Synced

Transformers Scale to Long Sequences With Linear Complexity Via Nyström-Based Self-Attention Approximation

Synced Feb 11 · 3 min read ★

In the early days of NLP research, establishing long-term dependencies brought with it the vanishing gradient problem, as nascent models handled input sequences one by one, without parallelization. More recently, revolutionary transformer-based architectures and their self-attention mechanisms have enabled interactions of token

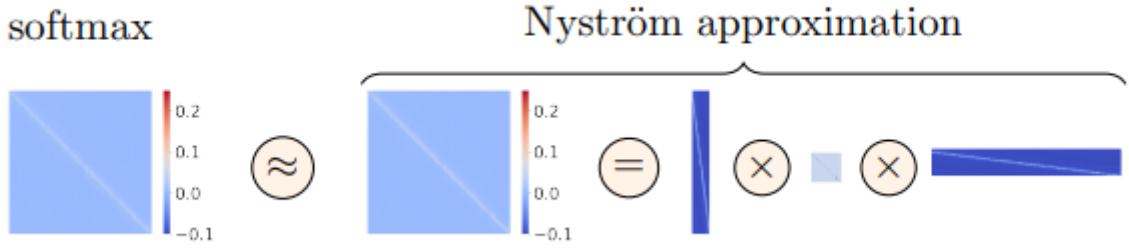
pairs across full sequences, modelling arbitrary dependencies in a constant number of layers to achieve state-of-the-art performance across many NLP tasks.

These advantages however came with a high cost, as transformer-based networks' memory and computational requirements grow quadratically with sequence length, resulting in major efficiency bottlenecks when dealing with long sequences. In the new paper *Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention,* researchers from the University of Wisconsin-Madison, UC Berkeley, Google Brain and American Family Insurance propose Nyströmformer, an O(n) approximation in both memory and time for self-attention designed to reduce the quadratic cost associated with long input sequences.

Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention

Yunyang Xiong [1] **Zhanpeng Zeng** [1] **Rudrasis Chakraborty** [2] **Mingxing Tan** [3]
Glenn Fung [4] **Yin Li** [1] **Vikas Singh** [1]
[1] University of Wisconsin-Madison [2] UC Berkeley [3] Google Brain [4] American Family Insurance
yxiong43@wisc.edu, zzeng38@wisc.edu, rudra@berkeley.edu, tanmingxing@google.com, gfung@amfam.com,
yin.li@wisc.edu, vsingh@biostat.wisc.edu

The Nyström method is an efficient technique for obtaining a low-rank approximation of a large kernel matrix. The researchers' proposed method leverages Nyström approximation tailored for a softmax matrix to reduce complexity from $O(n^2)$ to O(n) for self-attention computation.



A Nyström approximation of a softmax matrix in self-attention

Algorithm 1: Pipeline for Nyström approximation of softmax matrix in self-attention

Input: Query Q and Key K.
Output: Nyström approximation of softmax matrix.

Compute landmarks from input Q and landmarks from input K, $\tilde{Q}$ and $\tilde{K}$ as the matrix form ;

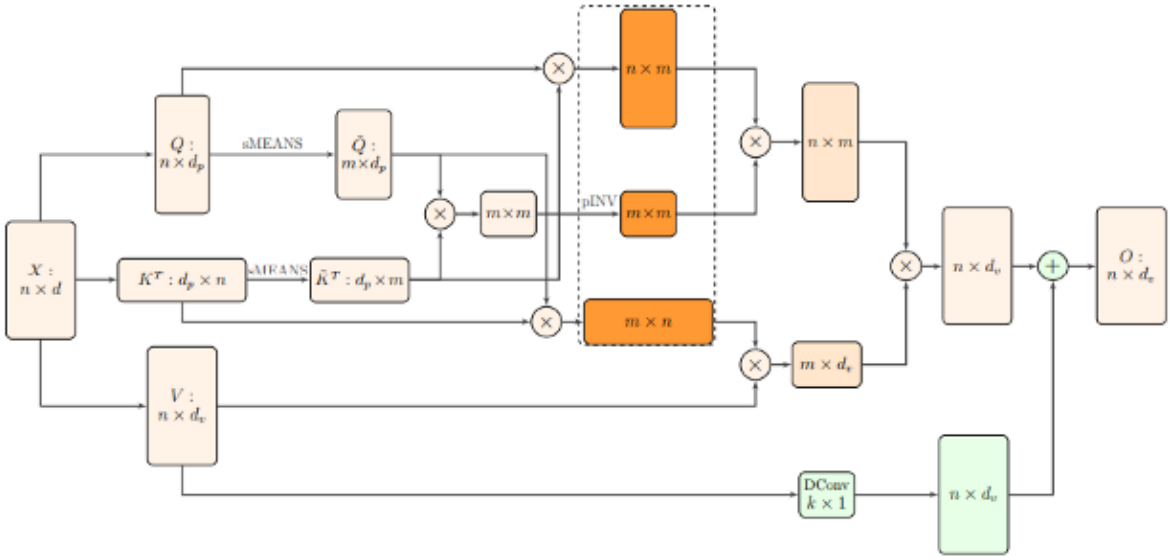
Compute $\tilde{F} = \text{softmax}(\frac{Q\tilde{K}^T}{\sqrt{d_q}})$, $\tilde{B} = \text{softmax}(\frac{\tilde{Q}K^T}{\sqrt{d_q}})$;

Compute $\tilde{A} = \text{softmax}(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_q}})^+$;

return $\tilde{F} \times \tilde{A} \times \tilde{B}$;

Pipeline for Nyström approximation of softmax matrix in self-attention

The basic idea behind the algorithm is to first define the matrix form of landmarks, then use these to form the three matrices needed for approximation. The landmarks are selected before the softmax operation to generate the approximation, which avoids calculating the full softmax matrix S. The Nystrom approximation thus scales linearly (O(n) complexity) with regard to input sequence length in terms of both memory and time.



Proposed architecture of efficient self-attention via Nyström approximation

Given an input key K and query Q, the proposed Nyströmformer first uses Segment-means to compute landmark points. Based on the landmark points, the architecture then calculates the Nyström approximation using approximate Moore- Penrose pseudoinverse.

To evaluate the model, the researchers conducted experiments in transfer learning setting in two stages. In the first, Nyströmformer was trained on BookCorpus and

English Wikipedia data. Next, the pretrained Nyströmformer was fine-tuned for different NLP tasks on the GLUE (General Language Understanding Evaluation) benchmark datasets (SST-2, MRPC, QNLI, QQP and MNLI) and IMDB reviews. For both stages, the baseline was popular transformer model BERT.

self-attention	input sequence length n									
	512		1024		2048		4096		8192	
	memory (MB)	time(ms)	memory (MB)	time (ms)	memory (MB)	time (ms)	memory (MB)	time (ms)	memory (MB)	time (ms)
Transformer	54 (1×)	0.8 (1×)	186 (1×)	2.4 (1×)	685 (1×)	10.0 (1×)	2620 (1×)	32.9 (1×)	10233 (1×)	155.4 (1×)
Linformer-256	41 (1.3×)	0.7 (1.1×)	81 (2.3×)	1.3 (1.8×)	165 (4.2×)	2.7 (3.6×)	366 (7.2×)	5.3 (6.2×)	635 (16.1×)	11.3 (13.8×)
Longformer-257	32.2 (1.7×)	2.4 (0.3×)	65 (2.9×)	4.6 (0.5×)	130 (5.3×)	9.2 (1.0×)	263 (10.0×)	18.5 (1.8×)	455 (22.5×)	36.2 (4.3×)
Nyströmformer-64	35 (1.5×)	0.7 (1.1×)	63 (3.0×)	1.3 (1.8×)	118 (5.8×)	2.7 (3.6×)	229 (11.5×)	5.9 (5.6×)	450 (22.8×)	12.3 (12.7×)
Nyströmformer-32	26 (2.1×)	0.6 (1.2×)	49 (3.8×)	1.2 (1.9×)	96 (7.1×)	2.6 (3.7×)	193 (13.6×)	5.6 (5.9×)	383 (26.7×)	11.5 (13.4×)

Memory consumption and running time results on various input sequence lengths

Model	SST-2	MRPC	QNLI	QQP	MNLI m/mm	IMDB
BERT-base	90.0	88.4	90.3	87.3	82.4/82.4	93.3
Nyströmformer	91.4	88.1	88.7	86.3	80.9/82.2	93.2

Results on natural language understanding tasks. F1 score for MRPC and QQP and accuracy for others.

The results show that Nyströmformer offers favourable memory and time efficiency over standard self-attention and Longformer self-attention, and performs competitively with the BERT-base model. Overall, Nyströmformer provides self-attention approximation with high efficiency, a big step towards running transformer models on very long sequences.

The paper *Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention* is on arXiv.

Author: Hecate He | **Editor**: Michael Sarazen

Transformers NLP Google