|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **文章信息** | **背景、目的及结论** | **结果与讨论** | **文章好在哪里** | **自我想法** |
| **作者**：Ashish Vaswani，Noam Shazeer  **单位**：Google  **来源**：arXiv  **题目**：Attention is all you need | **背景**：循序计算(RNN)不能并行化，而CNN又在两个token相距较远时需要多层卷积才能联系起来。而自注意力层通常又和RNN结合起来使用（难以并发）  **目的**：充分利用GPU，利用自注意力解决RNN不能并发，CNN需要多层卷积联系较远点的问题。  **结论**：基于纯注意力机制，抛弃了卷积和循环结构，使用常见的Encoder-Decoder结构为模型主体。并发性更好吗，且使用固定操作数来进行运算（CNN的缺陷） | 在WMT 2014数据集上，英德互译任务上的BLEU比之前的SOTA提升了两个点，最终分数为28.4.英法翻译任务中，大型模型达到了41.0的分数，比之前的所有模型表现都要好，而且训练cost不到之前SOTA模型的1/4。  模型在训练过程中表现出了对语法结构和句子框架相关的行为。可以将其运用到翻译之外的任务(音频，图像和视频)。 | 1. 利用常见的Encoder-Decoder结构使用纯注意力机制代替了之前的卷积，循环结构，从而使训练cost以及翻译任务上面的表现改善了许多。 2. 提出了Transformer应用于其他领域的展望。 | **一个思路**：  ViT中的将图片分割为多个Patch作为token送入Transformer提取特征，文本送入完整的Transformer提取特征，使用相似度函数衡量二者差距来进行分类。  **两个图表**： |