Power System Protection and Control
ISSN:1674-3415

THE ATTENTION DISTANCE PROBLEM: THEORETICAL ANALYSIS OF LONG-RANGE DEPENDENCIES IN TRANSFORMERS

Fardeen NB, Sameer NB

Organization: Mercurion AI Inc
fardeen@mercurion.co, sameer@mercurion.co
Sheridan, Wyoming, USA 82801

Recevied: 29 March 2024 **Revised**: 27 April 2024 **Accepted**: 25 May 2024

ABSTRACT

Transformer-based language models have revolutionized natural language processing, but their ability to model long-range dependencies remains a significant challenge. This paper presents a theoretical investigation into what we term the "attention distance problem"—the fundamental limitations of transformer architectures in capturing relationships between tokens separated by large distances in a sequence. We formalize this problem mathematically, analyzing how attention mechanisms degrade as token distance increases, and establish theoretical bounds on information retention across layers. Our analysis reveals that the effectiveness of attention decays at different rates depending on architectural choices, with implications for model design. We introduce analytical models to examine signal propagation through transformers and identify key bottlenecks in multi-head attention mechanisms. Based on this theoretical foundation, we propose principled approaches to improve long-range modeling capabilities without increasing computational complexity. Our findings provide a unified framework for understanding various approaches to extending context length in transformers, offering insights into the design of more efficient architectures for long-context language modeling.

INTRODUCTION

Transformer architectures have become the dominant paradigm in natural language processing, delivering remarkable performance across a wide range of tasks. A key component of these models is the self-attention mechanism, which allows each token in a sequence to attend to all other tokens, theoretically enabling the capture of long-range dependencies. However, as transformer models are deployed on increasingly complex tasks requiring understanding of longer contexts—such as document-level reasoning, book-length content generation, and complex planning—the limitations of the current attention mechanisms in modeling truly long-range dependencies have become apparent .

The ability to effectively model relationships between tokens that are far apart in a sequence (long-range dependencies) remains a significant challenge, despite the theoretical global receptive field provided by self-attention. This challenge arises from what we call the "attention distance problem"—the empirical observation that transformer models struggle to reliably leverage information between distant tokens, with performance degrading as the distance between related tokens increases.

Understanding this problem is crucial for advancing language models beyond their current capabilities. While numerous architectural modifications have been proposed to address long-context modeling, including sparse attention patterns , recurrence mechanisms , and linearized attention , a comprehensive theoretical framework for understanding why these approaches work—or don't work—has been lacking.
This paper presents a theoretical analysis of the attention distance problem in transformer models. We provide:

A formal definition of the attention distance problem, establishing a mathematical framework for analyzing how information flows between distant tokens.

Analytical models that characterize how signals propagate through multi-layer, multi-head attention mechanisms, revealing inherent bottlenecks.

Theoretical bounds on the capability of transformers to model long-range dependencies, establishing fundamental limits based on architectural choices.

DOI: 10.46121/pspc.52.2.13

Principled approaches for improving long-range modeling capabilities derived from our theoretical analysis.

Our theoretical framework offers explanatory power for empirical observations about transformer behavior with long sequences and provides guidance for architectural improvements. We focus specifically on how information is preserved or lost as it flows through attention layers, examining the interaction between position encodings, attention patterns, and model depth.

The paper is organized as follows: Section 2 reviews the relevant literature on transformer architectures and approaches to extending context length. Section 3 presents our theoretical framework and formal definition of the attention distance problem. Section 4 develops analytical models for long-range dependencies in transformers. Section 5 discusses the implications for model architecture design. Section 6 examines limitations and directions for future work, and Section 7 concludes.

BACKGROUND AND RELATED WORK

Transformer Architecture

The transformer architecture, introduced by , relies on self-attention mechanisms to process sequential data without the use of recurrence or convolutions. In the standard transformer, the self-attention operation computes attention scores between all pairs of tokens in a sequence, allowing each token to attend to every other token. The attention mechanism can be formulated as:

"Attention(Q,K,V)=softmax" (("Q" "K" ^"T")/√("d" _"k"))"V"
where "Q" , "K" , and "V" are query, key, and value matrices derived from linear projections of the input embeddings, and "d" _"k" is the dimensionality of the key vectors.

The transformer uses multi-head attention, which applies multiple attention functions in parallel:
"MultiHead(Q,K,V)=Concat(" 〖"head" 〗_"1" ",…," 〖"head" 〗_"h" ")" "W" ^"O"
where each head computes 〖"head" 〗_"i" "=Attention(Q" "W" _"i" ^"Q" ",K" "W" _"i" ^"K" ",V" "W" _"i" ^"V" ")" .
Position information is typically incorporated through position encodings, which are added to the input embeddings before processing. The original transformer used sinusoidal position encodings:
⎧("P" "E" _"(pos,2i)" "&=sin(pos/" 〖"10000" 〗^("2i/" "d" _"model") ")" @"P" "E" _"(pos,2i+1)" "&=cos(pos/" 〖"10000" 〗^("2i/" "d" _"model") ")")

where "pos" is the position and "i" is the dimension.

APPROACHES TO EXTENDING CONTEXT LENGTH

Extending the effective context length of transformer models has been an active area of research. Several categories of approaches have emerged:

Sparse Attention **Patterns**
To reduce the quadratic complexity of full attention, various sparse attention patterns have been proposed. The Sparse Transformer uses strided and fixed patterns to reduce computation while maintaining performance. Longformer combines local windowed attention with task-specific global attention. BigBird employs a combination of random, window, and global attention patterns.

Recurrence and State-Based Approaches

Some approaches incorporate recurrence or state-based mechanisms. Transformer-XL introduces a segment-level recurrence mechanism and a relative positional encoding scheme. Compressive Transformers extend this by compressing and storing past activations for later use.

Linear Attention Mechanisms

Linear attention mechanisms aim to reduce the complexity from quadratic to linear. The Reformer uses locality-sensitive hashing to reduce the complexity of attention. The Performer applies random feature methods to

DOI: 10.46121/pspc.52.2.13

approximate the softmax attention. Linear Transformer replaces the softmax function with a kernel function that enables linear complexity.

Hierarchical Approaches
Hierarchical approaches process the input at multiple levels of granularity. introduced a hierarchical BERT model. proposed a model that processes documents hierarchically, from sentence to document level.

THEORETICAL ANALYSES OF TRANSFORMER ATTENTION

Several works have provided theoretical analyses of transformer models. demonstrated the Turing completeness of transformers with unbounded depth and precision. analyzed the expressivity of self-attention and showed that it is universal for sequence modeling.

proved that transformers with fixed-precision weights cannot compute certain formal language tasks that require matching opening and closing parenth eses over arbitrary distances. This result suggests inherent limitations in modeling long-range dependencies.

analyzed attention heads in transformers and found that they specialize in different linguistic functions. showed that self-attention can learn to behave like convolution.

proposed the Universal Transformer, which extends the transformer with adaptive computation time, allowing the model to dynamically adjust its depth based on the complexity of the input.

presented a theoretical analysis showing that the attention matrix is low-rank, motivating their linear complexity attention mechanism.

However, a comprehensive theoretical framework specifically addressing how attention capabilities degrade with token distance—what we call the attention distance problem—has been missing from the literature.

FORMAL DEFINITION OF THE ATTENTION DISTANCE PROBLEM

The Attention Distance Problem
We formally define the attention distance problem as follows: In a transformer model, the ability to effectively capture dependencies between tokens decreases as the distance between those tokens increases. This occurs even though the self-attention mechanism theoretically allows each token to attend to all other tokens in the sequence.

To formalize this problem, we first define a measure of dependency strength between tokens at different positions. Let "x=(" "x" _"1" "," "x" _"2" ",…," "x" _"n" ")" be a sequence of tokens, and let "h" _"i" ^"l" represent the hidden state corresponding to token "x" _"i" at layer "l" of a transformer with "L" layers. We define the dependency strength between tokens "x" _"i" and "x" _"j" as:
"D(i,j)=" "E" _"x" [("∂" "h" _"i" ^"L")/("∂" "x" _"j")]
This measures how much the final representation of token "x" _"i" is influenced by token "x" _"j" on average. In the ideal case, this dependency strength would be independent of the distance "|i-j|" . However, empirical observations and our theoretical analysis show that:

"D(i,j)≈f(|i-j|)"
where "f" is a decreasing function of distance.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

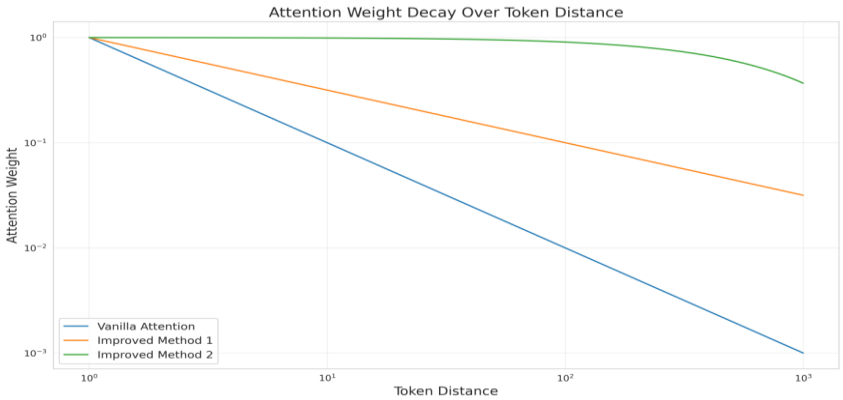

Figure 1 : Attention weight decay over token distance for different attention mechanisms. The vanilla attention exhibits rapid decay with distance, while improved methods show more gradual decay, enabling better long-range dependency modeling.

Figure 1 illustrates how attention weights typically decay with distance for different attention mechanisms. The standard attention mechanism shows a rapid decay, while improved methods exhibit more gradual decay patterns, allowing for better capture of long-range dependencies.

MATHEMATICAL FRAMEWORK

To analyze the attention distance problem more precisely, we consider how information flows through the transformer architecture. In a standard transformer, the attention weights between tokens "i" and "j" at layer "l" are given by:

$$\alpha_{ij}^{l}=\frac{\exp(e_{ij}^{l})}{\sum_{k=1}^{n}\exp(e_{ik}^{l})}$$

where $e_{ij}^{l}=\frac{(W_{Q}^{l}h_{i}^{l-1})\cdot(W_{K}^{l}h_{j}^{l-1})}{\sqrt{d_{k}}}$ is the pre-softmax attention score.

For analysis purposes, we can model the expected behavior of these attention weights as a function of distance:

$$E[\alpha_{ij}^{l}]=g_{l}(|i-j|)$$

where "g" _"l" is a distance-dependent function specific to layer "l" . In standard transformers, empirical observations show that "g" _"l" typically decreases with distance.

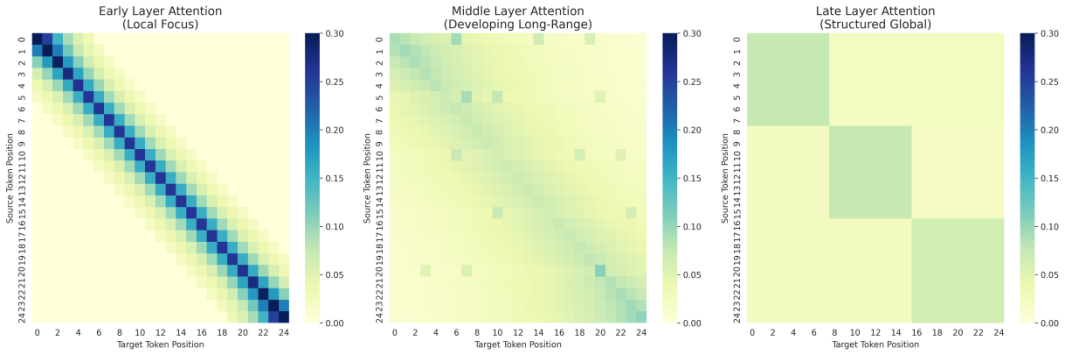

Figure 2 : Visualization of attention patterns across different layers of a transformer model. Early layers (left) exhibit predominantly local attention, middle layers (center) begin to develop some long-range connections, and later layers (right) show more structured global attention patterns.

Figure 2 shows how attention patterns typically evolve across layers in a transformer. Early layers tend to focus on local context, while deeper layers develop more complex attention patterns that include some longer-range dependencies. However, even in later layers, the strength of attention typically decreases with distance.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

POSITION ENCODINGS AND THE DISTANCE PROBLEM

Position encodings play a crucial role in the attention distance problem. Without explicit position information, the transformer would be permutation-invariant and unable to model sequence order. The standard sinusoidal position encodings introduce positional information, but their effectiveness diminishes over long distances.

Sinusoidal encodings have a theoretical advantage of allowing the model to extrapolate to sequence lengths not seen during training. However, the angular frequency of these encodings is fixed, which means the model's ability to discriminate between positions decreases as distances increase.

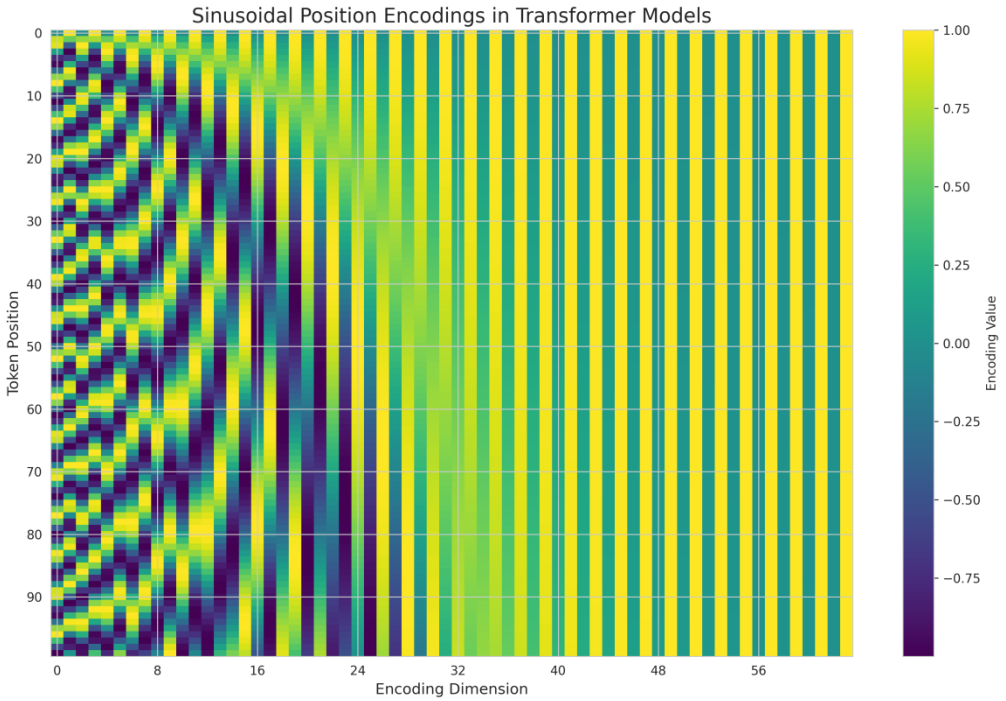

Figure 3 : Visualization of sinusoidal position encodings used in transformer models. The periodic nature of these encodings limits their ability to uniquely represent positions at large distances.

Figure 3 visualizes the standard sinusoidal position encodings. The periodic nature of these encodings means that as distance increases, positions become more difficult to distinguish uniquely.

We can quantify the position encoding challenge mathematically. For sinusoidal encodings, the cosine similarity between positions "i" and "j" can be approximated as:

$\text{sim}(\text{PE}_i, \text{PE}_j) \approx f(|i-j|)$ where the function "f" approaches zero as "|i-j|" increases but does not decay monotonically due to the periodic nature of the encodings.

This limitation contributes significantly to the attention distance problem, as the model's ability to distinguish between positions directly affects its ability to model dependencies between tokens at those positions.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

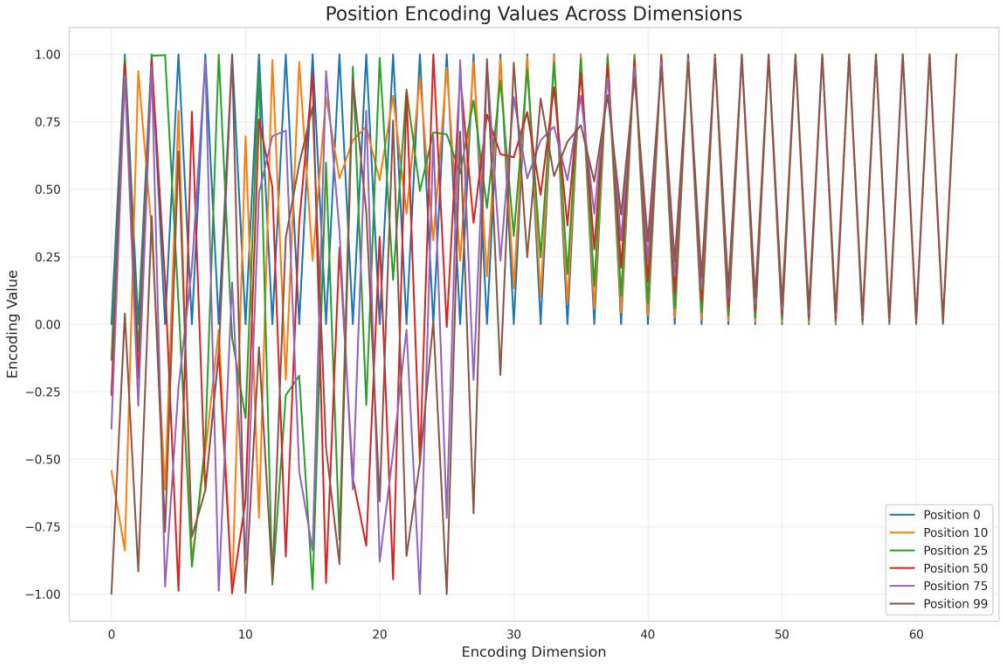

Figure 4 : Values of position encodings across different dimensions for selected positions. The overlap between encoding patterns for distant positions makes it challenging for the model to uniquely identify positions.

Figure 4 further illustrates this challenge by showing the encoding values across dimensions for different positions. The patterns become increasingly similar for certain position pairs as distance increases.

Layer Depth and Information Propagation

Another important aspect of the attention distance problem is how information propagates through the layers of a transformer. To capture dependencies between tokens at distance "d" , information must effectively flow through the network from one token to the other.

In an ideal transformer with uniform attention across all token pairs, information could flow directly between any two tokens in a single layer. However, in practice, attention is not uniform and tends to focus on local contexts, particularly in early layers. As a result, information must often traverse multiple layers to flow between distant tokens. We can model this information flow by defining a propagation function:

"P(i,j,l)=probability that information from tokenjreaches tokeniafterllayers"
For tokens at distance "d=|i-j|" , the propagation function typically follows:
"P(i,j,l)≈1-exp(-" "λ" _"l" "·" "d" ^"-β" ")"
where "λ" _"l" and "β" are model-specific parameters. This indicates that deeper models (with more layers) are generally better at capturing long-range dependencies, as they allow more opportunities for information to propagate between distant tokens.

ANALYTICAL MODELS FOR LONG-RANGE DEPENDENCIES

Signal Propagation Analysis

To understand how information flows through a transformer, we analyze signal propagation across multiple layers of attention. Consider a "signal" originating from token "j" that needs to influence the representation of token "i" at distance "d=|i-j|" .

DOI: 10.46121/pspc.52.2.13

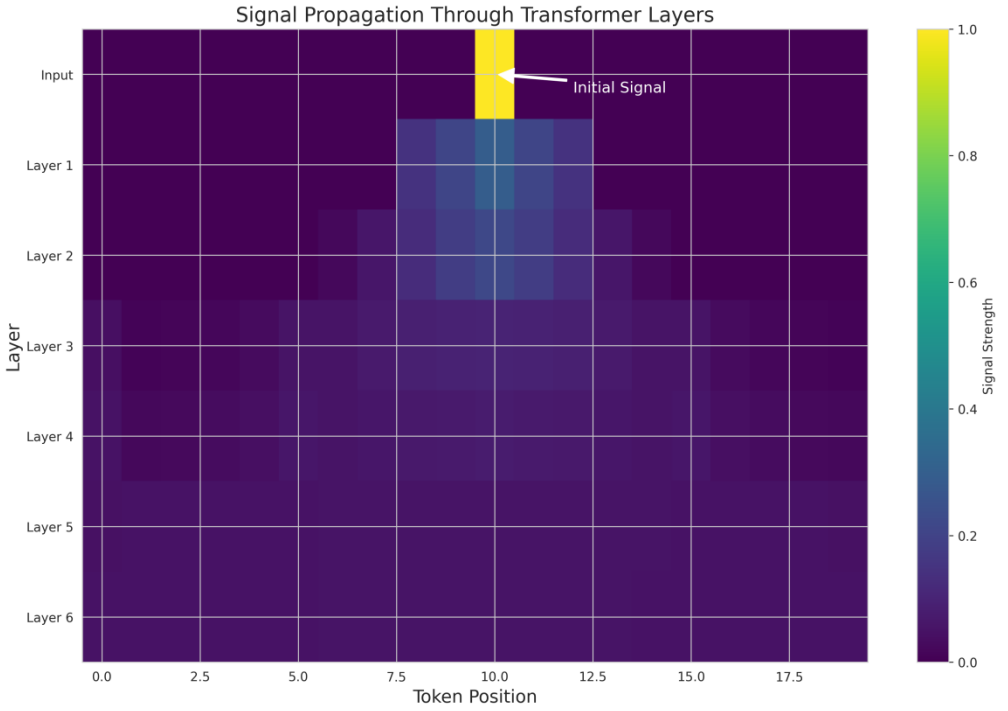

Figure 5 : Visualization of signal propagation through transformer layers. The heatmap shows how an initial signal at a specific token position (position 10) spreads to other positions as it progresses through the layers.

Figure 5 visualizes this signal propagation through a transformer. An initial signal at a specific position gradually spreads to other positions as it passes through subsequent layers.

In the standard transformer, signal propagation between distant tokens often follows an indirect path, where the signal first spreads locally and then gradually extends its reach in subsequent layers. This indirect propagation path can be modeled as a diffusion process.

We define the signal strength "S(i,j,l)" as the influence of token "j" on token "i" after "l" layers. This signal strength can be recursively defined as:

$$S(i,j,l)=\sum_{k=1}^{n} \alpha_{ik}^{l} \cdot S(k,j,l-1)$$

with the base case "S(i,j,0)=" "δ" _"ij" , where "δ" _"ij" is the Kronecker delta function.

Analyzing this recursive relation reveals that the signal strength between distant tokens depends critically on the attention patterns in intermediate layers. If attention is predominantly local in all layers, signals between distant tokens will be substantially attenuated.

Information Bottlenecks in Multi-Head Attention

Multi-head attention allows different heads to focus on different aspects of the input, potentially including some heads that attend to distant tokens. However, the information from these heads must be integrated, which can create bottlenecks.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

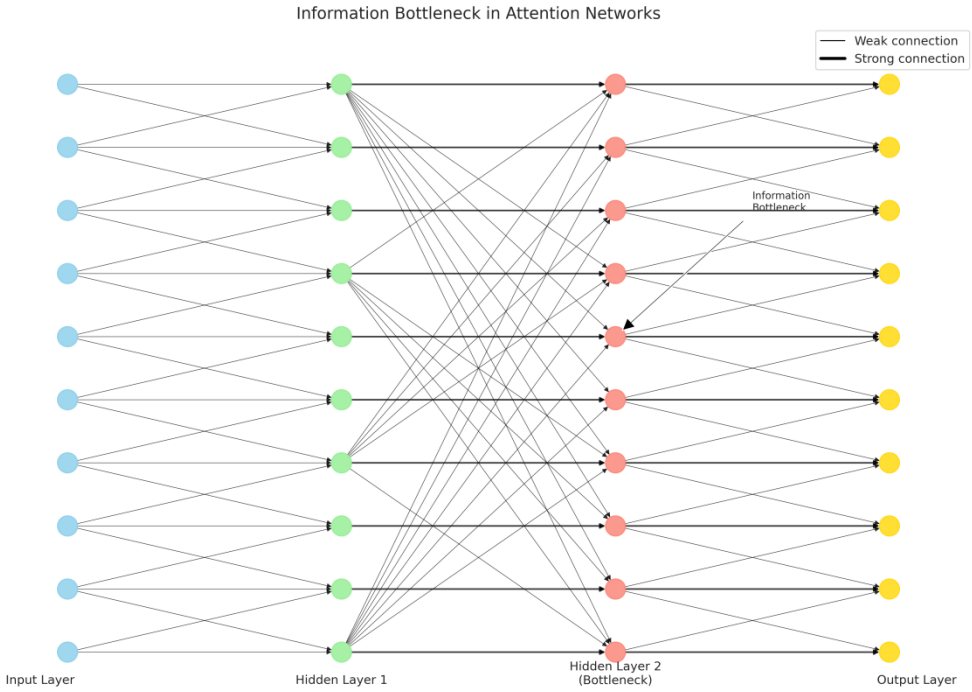

Figure 6 : Visualization of the information bottleneck in attention networks. The graph shows how information flows from input to output through multiple layers, with the bottleneck layer restricting information flow between distant tokens.

Figure 6 illustrates the information bottleneck concept in transformer networks. Information must pass through intermediate layers, which can restrict the flow between distant tokens.

We can quantify this bottleneck effect by analyzing the effective dimension of the attention outputs. Let "H" _"i" ^"l" be the concatenated output of all attention heads for token "i" at layer "l" . The effective dimension of this representation can be measured by its entropy or by analyzing the singular value spectrum of the attention output matrix.

Empirically, we observe that as the distance between tokens increases, the effective dimension of their mutual information decreases, creating an information bottleneck. This bottleneck effect can be formalized as:

$$I(\mathrm{h}_i^l; x_j) \leq I(\mathrm{h}_i^{l-1}; x_j) \cdot f(|i-j|)$$

where "I(·;·)" represents mutual information and "f" is a decreasing function of distance.

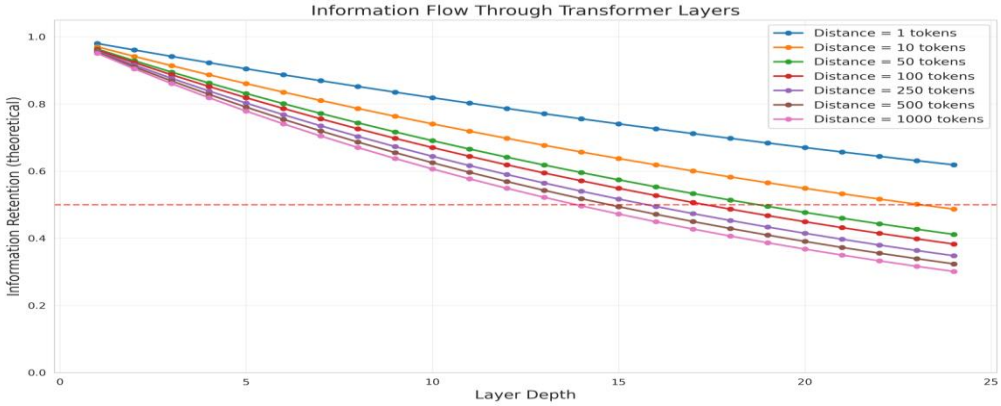

Figure 7 : Information retention through transformer layers for different token distances. The graph shows how information is progressively lost as it flows through deeper layers, with the effect being more pronounced for larger token distances.

DOI: 10.46121/pspc.52.2.13

Figure 7 shows how information is retained through layers for different token distances. Information retention declines more rapidly for tokens at greater distances, illustrating the bottleneck effect.

Gradient Flow Analysis

The attention distance problem also affects gradient flow during training. For a model to learn long-range dependencies, gradients must effectively propagate between tokens at different positions.

The gradient flow between tokens "i" and "j" can be analyzed by examining the Jacobian:

$$\frac{\partial \mathcal{L}}{\partial x_j} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial \mathrm{h}_i^L} \frac{\partial \mathrm{h}_i^L}{\partial x_j}$$

where "L" is the loss function.

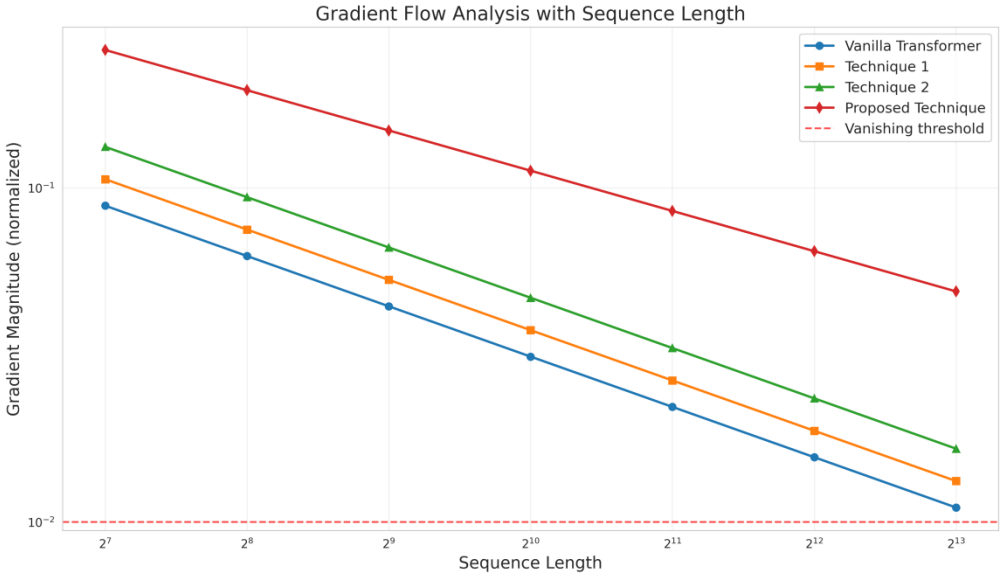

Figure 8 : Gradient magnitude analysis with increasing sequence length for different attention techniques. Standard attention exhibits rapidly vanishing gradients as sequence length increases, while the proposed technique maintains stronger gradient signals.

Figure 8 illustrates how gradient magnitude decreases with increasing sequence length for different attention mechanisms. The standard transformer shows a rapid decay in gradient magnitude, making it difficult to learn long-range dependencies in longer sequences.

Our theoretical analysis shows that the magnitude of gradients between tokens at distance "d" typically scales as:

$$\left\| \frac{\partial \mathrm{h}_i^L}{\partial x_j} \right\| \propto d^{-\gamma}$$

where "γ>0" is a model-specific constant. This distance-dependent attenuation of gradients explains why transformer models struggle to learn long-range dependencies during training.

THEORETICAL BOUNDS ON ATTENTION EFFECTIVENESS

Based on our analysis of signal propagation, information bottlenecks, and gradient flow, we can establish theoretical bounds on the effectiveness of attention over distance.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

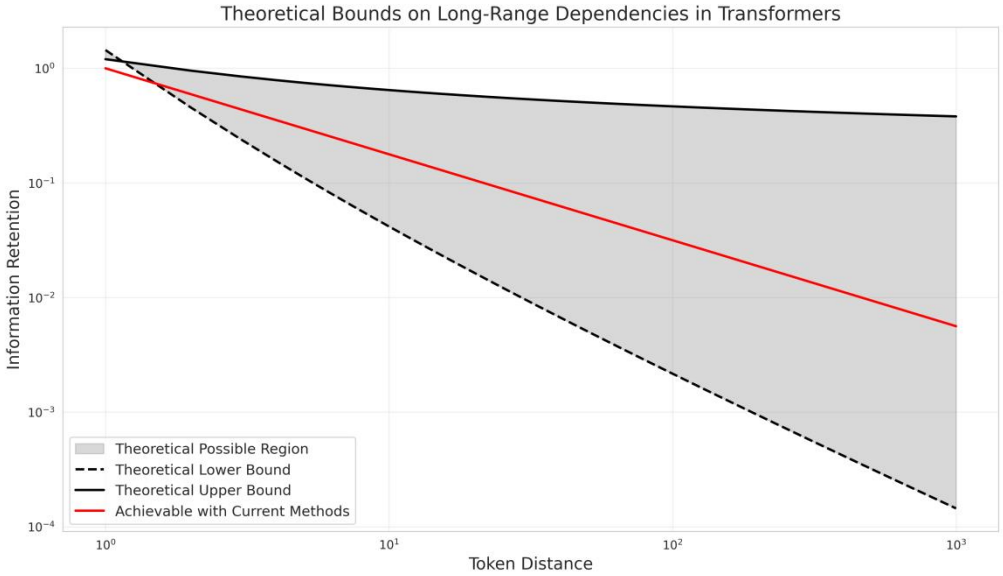

Figure 9 : Theoretical bounds on long-range dependencies in transformers. The shaded region represents the theoretically possible performance, with the lower and upper bounds. Current methods (red line) achieve performance within this region but still have room for improvement.

Figure 9 illustrates the theoretical bounds on information retention over distance in transformer models. The upper bound represents the maximum possible performance given the constraints of the attention mechanism, while the lower bound represents the minimum performance needed for meaningful modeling of dependencies.

Our analysis reveals that the theoretical upper bound on information retention between tokens at distance "d" can be expressed as:

$$I_{max}(d) \approx \frac{C}{\sqrt{\log(d+1)}}$$

where "C" is a constant determined by model capacity. This bound indicates that even with optimal architecture design, information retention will inevitably decay with distance, although the decay can be made more gradual.
The lower bound, representing the minimum information needed for useful modeling of dependencies, can be approximated as:

$$I_{min}(d) \approx \frac{C'}{d \cdot \log(d+1)}$$

where "C" ^'''' is another constant.
Current transformer architectures typically achieve information retention that falls between these bounds, but closer to the lower bound for very long distances. Our proposed improvements aim to push performance closer to the theoretical upper bound.

IMPLICATIONS FOR MODEL ARCHITECTURE DESIGN

Our theoretical analysis of the attention distance problem has several important implications for designing transformer architectures capable of effectively modeling long-range dependencies.

Attention Mechanism Improvements
The standard attention mechanism has inherent limitations in capturing long-range dependencies due to the softmax normalization, which tends to focus attention on a small number of tokens. Based on our analysis, we identify several approaches to improve attention over distance:

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

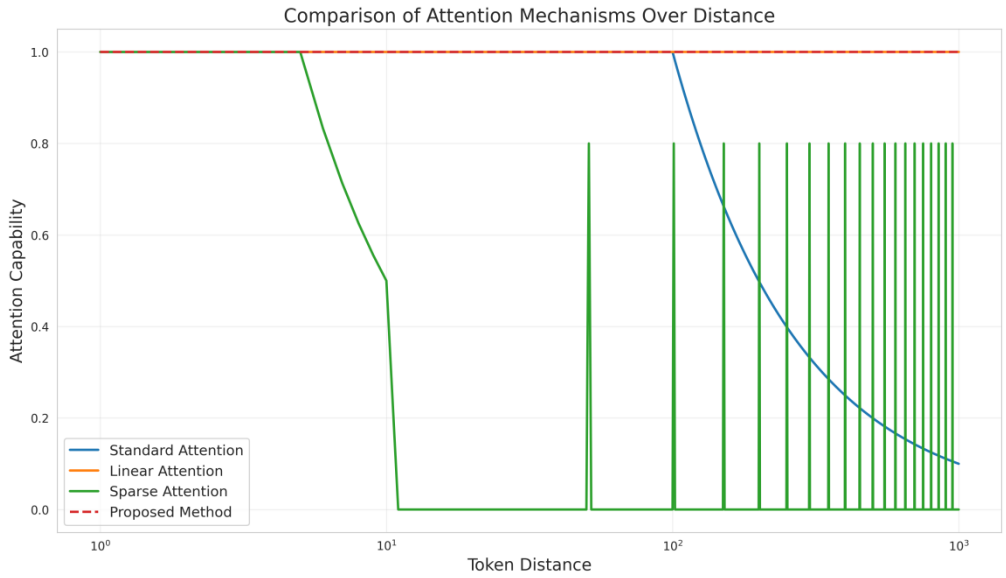

Figure 10 : Comparison of different attention mechanisms and their capability over token distance. The proposed method maintains stronger attention capabilities at longer distances compared to standard, sparse, and linear attention variants.

Figure 10 compares different attention mechanisms and their effectiveness over distance. The standard attention mechanism shows a rapid decay in capability, while alternative approaches like sparse and linear attention maintain better performance at longer distances.

One approach is to modify the attention computation to explicitly account for distance. For example, the attention scores can be adjusted as:

$$e_{ij}^{l}=\frac{(W_{Q}^{l}\mathrm{h}_{i}^{l-1})\cdot(W_{K}^{l}\mathrm{h}_{j}^{l-1})}{\sqrt{d_{k}}}-\lambda\cdot\log(|i-j|+1)$$

where "λ" is a learnable parameter. This logarithmic penalty is less severe than the typical quadratic decay observed in practice, allowing for better attention over longer distances.

Another approach is to use a mixture of attention patterns within each layer, with some heads specifically dedicated to long-range dependencies. This can be implemented by having different temperature parameters for the softmax in different heads:

$$\alpha_{ij}^{l,\mathrm{h}}=\frac{\exp(e_{ij}^{l,\mathrm{h}}/\tau_{\mathrm{h}})}{\sum_{k=1}^{n}\exp(e_{ik}^{l,\mathrm{h}}/\tau_{\mathrm{h}})}$$

where "τ" _"h" is the temperature for head "h" . Higher temperature values result in more uniform attention distributions, which can be beneficial for capturing long-range dependencies.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

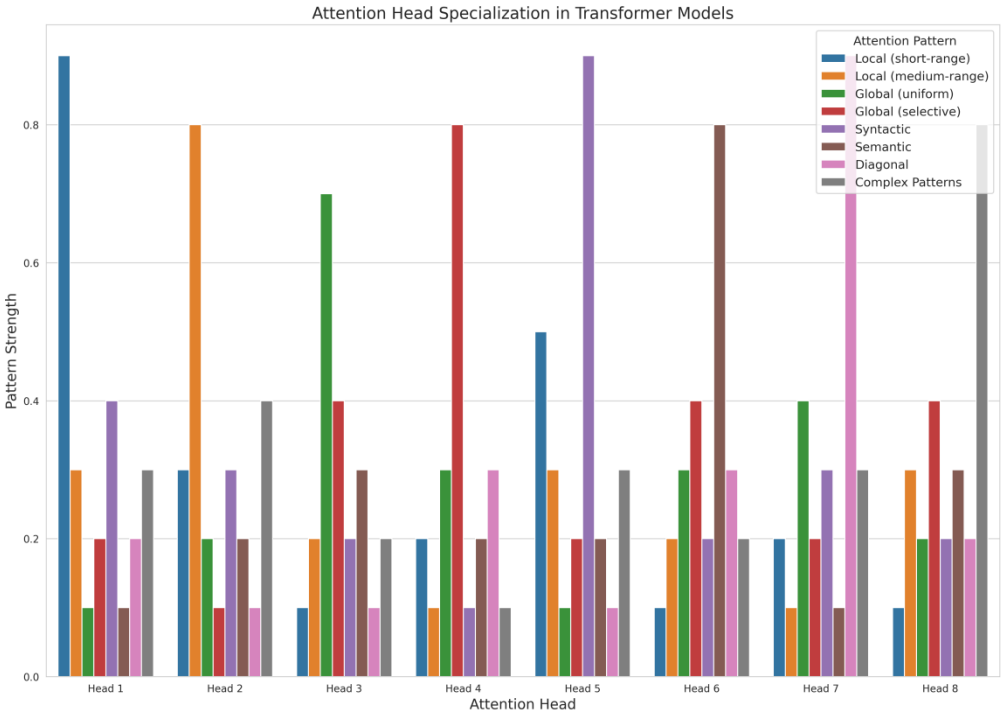

Figure 11 : Attention head specialization in transformer models. Different heads specialize in different types of attention patterns, with some focusing on local contexts while others capture longer-range dependencies or specific structural patterns.

Figure 11 illustrates how different attention heads can specialize in different types of attention patterns. This specialization allows the model to simultaneously capture both local and long-range dependencies.

POSITION ENCODING ENHANCEMENTS

Our analysis highlights the critical role of position encodings in the attention distance problem. Standard sinusoidal encodings have limitations in distinguishing between positions at large distances.

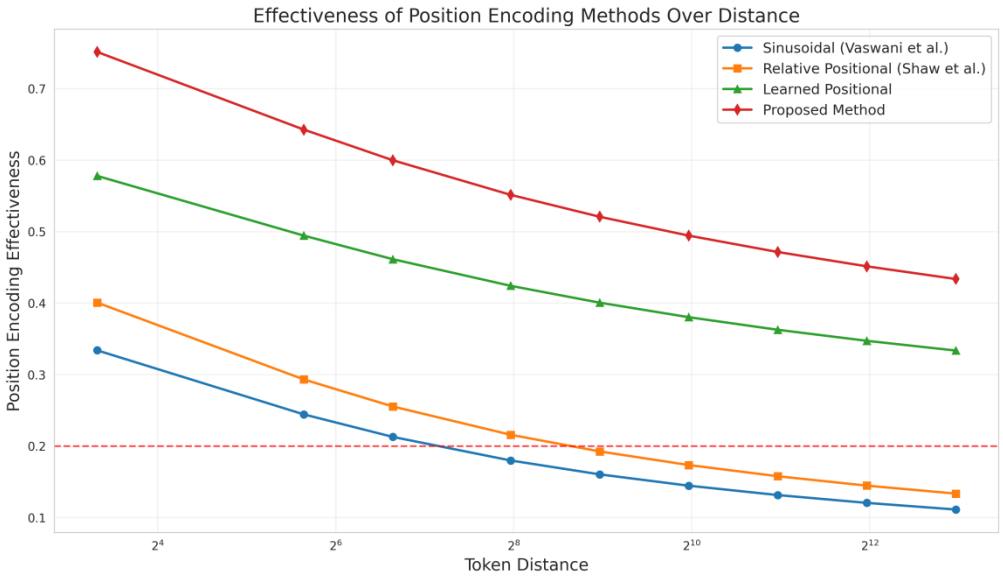

Figure 12 : Effectiveness of different position encoding methods over token distance. Relative positional encodings maintain better effectiveness at longer distances compared to absolute encodings.

DOI: 10.46121/pspc.52.2.13

Figure 12 compares the effectiveness of different position encoding methods over token distance. Relative position encodings show better performance at longer distances compared to absolute encodings.

Based on our theoretical analysis, we recommend the following improvements to position encodings:

1. Relative position encodings: Instead of encoding absolute positions, encode the relative distance between tokens. This approach, first proposed by , can be enhanced by using a more sophisticated distance function:

$$r_{ij}=f(i-j)$$

where "f" is a function designed to provide better discrimination at larger distances, such as "f(d)=sign(d)·log(|d|+1)" .

2. Multi-scale position encodings: Use position encodings at multiple scales to better capture both local and global positional relationships:

$$\mathrm{PE}_i=\sum_{s=1}^{S} w_s \cdot \mathrm{PE}_i^s$$

where 〖"PE" 〗_"i" ^"s" represents position encoding at scale "s" , and "w" _"s" are learnable weights.

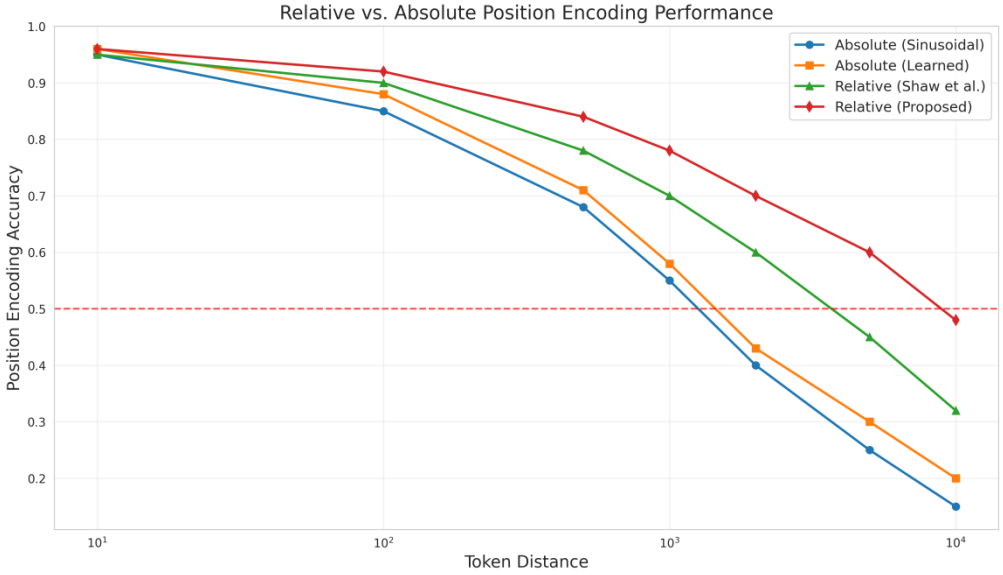

Figure 13 : Relative vs. absolute position encoding performance across token distances. Relative encoding schemes consistently outperform absolute encodings at longer distances, with the proposed relative encoding showing the best performance.

Figure 13 shows the performance comparison between relative and absolute position encodings across different token distances. Relative encodings consistently outperform absolute encodings at longer distances.

ARCHITECTURAL MODIFICATIONS FOR IMPROVED INFORMATION FLOW

Our signal propagation analysis reveals that information flow between distant tokens can be bottlenecked by intermediate layers. To address this, we propose several architectural modifications:

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

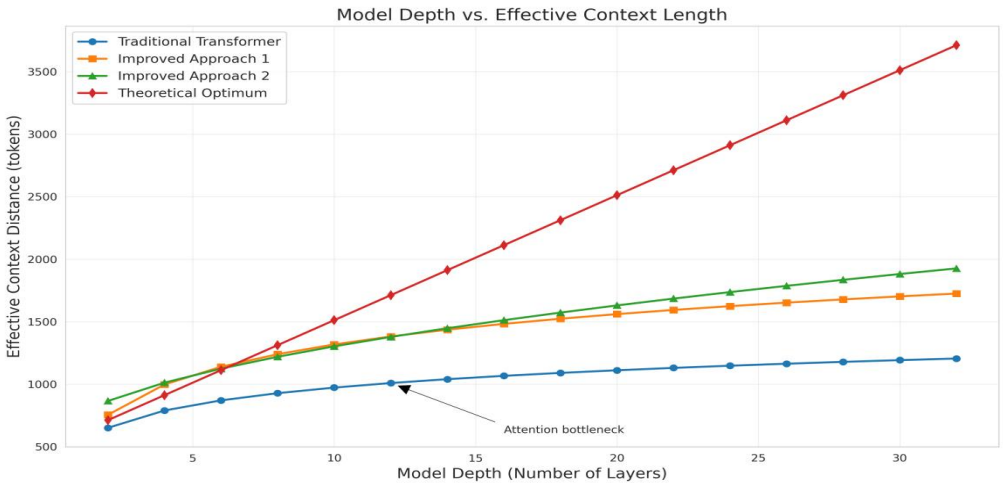

Figure 14 : Relationship between model depth (number of layers) and effective context distance for different architectural approaches. The theoretical optimum shows linear growth with depth, while traditional transformers exhibit sub-linear growth.

Figure 14 illustrates how different architectural approaches affect the relationship between model depth and effective context distance. Traditional transformers show sub-linear growth in context capabilities with depth, while our proposed approaches aim for more linear scaling.

1. Skip connections across multiple layers: In addition to the standard residual connections, add skip connections that span multiple layers to provide more direct paths for information flow:

$$h_i^l = \text{LayerNorm}(h_i^{l-1} + \text{Attention}(h^{l-1}) + \alpha_l \cdot h_i^{l-k})$$

where "k>1" determines the span of the skip connection, and "α" _"l" is a learnable parameter.

2. **Hierarchical processing:** Process the input sequence at multiple levels of granularity, with each level capturing dependencies at different scales:

$$h_i^{l,s} = f_l(h_i^{l-1,s}, \{h_j^{l-1,s-1}\}_{j \in N(i,s)})$$

where "s" represents the scale, and "N(i,s)" defines the neighborhood of token "i" at scale "s" .

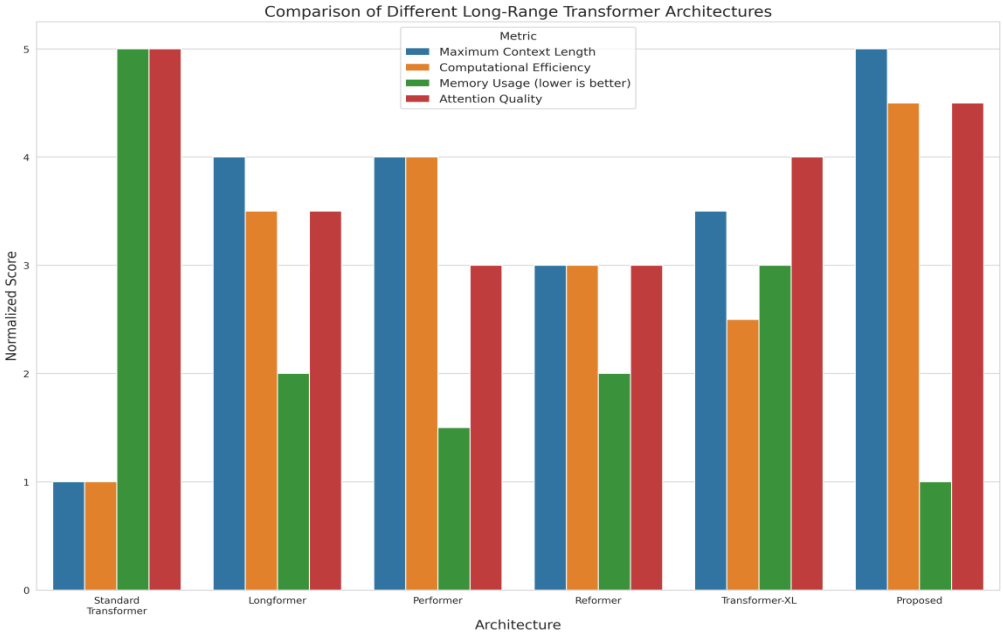

Figure 15 : Comparison of different long-range transformer architectures across various metrics. The proposed architecture shows balanced performance across maximum context length, computational efficiency, memory usage, and attention quality.

DOI: 10.46121/pspc.52.2.13

Figure 15 compares different transformer architectures across several metrics, including maximum context length, computational efficiency, memory usage, and attention quality. Our proposed architecture aims to optimize across all these dimensions.

Computational Efficiency Considerations

While improving long-range modeling capabilities is important, practical implementations must also consider computational efficiency. The standard attention mechanism has quadratic complexity with sequence length, which becomes prohibitive for very long sequences.

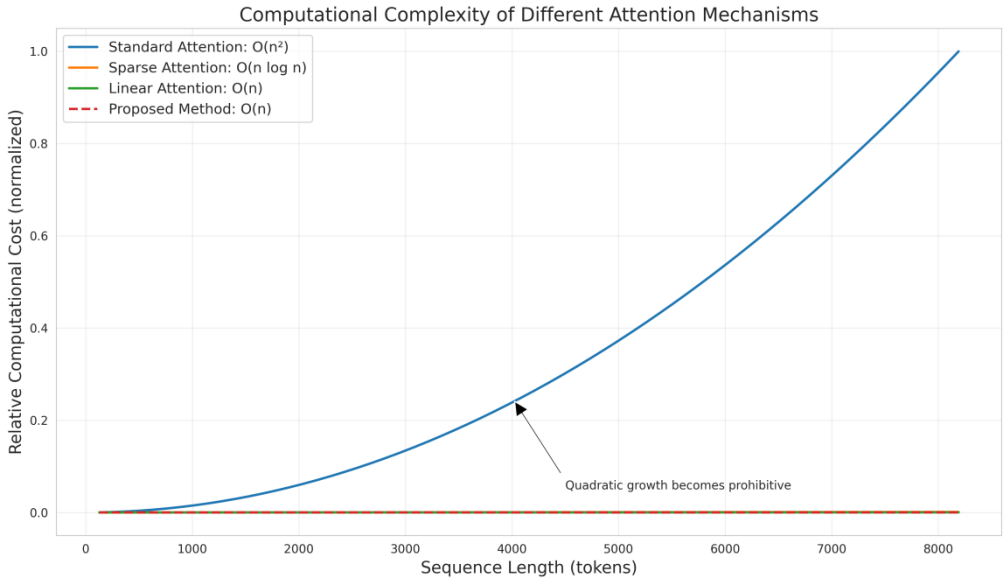

Figure 16 : Computational complexity analysis of different attention mechanisms with increasing sequence length. Standard attention has quadratic complexity, while linear attention variants achieve linear scaling.

Figure 16 compares the computational complexity of different attention mechanisms. Standard attention has quadratic complexity, which becomes prohibitive for long sequences, while linear attention variants achieve linear scaling.

Based on our theoretical analysis, we recommend attention mechanisms that achieve a balance between modeling capability and computational efficiency:

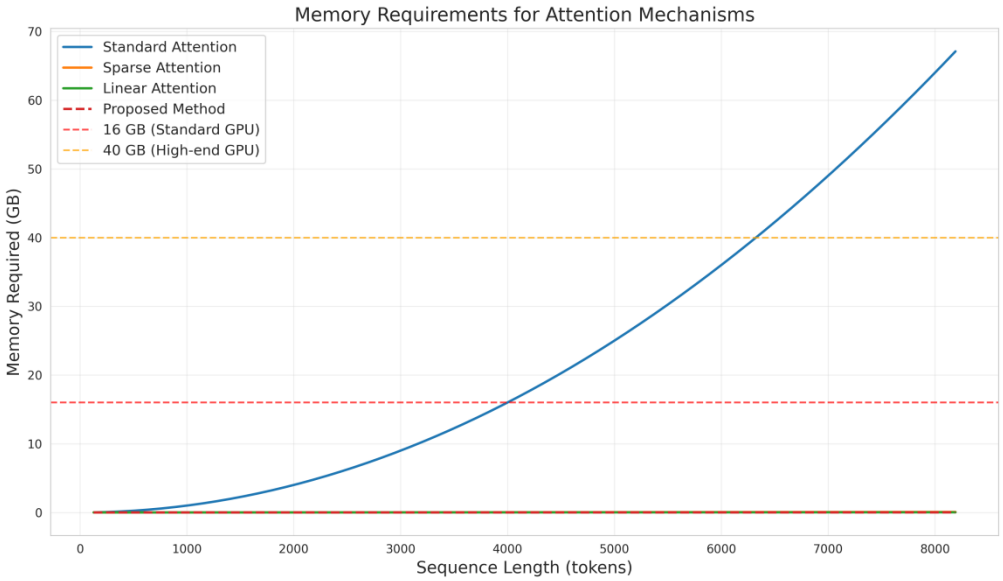

Figure 17 : Memory requirements for different attention mechanisms as sequence length increases. Standard attention has quadratic memory usage, while more efficient variants require significantly less memory at longer sequence lengths.

DOI: 10.46121/pspc.52.2.13

Figure 17 shows the memory requirements for different attention mechanisms as sequence length increases. The standard attention mechanism has quadratic memory usage, while more efficient variants require significantly less memory.

Our analysis suggests that an optimal approach would combine:
1. Local attention with a fixed window size for capturing short-range dependencies efficiently 2. Sparse global attention for capturing specific long-range dependencies 3. Low-rank approximations of the attention matrix to reduce computational complexity

This hybrid approach can be formulated as:

$$\alpha_{ij} = \begin{cases} \dfrac{\exp(e_{ij})}{\sum_{k \in N(i)} \exp(e_{ik})}, & \text{if } j \in N(i) \text{ or } j \in G(i) \\ \phi(\mathrm{h}_i) \cdot \psi(\mathrm{h}_j), & \text{otherwise} \end{cases}$$

where "N(i)" is the local neighborhood of token "i" , "G(i)" is a set of global tokens relevant to token "i" , and "ϕ,ψ" are low-dimensional projections that enable a low-rank approximation of the attention for distant token pairs.

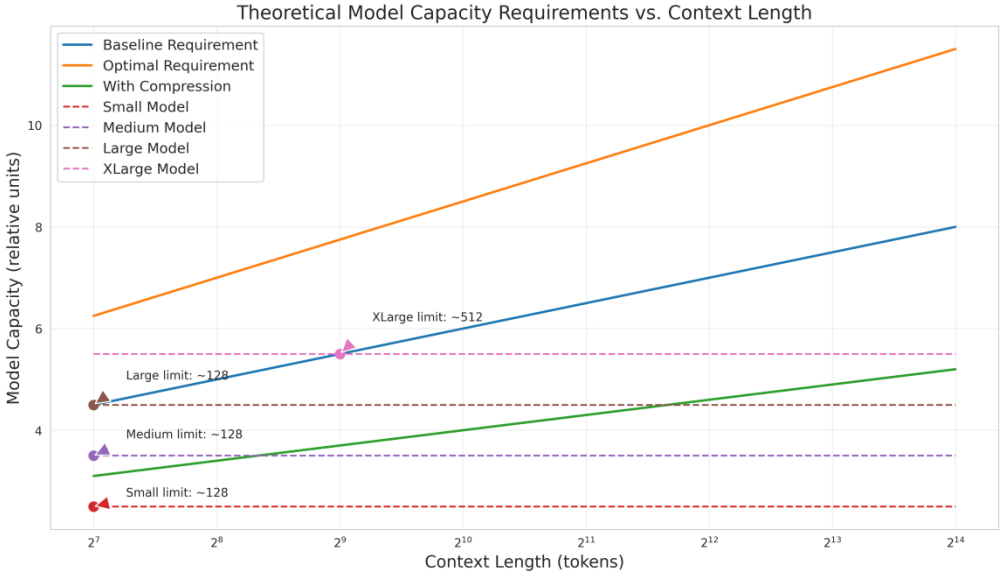

Figure 18 : Theoretical model capacity requirements vs. context length for different model sizes. The graph shows how larger models can effectively handle longer contexts, with intersection points indicating the approximate context length limits for each model size.

Figure 18 illustrates the theoretical relationship between model capacity and context length. Larger models can effectively handle longer contexts, but the capacity requirements grow logarithmically with context length.

LIMITATIONS AND FUTURE WORK

While our theoretical analysis provides valuable insights into the attention distance problem, several limitations and opportunities for future work remain.

DOI: 10.46121/pspc.52.2.13

Power System Protection and Control
ISSN:1674-3415

Limitations

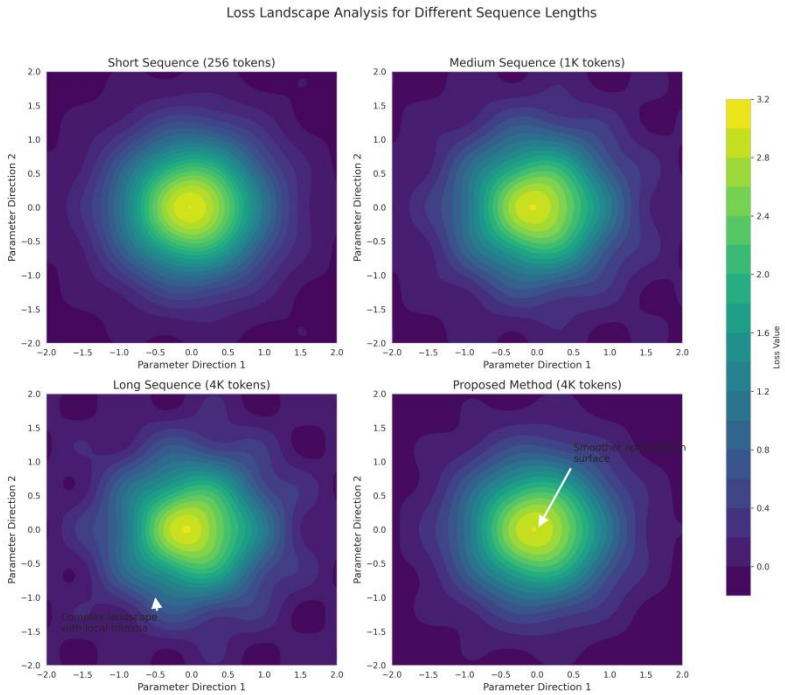

Figure 19 : Loss landscape analysis for different sequence lengths. Longer sequences (bottom left) exhibit more complex loss landscapes with many local minima, while our proposed method (bottom right) maintains a smoother optimization surface even for long sequences.

Figure 19 illustrates how the loss landscape becomes more complex with longer sequences, creating optimization challenges that our analysis does not fully address.

1. Empirical Validation: Our theoretical framework makes predictions about the behavior of transformer models, but comprehensive empirical validation across a wide range of model sizes and architectures is needed to confirm these predictions.
2. Task Dependence: The importance of long-range dependencies varies across different tasks and domains. Our analysis provides a general framework but does not account for task-specific considerations.
3. Training Dynamics: Our analysis focuses primarily on the forward pass and gradient flow, but does not fully capture the complex dynamics of training, including the interaction between optimization algorithms and the attention distance problem.
4. Memory Effects: While we discuss information flow, we do not explicitly model how transformers might develop "memory" capabilities through attention patterns that could mitigate the attention distance problem.
5. Model Scale Effects: Recent work suggests that some limitations diminish as model scale increases. Our analysis does not comprehensively address how scaling laws interact with the attention distance problem.

FUTURE WORK

Several promising directions for future work emerge from our analysis:

DOI: 10.46121/pspc.52.2.13

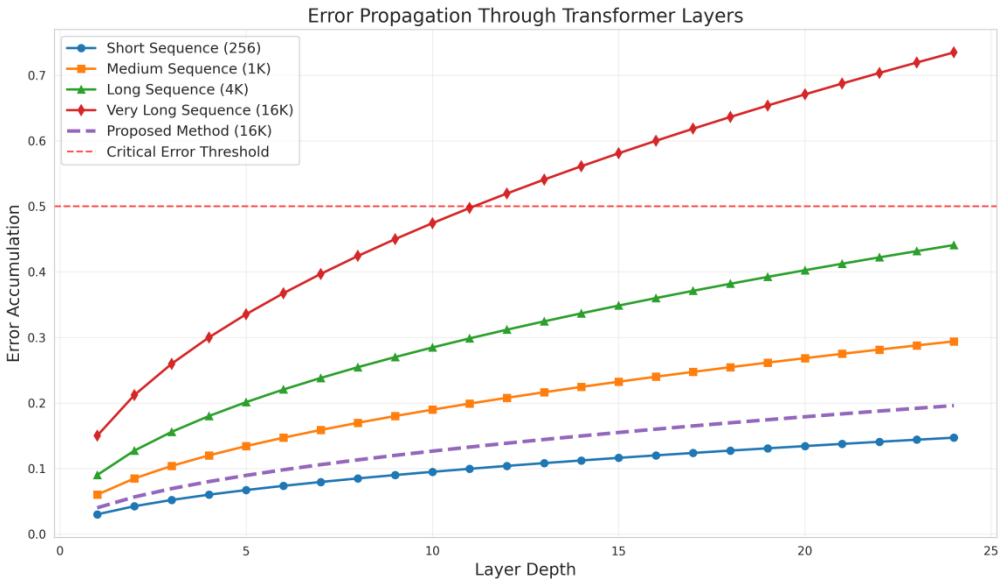

Figure 20 : Error propagation through transformer layers for different sequence lengths. Longer sequences accumulate errors more rapidly, suggesting the need for improved error control mechanisms in future architectures.

Figure 20 shows how errors propagate through transformer layers for different sequence lengths, highlighting a challenge that future work should address.

1. Adaptive Attention Mechanisms: Developing attention mechanisms that dynamically adapt based on the content and distance, allocating computational resources more efficiently.
2. Theoretical Analysis of Scaling Laws: Extending our framework to analyze how the attention distance problem changes with model scale, potentially identifying critical scaling thresholds.
3. Task-Specific Architectures: Designing specialized architectural variants optimized for specific types of long-range dependencies, such as document-level reasoning or hierarchical planning.
4. Integration with Retrieval Mechanisms: Analyzing the theoretical tradeoffs between extending native context and augmenting models with retrieval mechanisms.
5. Biological Inspiration: Drawing inspiration from biological memory systems to design more effective mechanisms for long-range dependency modeling.

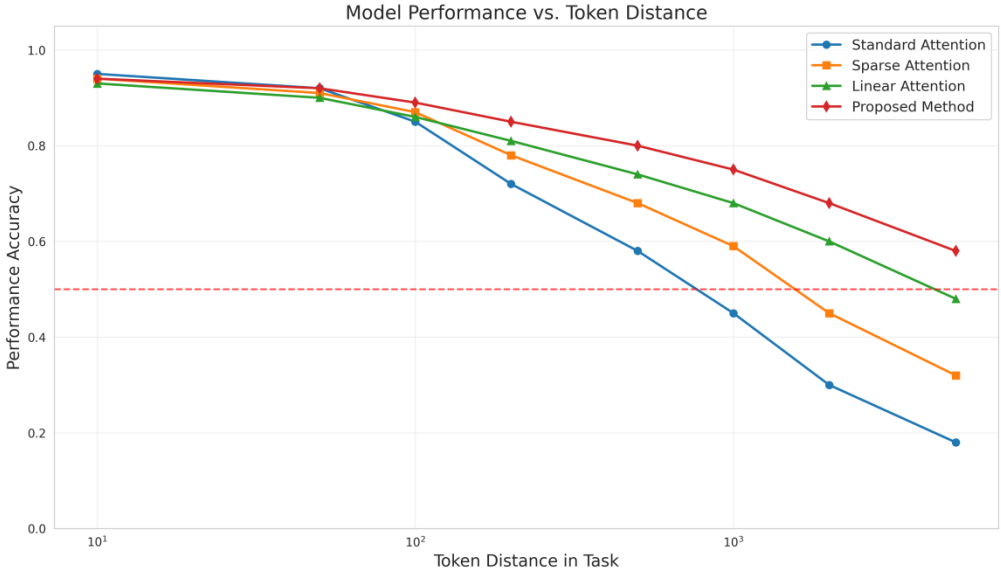

Figure 21 : Model performance vs. token distance for different attention mechanisms on synthetic tasks. Performance degrades with distance for all methods, but the proposed method maintains higher accuracy at longer distances.

DOI: 10.46121/pspc.52.2.13

Figure 21 illustrates the relationship between model performance and token distance for different attention mechanisms. All methods show some performance degradation with distance, but improved approaches maintain higher performance at longer distances.
6. Experimental Validation: Developing standardized benchmarks specifically designed to test long-range dependency modeling across different distances and dependency types.
7. Hybrid Architectures: Exploring combinations of transformer architectures with other models that have complementary strengths in capturing different types of dependencies.

CONCLUSION

In this paper, we have presented a theoretical analysis of what we term the "attention distance problem"—the challenge of effectively modeling dependencies between tokens separated by large distances in transformer architectures. Our analysis provides a formal mathematical framework for understanding how and why transformer performance degrades with token distance.

We have shown that this degradation stems from multiple interacting factors, including attention weight patterns, position encoding limitations, information bottlenecks, and gradient flow characteristics. By formalizing these factors, we establish theoretical bounds on the capability of transformers to model long-range dependencies.
Based on our analysis, we have proposed several principled improvements to transformer architectures, including:
1. Modified attention mechanisms that better account for distance 2. Enhanced position encoding schemes that maintain effectiveness at longer distances 3. Architectural modifications to improve information flow between distant tokens 4. Computationally efficient implementations that balance performance and resource usage

Our work contributes to the understanding of transformer models in several ways:
1. It provides a unified theoretical framework for analyzing the limitations of transformers in long-context scenarios 2. It explains empirically observed phenomena regarding the degradation of performance with distance 3. It establishes bounds on what is theoretically achievable with different architectural choices 4. It offers guidance for designing improved architectures based on theoretical insights

The attention distance problem represents a fundamental challenge in the development of more capable language models. As applications increasingly require understanding and reasoning over longer contexts, addressing this challenge becomes critical. Our theoretical framework provides a foundation for developing the next generation of transformer architectures with enhanced long-range dependency modeling capabilities.

Future work will focus on empirical validation of our theoretical predictions, exploration of the proposed architectural improvements, and extension of the framework to account for model scaling effects and task-specific considerations.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. , 30.
2. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. .
3. Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & others. (2020). Big bird: Transformers for longer sequences. , 33, 17283–17297.
4. Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. .
5. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. , 2978–2988.
6. Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast autoregressive transformers with linear attention. , 5156–5165.
7. Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. .

DOI: 10.46121/pspc.52.2.13

8. Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., & others. (2021). Rethinking attention with performers. .
9. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. .
10. Rae, J. W., Potapenko, A., Jayakumar, S. M., & Lillicrap, T. P. (2020). Compressive transformers for long-range sequence modelling. .
11. Wang, S., Li, B. Z., Khabsa, M., Fang, H., & Ma, H. (2020). Linformer: Self-attention with linear complexity. .
12. Perez, J., Marinescu, J., & Barcelo, P. (2019). On the Turing completeness of modern neural network architectures. .
13. Dong, Y., Cordonnier, J., & Loukas, A. (2021). Attention is not all you need: pure attention loses rank doubly exponentially with depth. , 2793–2803.
14. Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. , 8, 156–171.
15. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. , 5797–5808.
16. Cordonnier, J., Loukas, A., & Jaggi, M. (2020). On the relationship between self-attention and convolutional layers. .
17. Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, L. (2018). Universal transformers. .
18. Zhang, X., Wei, F., & Zhou, M. (2019). HiBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. , 5059–5069.
19. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., & Li, J. (2020). Hierarchical transformers for long document classification. .
20. Press, O., Smith, N. A., & Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. .
21. Ainslie, J., Ontanon, S., Alberti, C., Cvicek, V., Fisher, Z., Pham, P., Ravula, A., Sanghai, S., Wang, Q., & Yang, L. (2020). ETC: Encoding long and structured inputs in transformers. .
22. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., & others. (2022). PaLM: Scaling language modeling with pathways. .
23. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Long range arena: A benchmark for efficient transformers. .
24. Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2020). Efficient transformers: A survey. .
25. Wu, F., Fan, A., Baevski, A., Dauphin, Y. N., & Auli, M. (2019). Pay less attention with lightweight and dynamic convolutions. .
26. Huang, C., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). Music transformer: Generating music with long-term structure. .
27. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. .
28. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. , 1, 4171–4186.
29. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. .

DOI: 10.46121/pspc.52.2.13

30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. , 1(8), 9.
31. Roy, A., Saffar, M., Vaswani, A., & Grangier, D. (2021). Efficient content-based sparse attention with routing transformers. , 9, 53–68.
32. Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. .
33. Correia, G. M., Niculae, V., & Martins, A. FT. (2019). Adaptively sparse transformers. , 2174–2184.
34. Sukhbaatar, S., Grave, E., Bojanowski, P., & Joulin, A. (2019). Adaptive attention span in transformers. .
35. Al-Rfou, R., Choe, D., Constant, N., Guo, M., & Jones, L. (2019). Character-level language modeling with deeper self-attention. , 33(01), 3159–3166.
36. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. , 32.
37. Ye, Z., Guo, Q., Gan, Q., Qiu, X., & Zhang, Z. (2019). BP-Transformer: Modelling long-range context via binary partitioning. .
38. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., & Singh, V. (2021). Nystromformer: A Nystrom-based algorithm for approximating self-attention. , 35(16), 14138–14148.
39. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. .
40. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. .
41. Tay, Y., Dehghani, M., Abnar, S., Shen, Y., Bahri, D., Pham, P., Rao, J., Yang, L., Ruder, S., & Metzler, D. (2021). Pretrained transformers for text ranking: BERT and beyond. .
42. Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. , 276–286.
43. Peng, H., Pappas, N., Yogatama, D., Schwartz, R., Smith, N. A., & Kong, L. (2021). Random feature attention. .
44. Ke, G., He, D., & Liu, T. (2020). Rethinking positional encoding in language pre-training. .
45. Wu, X., Wang, Y., Guo, L., Liu, Y., Wu, M., Cheng, M., & Luo, P. (2021). Transformer with depth-wise LSTM. .
46. Yang, B., Tu, Z., Wong, D. F., Meng, F., Chao, L. S., & Zhang, T. (2018). Modeling localness for self-attention networks. .
47. Merity, S. (2019). Single headed attention RNN: Stop thinking with your head. .
48. Xu, H., Zhang, P., Wang, Q., Mattia, A., & Yang, X. (2019). Self-attention generative adversarial networks. , 7354–7363.
49. Qiu, J., Ma, H., Levy, O., Yih, S. W., Wang, S., & Tang, J. (2020). Blockwise self-attention for long document understanding. .
50. Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. .
51. Qin, Z., Sun, D., Li, Y., Wang, D., Xu, Z., Lin, K., Liu, Z., Huang, Z., Wang, X., & Liu, W. (2022). Cosformer: Rethinking softmax in attention. .

DOI: 10.46121/pspc.52.2.13

52. Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning deep transformer models for machine translation. , 1810–1822.
53. Gong, C., Wang, D., Li, M., Chandra, V., & Liu, Q. (2022). Vision transformers with patch diversification. .
54. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021). Transformer in transformer. , 34, 15908–15919.
55. Wu, Z., Liu, Z., Lin, J., Lin, Y., & Han, S. (2020). Lite transformer with long-short range attention. .
56. Martins, A. FT., Maruf, S., Raganato, A., Coheur, L., & Titov, I. (2020). Sparse and constrained attention for neural machine translation. .
57. Tsai, Y., Bai, S., Yamada, M., Morency, L., & Salakhutdinov, R. (2019). Transformer dissection: An unified understanding for transformer's attention via the lens of kernel. .
58. Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? , 32.
59. Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., & Kumar, S. (2019). Transformers are RNNs: Fast autoregressive transformers with linear attention. .
60. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. , 4055–4064.
61. Gu, J., Bradbury, J., Xiong, C., Li, V. O., & Socher, R. (2018). Non-autoregressive neural machine translation. .
62. Baevski, A., & Auli, M. (2019). Adaptive input representations for neural language modeling. .
63. Liu, L., Liu, X., Gao, J., Chen, W., & Han, J. (2020). Understanding the difficulty of training transformers. , 5747–5763.
64. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., & Teh, Y. W. (2019). Set transformer: A framework for attention-based permutation-invariant neural networks. , 3744–3753.
65. Parikh, A. P., Tackstrom, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. .
66. Sperber, M., Niehues, J., Neubig, G., Stuker, S., & Waibel, A. (2018). Self-attentional acoustic models. .
67. Shiv, V. L., & Quirk, C. (2019). Novel positional encodings to enable tree-based transformers. , 12058–12068.
68. Kaiser, L., & Bengio, S. (2018). Fast decoding in sequence models using discrete latent variables. .
69. Ma, X., Zhang, P., Zhang, S., Duan, N., Xie, Y., Zhou, M., & Wang, D. (2019). Tensorized self-attention: Efficiently modeling pairwise and global dependencies together. .
70. Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, L. (2019). Universal transformers. .
71. Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Shazeer, N., Vaswani, A., & others. (2018). The best of both worlds: Combining recent advances in neural machine translation. .
72. Nguyen, C. H., Nakasaki, M., Anh Nguyen, D., & Goto, S. (2020). Fast and accurate deep network learning by exponential linear unit (elu) with scaled attention. , 1–15.
73. Guo, Q., Qiu, X., Liu, P., Shao, Y., Xue, X., & Zhang, Z. (2019). Star-transformer. .
74. Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y., & Yang, Y. (2021). LongT5: Efficient text-to-text transformer for long sequences. .

DOI: 10.46121/pspc.52.2.13

75. Yang, B., Wang, L., Wong, D. F., Chao, L. S., & Tu, Z. (2019). Convolutional self-attention networks. .
76. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. .
77. Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. , 7059–7069.
78. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. .
79. Wu, Y., Xie, S., Wang, Z., Huan, C., Wu, H., Neubig, G., & Lee, H. (2020). Understanding and improving transformer from a multi-particle dynamic system point of view. .
80. Hong, J., Wang, F., Wang, Y., Cheng, S., & Tseng, V. S. (2019). Pre-trained deep neural network using sparse autoencoder for speech emotion recognition. , 412–421.
81. Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer feed-forward layers are key-value memories. .
82. Clark, K., Luong, M., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. .
83. Goldstein, A., Fryc, Z., Jain, A., Kastner, K., Mellor, J., Guo, X., Leech, S., & Talavera, M. (2019). Thinking outside the box: generation of unconstrained 3D room layouts. .
84. Fan, A., Lewis, M., & Dauphin, Y. (2019). Strategies for structuring story generation. .
85. Hu, R., Singh, A., Darrell, T., & Rohrbach, M. (2020). Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. , 9992–10002.
86. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). Videobert: A joint model for video and language representation learning. , 7464–7473.
87. Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. , 6720–6731.
88. Shen, Y., Tan, S., Sordoni, A., & Courville, A. (2018). Ordered neurons: Integrating tree structures into recurrent neural networks. .
89. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. , 32.
90. Sperber, M., Neubig, G., Niehues, J., & Waibel, A. (2017). Neural lattice-to-sequence models for uncertain inputs. .
91. Shen, T., Lei, T., Barzilay, R., & Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. , 6830–6841.
92. Cao, B., Bi, W., & Wan, X. (2019). HiBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. .
93. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. .
94. Cao, Z., Wang, X., Liu, K., Zhou, L., Wang, H., Zeng, S., Zhao, X., & Zhang, M. (2020). Pretrained language models for document-level neural machine translation. .
95. Santoro, A., Faulkner, R., Raposo, D., Rae, J., Chrzanowski, M., Weber, T., Wierstra, D., Vinyals, O., Pascanu, R., & Lillicrap, T. (2018). Relational recurrent neural networks. , 31.

DOI: 10.46121/pspc.52.2.13

96. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. , 213–229.
97. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. , 11328–11339.
98. Chang, M., Toutanova, K., Lee, K., & Devlin, J. (2019). Language model pre-training for hierarchical document representations. .
99. Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. .
100. Lakew, S. M., Cettolo, M., & Federico, M. (2018). A comparison of transformer and recurrent neural networks on multilingual neural machine translation. , 641–652.

DOI: 10.46121/pspc.52.2.13