Base of RoPE Bounds Context Length

Xin Men*, Mingyu Xu*, Bingning Wang,
Hongyu Lin, Yaojie Lu, Xianpei Han
Weipeng Chen

NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

Position embedding is a core component of current Large Language Models (LLMs). Rotary position embedding (RoPE), a technique that encodes the position information with a rotation matrix, has been the de facto choice for position embedding in many LLMs, such as the Llama series. RoPE has been further utilized to extend long context capability, which is roughly based on adjusting the \textit{base} parameter of RoPE to mitigate out-of-distribution (OOD) problems in position embedding. However, in this paper, we find that LLMs may obtain a superficial long-context ability based on the OOD theory. We revisit the role of RoPE in LLMs and propose a novel property of long-term decay, we derive that the \textit{base of RoPE bounds context length}: there is an absolute lower bound for the base value to obtain certain context length capability. Our work reveals the relationship between context length and RoPE base both theoretically and empirically, which may shed light on future long context training.

Contributions

- Theoretical perspective: we derive a novel property of long-term decay in RoPE, indicating the model's ability to attend more to similar tokens than random tokens, which is a new perspective to study the long context capability of the LLMs.
- Lower Bound of RoPE's Base: to achieve the expected context length capability, we derive an absolute lower bound for RoPE's base according to our theory. In short, the base of RoPE bounds context length.
- •Superficial Capability: we reveal that if the RoPE's base is smaller than a lower bound,the model may obtain superficial long context capability, which can preserve low perplexity but lose the ability to retrieve information from long context.

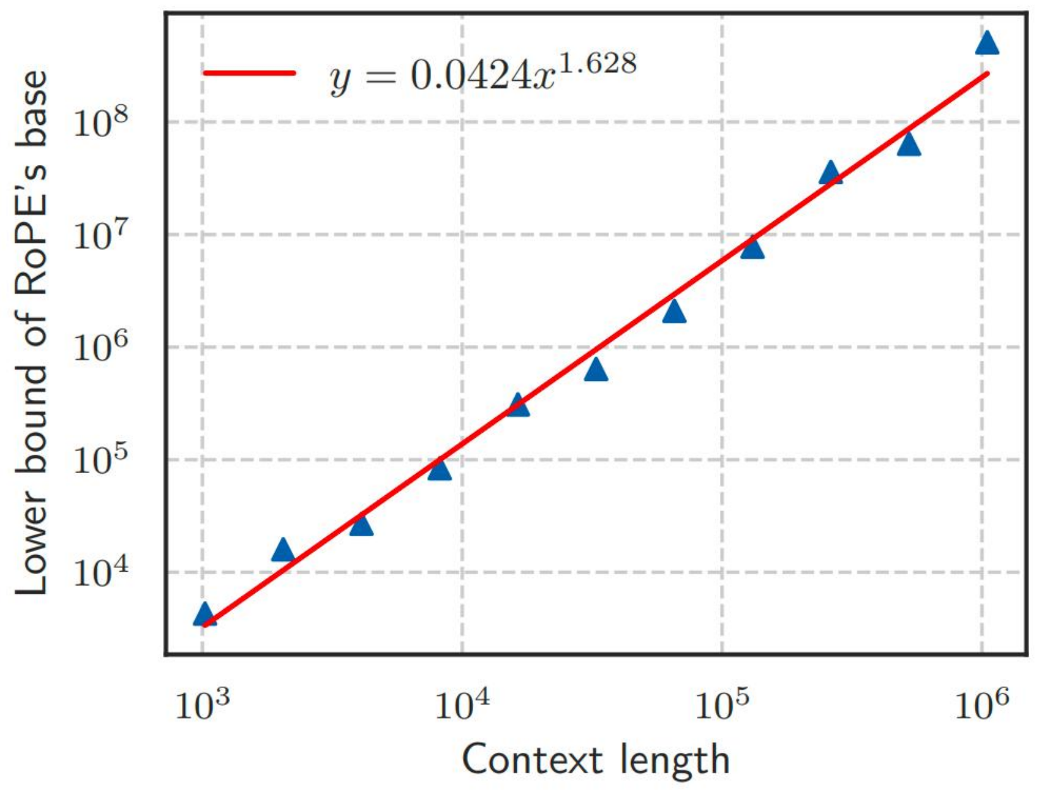
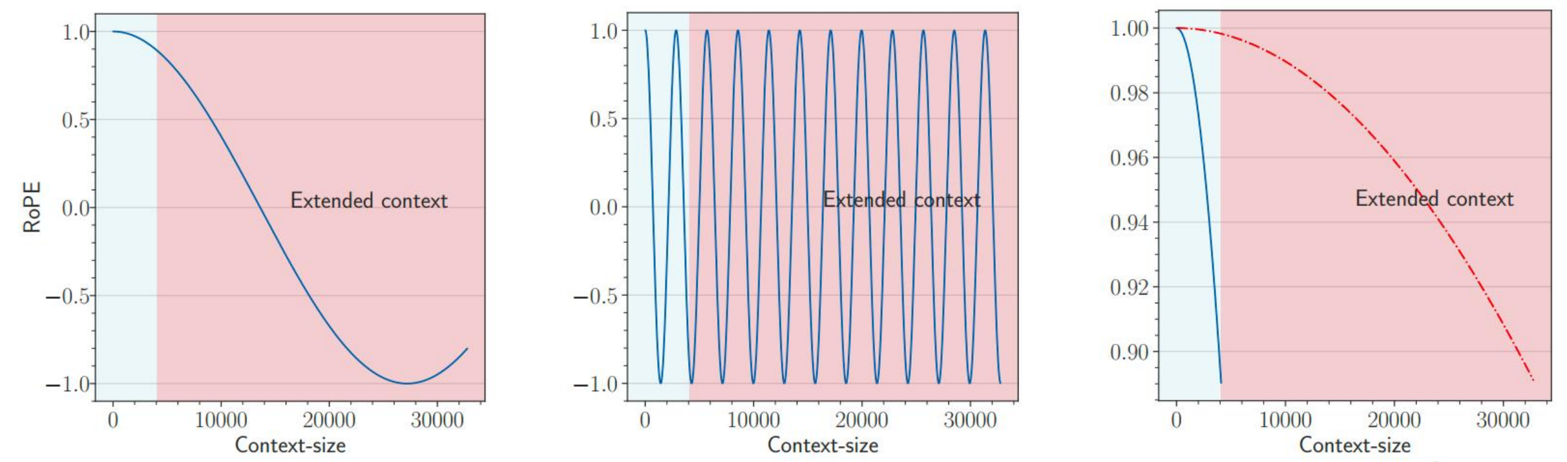


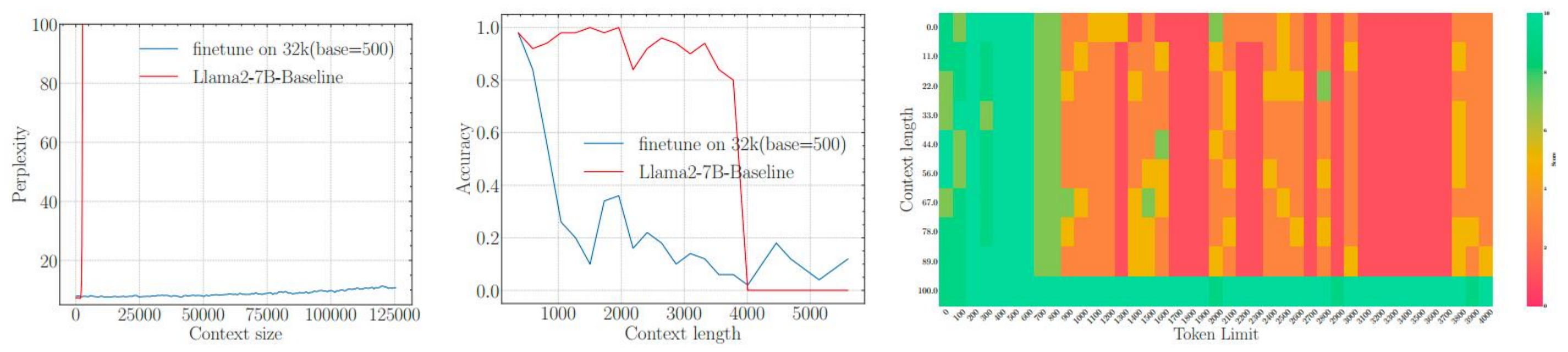
Figure 1: Context length and its corresponding lower bound of RoPE's base value.

Motivation



(a) base=1e4 (b) base=500 (c) base=$b \cdot s^{\frac{d}{d-2}}$

Figure 2: An illustration of OOD in RoPE when we extend context length from 4k to 32k, and two solutions to avoid the OOD. We show the last dimension as it is the lowest frequency part of RoPE, which suffers OOD mostly in extrapolation. (a) For a 4k context-length model with base value as 1e4, when we extend the context length to 32k without changing the base value, the context length from 4k to 32k is OOD for RoPE (red area in the figure). (b) OOD can be avoided with a small base value like 500 [14], since the full period has been fitted during fine-tuning stage. (c) We set base as $b \cdot s^{\frac{d}{d-2}}$ from NTK [10].The blue line denotes the pre-training stage (base=1e4) and the red dashed line denotes the fine-tuning stage (base=$b \cdot s^{\frac{d}{d-2}}$), we can observe that the RoPE's rotation angle of extended positions is in-distribution.



(a) Perplexity (b) Long-eval [19] (c) Needle in Haystack [20]

Figure 3: The superficial long context capability of avoiding OOD by the smaller base. Following the recent work [14], we fine-tune Llama2-7B with a small base (500) to a context length of 32k.

Theory Perspective

- Desiderata 1 The closer token gets more attention: the current token tends to pay more attention to the token that has a smaller relative distance.
- Desiderata 2 The similar token gets more attention: the token tends to pay more attention to the token whose key value is more similar to the query value of the current token.

$$\mathbb{E}_{q,k^*}\left[q^T R_{m,\theta} k^*\right] - \mathbb{E}_{q,k}\left[q^T R_{m,\theta} k\right], \tag{8}$$

where $q \in R^d$ is the query vector for the current token, $k^* = q + \epsilon$ is the key value of a similar token,

Theorem 1 *Assuming that the components of query $q \in R^d$ and key $k \in R^d$ are independent and identically distributed, their standard deviations are denoted as $\sigma \in R$. The key $k^* = q + \epsilon$ is a token similar to the query, where ϵ is a random variable with a mean of 0. Then we have:*

$$\frac{1}{2\sigma^2}\left(\mathbb{E}_{q,k^*}\left[q^T R_{m,\theta} k^*\right] - \mathbb{E}_{q,k}\left[q^T R_{m,\theta} k\right]\right) = \sum_{i=0}^{d/2-1} \cos(m\theta_i) \tag{9}$$

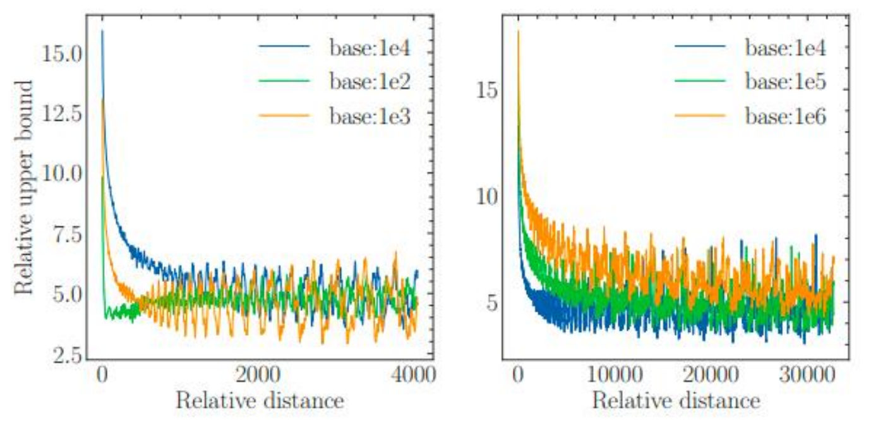


Figure 4: The upper bound of attention score with respect to the relative distance.

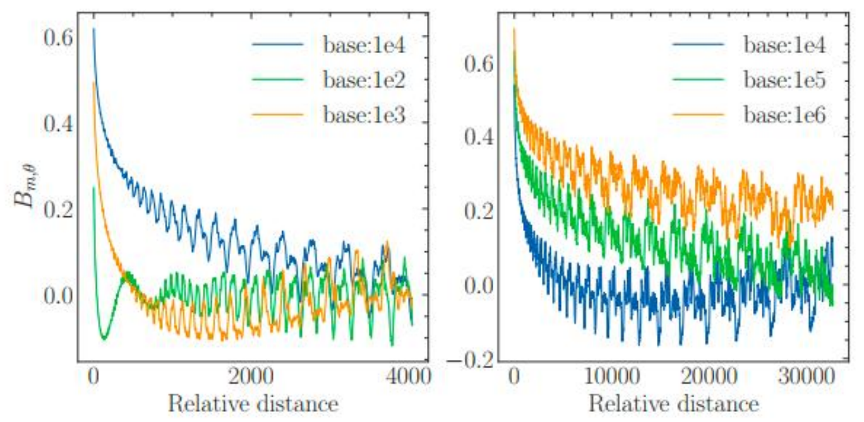
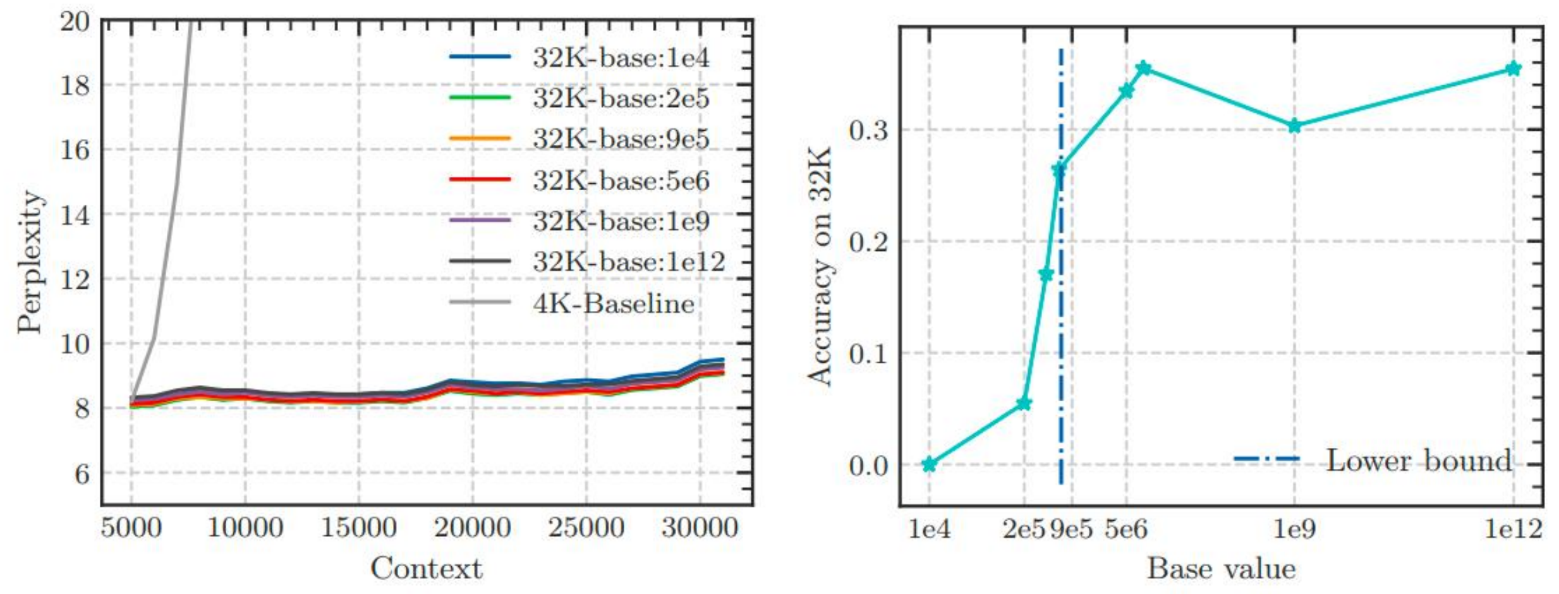


Figure 5: The ability to attend more to similar tokens than random tokens.



(a) Perplexity (b) Long-eval 32k

Figure 6: Fine-tuning Llama2-7B-Base on 32k context length with varying RoPE's base. Although the perplexity remains low with varying bases, the Long-eval accuracy reveals a discernible bound for the base value, below which the Long-eval accuracy declines significantly. The dotted line denotes the lower bound derived from Eq. 11 and code is provided in Appendix E

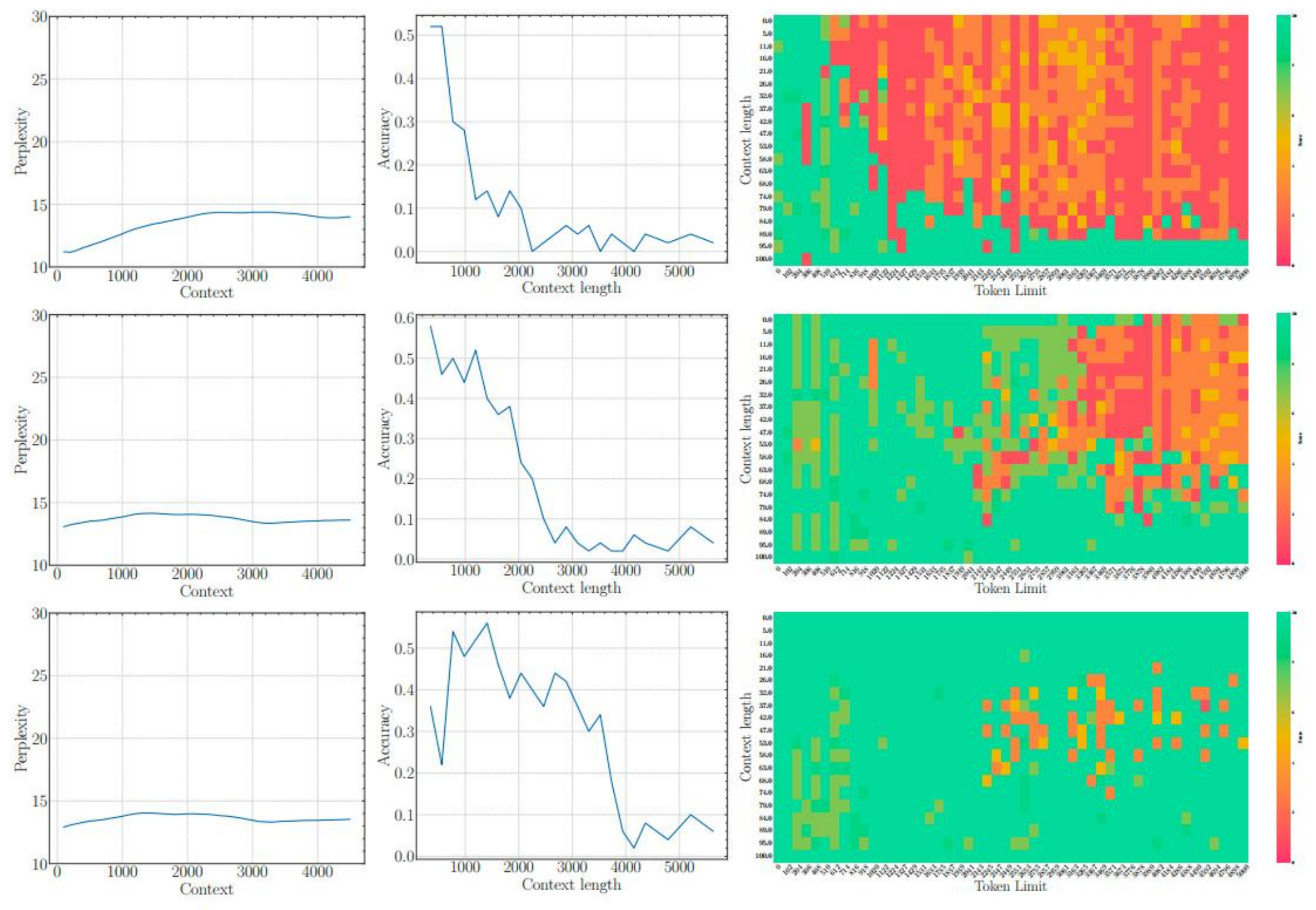


Figure 7: The first row: the results of a 2B model training from scratch with base=1e2. The second row: The results of fine-tuning the 2B model with base=1e4. The third row: The results of fine-tuning the 2B model with base=1e6.