Position Embedding and Evaluation of Long Context Language Models

汪杰

绝对位置编码

加性位置编码： $x_k + p_k$

- 训练式： p_k 是可训练的。没有外推性。
- 三角式（sinusoidal）：
$$\begin{cases} \boldsymbol{p}_{k,2i} = \sin\left(k/10000^{2i/d}\right) \\ \boldsymbol{p}_{k,2i+1} = \cos\left(k/10000^{2i/d}\right) \end{cases}$$
- 递归式：用RNN生成， $p_{k+1} = f(p_k)$

乘性位置编码： $x_k \otimes p_k$

相对位置编码

- 把绝对位置编码拆开，改写位置相关的项。（XLNET、DeBERTa……）

$$
\begin{aligned}
q_i k_j^T &= (x_i + p_i) W_Q W_K^T (x_j + p_j)^T \\
&= x_i W_Q W_K^T x_j^T + x_i W_Q W_K^T p_j^T + p_i W_Q W_K^T x_j^T + p_i W_Q W_K^T p_j^T \\
&\Rightarrow x_i W_Q W_K^T x_j^T + x_i W_Q W_K^T R_{i,j}^T + R_{j,i} W_Q W_K^T x_j^T \qquad \text{(DeBERTa)}
\end{aligned}
$$

- T5 Bias（加性）： $x_i W_Q W_K^T x_j^T + \beta_{i,j}$
  - 可训练，分桶

$i-j$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$f(i-j)$	0	1	2	3	4	5	6	7	8	8	8	8	9	9	9	9
$i-j$	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	$\cdots$
$f(i-j)$	10	10	10	10	10	10	10	11	11	11	11	11	11	11	11	$\cdots$

让研究人员绞尽脑汁的Transformer位置编码 - 科学空间|Scientific Spaces

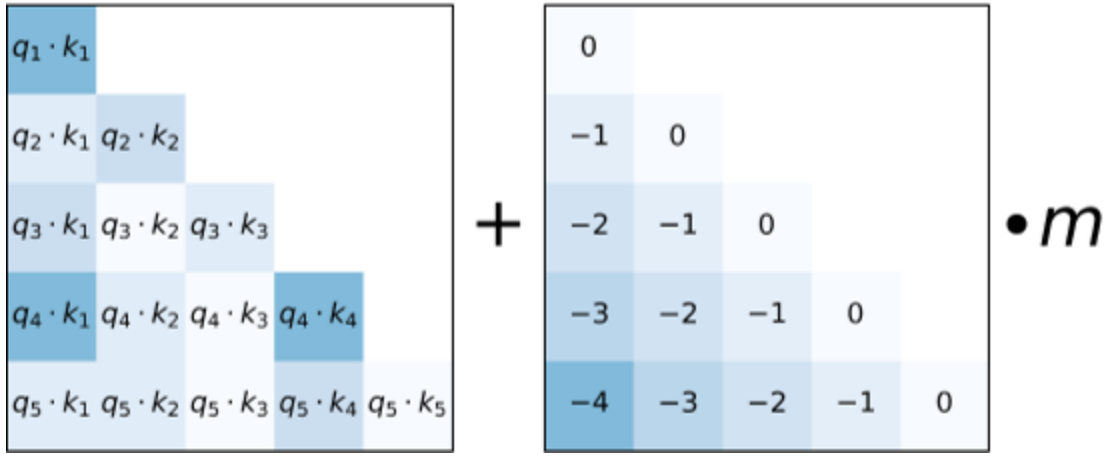
[JMLR20] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5)

相对位置编码

- RoPE（乘性）：

$$\begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \\ \vdots \\ q_{d-2} \\ q_{d-1} \end{pmatrix}$$

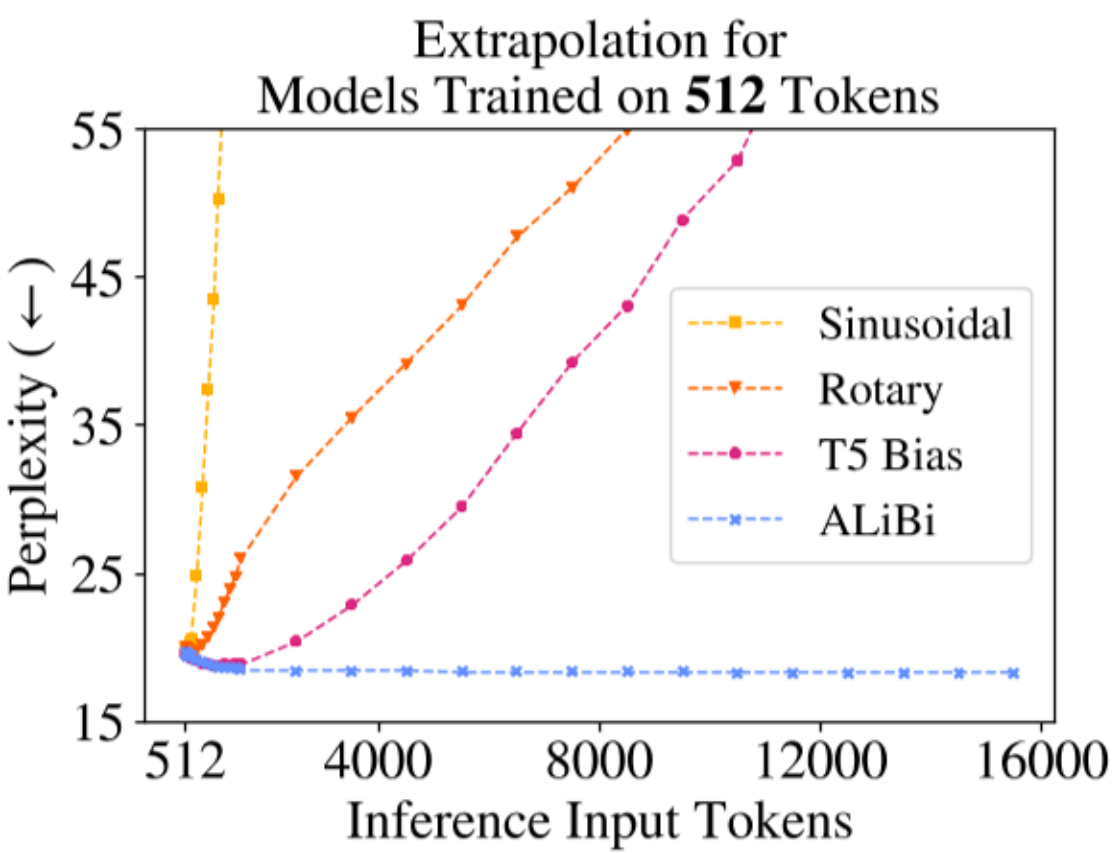
- ALiBi (Attention with Linear Biases)

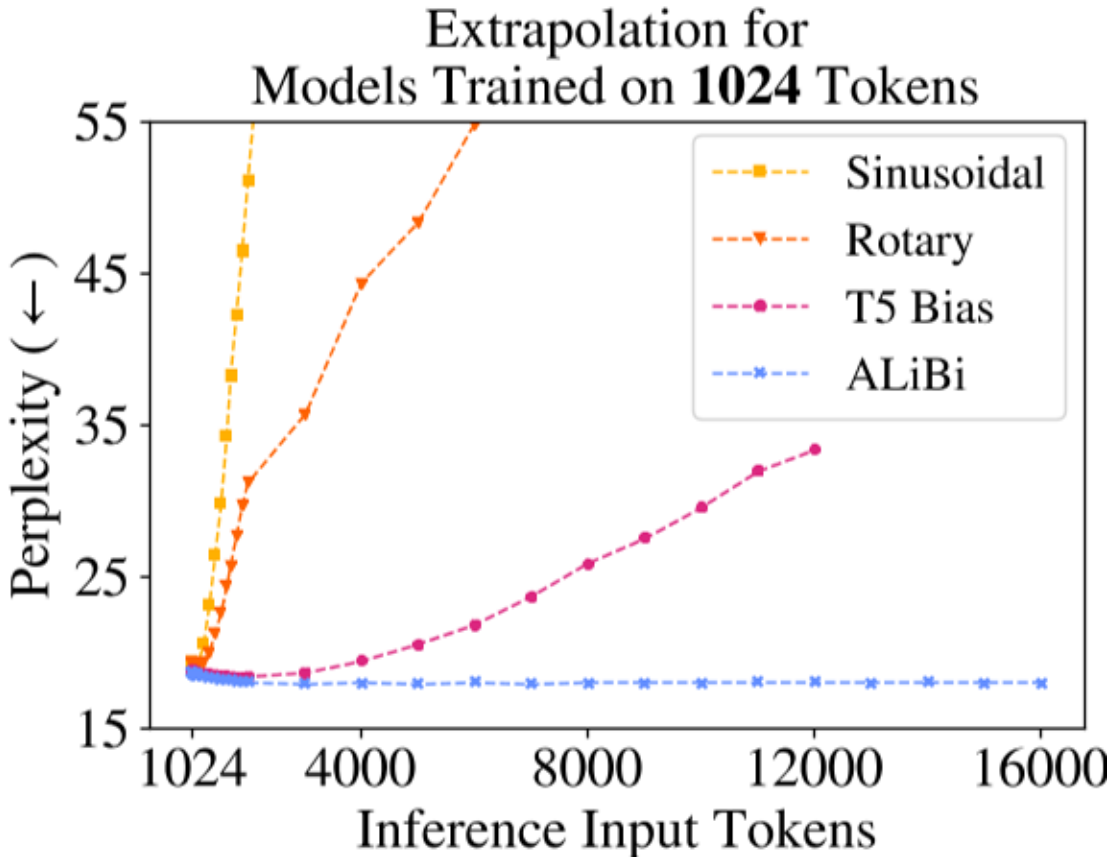


Transformer升级之路：2、博采众长的旋转式位置编码 - 科学空间|Scientific Spaces

[ICLR22] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (ALiBi)

外推性（Extrapolation）





Dataset: WikiText-103

[ICLR2022] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (ALiBi)

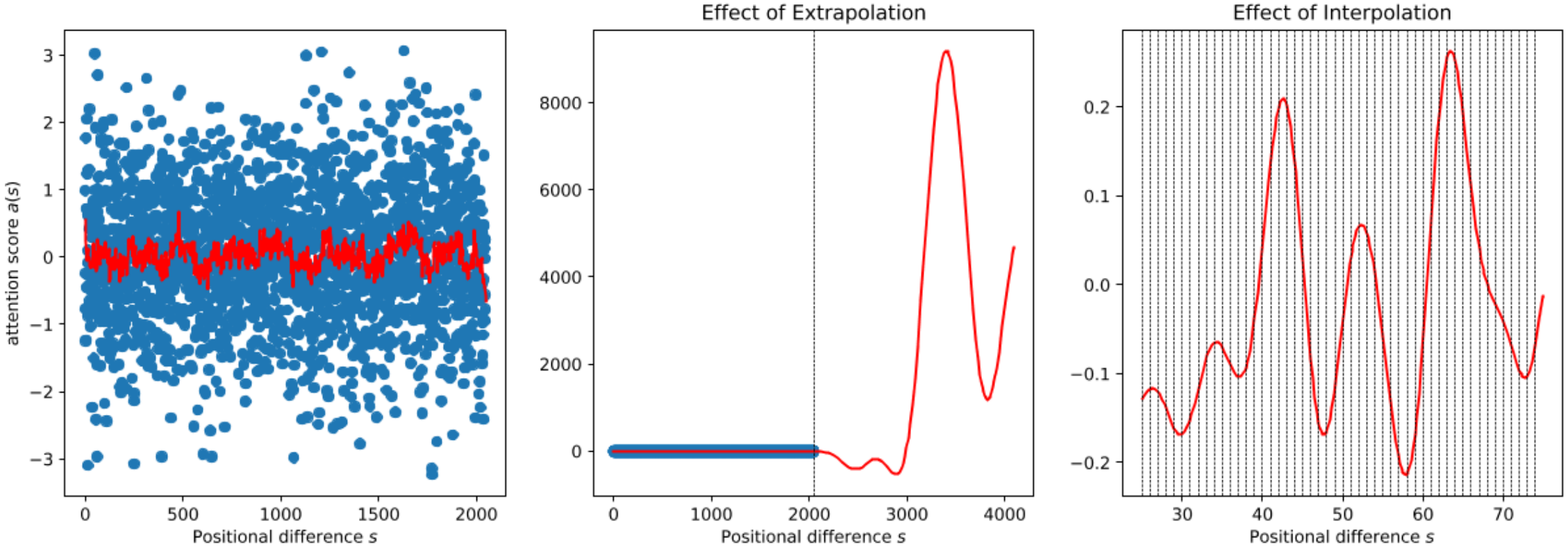
为什么 xformers 无法外推?

- Transformers **overfit** to positions [1]
  - token @ 1000 $\neq$ token @ 1001
- 用到了没训练过的位置编码 [2]
- 预测的时候注意力机制所处理的token数量远超训练时的数量 [2]
  - 应对：softmax之前，attention score 乘 $\log n$ [3]

---

[1] ALiBi enables transformer language models to handle longer inputs - YouTube
[2] Transformer升级之路：7、长度外推性与局部注意力 - 科学空间|Scientific Spaces
[3] 从熵不变性看Attention的Scale操作 - 科学空间|Scientific Spaces

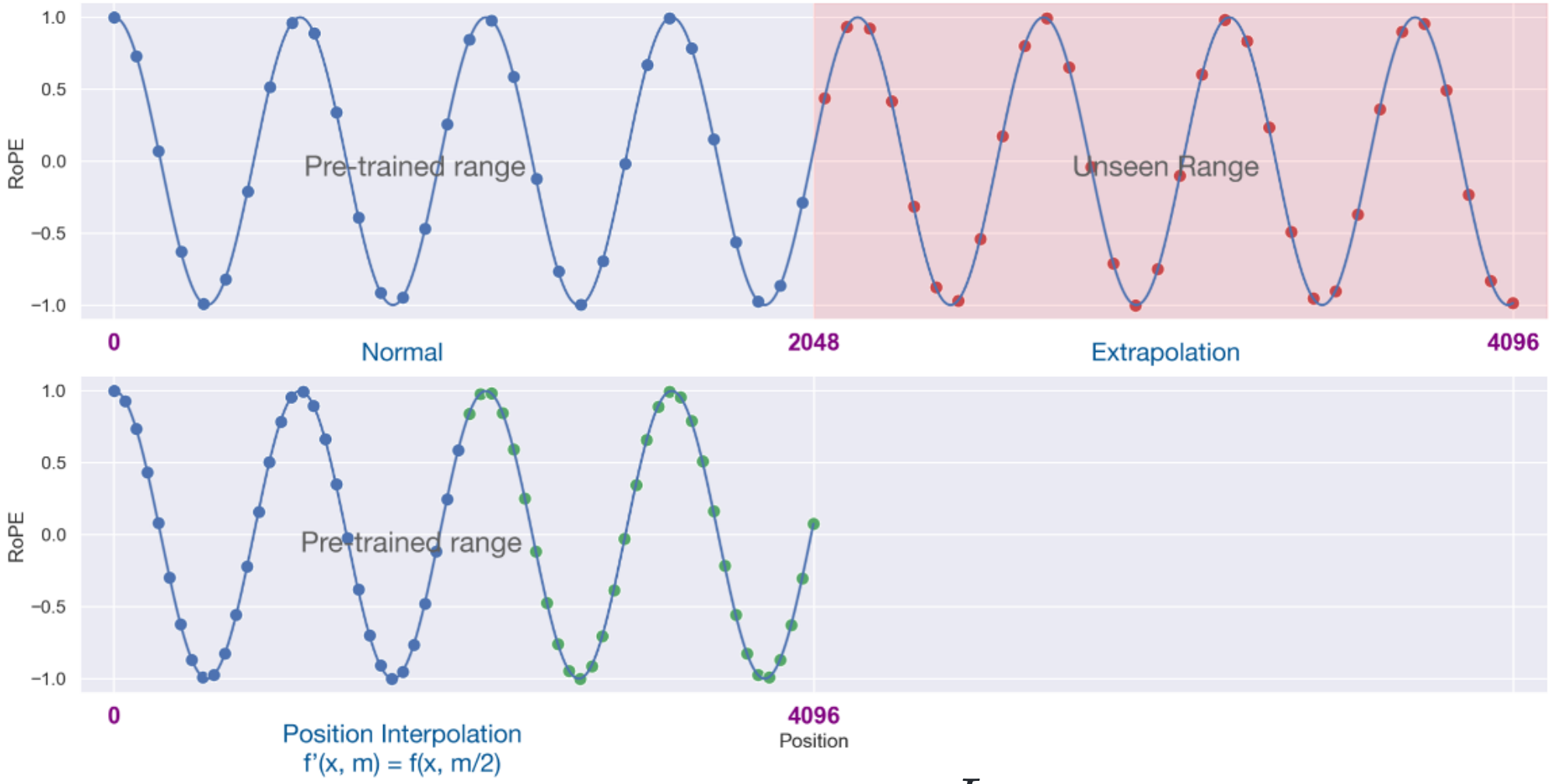
Catastrophic values in attention computation



RoPE: $a(s) = \mathrm{Re}\left[\sum_{j=0}^{d/2-1} h_j e^{\mathrm{i}s\theta_j}\right]$

其中，$h_j = \boldsymbol{q}_{[2j:2j+1]}\boldsymbol{k}^*_{[2j:2j+1]} = (q_{2j} + \mathrm{i}q_{2j+1})(k_{2j} - \mathrm{i}k_{2j+1}), \theta_j = 10000^{-2j/d}$

Position Interpolation



$$\text{Position Interpolation}(x, m) = \text{RoPE}(x, m\frac{L}{L'}) \qquad \text{for } L' > L$$

Language Modeling (Perplexity)

- 和 long context 没有直接关联：相距越远的 token，相关性越弱[1]
- 作为基础测试，保证模型不崩溃。

Dataset:

- Pretraining datasets (BERT, RoBERTa, ...)
- CC100 [2]
- book corpus (PG-19) [3]
- Arxiv Math proof-pile [3]
- Stack [4]

---

[1] Transformer 升级之路：1、Sinusoidal 位置编码追根溯源 - 科学空间|Scientific Spaces

[2] [ICLR22] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (ALiBi)

[3] Extending Context Window of Large Language Models via Positional Interpolation

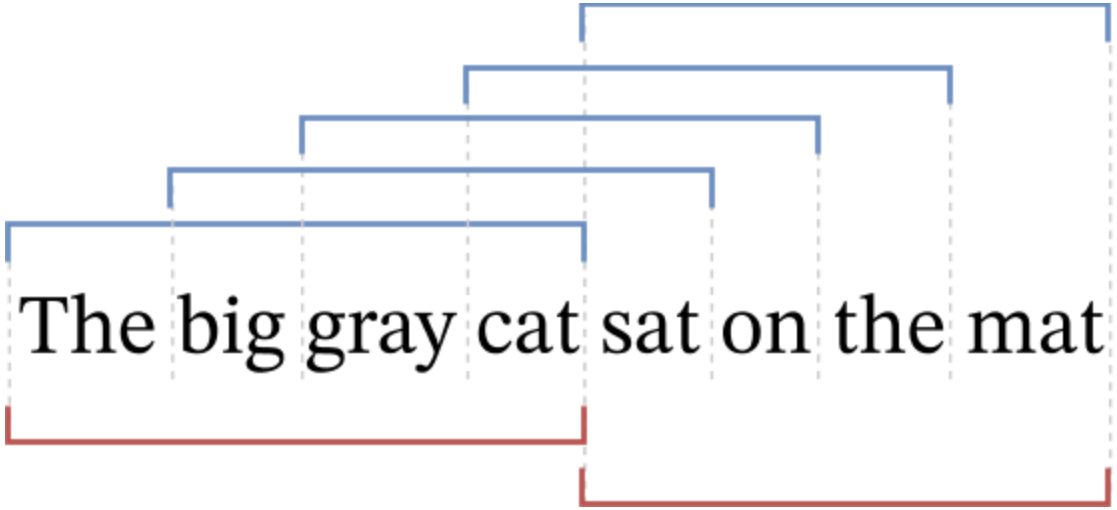
[4] LongNet: Scaling Transformers to 1,000,000,000 Tokens

Position Interpolation Results

Model Size	Model Context Window	Method	Evaluation Context Window Size 2048	4096	8192	16384	32768
7B	2048	None	7.20	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^3$
7B	8192	FT	7.21	7.34	7.69	-	-
7B	8192	PI	7.13	6.96	6.95	-	-
7B	16384	PI	7.11	6.93	6.82	6.83	-
7B	32768	PI	7.23	7.04	6.91	6.80	6.77
13B	2048	None	6.59	-	-	-	-
13B	8192	FT	6.56	6.57	6.69	-	-
13B	8192	PI	6.55	6.42	6.42	-	-
13B	16384	PI	6.56	6.42	6.31	6.32	-
13B	32768	PI	6.54	6.40	6.28	6.18	6.09
33B	2048	None	5.82	-	-	-	-
33B	8192	FT	5.88	5.99	6.21	-	-
33B	8192	PI	5.82	5.69	5.71	-	-
33B	16384	PI	5.87	5.74	5.67	5.68	-
65B	2048	None	5.49	-	-	-	-
65B	8192	PI	5.42	5.32	5.37	-	-

Early Token Curse on Perplexity

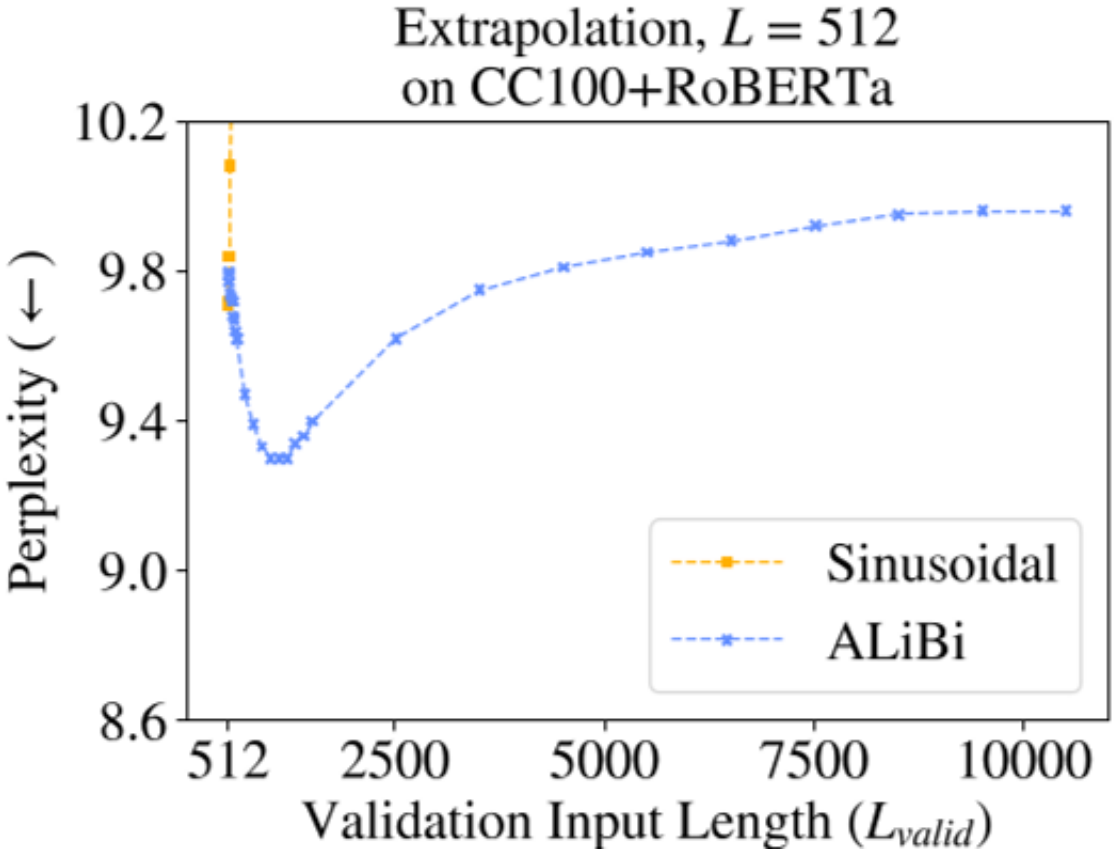
Sliding window evaluation (top; blue)

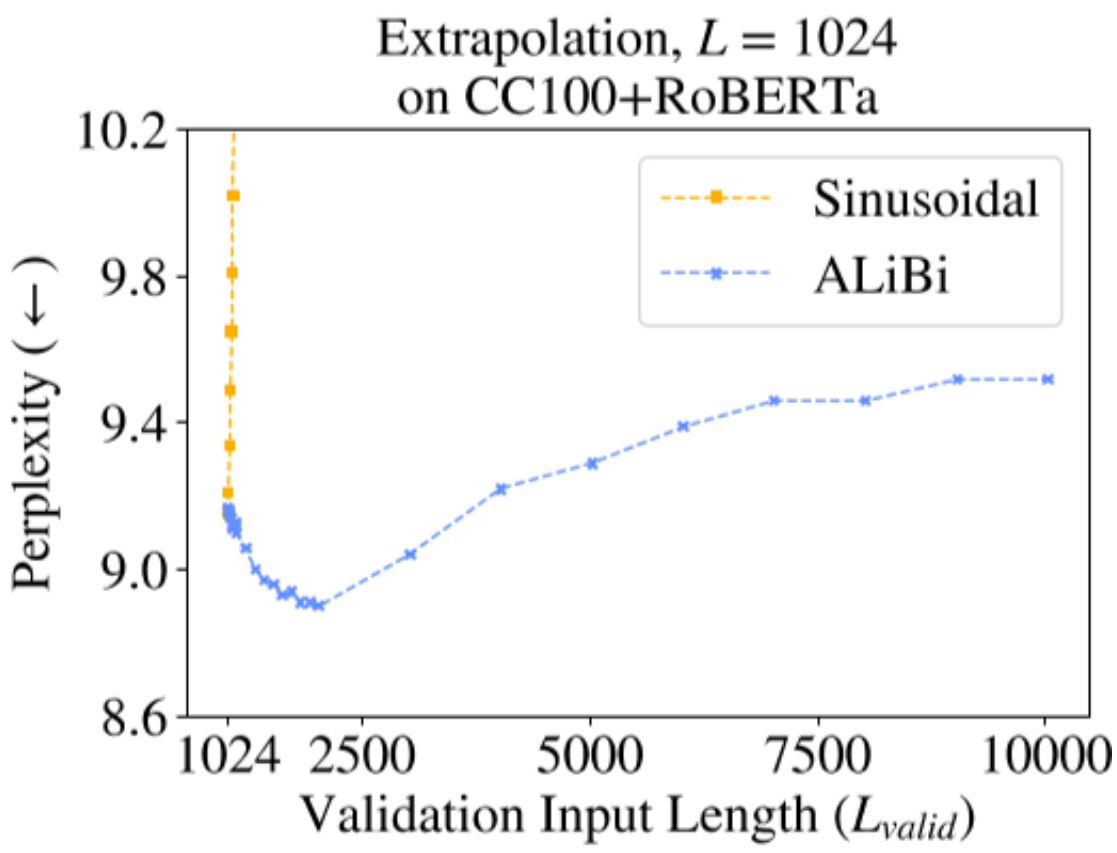


non-overlapping evaluation (bottom; red)

[ICLR22] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (ALiBi)

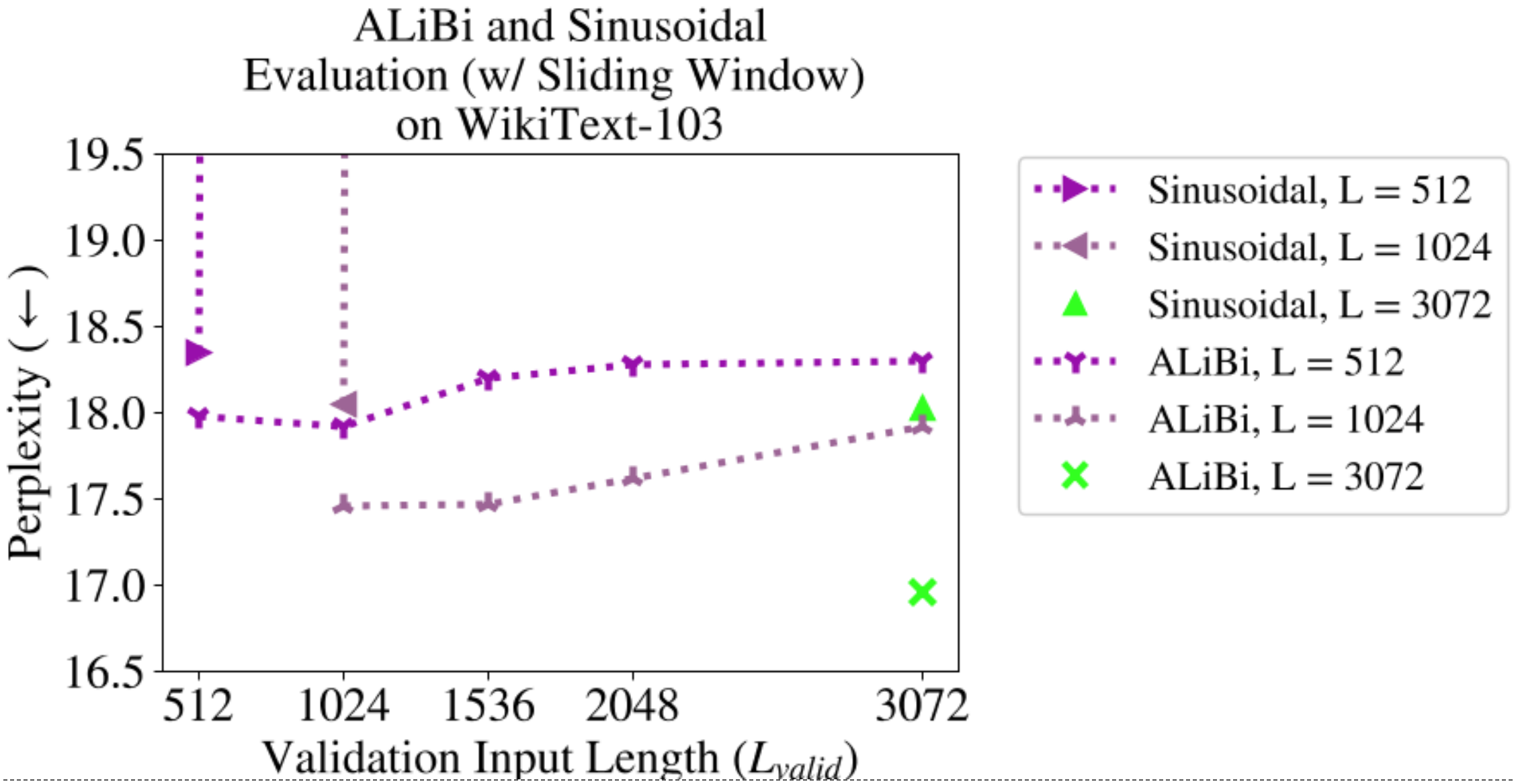
Results on non-overlapping evaluation





[ICLR22] Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation (ALiBi)

Results on sliding window evaluation



Synthetic Retrieval Tasks

基于LLM对话能力

- Passkey Retrieval [1] [2] [3]
- LongEval [4]

[1] Landmark Attention: Random-Access Infinite Context Length for Transformers
[2] Extending Context Window of Large Language Models via Positional Interpolation
[3] Focused Transformer: Contrastive Training for Context Scaling
[4] How Long Can Open-Source LLMs Truly Promise on Context Length? | LMSYS Org

Passkey Retrieval

```
There is an important info hidden inside a lot of irrelevant text.  Find
it and memorize them.  I will quiz you about the important information
there.
The grass is green.  The sky is blue.  The sun is yellow.  Here we go.
There and back again.  (repeat X times)
The pass key is 12345.  Remember it.  12345 is the pass key.
The grass is green.  The sky is blue.  The sun is yellow.  Here we go.
There and back again.  (repeat Y times)
What is the pass key?  The pass key is
```

Position Interpolation Results

We report k_{max} : the maximum k such that, for all $k' \leq k$, the model has a success rate of at least 20% on k'.

Model			Fine-tuning steps					
Size	Context Window	Method	200	400	600	800	1000	10000
7B	8192	FT	1792	2048	2048	2048	2304	2560
33B	8192	FT	1792	2048	1792	2048	2304	-
7B	8192	PI	8192	8192	8192	8192	8192	-
7B	16384	PI	16384	16384	16384	16384	16384	-
7B	32768	PI	32768	32768	18432	32768	32768	-
33B	8192	PI	8192	8192	8192	8192	8192	-
33B	16384	PI	16384	16384	16384	16384	16384	-

LongEval

Coarse-grained Topic Retrieval (conversation length: 400 ~ 600 tokens)

```
… (instruction of the task)
USER: I would like to discuss <TOPIC-1>
ASSISTANT: Sure! What about xxx of <TOPIC-1>?
… (a multi-turn conversation of <TOPIC-1>)
USER: I would like to discuss <TOPIC-k>
…
USER: What is the first topic we discussed?
ASSISTANT:
```

Fine-grained Line Retrieval

```
line torpid-kid: REGISTER_CONTENT is <24169>
line moaning-conversation: REGISTER_CONTENT is <10310>
…
line tacit-colonial: REGISTER_CONTENT is <14564>
What is the <REGISTER_CONTENT> in line moaning-conversation?
```

Synthetic Reasoning Tasks

- Long ListOps（from LRA）

```
INPUT: [MAX 4 3 [MIN 2 3 ] 1 0 [MEDIAN 1 5 8 9, 2]]        OUTPUT: 5
```

- Neural Networks and the Chomsky Hierarchy：需要从零开始训练

Level	Name	Example Input	Example Output
R	Even Pairs	$aabba$	True
	Modular Arithmetic (Simple)	$1 + 2 - 4$	4
	Parity Check†	$aaabba$	True
	Cycle Navigation†	011210	2
DCF	Stack Manipulation	$abbaa$ POP PUSH a POP	$abba$
	Reverse String	$aabba$	$abbaa$
	Modular Arithmetic	$-(1-2)\cdot(4-3\cdot(-2))$	0
	Solve Equation°	$-(x-2)\cdot(4-3\cdot(-2))$	1
CS	Duplicate String	$abaab$	$abaababaab$
	Missing Duplicate	10011021	0
	Odds First	$aaabaa$	$aaaaba$
	Binary Addition	$10010 + 101$	10111
	Binary Multiplication×	$10010 * 101$	1001000
	Compute Sqrt	100010	110
	Bucket Sort†★	421302214	011222344

Results on Chomsky Hierarchy

Level	Task	RNN	Stack-RNN	Tape-RNN	Transformer	LSTM
R	Even Pairs	**100.0**	**100.0**	**100.0**	**96.4**	**100.0**
	Modular Arithmetic (Simple)	**100.0**	**100.0**	**100.0**	24.2	**100.0**
	Parity Check†	**100.0**	**100.0**	**100.0**	52.0	**100.0**
	Cycle Navigation†	**100.0**	**100.0**	**100.0**	61.9	**100.0**
DCF	Stack Manipulation	56.0	**100.0**	**100.0**	57.5	59.1
	Reverse String	62.0	**100.0**	**100.0**	62.3	60.9
	Modular Arithmetic	41.3	**96.1**	**95.4**	32.5	59.2
	Solve Equation°	51.0	56.2	64.4	25.7	67.8
CS	Duplicate String	50.3	52.8	**100.0**	52.8	57.6
	Missing Duplicate	52.3	55.2	**100.0**	56.4	54.3
	Odds First	51.0	51.9	**100.0**	52.8	55.6
	Binary Addition	50.3	52.7	**100.0**	54.3	55.5
	Binary Multiplication×	50.0	52.7	58.5	52.2	53.1
	Compute Sqrt	54.3	56.5	57.8	52.4	57.5
	Bucket Sort†★	27.9	78.1	70.7	**91.9**	**99.3**

real-world long-context tasks

- Summarization
  - GovReport [1]
- QA
  - TriviaQA [2]
  - QasperQA [3]
- classification
  - Hyperpartisan [2]
- MT-bench [3]

---

[1] Extending Context Window of Large Language Models via Positional Interpolation

[2] [NAACL22] Simple Local Attentions Remain Competitive for Long-Context Tasks

[3] How Long Can Open-Source LLMs Truly Promise on Context Length? | LMSYS Org

Evaluation Methods

- language modeling (PPL)
- synthetic retrieval tasks
  - Passkey Retrieval
  - LongEval
- synthetic reasoning tasks：
  - Long ListOps（LRA）
  - Neural Networks and the Chomsky Hierarchy
- real-world long-context tasks
  - Summarization： GovReport
  - QA： TriviaQA, QasperQA
  - classification： Hyperpartisan
  - MT-bench