The Impact of Positional Encoding on Length Generalization in Transformers

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, Siva Reddy

1 Introduction

- Classically, Transformers are poor at **Length Extrapolation (OOD)** [1]. But, how different **Positional Encodings (PEs)** affect length generalization?
- Early studies show decoder-only Transformers without PE **(NoPE)** perform fine in IID [2][3], but how about length generalization?
- Our analysis shows NoPE's unexpected superiority over other positional encodings in length generalization tasks.
- How does NoPE can recover the order without explicit position info? We attempt to answer this both theoretically and empirically.

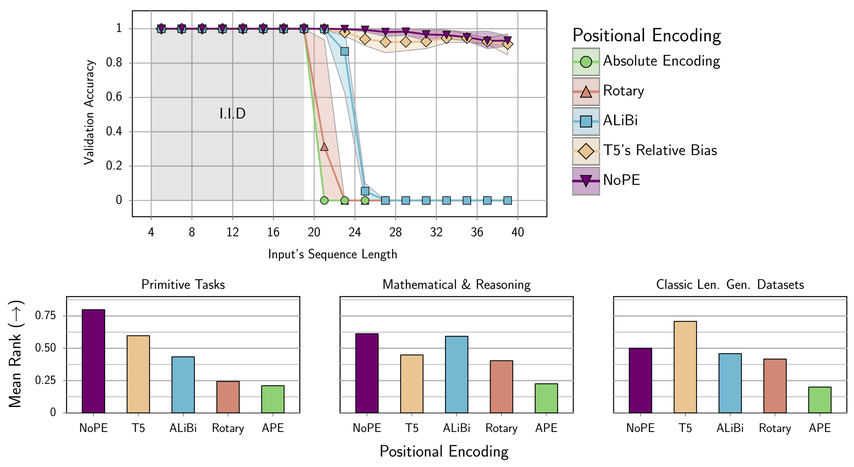
2 Evaluation Framework

- Evaluated on 10 synthetic mathematical & reasoning tasks.
- **Train** on sequence lengths ~ $U(1, L)$ & **Test** on lengths ~ $U(L+1, 2L)$.
- 100M decoder-only Transformers trained from scratch.

3 Results

Transformer with No Positional Encoding (NoPE) performs on par or better than SOTA encoding schemes.

Performance of models trained on the Copy task [4]



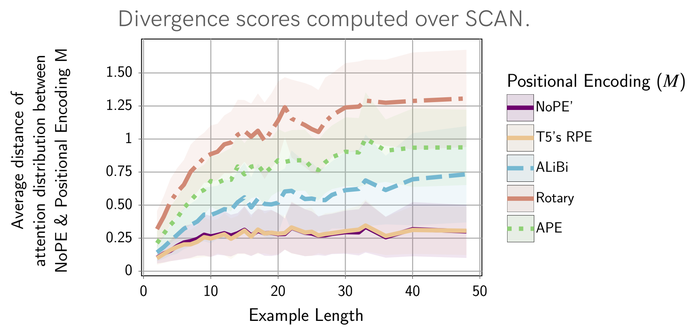
Theoretically, we show that a decoder-only Transformer with NoPE can recover both absolute and relative encoding.

Theorem 1 (Absolute Encoding):

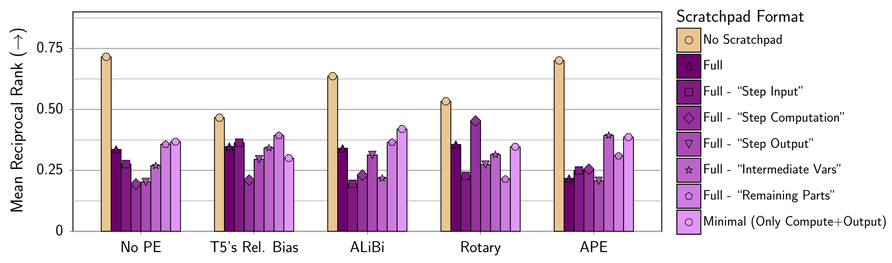
Let x = [<bos>, x_1, ..., x_T] be an input sequence of length T + 1 to the model. Then, there exists a set of weight such that **first layer** can recover absolute positions [1, . . . , T + 1] in the hidden state H.

Empirically, on the same inputs, NoPE attention pattern is more similar to relative encoding, rather than absolute.

Divergence scores computed over SCAN.



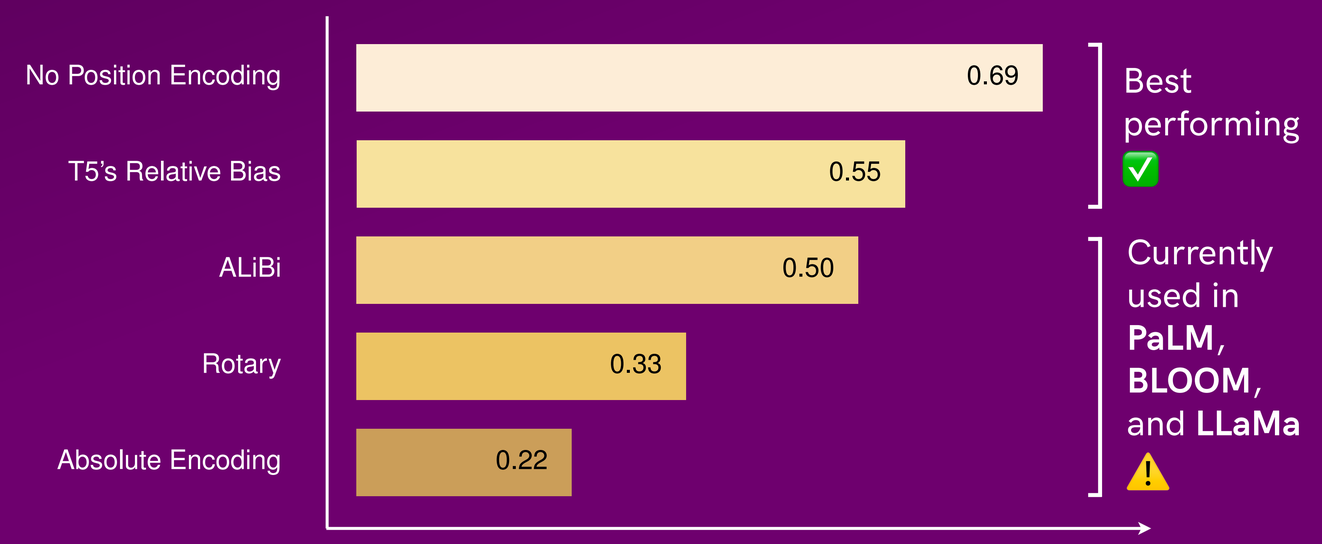
When we use scratchpad, the optimal format differs for each PE, suggesting each look at different part of the input.



4 Take-home messages

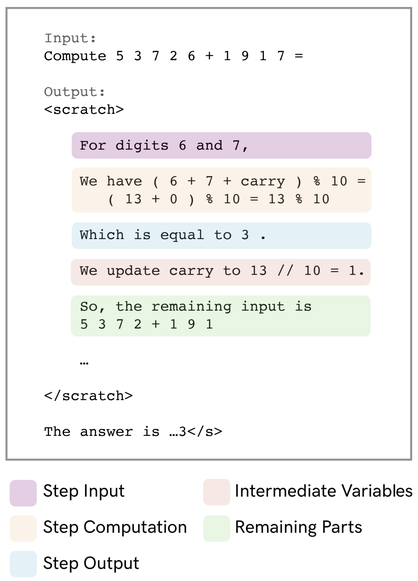
- Most popular positional encoding technique (Rotary and ALiBi) do not perform well on length extrapolation.
- Length Extrapolation on downstream tasks is a suitable test bed for PEs.
- NoPE holds promise as a modification to decoder-only Transformers.
- Scratchpad is not always helpful for length generalization and its format highly impacts the performance and interacts differently with different PEs.

Positional Encodings used in current LLMs **are NOT well-suited** for **Length Extrapolation**.

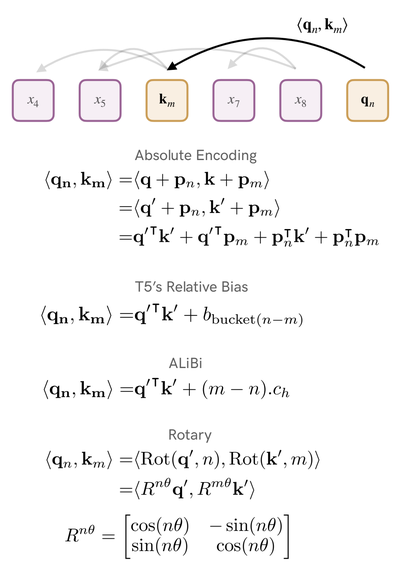


Use your camera to scan our QR codes for the **Twitter thread**, **GitHub repository**, and **ArXiv paper**.

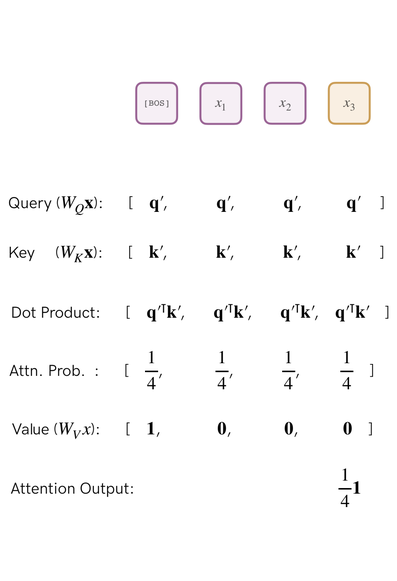
Additional Figures



Example of Full Scratchpad format.



How position is encoded in Transformers.



Proof sketch of Theorem 1: NoPE can represents absolute order.

References

1. *Anil, Cem et al. "Exploring Length Generalization in Large Language Models."*
2. *Tsai, Yao-Hung Hubert et al. "Transformer Dissection: An Unified Understanding for Transformer's Attention via the Lens of Kernel."*
3. *Haviv, Adi et al. "Transformer Language Models without Positional Encodings Still Learn Positional Information."*
4. *Ontan'on, Santiago et al. "Making Transformers Solve Compositional Tasks."*