Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention

Attention

(Self)-attention is the main component of the transformer architecture. Given a input sequence $x \in \mathbb{R}^{L \times d_{\text{model}}}$ a single head, vanilla self-attention is computed as follows:

1. Compute query, key, value with the trainable weights $W_q, W_k \in \mathbb{R}^{d_k \times d_{\text{model}}}$, and $W_v \in \mathbb{R}^{d_v \times d_{\text{model}}}$ by

$$Q = xW_q, \quad K = xW_k, \quad V = xW_v$$

2. Compute the *mask* by

$$\tilde{A}_{\text{dot}} = \text{softmax}\left(\frac{QK^\intercal}{\sqrt{d_{\text{model}}}}\right)$$

3. Compute the attention

$$A_{\text{dot}} = \tilde{A}_{\text{dot}} V$$

Problem with Attention

Note that computing $\tilde{A}_{\text{dot}}$ requires $\mathcal{O}(L^2)$ computation and memory complexity. Informally, since each element of sequence is compared with every other element, the amount of computation and memory increase quadratic to the length of sequence. To overcome this drawback:

- Within attention mechanism: use window size $N < L$ for attention.
- Use alternatives: e.g. State Space Model (SSM)

Problem with Attention

Note that computing $\tilde{A}_{\text{dot}}$ requires $\mathcal{O}(L^2)$ computation and memory complexity. Informally, since each element of sequence is compared with every other element, the amount of computation and memory increase quadratic to the length of sequence. To overcome this drawback:

- Within attention mechanism: use window size $N < L$ for attention.

  $\Rightarrow$ *Loss global context!*

- Use alternatives: e.g. State Space Model (SSM)

Infini-attention

This paper proposes *Infini-attention* to capture both local and global context states.

Compressive Memory

Memory retrieval (query). Given a memory $M_{s-1} \in \mathbb{R}^{d_k \times d_v}$, compressed memory A_{mem} is computed by

$$A_{\text{mem}} = \frac{\sigma(Q)M_{s-1}}{\sigma(Q)z_{s-1}},$$

where $Q \in \mathbb{R}^{L \times d_k}$ is shared with A_{dot} and z_{s-1} is the normalization term.

Memory update (key-value).

$$M_s = M_{s-1} + \sigma(K)^\intercal \left(V - \frac{\sigma(Q)M_{s-1}}{\sigma(Q)z_{s-1}} \right)$$

$$z_s = z_{s-1} + \sum_{t=1}^{N} \sigma(K_t)$$

Compressive Memory

High-level interpretation

- Local information: Q, K, V computed for A_{dot}
  $\Rightarrow A_{\text{dot}}$ queries for given sequence x to the current local KV
- Global information: M_s containing key-value entries
  $\Rightarrow A_{\text{mem}}$ queries for given sequence x to the global KV

Infini-attention

$$A = \text{sigmoid}(\beta) \odot A_{\text{mem}} + (1 - \text{sigmoid}(\beta)) \odot A_{\text{dot}}$$

Experiments

- Task: PG19, Arxiv-math
- $N = 2048$, input sequence length 32768

Model	Memory size (comp.)	XL cache	Segment length	PG19	Arxiv-math
Transformer-XL	50M (3.7x)	2048	2048	11.88	2.42
Memorizing Transformers	183M (1x)	2048	2048	11.37	2.26
RMT	2.5M (73x)	None	2048	13.27	2.55
Infini-Transformer (L)	1.6M (114x)	None	2048	9.65	2.24
Infini-Transformer (L + D)	1.6M (114x)	None	2048	9.67	2.23

Table 1: Comparison of different long-context language modeling.

Experiments

- Two types of heads
  - Specialized heads: gating score $\approx 0, 1$
  - Mixer heads: gating score ≈ 0.5
- Each layer has at least one short-range head (gating score ≈ 0)

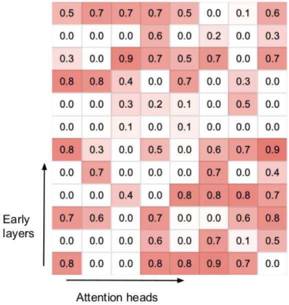

Figure 1: Visualization of gating scores

Experiments

	Zero-shot				
	32K	**128K**	**256K**	**512K**	**1M**
Infini-Transformer (L)	14/13/98	11/14/100	6/3/100	6/7/99	8/6/98
Infini-Transformer (L + D)	13/11/99	6/9/99	7/5/99	6/8/97	7/6/97
	FT (400 steps)				
Infini-Transformer (L)	100/100/100	100/100/100	100/100/100	97/99/100	96/94/100
Infini-Transformer (L + D)	100/100/100	100/100/100	100/100/100	100/100/100	100/100/100

Table 2: 1M passkey retrieval

Experiments

Model	Rouge-1	Rouge-2	Rouge-L	Overall
BART	36.4	7.6	15.3	16.2
BART + Unlimiformer	36.8	8.3	15.7	16.9
PRIMERA	38.6	7.2	15.6	16.3
PRIMERA + Unlimiformer	37.9	8.2	16.3	17.2
Infini-Transformers (Linear)	37.9	8.7	17.6	18.0
Infini-Transformers (Linear + Delta)	40.0	8.8	17.9	18.5

Table 3: 500k length book summary

Thank You

Q & A