【手撕LLM】长文本的Position Encoding的衰减性证明

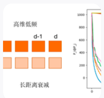
小冬瓜AIGC
原创课程➡️ 公众号：手撕LLM

来自专栏 · 手撕LLM

> 我是小冬瓜AIGC，原创超长文知识分享，原创课程已帮助多名同学速成上岸LLM赛道。
> 研究方向：LLM、RL、RLHF

1. Preview

在过去我们分析了Attention建模和位置编码的衰减特性，我们讨论了在 $64k$ 长度下的衰减，并出现大范围的震荡。

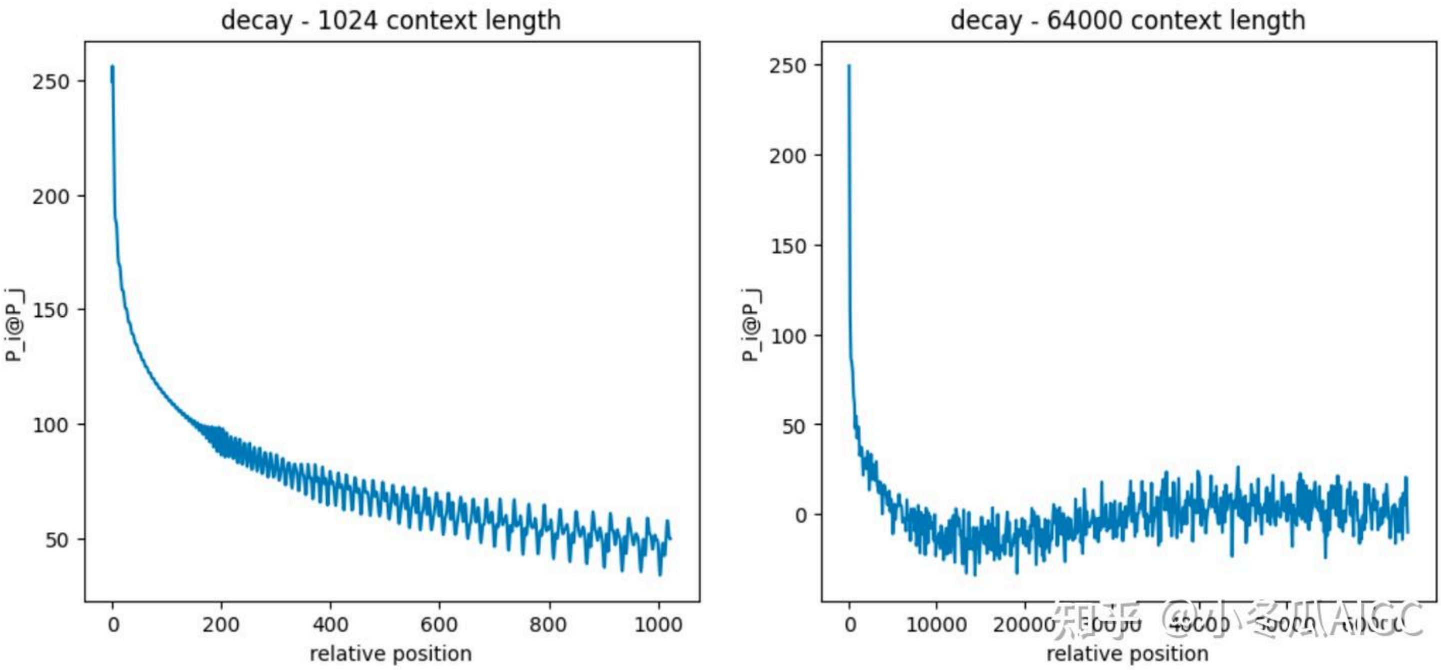


小冬瓜AIGC：【手撕LLM-NTK RoPE】长文本“高频外推、低频内插“从衰减性视...
114 赞同 · 12 评论 文章

如果我们仔细观察会发现, 位置编码之间内积会趋近于0, 而且由于震荡，内积也可能小于0

引出新的疑问：

1. 如何证明位置编码会衰减?
2. 为什么内积会震荡和小于0?
3. 无限长度下的位置编码的内积是什么?



2. 位置编码衰减性证明

2.1 问题定义

位置编码p_i 定义如下：

$$p_i = (\sin(\frac{i}{\frac{0}{d}}), \cos(\frac{i}{b^{\frac{0}{d}}}), \ldots, \sin(\frac{i}{b^{\frac{2l}{d}}}), \cos(\frac{i}{b^{\frac{2l}{d}}}))^T, l = d/2 - 1$$

其中, b为底数(base)

记$\theta_t = b^{-\frac{2t}{d}}$， $t = \{0, 1, \cdots, l\}$，则位置编码可表示为:

$$p_i = (\sin(i\theta_0), \cos(i\theta_0), \ldots, \sin(i\theta_l), \cos(i\theta_l))^T$$

给定两个位置i, j 的位置编码可以计算内积，可以得到相对距离系数$f_n, n = j - i$

$$\begin{aligned} f_n &= p_j^T p_i \\ &= \sum_{t=0}^{l} \sin(j\theta_t)\sin(i\theta_t) + \cos(j\theta_t)\cos(i\theta_t) \\ &= \sum_{t=0}^{l} \cos(j\theta_t - i\theta_t) \\ &= \sum_{t=0}^{l} \cos((j-i)(\theta_t)) = \sum_{t=0}^{l} \cos(n\theta_t) \end{aligned}$$

根据相对距离的衰减性，我们需要求证$f_n \to 0$, 当$n \to \infty$

2.2 数学证明

给定区间

$$\beta = \theta_0 > \ldots > \theta_l = \alpha$$

这里的$d \to \infty$时，我们将离散求和的形式逼近积分形式

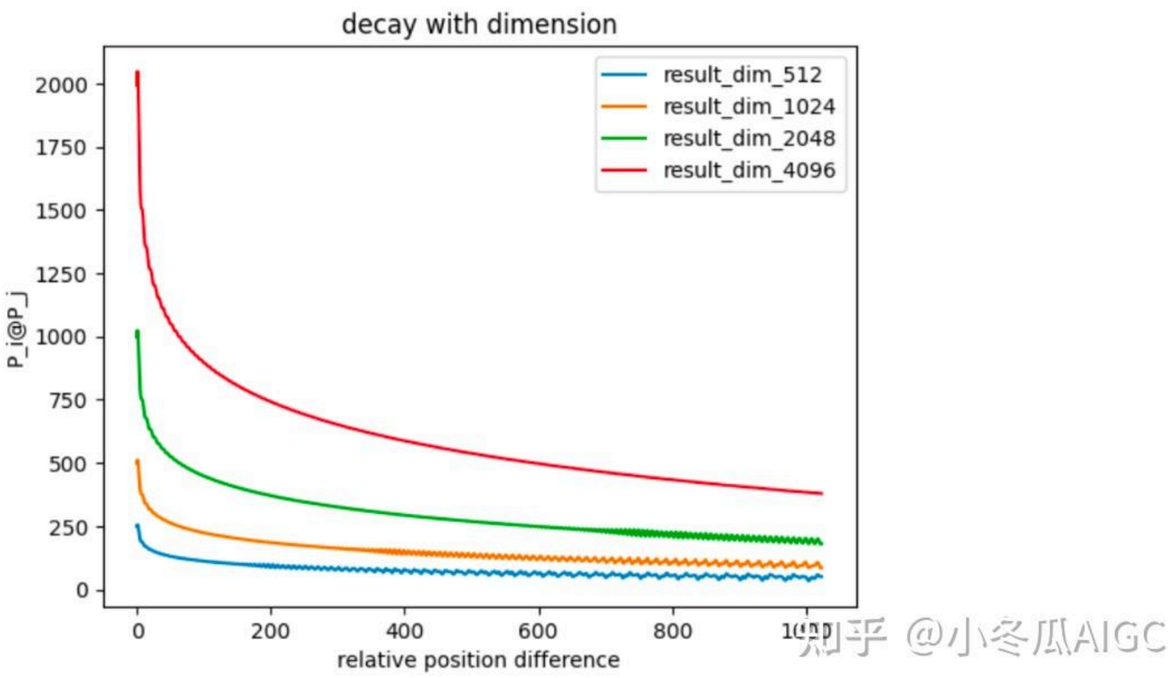
$$\begin{aligned} f_n &= \sum_{t=0}^{l} \cos(n\theta_t) \\ &= \int_{\alpha}^{\beta} \cos(nx)dx \end{aligned}$$

计算积分:

$$\begin{aligned} |f_n| &= |\int_{\alpha}^{\beta} \cos(nx)dx| \\ &= \sum_{t=0}^{l} |\int_{\theta_t}^{\theta_{t+1}} \cos(nx)dx| \\ &= \sum_{t=0}^{l} |[\frac{1}{n}\sin(nx)]_{\theta_t}^{\theta_{t+1}}| \\ &= \sum_{t=0}^{l} \frac{1}{n}|(\sin(n\theta_{t+1}) - sin(n\theta_t))| \\ &\le \sum_{t=0}^{l} \frac{1}{n} \cdot 2 \to 0 \end{aligned}$$

当$n \to \infty$ 时, $f_n \to 0$ 成立，证毕。

根据以上推导结论，我们可以观察到, 位置编码衰减曲线呈$y = \frac{l}{n}$ 形状



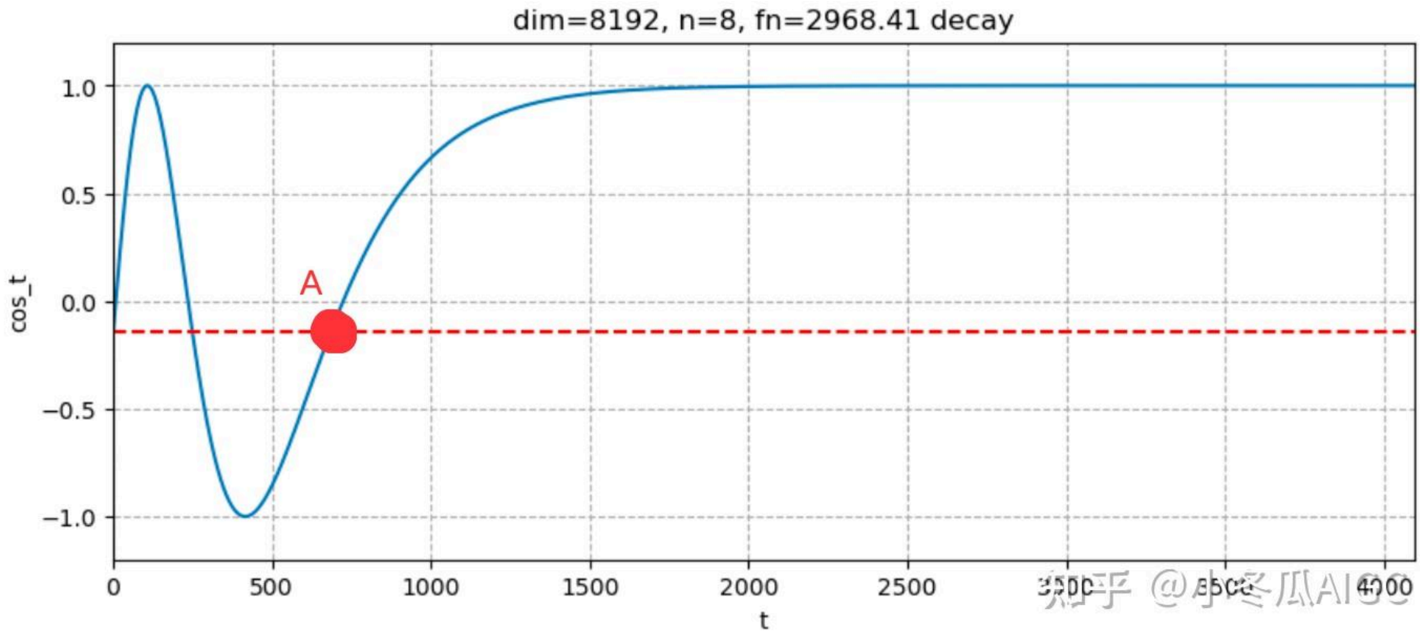
3. 代码验证

我们根据内积公式，编写以下代码

$f_n = \sum_{t=0}^{l} \cos(n\theta_t)$

```
def get_fn(n=1024, d=512, base=10000):
    t = torch.arange(0, d, step=2).float()
    theta_t = base ** (-t/d)
    cos_t = torch.cos(n * theta_t )
    fn = torch.sum(cos_t)
    return fn, cos_t
```

对应绘制图像, 我们以 `y=cos_t[0]` 绘制一条红色横向虚线, 在下图 `[0,A]` 处位置内的积分为 `0` , 此时自 `[A,4096]` 后的曲线与 `y=0` 所围成的面积即是，位置编码的内积



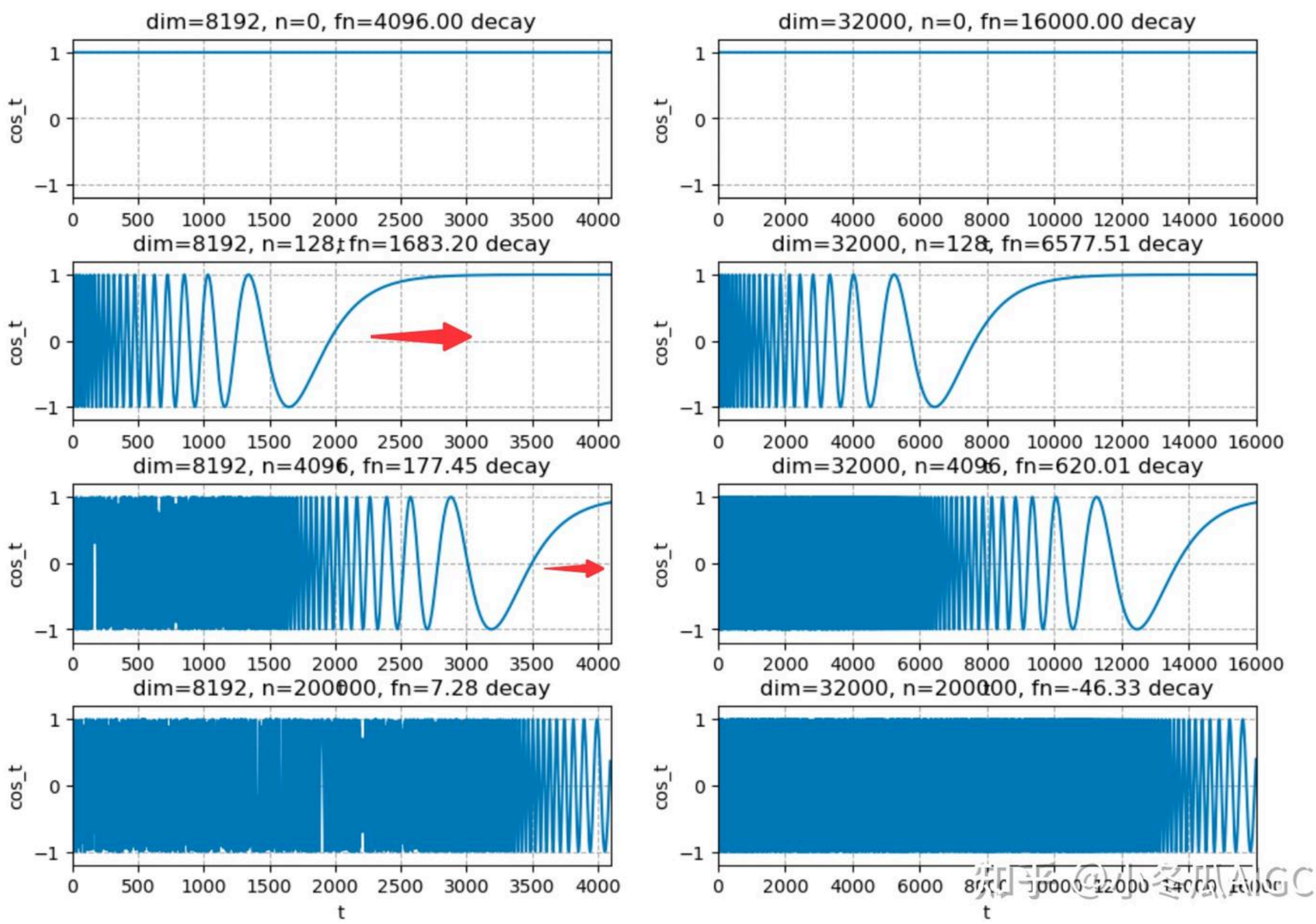
我们测试更长的位置编码距离, 配置如下

```
n_list = [0, 128, 4096, 200000]
d = 8192
x=range(d//2)
fn_list, fn_v_list = get_batch(n_list, d)
```

实验说明

- 在相对距离为0时，所有的 `cos_t` 的值都为1, 此时是最大内积
- 随着相对距离增加，左侧高频跳动更大，计算积分时正负抵消，最后一个周期的出现越往右的维度移动
- 左右栏对比，维度越高, 采样更加逼近连续函数(所绘制图像不出现空缺)

随着 `n` 增加， `fn` 变小。



注意到最右下角出现, fn<0, 观察最后一段曲线，明显能看到小于0的区域更大。 位置编码的震荡也是由此引起。

4. 其他证明

4.1 Squeeze Rule

对于任意$a > b$, 当$n \to \infty$, 证$\int_a^b \cos(nx)dx \to 0$

直接求积分和

$\int_a^b \cos(nx)dx = \frac{1}{n}\sin(nb) - \frac{1}{n}\sin(na) \to 0$

使用squeeze rule

$$
\begin{aligned}
&-1 \leq \sin(nx) \leq 1 \\
\Leftrightarrow &\frac{-1}{n} \leq \frac{\sin(nx)}{n} \leq \frac{1}{n} \\
&\Leftrightarrow 0 \leq \frac{\sin(nx)}{n} \leq 0
\end{aligned}
$$

此时 当 $n \to \infty \Rightarrow \frac{1}{n}\sin(nx) = 0$，即 $\int_a^b \cos(nx)dx \to 0$ 证毕

4.2 Riemann-Lebesgue Lemma

推广到更一般的数学问题，可以追溯到Riemann-Lebesgue引理⁺, 感兴趣的朋友可以继续证明，受限篇幅不再过多展开，详见ref

定理1.1（Riemann-Lebesgue引理） 假设$f : \mathbb{R} \to \mathbb{R}$是一个Lebesgue可积函数。那么我们有：

$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(x)\cos(nx)\,dx = 0.$

定理2.2 假设f是在$[a, b]$上的黎曼可积函数。那么我们有：

$\lim_{n\to\infty} \int_a^b f(x)\cos(nx)\,dx = 0.$

5. 总结

- 本文受Riemann-Lebesgue引理证明启发，将证明技巧运用于位置编码衰减性证明，仅通过积分就可以完成推导，不引入复数和傅立叶变换等相关知识
- 我们通过图示cos函数, 高频由于积分正负抵消，可以仅计算最后一个周期的积分，同时也能够解答，位置编码为什么会震荡和存在负号
- 不得不感慨位置编码设计的如此巧妙，能从motivation里想到如此match的数学表达且有效

Reference

Transformer升级之路：2、博采众长的旋转式位置编码 - 科学空间|Scientific Spaces

math.cuhk.edu.hk/course...

prove the Riemann-Lebesgue lemma

---

*《手撕RLHF》*解析如何系统的来做LLM对齐工程

小冬瓜AIGC：【手撕RLHF-DPO】step-by-step公式推导及实验分析

小冬瓜AIGC：【手撕RLHF-Aligner】7B模型外挂，暴涨GPT4安全性26.9%

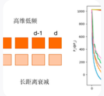
小冬瓜AIGC：【手撕RLHF_Weak-to-Strong】OpenAI超级对齐新思路（含代码解析）

小冬瓜AIGC：【手撕RLHF-Safe RLHF】带着脚镣跳舞的PPO

小冬瓜AIGC：【手撕RLHF-Rejection Sampling】如何优雅的从SFT过渡到PPO

小冬瓜AIGC：【手撕RLHF-LLaMA2】 Reward Model PyTorch实现

*《手撕LLM》*系列文章+原创课程：LLM原理涵盖Pretrained/PEFT/RLHF/高性能计算



小冬瓜AIGC：【手撕LLM-NTK RoPE】长文本“高频外推、低频内插“从衰减性视...
114 赞同 · 12 评论 文章

小冬瓜AIGC：【手撕LLM - Mixtral-8x7B】Pytorch 实现

小冬瓜AIGC：【手撕LLM-Medusa】并行解码范式: 美杜莎驾到, 通通闪开！！

小冬瓜AIGC：【手撕LLM-Speculative Decoding】大模型迈向"并行"解码时代

小冬瓜AIGC：【手撕LLM-FlashAttention2】只因For循环优化的太美

小冬瓜AIGC：【手撕LLM-FlashAttention】从softmax说起，保姆级超长文！！

小冬瓜AIGC:【手撕LLM-Generation】Top-K+重复性惩罚

小冬瓜AIGC：【手撕LLM-KVCache】显存刺客的前世今生--文末含代码

> 我是小冬瓜AIGC，原创超长文知识分享，原创课程已帮助多名同学速成上岸LLM赛道。
> 研究方向：LLM、RL、RLHF

编辑于 2024-08-28 14:59・中国香港

LLM 大模型 GPT