背景

deepseek v3 使用了transformer架构的moe模型，其中一个环节，attention环节，使用了Multi-head latent attention技术，对attention进行压缩，压缩后和其他gpu进行communication，减小了communication的带宽，其中位置编码使用的是rotary position encoding, https://arxiv.org/pdf/2104.09864, 对相对位置编码的效果最好。

project内容

选择第二个benchmark paper, 老师今天课堂最后说了这个benchmark paper目的是，有些论文实验是没有公开的，只有结果，所以你去复现他，然后在更多数据集上复现，另外可以进行方法优化。

我可以对rotary position encoding这篇论文进行复现，横向对比其他方法，另外这篇论文忽略了一个实验，就是在相同参数量，相同计算量，相同数据精度（比如fp16, fp32）的情况下，对比其他position encoding方法是否效果最好。

实验也可以升级，论文中实验是对基础的transformer模型，改变其中的position encoding,其他不变，进行downstream 的不同任务（具体什么任务没说）进行比较，我们可以复现。另外加一种任务，就是还原相对位置的accuracy度量。

特点

我提出的project,是需要学习transformer架构，探讨其中一些数学问题的，不属于拿来即用的应用层面，但是方便在，如果有对架构的了解（deep learning这节课后面会学transformer,我觉得任何和ai相关的工作者，最起码得先了解一下基本的transformer架构），我们不需要对应用层面，比如医学，生物学，这些背景知识，花大量时间去了解，上手即可做实验。

学习

1. transformer： https://d2l.ai/chapter_attention-mechanisms-and-transformers/transformer.html, B站`李沐`有对应的视频教学