

Kimi

背景

在开发 Kimi 的过程中，我们遇到了一个挑战：如何高效地处理长上下文。传统的 Transformer 模型在处理长上下文时，由于注意力机制的计算复杂度随序列长度平方增长，导致性能下降。为了解决这个问题，我们引入了 **LongRoPE** 技术，它通过插值和 extrapolation 来扩展模型的上下文长度。

原理

LongRoPE 的核心思想是利用插值和 extrapolation 来扩展模型的上下文长度。具体来说，我们将输入序列分为两部分：已知部分和未知部分。已知部分的长度在 $[0, 2048)$ 范围内，而未知部分的长度在 $[2048, 4096)$ 范围内。对于未知部分，我们使用 extrapolation 技术来生成新的 token，从而使得模型能够处理更长的序列。

LongRoPE 的实现依赖于 Transformer-XL 的 LongRoPE 技术，它通过插值和 extrapolation 来扩展模型的上下文长度。ReRoPE 是一种改进的 LongRoPE 实现，它通过更精细的插值来进一步提高模型的长上下文处理能力。

实现

LongRoPE 的实现基于 First Principles，即从基本原理出发。我们首先定义了一个 **LongRoPE** 函数，该函数接收输入序列并返回处理后的序列。该函数内部实现了对输入序列的插值和 extrapolation 操作。

LongRoPE 的实现依赖于以下三个步骤：

1. 将输入序列分割为已知部分和未知部分。
2. 对已知部分进行插值操作，生成新的 token。
3. 对未知部分进行 extrapolation 操作，生成新的 token。
4. 将已知部分和未知部分拼接起来，得到最终的输出序列。

