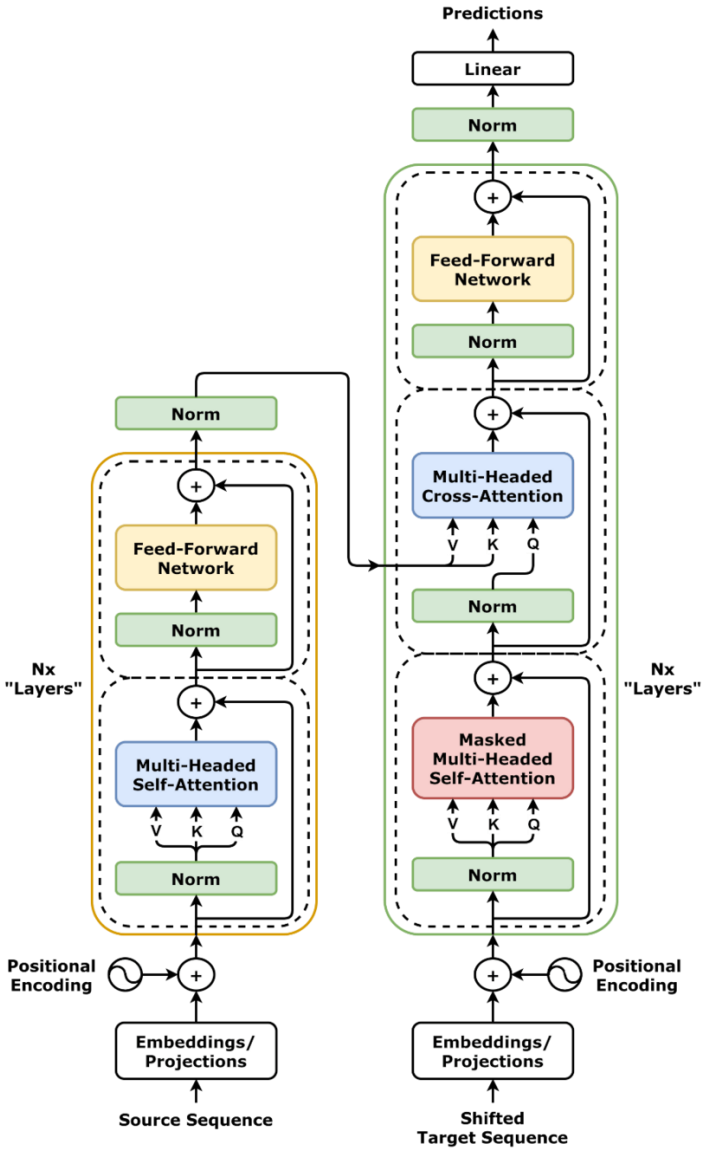
寻找文献

Exploring_KV_Cache_Quantization_in_Multimodal_Large_Language_Model_Inference

这篇文献与 refrag 相同，优化 prefilll 到 decode 过程，不同点在于优化内容为 KV Cache。通过压缩-重载 KV Cache 获得内存优化与时间优化。



Nested Learning: The Illusion of Deep Learning Architectures

谷歌的论文提出了 HOPE 框架，其中包含连续记忆系统，与自我更新的算法，使得训练的模型可以真正的学习与保存记忆。

测试公式

$$e^{i\pi} + 1 = 0$$