AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Abstract

Transformer在NLP任务上得到了很优秀的效果，然而视觉任务上依然使用CNN等技术。本文试图使用了Transformer架构应用于视觉任务上，并且取得了非常好的效果。

Conclusion

不同于以往的工作，本文没有引入传统的归纳偏置，相反，我们使用了将图像分割为patch作为序列输入的思想，直接应用Transformer而尽量不做改动。结果证明效果非常好。CNN老登爆金币啦！

Introduction

在NLP任务上，Transformer作为一颗冉冉升起的新星，其优越的拓展性和支持预训练等特性使其占据了统治地位。然而，在CV任务上还是CNN主导。虽然也有人曾试图将CNN部份甚至全部替换为self-attention，但是训练结果始终差强人意。本文另辟蹊径，直接应用了完整的Transformer替换架构而不是

仅仅替代CNN的一部份。实验证明，对于视觉任务，在小型数据集上使用Transformer没有CNN那么优秀，然而在中大型数据集上Transformer效果超越了CNN，展现其规模性和拓展性。

Related Work

先前曾经有以下试图应用Transformer的工作：

1）仅查询图像的局部，不使用全局注意力；应用muti-head attention取代CNN

2）全局可拓展注意力（我也没看懂）

3）将图像分割为2x2的块，这个是和本文方法最相近的

Method

本文不同于上述的将图像切割为2x2 patch，而是切成了16x16 patch（个人理解是减少了序列长度，从而减少计算量，训练更快）。同时，为了表示出每个patch在原图像的位置，也是为了和Transformer中的做法相一致，ViT使用了图像的2d位置编码（还没细看，有空了回来补充），虽然实验人员惊讶地发现2d位置编码和1d位置编码好像效果差不多（我也没搞懂，懒，以后回来补充）。最后值得一提的是，在每一张图像patch构成的序列首部，ViT还模仿BERT加了个cls token，作用类似于情感分析任务中的标签，ViT认为cls token会聚合图

像patch序列的特征，并且作为该图像的标签。同时，作者发现由于Transformer有全局注意作用，因此不太用考虑不同分块之间的归纳偏置了。

最近时间很紧，等我看完Coco再回来补充具体的网络和实验啦～