Non-square Matrix Product Exponentiation-based Attention

Nicolas Ortiz Valencia,
Ruben Dario Hernandez Beleño,
Mario Aldape Pérez

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, yet their deployment is often hindered by substantial computational and memory demands, particularly due to the multi-head attention mechanism. This paper introduces NSEA-Attention, a novel exponentiation-based attention mechanism designed to capture high-order feature interactions while significantly reducing parameter count. We formulate a generalized matrix exponentiation operation for non-square matrices, constructing an attention mechanism, which can emulates the information mixing behavior of standard attention with enhanced efficiency. Our experiments demonstrate a substantial reduction in model size and computation overhead of about 257×. This work paves the way for more efficient LLMs, facilitating their deployment in resource-constrained environments without compromising performance.

1 Introduction

Attention-based mechanisms have 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 in the development of LLMs. Its canonical formulation computes a pairwise affinity matrix over all input tokens, leading to quadratic complexity $O(n^2)$ in both computation and memory with respect to sequence length n (Niu, Zhong, & Yu, 2021). Despite modern advances in efficient attention computation, which lead to linear complexity (A. Liu et al., 2025), the number of parameters and overall compute remains a critical impediment for training on long contexts and, more pertinently, for the deployment of these models in

resource-constrained environments.

In response, compression and acceleration techniques have emerged as a strategy for compressing large LLMs into smaller, faster, and computationally efficient models, like knowledge distillation in TinyBERT, which achieves more than 96.8% performance of its teacher BERTBASE on GLUE benchmark (Wang et al., 2018), while being **7.5x** smaller and **9.4x** faster on inference (Jiao et al., 2020). Common and powerful model compression and acceleration techniques include: parameter prunning and sharing; removing inessential parameters without any significant effect on the model performace, low-rank factorization; identify redundant parameters and decompose then into smaller matrices, convolutional filters transferring; removing inessential parameters by transferring or compressing convolutional filters, and knowledge distillation (KD); distilling knowledge from larger models into small models (Gou, Yu, Maybank, & Tao, 2021). However, these strategies remain limited by the inherent inefficiencies of the attention mechanism itself, which is typically retained in its original form within the student model (Jiao et al., 2020; Pan et al., 2021). Additionally, modeling different types of knowledge in a unified and complementary framework is still challenging; small models do not capture all complex language relationships associated with their homologous large versions (Gou et al., 2021).

To address these challenges, 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. We propose a fundamentally new approach to attention, which we term NSEA-Attention for Non-square Matrix Product Exponentiation-based Attention. Our key insight is to reconceptualize attention not as a static affinity computation, but as a parameterized sequence of non-square matrix multiplications that can effectively emulate the information mixing behavior of standard attention and capture complex relationships, while being significantly more efficient in terms of parameters. For this end, we design a matrix product exponentiation process, $\exp\left((\mathbf{XW}+\mathbf{B})^2\mathbf{\Lambda}\right)$, where W and $\mathbf{\Lambda}$ are deliberately non-square matrices that project input tokens into asymmetric latent spaces. This exponentiation process inherently encodes a form of exponential decay or amplification of learned relationships over k steps (Taylor series expansion approximation $\exp(\mathbf{A}) = I_{m\times n} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!}$), mimicking the way standard attention propagates influence, but doing so through a deep, compositional parameterized pathway. We hypothesize that this allows controlling specialized representation shaping, enabling the model to learn a compressed yet semantically rich attention mechanism. By doing so, we can reduce the number of Attention layers and,

in consequence, reduce the number of model parameters without lossing model's modeling capabilities.

The remainder of this article is structured as follows: We first provide necessary background on attention. We then detail the mathematical formulation and architecture of NSEA-Attention. Following this, we present our proposed methodology. Experimental results demonstrate the efficacy of our approach in creating a highly efficient, high-performing attention mechanism. We conclude with an analysis and a discussion of future work.

2 Background and Rationale

Transformer architectures, built on attention mechanisms, are now fundamental to modern deep learning, particularly in the fields of natural language processing and computer vision. In particular, the multi-head attention mechanism, as introduced in the seminal work by Vaswani et al. (2017), allows models to focus on different parts of the input data simultaneously, enhancing their ability to capture complex patterns and dependencies. Nonetheless, pre-trained attention-based models are computationally inefficient, require substantial amounts of data, and need to be fine-tuned for specific tasks (Lin, Wang, Liu, & Qiu, 2022).

In this section, we provide a concise overview of the standard attention mechanism, emphasizing its mathematical formulation and the role of exponentiation within it. This sets the stage for our proposed methodology, which seeks to redefine attention through the lens of non-square matrix exponentiation.

2.1 The standard attention mechanism

The standard attention mechanism

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

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

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.

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 as showed in Equation 2.

$$H_i \stackrel{\texttt{def}}{=} \begin{pmatrix} \frac{\exp^z_{1,1}}{\sum_{j=1}^{m} \exp^z_{1,j}} & \cdots & \frac{\exp^z_{1,m}}{\sum_{j=1}^{m} \exp^z_{1,j}} \\ \vdots & \ddots & \vdots \\ \frac{\exp^z_{m,1}}{\sum_{j=1}^{m} \exp^z_{m,j}} & \cdots & \frac{\exp^z_{m,m}}{\sum_{j=1}^{m} \exp^z_{m,j}} \end{pmatrix} \mathbf{V}_i, \tag{2}$$

Finally, the model then concatenates all the outputs and projects them back to a m-dimensional representation as follows:

$$MHA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concat}(H_1, \ldots, H_h)\mathbf{W}_o, \tag{3}$$

where $\mathbf{W}_o \in \mathbb{R}^{d_v \times d_m}$. In essence, the attention mechanisms select, modulate, and focus on the information most relevant to behavior. According to de Santana Correia and Colombini (2022), attention has existed for at least three decades and has been applied in various domains, including computer vision, natural language processing, and speech recognition, among others. Modern approaches to attention mechanisms are primarily based on the Transformer architecture (Vaswani et al., 2017), which relies on self-attention to model relationships between different parts of the input data. However, as mentioned earlier, the computational and parametric cost present significant challenges, and in response, other attention variants have been proposed, such as sparse attention, linear attention, prototype and memory Compression, low-rank attention, and attention with prior (Lin et al., 2022).

Although several variants of attention mechanisms exist, the multi-head attention mechanism remains the most widely adopted due to its effectiveness in capturing diverse patterns (Bahdanau, Cho, & Bengio, 2015a, 2015b; Hasan et al., 2024; Li et al., 2025; Y. Liu, He, & Hui, 2025; Sharifi & Safari, 2025; Xiong et al., 2025; Zhang, Feng, Wang, Lu, & Mei, 2025).

2.2 Attention Variants Taxonomy

2.2.1 Sparse Attention

Sparse attention mechanisms aim to reduce the computational burden of standard attention by limiting the number of interactions between tokens. Techniques such as local attention, where each token attends only to its neighboring tokens, and global attention, which allows certain tokens to attend to all others, are common strategies (Child, Gray, Radford, & Sutskever, 2019; Y. Guo et al., 2019; Xu et al., 2021). These methods significantly decrease the number of computations required, making them more efficient for long sequences.

2.2.2 Linearized atttention

Let $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n \times d_k}$, the complexity of computing $softmax(\mathbf{Q}\mathbf{K}^T)\mathbf{V}$ is quadratic, then we can try to disentangle the equation into $\mathbf{Q}'\mathbf{K}'^T$, and we can compute $\mathbf{Q}\mathbf{K}^T$ in reverse order ($\mathbf{Q}(\mathbf{K}^T\mathbf{V})$) leading to linear complexity. This is the main idea behind linearized attention mechanisms, which approximate the attention computation using kernel methods or low-rank approximations (Choromanski et al., 2022; Feng, Xu, Jiang, Liu, & Zheng, 2022; Katharopoulos, Vyas, Pappas, & Fleuret, 2020). This could be especially beneficial for autoregressive attention, where the model generates sequences token by token, as it allows for efficient computation without sacrificing performance, and also enables Transformer decoders to run like RNNs (Lin et al., 2022)

2.2.3 Query Prototyping and Memory Compression

Apart from the variants mentioned above, other approaches focus on compressing the attention mechanism itself by reducing the number of queries or keys used in the attention computation. Either the queries are selected from a subset of representative tokens (prototyping) or the keys and values are compressed into a smaller set of memory slots. These methods aim to retain the most salient information while reducing the overall computational load. They can lower the quadratic complexity of self-attention to linear or near-linear scaling, enabling more efficient processing of long inputs without substantial loss in performance.

2.2.4 Low-Rank Attention

Consider the attention matrix $\mathbf{A} = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{(d_k)}}\right)$. Theoretical analysis and empirical evidence suggest that $\mathbf{A}$ often exhibits low-rank properties (Q. Guo, Qiu, Xue, & Zhang, 2019), this means that the matrix can be approximated by a product of two smaller matrices, $\mathbf{A} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$

with $r << min(m, n)$. Decomposing the attention matrix into smaller components reduces the number of parameters and computation cost while preserving complex relationships in the data.

2.2.5 Attention with prior

Attention with prior enhances the standard self-attention mechanism by incorporating external or pre-existing knowledge into the attention distribution, rather than relying solely on query-key similarity. This prior can take various forms: it may encode structural information such as positional relationships (e.g., via trainable positional biases or Gaussian locality biases) (Raffel et al., 2020), reuse attention patterns from previous layers (acting as a form of residual or convolutional prior) (Esearch & Koppius, 2011), or even serve as a task-specific adapter in transfer-learning setups (Ying, Ke, He, & Liu, 2021). In some cases, the prior can entirely replace the dynamically generated attention—for example, using a fixed uniform or Gaussian distribution—which simplifies computation and can improve efficiency. Experiments show that the resulting model remains effective while being much more efficient to compute (Lin et al., 2022).

2.2.6 Rationale

While these variants significantly improve atttention-based models efficiency while preserving or even enhancing performance, they often introduce additional complexity in terms of implementation and hyperparameter tuning. Moreover, many of these methods still rely on the fundamental formulation of the standard attention mechanism, which may limit their ability to fully exploit the potential of alternative mathematical formulations, and even more, they enhance performance but lose the relationship inference power of the original attention mechanism. This motivates our exploration into a novel exponentiation-based attention mechanism that seeks to redefine the attention operation itself, aiming for a more parameter-efficient and theoretically grounded approach.

3 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:

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

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:*

$$\mathbf{A} \circledast \mathbf{B} \stackrel{\texttt{def}}{=} \sum_{k=1}^{n} \begin{bmatrix} \mathbf{A}_k \odot \mathbf{B}_1 \\ \vdots \\ \mathbf{A}_k \odot \mathbf{B}_m \end{bmatrix} \tag{5}$$

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 5 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 5 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:

$$(\mathbf{A} \circledast \mathbf{B}) \circledast \mathbf{C} = \sum_{p=1}^{n} \begin{bmatrix} \mathbf{A}_p \odot \mathbf{B}_1 \\ \vdots \\ \mathbf{A}_p \odot \mathbf{B}_m \end{bmatrix} \circledast \mathbf{C} = \begin{bmatrix} \sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_1 \\ \vdots \\ \sum_{p=1}^{n} \mathbf{A}_p \odot \mathbf{B}_m \end{bmatrix} \circledast \mathbf{C}$$

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

(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 ($(\mathbf{0}_{m\times n})_{i,j} = 0$). 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:*

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

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

$$\exp(\mathbf{A}) \stackrel{\texttt{def}}{=} I_{m\times n} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!}, \tag{6}$$

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

For practical implementation, we approximate the infinite series in Equation 6 by truncating it at a finite number of terms K:

$$\exp(\mathbf{A}) \approx I_{m\times n} + \sum_{k=1}^{K} \frac{\mathbf{A}^k}{k!}. \tag{7}$$

3.1 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. Similarly to the linearized attention mechanism, we aim to capture complex relationships by preserving a radial basis function kernel structure ($e^{-\|x\|^2\lambda}$), and preserving attention's exponential nature and dimensionality.

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:

$$H \overset{\texttt{def}}{=} \exp\left((\mathbf{X}\mathbf{W} + \mathbf{B})^2 \mathbf{\Lambda}\right) \tag{8}$$

where $\mathbf{W} \in \mathbb{R}^{n \times d_k}$, $\mathbf{\Lambda} \in \mathbb{R}^{d_k \times d_v}$. Where $\mathbf{\Lambda}$ is a learnable diagonal matrix that scales the contributions of each dimension in the transformed space, allowing the model to adaptively focus on different aspects of the input features. The bias term $\mathbf{B} \in \mathbb{R}^{m \times d_k}$ is included to provide additional flexibility in the transformation, enabling the model to better capture complex patterns in the data. The output $H \in \mathbb{R}^{m \times d_v}$ represents the attention-weighted representations of the input tokens, where d_v is the dimensionality of the value vectors.

4 Experiments and Results

To evaluate the efficiency of the NSEA mechanism, we compare its performance against traditional multi-head attention.

4.0.1 Computation Effeciency

We analyszed the computation effeciency of NSEA compared against Multi-head attention (MHA), by measuring the FLOPs taken for forward passes on varying vocab sizes and batch length for a random input sequence of lengh $16384 = (4096 \times 4)$, an wmbedding size of 2048 (DeeepSeek's maximm sequence length and embedding size (A. Liu et al., 2025)), and a 4-degree Taylor aproxximation on a computer with 24 GB of RAM with a speed of $4800MT/s$, and a 13^{th} Generation Intel(R) Core(TM) i5-3420H. Figure 1 illustrates the testing setup.

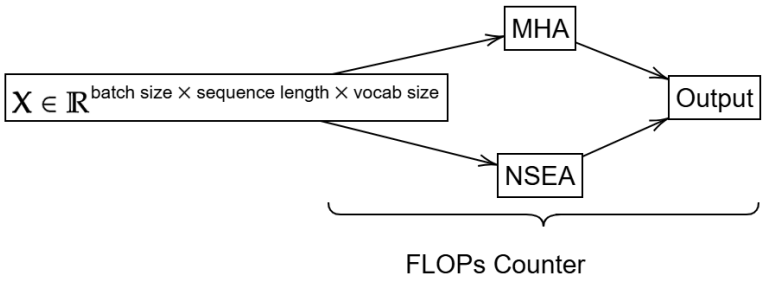

Figure 1: Computation Efficiency Testing Setup

Model	Vocab Size	Total Batches	FLOPs Per Batch	FLOPs Per Token	compression Ratio
MHA	131072	64	8623489024	33685504	
NSEA	131072	64	33554432	131072	257

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.

Table 1 indicates that HEA outperforms MHA. Results prove that NSEA is significantly more efficient than MHA in terms of FLOPs, with a compression ratio of approximately 257 times. This efficiency gain is attributed to the reduced computational complexity of the non-square exponential operation compared to the traditional attention mechanism, which involves more intensive matrix multiplications and softmax operations.

5 Conclusion and future directions

The fundamental goal of this work was to design an efficient attention mechanism without losing the state-of-the-art complexity abstraction of standard multi-head attention. By designing a new matrix exponentiation formulation, we were able to reduce the computational complexity compuation of attention while maintaining expressivity. Our experiments demonstrate that the proposed method is able to compress attention computation 257× more than standard attention, making it suitable for long sequence tasks on transformer architectures. We also prove some useful properties that can be used to define the exponential function and other related functions for non-square matrices, opening new research directions in the field of matrix functions and attention-based architectures. Moreover, we believe that our work can be extended in several ways. Future research could explore the integration of our efficient attention mechanism with other transformer variants, such as sparse attention or adaptive attention mechanisms. Additionally, investigating the impact of our method on different modalities, such as vision or audio, could provide further insights into its versatility and effectiveness. Finally, exploring theoretical aspects of matrix functions in the context of deep learning could lead to new advancements in model design and optimization techniques. For example, one could think of exploring whether the proposed attention gradient is bounded, such that gradient explosions are avoided during training. Also, investigating the convergence properties of models utilizing our attention mechanism could yield valuable theoretical insights. Overall, our work lays a solid foundation for future explorations in efficient attention mechanisms and their applications

across various domains.

In conclusion, we have presented a novel approach to efficient attention computation that significantly reduces complexity while preserving the expressiveness of traditional multi-head attention mechanisms. Our method not only advances the state-of-the-art in attention mechanisms but also opens up new avenues for research in matrix functions and their applications in deep learning.

Acknowledgements

We would like to express our sincere gratitude to all those who contributed to this research. Special thanks to our colleagues and mentors for their invaluable feedback and support throughout the development of this work. We also acknowledge a PhD. Santiago Herce Castañón and PhD Student Fernando Avitúa Varela for the opportunity to discuss our work at 'Centro de la Complejidad UNAM', which provided us with new perspectives and insights. Finally, we extend our appreciation to the National Polytechnic Intitute and CONACYT funding that supported this research, enabling us to pursue our goals and contribute to the field of efficient attention mechanisms.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015a). Rnn encoder-decoder with attention. *Computer Science*.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015b). seq2seq + attention. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*.

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). *Generating long sequences with sparse transformers.* Retrieved from `https://arxiv.org/abs/1904.10509`

Choromanski, K., Lin, H., Chen, H., Zhang, T., Sehanobish, A., Likhosherstov, V., ... Weingarten, T. (2022). From block-toeplitz matrices to differential equations on graphs: towards a general theory for scalable masked transformers. In *Proceedings of machine learning research* (Vol. 162).

de Santana Correia, A., & Colombini, E. L. (2022). Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, *55*. doi: 10.1007/s10462-022-10148-x

Esearch, S. Y. R., & Koppius, O. R. (2011). Predictive attention transformer: Improving transformer with attention map prediction. *MIS Quarterly*, *35*.

Feng, Y., Xu, H., Jiang, J., Liu, H., & Zheng, J. (2022). Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*. doi: 10.1109/TGRS.2022.3168331

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*. doi: 10.1007/s11263-021-01453-z

Guo, Q., Qiu, X., Xue, X., & Zhang, Z. (2019). Low-rank and locality constrained self-attention for sequence modeling. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *27*. doi: 10.1109/TASLP.2019.2944078

Guo, Y., Ji, J., Lu, X., Huo, H., Fang, T., & Li, D. (2019). Global-local attention network for aerial scene classification. *IEEE Access*, *7*. doi: 10.1109/ACCESS.2019.2918732

Hasan, M. A., Haque, F., Roy, T., Islam, M., Nahiduzzaman, M., Hasan, M. M., ... Haider, J. (2024). Prediction of fetal brain gestational age using multihead attention with xception. *Computers in Biology and Medicine*, *182*. doi: 10.1016/j.compbiomed.2024.109155

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... Liu, Q. (2020). Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics findings of acl: Emnlp 2020.* doi: 10.18653/v1/2020.findings-emnlp.372

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are rnns: Fast autoregressive transformers with linear attention. In *37th international conference on machine learning, icml 2020* (Vol. PartF168147-7).

Li, D., Neira-Molina, H., Huang, M., Syam, M. S., Zhang, Y., Junfeng, Z., ... Awwad, E. M. (2025). Cstfnet: A cnn and dual swin-transformer fusion network for remote sensing hyperspectral data fusion and classification of coastal areas. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*. doi: 10.1109/JSTARS.2025.3530935

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*. doi: 10.1016/j.aiopen.2022.10.001

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... Pan, Z. (2025). *Deepseek-v3 technical report.* Retrieved from `https://arxiv.org/abs/2412.19437`

Liu, Y., He, M., & Hui, B. (2025). Eso-detr: An improved real-time detection transformer model for enhanced small object detection in uav imagery. *Drones*, *9*. doi: 10.3390/drones9020143

Niu, Z., Zhong, G., & Yu, H. (2021, 9). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62. Retrieved from `https://www.sciencedirect.com/science/article/abs/`

pii/S092523122100477X?via%3Dihub doi: 10.1016/J.NEUCOM.2021.03.091

Pan, H., Wang, C., Qiu, M., Zhang, Y., Li, Y., & Huang, J. (2021). Meta-kd: A meta knowledge distillation framework for language model compression across domains. In *Acl-ijcnlp 2021 - 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, proceedings of the conference* (Vol. 1). doi: 10.18653/v1/2021.acl-long.236

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*.

Sharifi, A., & Safari, M. M. (2025). Enhancing the spatial resolution of sentinel-2 images through super-resolution using transformer-based deep-learning models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*. doi: 10.1109/JSTARS.2025.3526260

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 2017-December). doi: 10.1201/9781003561460-19

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Emnlp 2018 - 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp, proceedings of the 1st workshop.* doi: 10.18653/v1/w18-5446

Xiong, X., Zhang, X., Jiang, W., Liu, T., Liu, Y., & Liu, L. (2025). Lightweight dual-stream sar-atr framework based on an attention mechanism-guided heterogeneous graph network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*. doi: 10.1109/JSTARS.2024.3498327

Xu, X., Li, J., Guan, Y., Zhao, L., Zhao, Q., Zhang, L., & Li, L. (2021). Glanet: A global-local attention network for automatic cataract classification. *Journal of Biomedical Informatics*, *124*. doi: 10.1016/j.jbi.2021.103939

Ying, C., Ke, G., He, D., & Liu, T.-Y. (2021). *Lazyformer: Self attention with lazy update.* Retrieved from https://arxiv.org/abs/2102.12702

Zhang, X., Feng, Y., Wang, N., Lu, G., & Mei, S. (2025). Transformer-based person detection in paired rgb-t aerial images with vtsar dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*. doi: 10.1109/JSTARS.2025.3526995