VIT

1标题

2摘要

3引言

4结论

5相关工作

6模型

7实验

8评论

Inductive biases 归纳偏置 locality translation equivalent

在计算机视觉中，卷积架构仍然占主导地位,受到NLP成功的启发，许多作品尝试将CNN类架构与自我注意结合，一些完全取代卷积。后一种模型虽然理论上是有效的，但由于使用了专门的注意力模式，在现代硬件加速器上还没有得到有效的扩展。

为了将Transformer应用于图像处理，过去已经尝试了几种近似方法。Parmar等人对每个查询像素只应用局部邻域的自我关注，而不是全局的。这种局部多头点积自我注意块可以完全替代卷积。在另一项工作中，稀疏Transformer采用了可扩展的近似的全局自注意力机制，以便适用于图像。

受NLP中Transformer缩放成功的启发，我们尝试将标准Transformer直接应用于图像，并尽可能减少修改。为此，我们将图像分割为多个小块，并将这些小块的线性嵌入序列作为Transformer的输入。图像补丁的处理方式与NLP应用程序中的标记(单词)相同。我们以监督的方式训练模型进行图像分类。

然而，如果模型在更大的数据集(14M-300M图像)上训练，图像就会发生变化。我们发现大规模训练战胜了归纳偏 置。我们的视觉转换器(Vision Transformer, ViT)在足够规模的预训练和转移到具有更少数据点的任务时取得了优异的结果。



