Rethinking Addressing in Language Models via Contexualized Equivariant Positional Encoding

Jiajun Zhu[1]*, Peihao Wang[1]*, Ruisi Cai[1], Jason D. Lee[2], Pan Li[3], Zhangyang Wang[1]
[1]The University of Texas at Austin, [2]Princeton University, [3]Georgia Tech

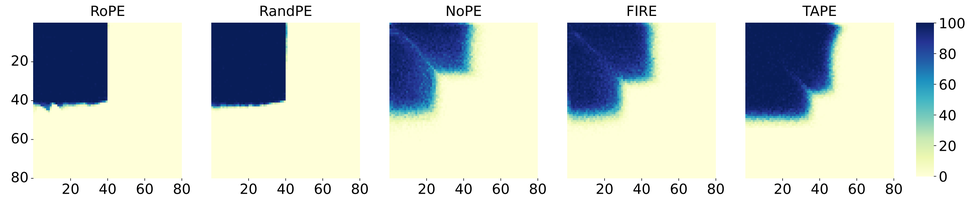
ICML International Conference On Machine Learning

Background

Existing positional embeddings usually impose a fixed distance-decay pattern on attention maps, thereby enforcing a locality bias. Their rigid structure hinders effective modeling of complex tasks like long-context retrieval and arithmetic computations [1]. We argue that contextualizing positional embeddings with sequence content is essential.

Experiments

Arithmetic Learning.

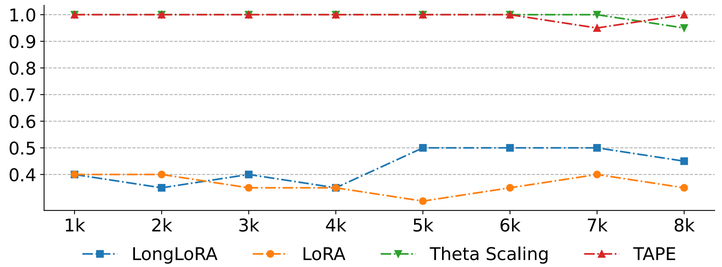


Accuracy on addition task on 2× context length. The x- and y-axes represent the sequence lengths of the two operands respectively. The average accuracy across the heatmap is 26.32%, 26.56%, 22.45%, 26.98% and 32.82% respectively for RoPE, RandPE, NoPE, FIRE, and our TAPE.

Long Context Modeling.

Performance comparison on SCROLLS benchmark.

	QAS	CNLI	NQA	QuAL	QMS	SumS	GovR
Metric (%)	F1 (↑)	EM (↑)	F1 (↑)	EM (↑)	Rgm (↑)	Rgm (↑)	Rgm (↑)
Median length	5472	2148	57829	7171	14197	9046	8841
RoPE	8.39	65.00	1.77	0.04	6.34	5.63	9.71
ALiBi	8.25	69.62	4.11	0.0	9.92	9.78	**18.81**
RandPE	**13.44**	62.01	4.63	0.38	8.43	8.31	8.93
FIRE	3.41	71.26	0.48	1.25	8.78	7.42	11.03
xPos	9.02	71.75	4.83	0.24	10.73	9.38	16.38
TAPE (Ours)	11.52	**72.80**	**6.79**	**11.60**	**12.42**	**10.34**	15.18



Accuracy on passkey retrieval. Model is based on Llama2 7B.

Perplexity evaluation on two datasets. Lower means better(↓).

Method	1024	2048	4096	8192	Method	1024	2048	4096	8192
	Proof-pile					PG-19			
LoRA	3.828	3.369	3.064	2.867	LoRA	9.791	9.098	8.572	8.199
LongLoRA	3.918	3.455	3.153	2.956	LongLoRA	9.989	9.376	8.948	8.645
Theta Scaling	3.864	3.415	3.121	2.934	Theta Scaling	9.257	8.640	8.241	7.999
TAPE (Ours)	**3.641**	**3.196**	**2.901**	**2.708**	TAPE (Ours)	**8.226**	**7.642**	**7.278**	**7.063**

Efficiency Analysis.

Speed measurement. We report execution time per step and iteration per second.

Method	TAPE		RoPE	FIRE	T5's relative bias
	w/ Fusion	w/o Fusion			
Time ($\times 10^{-4}$)	2.56	5.63	2.08	5.56	6.90
Throughput	3910	1775	4810	1799	1449
Flash Attention	✓	✓	✓	✗	✗

Our Approach: Contexualized Equivariant Positional Encoding (TAPE)

General Formulation. Let a tuple $(\boldsymbol{X}, \boldsymbol{E})$ represent a language sequence, where $\boldsymbol{X} \in \mathscr{X} \subseteq \mathbb{R}^{N \times C}$ are token features, and $\boldsymbol{E} \in \mathscr{E} \subset \mathbb{R}^{N \times D}$ are positional embeddings.

A transformer block consists of two separate operations:

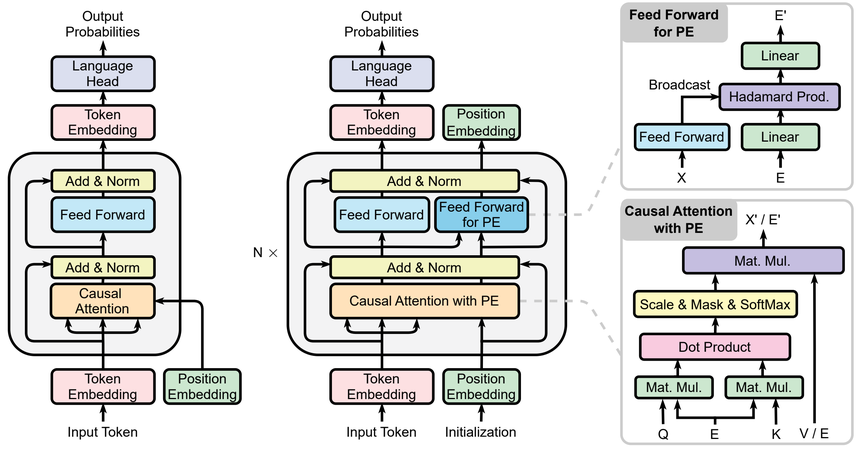
- Token Mixing: $f : \mathscr{X} \times \mathscr{E} \to \mathbb{R}^{N \times C}$ encodes positional embeddings to represent token features.
- Position Contextualization: $g : \mathscr{X} \times \mathscr{E} \to \mathbb{R}^{N \times D}$ encodes the context information into the positional embeddings.

Equivariance Principles. Let $\Pi(\cdot)$ denote permutation group and $O(\cdot)$ denote an orthogonal group. For $\forall \boldsymbol{P} \in \Pi(N), \boldsymbol{R} \in O(R)$, the two principles is required for f and g:

- Permutation Invariance: $f(\boldsymbol{PX}, \boldsymbol{PER}) = \boldsymbol{P} f(\boldsymbol{X}, \boldsymbol{E})$, an intrinsic property of standard transformer.
- Rotation Equivariance: $g(\boldsymbol{PX}, \boldsymbol{PER}) = \boldsymbol{P} g(\boldsymbol{X}, \boldsymbol{E}) \boldsymbol{R}$, important to keep transformer invariant to shift on token indices and thus can generalize to long sequence (Proposition 4.3).

Our proposed TAPE implements the two principles. (Proposition 3.1)

Model Architecture.



(a) Traditional position embedding. (b) TAPE with enhanced causal attention and feed forward layers.

Reference

[1] Ebrahimi, M., Panchal, S., & Memisevic, R. Your Context Is Not an Array: Unveiling Random Access Limitations in Transformers. In *First Conference on Language Modeling*.