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：

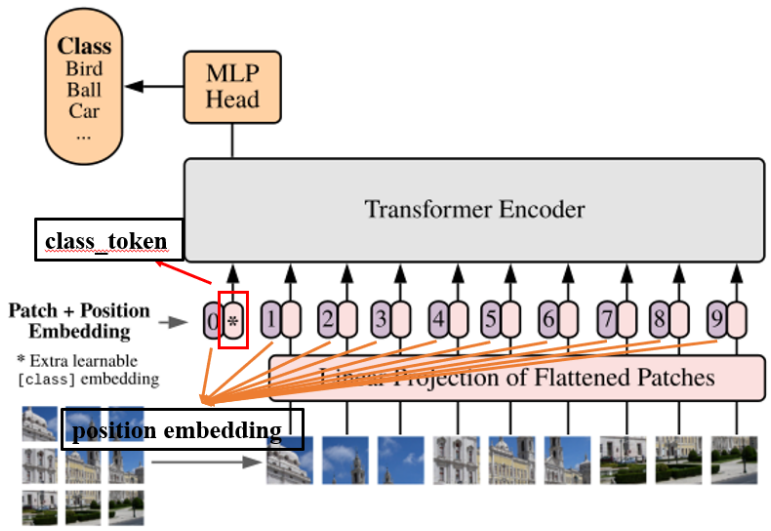

图 1

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]，$N = \frac{H \times W}{P^2}$，此时，patch 的尺寸为[$P^2$, C]，再通过一个线性映射将每个 patch 映射到一维向量中，即将[P^2, 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

Pos. Emb.	Default/Stem	Every Layer	Every Layer-Shared
No Pos. Emb.	0.61382	N/A	N/A
1-D Pos. Emb.	0.64206	0.63964	0.64292
2-D Pos. Emb.	0.64001	0.64046	0.64022
Rel. Pos. Emb.	0.64032	N/A	N/A

Table 8: Results of the ablation study on positional embeddings with ViT-B/16 model evaluated on ImageNet 5-shot linear.

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。

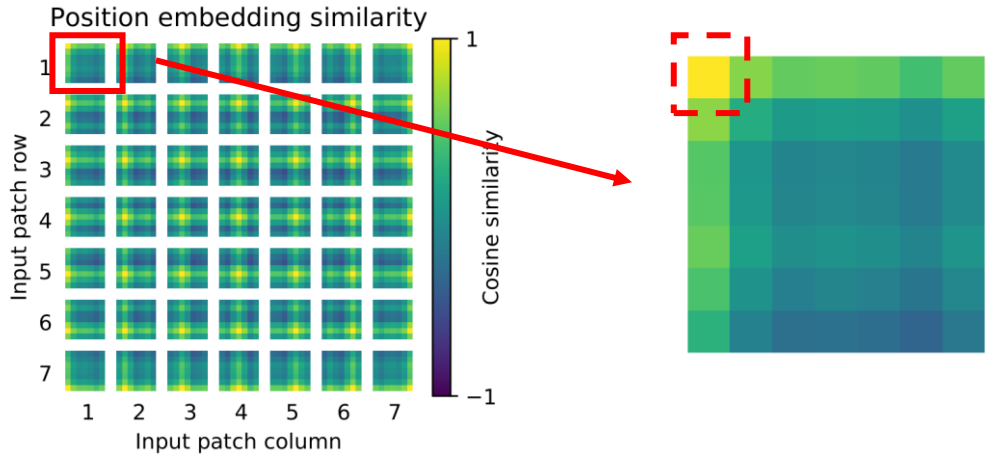

图 2

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

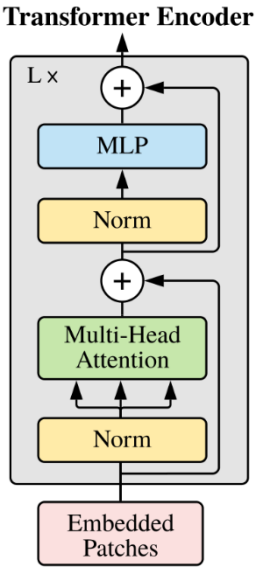

图 3

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

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.

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。

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0)$$

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：

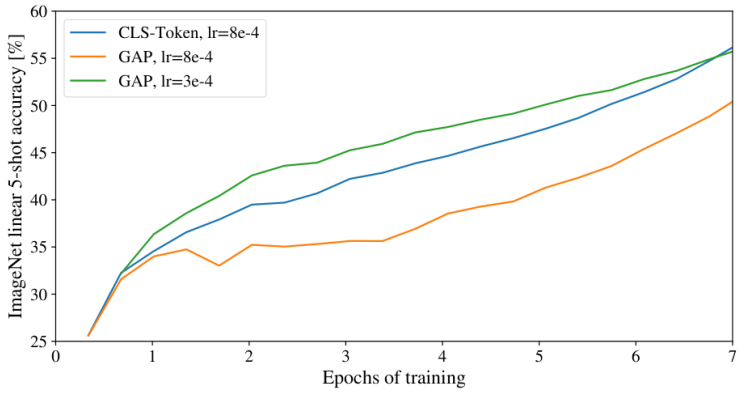

图 4

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

name	Epochs	ImageNet	ImageNet ReaL	CIFAR-10	CIFAR-100	Pets	Flowers	exaFLOPs
ViT-B/32	7	80.73	86.27	98.61	90.49	93.40	99.27	55
ViT-B/16	7	84.15	88.85	99.00	91.87	95.80	99.56	224
ViT-L/32	7	84.37	88.28	99.19	92.52	95.83	99.45	196
ViT-L/16	7	86.30	89.43	99.38	93.46	96.81	99.66	783
ViT-L/16	14	87.12	89.99	99.38	94.04	97.11	99.56	1567
ViT-H/14	14	88.08	90.36	99.50	94.71	97.11	99.71	4262
ResNet50x1	7	77.54	84.56	97.67	86.07	91.11	94.26	50
ResNet50x2	7	82.12	87.94	98.29	89.20	93.43	97.02	199
ResNet101x1	7	80.67	87.07	98.48	89.17	94.08	95.95	96
ResNet152x1	7	81.88	87.96	98.82	90.22	94.17	96.94	141
ResNet152x2	7	84.97	89.69	99.06	92.05	95.37	98.62	563
ResNet152x2	14	85.56	89.89	99.24	91.92	95.75	98.75	1126
ResNet200x3	14	87.22	90.15	99.34	93.53	96.32	99.04	3306
R50x1+ViT-B/32	7	84.90	89.15	99.01	92.24	95.75	99.46	106
R50x1+ViT-B/16	7	85.58	89.65	99.14	92.63	96.65	99.40	274
R50x1+ViT-L/32	7	85.68	89.04	99.24	92.93	96.97	99.43	246
R50x1+ViT-L/16	7	86.60	89.72	99.18	93.64	97.03	99.40	859
R50x1+ViT-L/16	14	87.12	89.76	99.31	93.89	97.36	99.11	1668

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。