Nystromformer Description

Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention, AAAI 2021

Nystromformer official repository: https://github.com/mlpen/Nystromformer

Nystromformer paper link: https://arxiv.org/pdf/2102.03902v3.pdf

Motivation

Transformer has O(n^2) complexity which limits its application to longer sequences. The paper adapts the Nyström method to approximate standard self-attention with O(n) complexity, in the sense of both time and memory.

Algorithm

Standard Self-attention

Formally, an input sequence of n tokens of dimensions d, $X \in \mathbf{R}^{n\times d}$, is projected using three matrices $W_Q \in \mathbf{R}^{n\times d_q}$, $W_K \in \mathbf{R}^{n\times d_k}$ and $W_V \in \mathbf{R}^{n\times d_v}$ to extract feature representations Q, K, and V , referred to as query, key, and value respectively with $d_k = d_q$. The outputs Q, K, V are computed as $Q = XW_Q$, $K = XW_K$, $V = XW_V$. So self-attention can be written as

$$D(Q,K,V) = SV = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{1}$$

The self-attention mechanism requires calculating n^2 similarity scores between each pair of tokens, leading to a complexity of $O(n^2)$ for both memory and time.

Nyström Method for Matrix Approximation

Denote the softmax matrix used in self-attention $S = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right)$ can be written as

$$S = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right) = \begin{bmatrix} A_S & B_S \\ F_S & C_S \end{bmatrix} \tag{2}$$

where $A_S \in \mathbf{R}^{m \times m}$, $B_S \in \mathbf{R}^{m \times (n-m)}$, $F_S \in \mathbf{R}^{(n-m) \times m}$, $C_S \in \mathbf{R}^{(n-m) \times (n-m)}$. A_S is designated to be the sample matrix by sampling m columns and rows from S. S can be approximated via the basic quadrature technique of the Nyström method. By doing singular value decomposition (SVD) on A_S, we can get its Moore-Penrose inverse A_S^+ and reconstruct the self-attention matrix S as

$$\hat{S} = \begin{bmatrix} A_S & B_S \\ F_S & F_S A_S^+ B_S \end{bmatrix} = \begin{bmatrix} A_S \\ F_S \end{bmatrix} A_S^+ \begin{bmatrix} A_S & B_S \end{bmatrix} \tag{3}$$

C_S is approximated by $F_S A_S^+ B_S$. Here, (3) suggests that the n × n matrix S can be reconstructed by sampling m rows $\begin{bmatrix} A_S & B_S \end{bmatrix}$ and m columns $\begin{bmatrix} A_S \\ F_S \end{bmatrix}$ from S and finding the Nyström approximation $\hat{S}$.

Unfortunately, (3) require calculating all entries in QK^T , even though the approximation only needs to access a subset of the columns of S. Thus, its complexity is also $O(n^2)$.

Linearized Self-Attention via Nyström Method

To avoid calculating all the entries in QK^T , the basic idea is to use landmarks $\tilde{K}$ and $\tilde{Q}$ from key K and query Q. Assuming that $Q = [q_1; q_2; \ldots; q_n]$ and $K = [k_1; k_2; \ldots; k_n]$, Nyströmformer select landmarks simply using Segment-means. As we can pad inputs to a length divisible to m, we assume n is divisible by m for simplicity. Let $l = n/m$, landmark points are computed as

$$\tilde{K} = [\tilde{k_1}; \tilde{k_2}; \ldots; \tilde{k_m}],\ (\tilde{k_j} = \sum_{i=(j-1)l+1}^{(j-1)l+m} \frac{k_i}{m},\ j = 1, 2, \ldots, m) \tag{4}$$

$$\tilde{Q} = [\tilde{q_1}; \tilde{q_2}; \ldots; \tilde{q_m}],\ (\tilde{q_j} = \sum_{i=(j-1)l+1}^{(j-1)l+m} \frac{q_i}{m},\ j = 1, 2, \ldots, m) \tag{5}$$

After selecting landmarks, the softmax matrix can be reconstructed as

$$\hat{S} = softmax\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right)\left(softmax\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_q}}\right)\right)^+ softmax\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) \tag{6}$$

Comparing to (3), the three matrices are represented by landmarks as
$\begin{bmatrix} A_S \\ F_S \end{bmatrix} = softmax\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right)$, $A_S = softmax\left(\frac{\tilde{Q}\tilde{K}^T}{\sqrt{d_q}}\right)$,
$\begin{bmatrix} A_S & B_S \end{bmatrix} = softmax\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right)$ respectively.

To accelerate the computation of Moore-Penrose inverse, Nyströmformer use an iterative method to approximate the Moore-Penrose
inverse via efficient matrix-matrix multiplications. A single iteration is calculated as

$$Z_{j+1} = \frac{1}{4} Z_j(13I - A_S Z_j(15I - A_S Z_j(7I - A_S Z_j))) \tag{7}$$

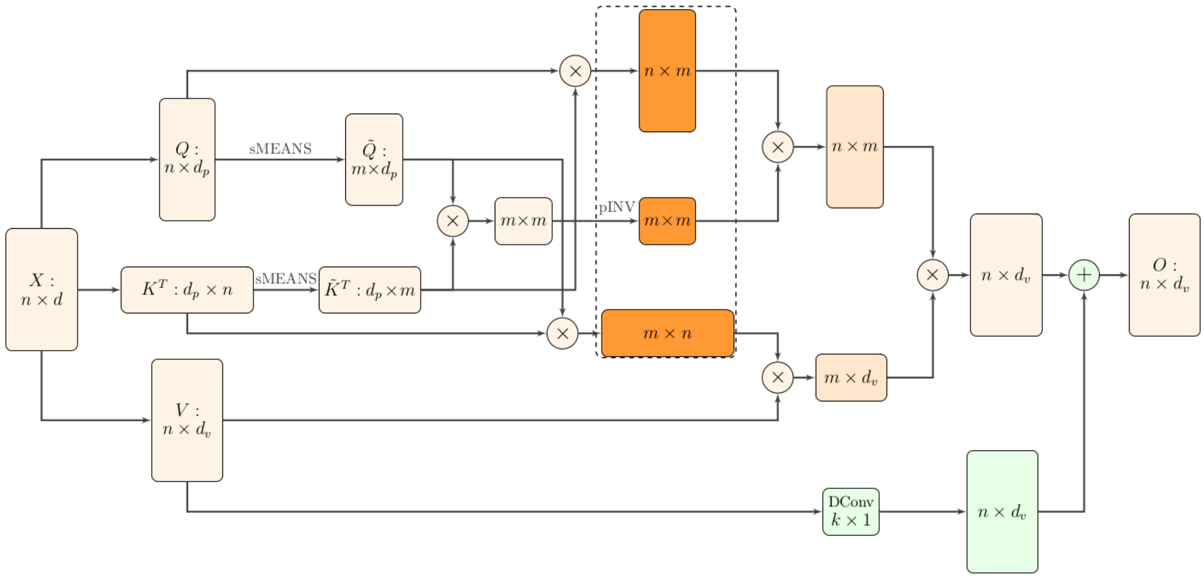
and Z_0 is chosen by

$$Z_0 = \frac{A_S^T}{||A_S||_1\ ||A_S||_\infty}, \tag{8}$$
$$||A_S||_1 = \max_j \sum_{i=1}^{m} |(A_S)_{ij}|,\ ||A_S||_\infty = \max_i \sum_{j=1}^{n} |(A_S)_{ij}|$$

For all the experiments, 6 iterations can achieve a good approximation of the pseudoinverse. Let A_S^+ be approximated by Z^* with (7). Nyströmformer approximation can be written as

$$\hat{S}V = softmax\left(\frac{Q\tilde{K}^T}{\sqrt{d_q}}\right) Z^* softmax\left(\frac{\tilde{Q}K^T}{\sqrt{d_q}}\right) V \tag{9}$$

The proposed Nyströmformer architecture is shown as



Complexity

Time complexity: $O(n + m^3 + nm^2 + mnd_v)$

Memory complexity: $O(md_q + nm + m^2 + mn + nd_v)$

When the number of landmarks m << n, the time and memory complexity of Nyströmformer is O(n).