Paper Reading Report

|  |  |
| --- | --- |
| Paper Title | **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale** |
| Conference/  Journal Name | **International Conference on Learning Representations 2021** |
| Authors &  Affiliation | **Alexey Dosovitskiy, Lucas Beyer, Alexander A. Kolesnikov and more**  **Google Brain Team** |
| Year | **2021** |
| Citations | **2957** |
| Q&A | **Q1: P5 为什么不用Decoder**  A1: 图像分类任务不用像标准Transformer一样使用Decoder对每个词进行预测，仅需一个分类头就可以完成分类任务  **Q2: 512是什么意思**  A2: 512是标准Transformer的输入序列中每一个词向量的长度  **Q3: 计算复杂度是平方级，可不可以下降？**  A3: 可以，如Swing Transformer的计算复杂度是线性增长的  **Q4: 训练过程长短？实验对比？作者提出改进？**  A4: 在大规模数据集上训练时，ViT的表现与训练成本要优于CNN，作者未提出改进，但是认为ViT能够应用于图像处理的其他方向 |