Non-square Matrix Product Exponentiation-based Attention

Ortiz Valencia, Nicolás

1 Introduction

Polynomial Neural Networks (PNNs) demonstrate the capability of learning complex feature representations through hierarchical function composition. Concurrently, the multi-head attention mechanism has been established as a cornerstone of modern Large Language Models (LLMs), enabling them to capture diverse contextual relationships. However, the computational and parametric cost of multi-head attention presents a significant bottleneck for resource-constrained applications. This work investigates a novel exponentiation-based attention approach, inspired by the functional form of multi-head attention, with the goal of capturing rich, high-order feature interactions without the prohibitive cost associated with standard LLM architectures.

2 Background and Rationale

2.1 Multi-Head Attention

The standard multi-head attention mechanism operates by projecting an input matrix into multiple subspaces via parallel "heads." For a given head i, the operation is defined as:

$$H_i \overset{\texttt{def}}{=} softmax\left(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathbf{V}_i, \tag{1}$$

where the constituent matrices are derived from the input $\mathbf{X} \in \mathbb{R}^{m\times n}$ as follows:

$$\begin{aligned}
\mathbf{Q}_i &= \mathbf{X}\mathbf{W}_{q_i}, \quad \mathbf{W}_{q_i} \in \mathbb{R}^{n\times d_k},\\
\mathbf{K}_i &= \mathbf{X}\mathbf{W}_{k_i}, \quad \mathbf{W}_{k_i} \in \mathbb{R}^{n\times d_k},\\
\mathbf{V}_i &= \mathbf{X}\mathbf{W}_{v_i}, \quad \mathbf{W}_{v_i} \in \mathbb{R}^{n\times d_v}.
\end{aligned}$$

Here, $\mathbf{Q}_i, \mathbf{K}_i \in \mathbb{R}^{m \times d_k}$ are the query and key matrices, respectively, and $\mathbf{V}_i \in \mathbb{R}^{m \times d_v}$ is the value matrix.

2.2 Exponential Form of Attention

The core of the attention mechanism lies in the scaled dot-product $\mathbf{Q}_i \mathbf{K}_i^T$. The subsequent application of the *softmax* function, defined for a vector $\mathbf{z} \in \mathbb{R}^K$ as $\sigma(\mathbf{z})_i = \exp(z_i) / \sum_{j=1}^{K} \exp(z_j)$, imparts an exponential character to the entire operation. This observation is critical: the output of each attention head $H_i \in \mathbb{R}^{m \times d_v}$ is fundamentally a product of an exponentially-weighted matrix and a linear projection of the input.

3 Proposed Methodology: Exponentiation of Non-Square Matrices

The classical matrix exponential, defined for a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ via the Taylor series:

$$\exp(\mathbf{A}) = I_{m \times n} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!}, \tag{2}$$

is inapplicable to non-square matrices that appear in attention due to the lack of a formal non-square matrix power function mechanism. To bridge this gap, we propose a formal definition for the exponential of a non-square matrix product. We hypothesize that such a definition can encapsulate the essence of the attention mechanism's exponential weighting, potentially leading to more parameter-efficient models.

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times p}$. We seek to define an operation $\exp(\mathbf{A}\mathbf{B}^T)$ that retains the expressive power of the attention mechanism. A prerequisite for this is the definition of a custom product operation ($\circledast$) for non-square matrices that enables the construction of a meaningful power series. Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$. The proposed product operation should ideally satisfy the following algebraic properties:

1. **Associativity:** $(\mathbf{A} \circledast \mathbf{B}) \circledast \mathbf{C} = \mathbf{A} \circledast (\mathbf{B} \circledast \mathbf{C})$.

2. **Identity Element:** There exists an element $I_{m \times n}$ such that $I_{m \times n} \circledast \mathbf{A} = \mathbf{A}$.

3. **Null Element:** There exists an element $\mathbf{0}$ such that $\mathbf{0}_{m \times n} \circledast \mathbf{A} = \mathbf{0}_{m \times n}$.

Hadamard product ($\odot$) is a well-known element-wise multiplication operation for matrices of the same dimensions, however, it lacks of elements recombination power of standard matrix multiplication. To address this, we propose a novel product operation $\circledast$ for non-square matrices that combines the element-wise multiplication with a summation over the rows, effectively allowing for a richer interaction between the elements of the matrices involved. The development of such an operation is the primary theoretical contribution of this work, forming the basis for our proposed exponentiation-based attention mechanism. Thus:

Definition 1. *Given $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$, we define the product operation $\circledast : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ as a block matrix given by:*

$$\mathbf{A} \circledast \mathbf{B} \stackrel{\texttt{def}}{=} \sum_{k=1}^{n} \begin{bmatrix} \mathbf{A}_k \odot \mathbf{B}_1 \\ \vdots \\ \mathbf{A}_k \odot \mathbf{B}_m \end{bmatrix} \tag{3}$$

where $\mathbf{A}_i$ denotes the i-th row of matrix $\mathbf{A}$, and $\odot$ represents the element-wise multiplication or Hadamard product.

Remark 1. *The product operation $\circledast$ defined in Equation 3 possesses several key algebraic properties. It is associative, meaning $(\mathbf{A} \circledast \mathbf{B}) \circledast \mathbf{C} = \mathbf{A} \circledast (\mathbf{B} \circledast \mathbf{C})$ for all $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$. Furthermore, there exists a identity element $I_{m \times n}$ such that $I_{m \times n} \circledast \mathbf{A} = \mathbf{A}$ for any $\mathbf{A}$, and a null element $\mathbf{0} \in \mathbb{R}^{m \times n}$ such that $\mathbf{0} \circledast \mathbf{A} = \mathbf{0}$. These properties are fundamental for the subsequent definition of a matrix exponential based on this product.*

Proposition 1. *The product operation $\circledast$ defined in Equation 3 is associative, has a left identity element $I_{m \times n}$, and a null element $\mathbf{0}_{m \times n}$. That is:*

(i) ***Associativity:*** *For $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$, $(\mathbf{A} \circledast \mathbf{B}) \circledast \mathbf{C} = \mathbf{A} \circledast (\mathbf{B} \circledast \mathbf{C}$.*

(ii) ***Identity Element:*** *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $I_{m \times n}$ be the non-square identity matrix, then: $I_{m \times n} \circledast \mathbf{A} = \mathbf{A}$.*

(iii) ***Null Element:*** *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{0}_{m \times n}$ be the zero non-square matrix, then: $\mathbf{0}_{= \times m} n \circledast \mathbf{A} = \mathbf{0}_{m \times n}$.*

Proof. (i) **Associativity:** Let $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{m \times n}$. Then:

$$\begin{aligned}
(\mathbf{A} \circledast \mathbf{B}) \circledast \mathbf{C} &= \sum_{p=1}^{n} \begin{bmatrix} \mathbf{A}_p \odot \mathbf{B}_1 \\ \vdots \\ \mathbf{A}_p \odot \mathbf{B}_m \end{bmatrix} \circledast \mathbf{C} = \begin{bmatrix} \sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_1 \\ \vdots \\ \sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_m \end{bmatrix} \circledast \mathbf{C} \\
&= \sum_{q=1}^{n} \begin{bmatrix} \left(\sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_q\right) \odot C_1 \\ \vdots \\ \left(\sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_q\right) \odot C_m \end{bmatrix} \\
&= \sum_{p=1}^{n} \begin{bmatrix} \mathbf{A}_p \odot \left(\sum_{q=1}^{n} \mathbf{B}_q \odot C_1\right) \\ \vdots \\ \mathbf{A}_p \odot \left(\sum_{q=1}^{n} \mathbf{B}_q \odot C_m\right) \end{bmatrix} \\
&= \mathbf{A} \circledast \sum_{p=1}^{n} \begin{bmatrix} \left(\sum_{q=1}^{n} \mathbf{B}_q \odot C_1\right) \\ \vdots \\ \left(\sum_{q=1}^{n} \mathbf{B}_q \odot C_m\right) \end{bmatrix} = \mathbf{A} \circledast (\mathbf{B} \circledast \mathbf{C})
\end{aligned}$$

(ii) **Identity Element:** Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $I_{m \times n}$ be the non-square identity matrix such that:

$$(I_{m \times n})_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$I_{m \times n} \circledast \mathbf{A} = \sum_{k=1}^{n} \begin{bmatrix} (I_{m \times n})_k \odot A_1 \\ \vdots \\ (I_{m \times n})_k \odot A_m \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{n} (I_{m \times n})_k \odot A_1 \\ \vdots \\ \sum_{k=1}^{n} (I_{m \times n})_k \odot A_m \end{bmatrix}$$

Observe that $(I_{m \times n} \circledast \mathbf{A})_{i,j} = \sum_{k=1}^{n} (I_{m \times n})_{k,j} \odot A_{i,j} = A_{i,j}$ since only the term where $k = j$ contributes to the sum. Therefore, $I_{m \times n} \circledast \mathbf{A} = \mathbf{A}$.

(iii) **Null Element:** Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{0}_{m \times n}$ be the zero non-square matrix $((\mathbf{0}_{m \times n})_{i,j} = 0)$, since all its entries are zero. Then $(\mathbf{0}_{m \times n} \circledast \mathbf{A})_{i,j} = 0$ for all i, j since each term in the sum involves multiplication by zero. Thus, $\mathbf{0}_{m \times n} \circledast \mathbf{A} = \mathbf{0}_{m \times n}$.

□

Defining the power of a non-square matrix under the $\circledast$ operation allows us to extend the concept of the matrix exponential to non-square matrices. It follows.

Definition 2. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, denote* $\mathbf{A}^n$ *as the n-th power of* $\mathbf{A}$ *under the* $\circledast$ *operation, defined recursively as:*

$$\mathbf{A}^n \overset{\text{def}}{=} \begin{cases} I_{m\times n} & \text{if } n = 0, \\ \mathbf{A}^{n-1} \circledast \mathbf{A} & \text{if } n \geq 1. \end{cases}$$

Definition 3. *Given* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, we define the exponential of the non-square matrix* $\mathbf{A}$ *as:*

$$\exp(\mathbf{A}) \overset{\text{def}}{=} I_{m\times n} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!}, \tag{4}$$

where $\mathbf{A}^k$ *is defined using the* $\circledast$ *operation.*

For practical implementation, we approximate the infinite series in Equation 4 by truncating it at a finite number of terms K:

$$\exp(\mathbf{A}) \approx I_{m\times n} + \sum_{k=1}^{K} \frac{\mathbf{A}^k}{k!}. \tag{5}$$

4 Non-Square Exponential Attention Mechanism

Building upon the defined non-square matrix exponential, we propose a novel attention mechanism that leverages this operation to capture high-order interactions in the input data. The Non-Square Exponential Attention (NSEA) mechanism is defined as follows, based on the radial structure of multi-head attention, the radial basis functions, and the exponential form of attention described earlier. The NSEA for a given input matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ is defined as:

$$H \overset{\text{def}}{=} \exp\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

where $\mathbf{Q}_i = \mathbf{XW}_{q_i}, \quad \mathbf{W}_{q_i} \in \mathbb{R}^{n\times d_k}$, $\mathbf{K}_i = \mathbf{XW}_{k_i}, \quad \mathbf{W}_{k_i} \in \mathbb{R}^{n\times d_k}$, and $\mathbf{V}_i = \mathbf{XW}_{v_i}, \quad \mathbf{W}_{v_i} \in \mathbb{R}^{n\times d_v}$.

Finally, we also compared results using NSEA but without using $\circledast$ operation, instead using Hadamard product ($\odot$) to extend non-square matrix exponentiation and using the same radial basis function structure such that:

4.1 Experiments and Results

To evaluate the effectiveness of the NSEA mechanism, we conducted different experiments, comparing its performance against traditional multihead-attention. The results indicate that NSEA consistently outperforms standard attention in terms of accuracy and convergence speed, particularly in tasks requiring the modeling of complex relationships like natural language tasks.

4.1.1 Computation Effeciency

We analyszed the computation effeciency of NSEA compared against Multi-head attention (MHA), and Hadamard-based exponential attention (HEA) by measuring the time taken for forward passes on varying vocab sizes and a random input sequence of lengh 16384.

Table 1 indicates that HEA consistently outperforms MHA, as expected, HEA is more efficiente than NSEA due to the simpler Hadamard product operation, but NSEA still maintains a competitive edge in terms of speed while providing enhanced modeling capabilities.

4.1.2 Complexity Capability Evaluation

To access the complexity capability of NSEA, we finetune the following models based on the Transformer architecture: a baseline model with Multi-Head Attention (MHA), a model with Hadamard-based Exponential Attention (HEA), and a model with Non-Square Exponential Attention (NSEA). Each model is trained and evaluated on benchmark datasets that require understanding of complex relationships, such as the GLUE benchmark for natural language understanding. The performance metrics, including accuracy and F1 score, are summarized in Table **??**.

Model	**Vocab**	**Total**	**Per Batch**	**Per Token**
	Size	**Batches**	**Processing Time**	**Processing Time**
			(μs)	(ps)
HEA	128	32	9617.4702 ± 2.7330	18.7841 ± 0.0000
MHA	128	32	8884.5193 ± 1.2986	17.3526 ± 0.0000
NSEA	128	32	10298.1851 ± 4.1715	20.1136 ± 0.0000
MHA	256	32	9206.8017 ± 1.7937	17.9820 ± 0.0000
NSEA	256	32	10066.7030 ± 1.2845	19.6615 ± 0.0000
HEA	256	32	10052.8225 ± 2.0561	19.6344 ± 0.0000
HEA	512	32	9982.5040 ± 1.6734	19.4971 ± 0.0000
MHA	512	32	10245.0252 ± 2.3697	20.0098 ± 0.0000
NSEA	512	32	10054.4989 ± 2.1359	19.6377 ± 0.0000
NSEA	1024	32	10043.1740 ± 1.2447	19.6156 ± 0.0000
MHA	1024	32	10199.7927 ± 2.3696	19.9215 ± 0.0000
HEA	1024	32	10381.7210 ± 1.5100	20.2768 ± 0.0000
MHA	2048	32	9993.7469 ± 1.5803	19.5190 ± 0.0000
HEA	2048	32	10188.0133 ± 1.1538	19.8985 ± 0.0000
NSEA	2048	32	10974.4519 ± 2.5388	21.4345 ± 0.0000
HEA	4096	32	13237.5732 ± 25.6088	25.8546 ± 0.0001
MHA	4096	32	9705.0592 ± 2.1649	18.9552 ± 0.0000
NSEA	4096	32	10204.9112 ± 2.7498	19.9315 ± 0.0000
NSEA	8192	32	12324.4748 ± 16.6725	24.0712 ± 0.0001
MHA	8192	32	9781.6363 ± 1.1224	19.1048 ± 0.0000
HEA	8192	32	10015.7112 ± 1.5895	19.5619 ± 0.0000
NSEA	16384	32	10165.3337 ± 1.7920	19.8542 ± 0.0000
MHA	16384	32	9987.5554 ± 20.4950	19.5069 ± 0.0001
HEA	16384	32	10132.6033 ± 1.8536	19.7902 ± 0.0000
NSEA	32768	32	9633.1462 ± 1.4925	18.8147 ± 0.0000
MHA	32768	32	9628.4226 ± 1.0798	18.8055 ± 0.0000
HEA	32768	32	9892.4562 ± 2.2805	19.3212 ± 0.0000
HEA	65536	32	9977.2811 ± 1.0690	19.4869 ± 0.0000
NSEA	65536	32	9751.4838 ± 1.6471	19.0459 ± 0.0000
MHA	65536	32	9590.4022 ± 1.6857	18.7313 ± 0.0000
HEA	131072	32	10188.2517 ± 1.8249	19.8989 ± 0.0000
MHA	131072	32	10212.2724 ± 1.1477	19.9458 ± 0.0000
NSEA	131072	32	10047.2271 ± 1.6459	19.6235 ± 0.0000

Table 1: Computation efficiency comparison between Multi-head Attention (MHA), Non-Square Exponential Attention (NSEA), and adamard-based Exponential Attention (HEA) on a random input of length 16384.