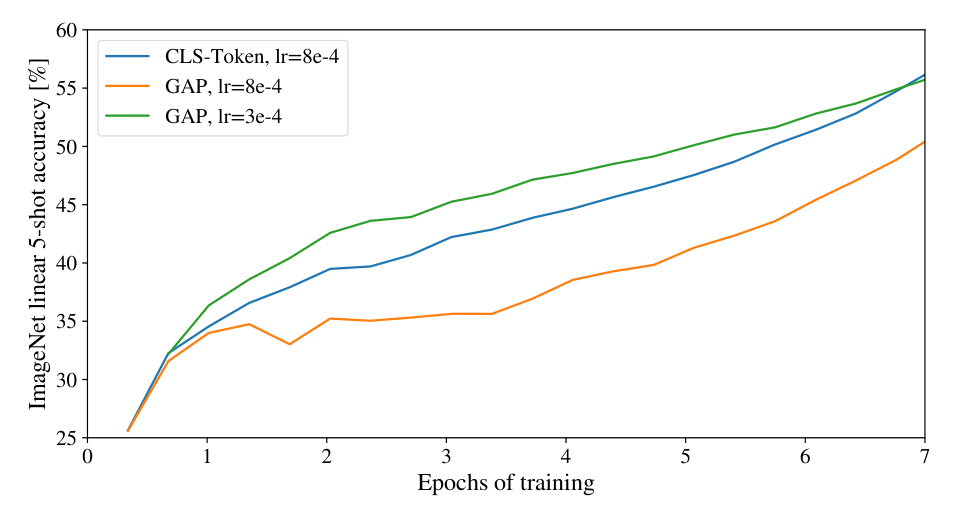


图 4

从图中可以看出，使用相同的学习率，CLS-Token的性能比使用GAP的性能要好很多，但是调整了学习率后可以看出使用GAP和CLS-Token的性能差不多。两者的性能差别在于学习率的不同。

# 实验

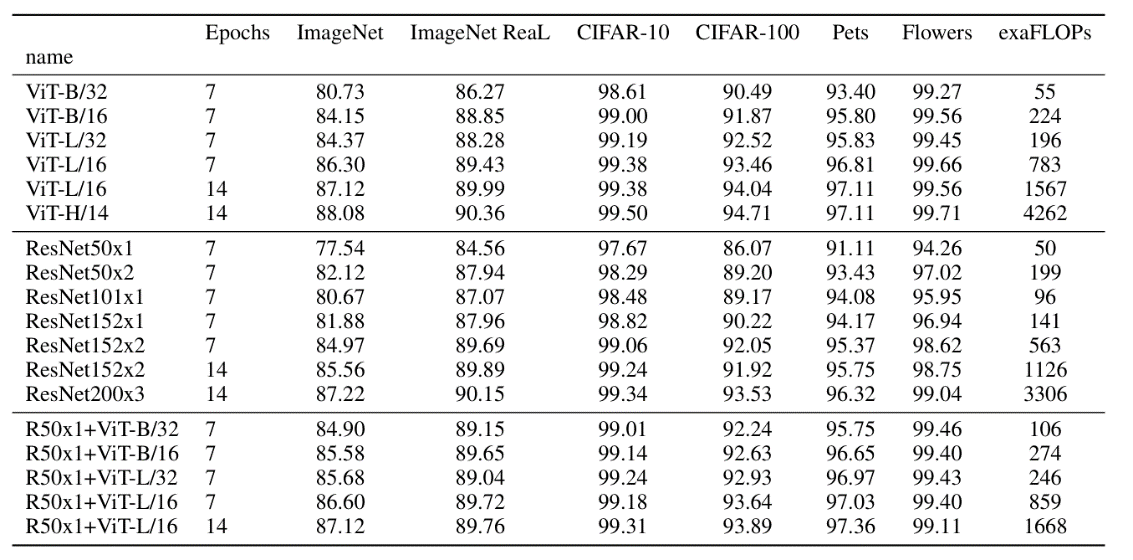


图 5

回到背景中提到的Hybrid和纯Transformer对比，从实验结果中可以看出，在训练轮数相同，使用ResNet50进行基础特征提取再结合ViT-L/16和纯ViT-L/16相比（图中红色框），Hybrid模型效果更好一些，但是在增加了训练轮数（图中蓝色框）后，纯Transformer模型的性能优于Hybrid模型，这就证明了在计算机视觉领域中使用纯Transformer模型其性能完全可以达到CNN模型的效果。