Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 3, 2025, 359-368.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Review paper

Attention Mechanisms in Transformers: A General Survey

Rasoul Hosseinzadeh* and Mehdi Sadeghzadeh

Department of Computer Engineering, Science and Research SR.C., Islamic Azad University, Tehran, Iran.

Article Info

Article History:
Received 19 January 2025
Revised 29 January 2025
Accepted 28 May 2025

DOI: 10.22044/jadm.2025.15584.2679

Keywords:
Attention mechanism, Transformer, Deep Learning.

**Corresponding author: rasoul.hosseinzadeh@iau.ac.ir (R. Hosseinzadeh).*

Abstract

The attention mechanisms have significantly advanced the field of machine learning and deep learning across various domains, including natural language processing, computer vision, and multimodal systems. This paper presents a comprehensive survey of attention mechanisms in Transformer architectures, emphasizing their evolution, design variants, and domain-specific applications in NLP, computer vision, and multimodal learning. We categorize attention types by their goals like efficiency, scalability, and interpretability, and provide a comparative analysis of their strengths, limitations, and suitable use cases. This survey also addresses the lack of visual intuitions, offering a clearer taxonomy and discussion of hybrid approaches, such as sparse-hierarchical combinations. In addition to foundational mechanisms, we highlight hybrid approaches, theoretical underpinnings, and practical trade-offs. The paper identifies current challenges in computation, robustness, and transparency, offering a structured classification and proposing future directions. By comparing state-of-the-art techniques, this survey aims to guide researchers in selecting and designing attention mechanisms best suited for specific AI applications, ultimately fostering the development of more efficient, interpretable, and adaptable Transformer-based models.

1. Introduction

The Transformer model has altered the whole deep learning landscape in 2017[1] when it was first introduced for translation in terms of natural language processing. Most of the success, is attributed to the introduction of the attention concept, which in turn depicts a view or weighting of the importance of several parts of an input sequence by the model. Attention was begotten for neural networks from the extension of sequence-to-sequence models for machine translation. The earliest attempts at such models, especially RNNs and Long Short-term Memory (LSTM) networks, find it impossible to handle long-term dependencies due to the vanishing gradient problem. The attention mechanism was introduced to solve this type of problem by allowing models to dynamically keep track of important parts of the input sequence, even if it is too long; without exceeding the resource bounds, they cannot give all input parts equal attention. Bahdanau et al. [2], wherein they studied machine translation techniques, introduced attention, which allowed a decoder to concentrate dynamically on some parts of the encoder's outputs. From the outset, human cognitive processes inspired that attention mechanism is now one of the cornerstones of applying deep learning in today's world as it can learn complex dependencies in large sized input sequences without relying on recurrent structures. Attention mechanisms have remarkably evolved starting from their introduction, with different advantages, taking on various problems in handling large-scale data, such as computational efficiency, long range dependencies, and multimodal inputs all of which were in need of design.

We present, in this article, a complete survey on of the many variants of attention mechanisms, proposed to improve Transformer-based models.

The reasons driving multiple attention mechanisms include performance, scalability, and flexibility in addressing the different task nature and domain variability. We discuss foundational attention mechanisms, self-attention, scaled dot-product attention, and multi-head attention, the major components of the original Transformer architecture. Each mechanism has its advantages, that is, parallelizing computations and being able to model relationships between distant tokens in a sequence.

1.1. Paper Organization

The rest of the paper is organized as follows. Section 2 reviews relevant literature on attention mechanisms in transformers. Section 3 outlines the core concepts and fundamental techniques in attention mechanisms. Section 4 presents our proposed classification of attention mechanisms based on efficiency, accuracy, scalability, and domain-specific applications. Section 5 discusses future directions and challenges in attention mechanism research. Finally, Section 6 concludes the paper.

2. Related Work

Attention mechanisms evolved from initial applications in neural machine translation to becoming central components in modern deep learning architectures. Bahdanau et al. [2] introduced a soft attention mechanism for neural machine translation, dynamically weighting source words during decoding.

The introduction of the Transformer architecture by Vaswani et al. [1] revolutionized the field with their "Attention Is All You Need" paper, which eliminated recurrence and convolutions entirely in favor of self-attention mechanisms. This architecture also introduced scaled dot-product attention and multi-head attention, allowing diverse contextual relationships to be captured in parallel.

This innovation allowed models to capture long-range dependencies while enabling parallel computation. Shaw et al. [4] further enhanced this approach by incorporating relative positional embeddings.

For handling longer contexts, Dai et al. [5] developed Transformer-XL, which incorporates segment-level recurrence to capture long-range dependencies beyond a single context window. Child et al. [6] introduced Sparse Transformers to process longer sequences efficiently. In language modeling, significant advances came with BERT by Devlin et al. [7], which used bidirectional pretraining for contextual understanding, and GPT models by Radford et al. [8], which focused on autoregressive modeling.

Recent efficiency-focused research includes linear attention techniques by Katharopoulos et al. [9], which reduce the quadratic scaling to linear complexity. Other approaches like Linformer by Wang et al. [3], BigBird by Zaheer et al. [10], and Switch Transformers by Fedus et al. [11], have addressed scalability challenges.

To enhance representation accuracy, mechanisms like relative positional encoding [5], Axial attention [12], and Dual attention [13] have been proposed. These mechanisms are particularly suited to fine-grained modeling, as they help capture positional and spatial relationships more precisely.

In computer vision, the Vision Transformer (ViT) by Dosovitskiy et al. [14] adapted the transformer architecture for image recognition. Subsequent improvements include hierarchical designs like Swin Transformer by Liu et al. [15] and hybrid models like CvT by Wu et al. [16], which combine CNNs with attention mechanisms. Domain-optimized models like SegFormer [17], Conformer [18], and Adavit [19] highlight the transformer’s capacity for vision-specific adaptation, often at the expense of requiring large-scale pretraining and domain-specific fine-tuning.

Multimodal learning has also benefited from attention mechanisms, with works like those by Fan et al. [20] and Lin et al. [21] implementing cross-attention to improve vision-language tasks. The Multimodal Transformer (MulT) by Tsai et al. [22] handles unaligned multimodal time-series data without explicit alignment.

Recently, Infini-attention [23] has addressed the problem of infinite context lengths by integrating compressive memory into attention, showing promise for long-context modeling.

A growing research branch also explores lightweight and flexible attention mechanisms such as Set Transformer [24], Roformer [25], and Generalized Attention [26], aiming to enhance interpretability and adaptability across input structures. However, these models often face performance drops on unstructured or noisy data and require significant tuning.

Furthermore, graph-based attention mechanisms like Hypergraph Attention [27] and GAAN [28] provide state-of-the-art solutions for graph representation learning, particularly in tasks like node classification and link prediction. Their performance in dynamic and large-scale graphs, however, remains a subject for further optimization.

To illustrate the current landscape, Table 1 summarizes several widely used attention mechanisms and their key characteristics.

Table1. Key Attention Mechanisms in Transformers.

Type of Attention Mechanism	Representative Mechanisms	Key Features
Classic Self-Attention	Self-Attention [1], Scaled Dot-Product Attention [1], Multi-Head Attention [1], Relative Position Encoding [4]	Foundational to Transformer architecture; captures global dependencies; parallelizable
Efficient Attention	Linear Attention [3,9], Linformer [3], Sparse Transformer [6], Big Bird [10], Performer [29], Reformer [30]	Reduces time and space complexity; scalable to long sequences
Multi-Dimensional & Structured Attention	Axial Attention [12], Dual Attention [13], Swin Transformer [15], Conformer [18]	Captures spatial, temporal, or channel-wise relations; suited for high-dimensional data
Memory-Augmented / Recurrence-Based Attention	Transformer-XL [5], Infini-Attention [23], Reformer [30]	Handles very long contexts; leverages recurrence or memory compression
Domain-Specific Architectures	ViT [14], Swin Transformer [15], SegFormer [17], Adavit [19], BERT [7], GPT [8], T5 [36]	Optimized for specific domains (vision, NLP); often pre-trained for performance
Cross-Modal & Graph-Based Attention	MulT [22], IMRAM [31], Generalized Attention [26], Set Transformer [24], Hypergraph Attention [27], GAAN [28]	Enables multimodal learning or structured data modeling (graphs, sets)

3. Core Concepts

3.1. Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. [1] in "Attention is All You Need," has become fundamental to modern deep learning. Unlike traditional sequence models such as RNNs and LSTMs, Transformers rely entirely on attention mechanisms without recurrence or convolutions. This approach offers several advantages:

- **Parallelization**: Enables simultaneous processing of all input tokens, drastically reducing training time on modern hardware.
- **Long-term Dependencies**: Self-attention effectively captures relationships between distant tokens, overcoming limitations of RNNs.
- **Scalability**: Performance improves with increased model size and data, making Transformers suitable for complex tasks.
- **Flexibility**: Successfully adapted beyond NLP to computer vision, speech processing, and multimodal learning.

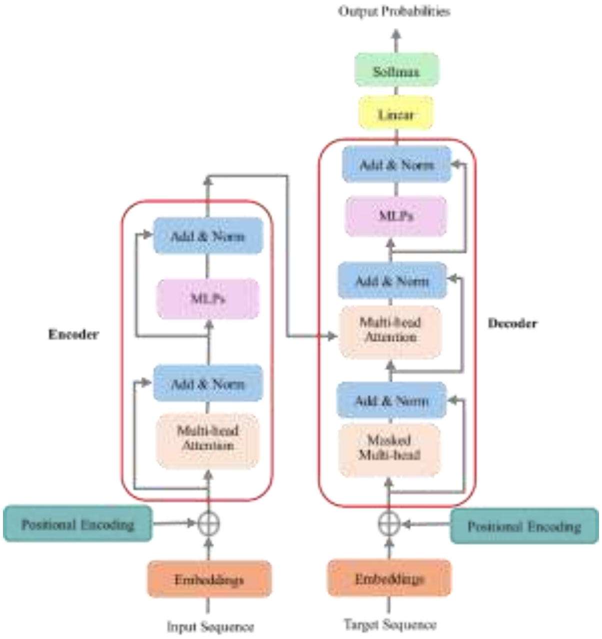

Figure1. Transformer architecture with encoder-decoder structure.

The architecture consists of encoder and decoder components, each comprising multiple identical layers with two main sub-layers: the self-attention layer and the feed-forward layer, as shown in Figure 1. The self-attention mechanism allows complex relationships between all input tokens to be learned regardless of their positions in the sequence.

3.2. Attention Mechanisms

3.2.1. What is Attention?

Attention in deep learning selectively focuses on relevant parts of input data based on context. Inspired by human cognitive processes, attention allows neural networks to assign different weights to input elements, emphasizing important information while disregarding less relevant parts. This capability is particularly valuable for processing sequential or unstructured data.

Mathematically, attention can be formulated using three key components:

- **Query (Q)**: The vector representing the current element seeking attention.
- **Key (K)**: Vectors representing input elements that the query attends to.
- **Value (V)**: The vector containing the actual information to be aggregated.

Equation (1) is the attention score computed as the dot product between query and key vectors, as follows:

$$score(Q.K) = Q.K^T \tag{1}$$

These scores are then normalized, usually by applying a SoftMax function, which ensures that the attention scores sum to one, as Equation (2):

$$\alpha_i = softmax(score(Q.K)) = (exp(score(Q.K_i)))/(\sum_j^n exp(score(Q.K_j))) \quad (2)$$

The output of the attention mechanism is the weighted sum of the values, where each value is weighted by its corresponding attention score α_i, as Equation (3):

$$AttentionOutput = \sum_{i=1}^n \alpha_i V_i \quad (3)$$

This output is then passed through further layers, depending on the architecture.

3.2.2. Self-Attention Mechanism

Self-attention is a specific type where elements in a sequence attend to all other elements in the same sequence. Each token determines attention scores with respect to all others, including itself. This enables the model to capture relationships between positions regardless of their distance. The self-attention operation involves:

1) Compute attention scores for each query-key pair.
2) Normalize the scores using SoftMax to obtain attention weights.
3) Compute the weighted sum of values based on these weights.

Self-attention offers efficient parallelization but has quadratic complexity with respect to sequence length, presenting challenges for long sequences.

3.2.3. Multi-Head Attention

Multi-head attention extends self-attention by applying multiple attention mechanisms in parallel. Rather than using a single set of attention weights, it employs several sets of parameters, creating different "heads" that can focus on various aspects of the inputs simultaneously.

This approach provides:

- **Representation Diversity:** Each head can focus on different parts of the input sequence, capturing varied relationships.
- **Enhanced Learning Capacity:** Multiple heads allow the model to learn richer representations.
- **Improved Stability:** The use of multiple heads reduces the risk of overfitting by distributing attention across different subspaces.

Multi-head attention concatenates outputs from all heads before projecting to the final output dimension, allowing the model to jointly attend to information from different representation subspaces.

4. Classifying Attention Mechanisms

Attention mechanisms can be classified based on various criteria, including design principles, computational efficiency, application domains, and adaptability. We present a comprehensive taxonomy to systematize the consideration of attention models used across different fields.

4.1. Efficiency vs. Accuracy

A fundamental trade-off in attention mechanism design involves balancing computational efficiency with model accuracy.

4.1.1. Efficiency-Focused Mechanisms

- **Scaled Dot-Product Attention** [1]: Uses matrix multiplications for efficiency but requires significant memory for large inputs.
- **Linear Attention** [9] **and Linformer** [3]: Reduces computational complexity to linear scaling, enabling processing of longer sequences with some accuracy trade-offs.
- **Performer** [29]: Approximates the SoftMax function using kernel methods to minimize complexity for long sequences.
- **Big Bird** [10]: Combines local, global, and random attention patterns, enabling efficient handling of extremely long sequences.
- **Switch transformers** [11]: Based on the complexity of the input, the system can adapt the computational effort enhancing both the accuracy and efficiency of the model.
- **Efficient Attention** [9,34]: Combines sparse and low-rank attention techniques to enhance efficiency.
- **Landmark Attention** [32]: Introduces a novel approach that enables transformers to handle arbitrarily long contexts while preserving random-access flexibility.
- **Lean Attention** [33]: Introduces a hardware-aware, scalable attention mechanism tailored for the decode phase of transformer-based models.
- **Sparse Attention** [34]: Applies structured sparsity to focus only on selected token subsets, reducing computation without significantly affecting performance.

4.1.1. Accuracy -Focused Mechanisms

- **Multi-Head Attention** [1]: Provides robustness by focusing on multiple aspects of the input, yielding higher accuracy in tasks like machine translation and language modeling.
- **Relative Position Representations** [4]: Improves position sensitive sequence modeling by incorporating relative positional information.
- **Axial Attention** [12]: Processes rows and columns separately, providing higher accuracy for spatial data and image tasks.
- **Dual Attention** [13]: Enhances scene segmentation through combined channel and spatial attention.

4.2. Scalability

Scalability is crucial for handling high-resolution inputs or long sequences:

- **Longformer** [35]: Uses local attention windows with global tokens to efficiently process long documents.
- **Swin Transformer** [15]: Implements a hierarchical attention mechanism with shifted windows for high-resolution images.
- **Linformer** [3]: Reduces self-attention dimensionality to improve scalability without significant accuracy loss.
- **Reformer** [30]: Leverages locality-sensitive hashing and reversible layers for enhanced scalability
- **Infini-attention** [23]: Processes infinitely long inputs using compressive memory integrated into the attention mechanism.

4.3. Adaptability

Cross task and domain generalization ability is observed through measures of adaptability of attention mechanisms.

- **Generalized Attention** [26]: Demonstrates flexibility by supporting diverse modalities and tasks.
- **IMRAM** [31]: Integrates information across different modalities, excelling in tasks such as image-text retrieval.
- **Roformer** [25]: While still using the rotary positional encodings developed rotor encodes are able to extend the reach of variability based on attainable lengths.
- **Set Transformer** [24]: Adapts attention to permutation-invariant tasks, such as clustering and summarization.
- **Lightweight Attention** [37]: This work proposes dynamic convolutions as a more efficient alternative to self-attention for sequence modeling, offering linear computational complexity by generating time-step-specific convolution kernels, unlike the quadratic cost of self-attention.

4.4. Domain-Specific Optimization

Attention mechanisms tailored to specific domains achieve superior performance:

4.3.1. Vision Transformers

- **ViT** [14]: Adapts transformers for image recognition by treating images as sequences of patches.
- **Swin Transformer** [15]: Uses a hierarchical structure with shifted windows for improved scalability in vision tasks.
- **Conformer** [18]: Couples local and global attention mechanisms for visual recognition.
- **SegFormer** [17]: Optimizes transformer architecture for semantic segmentation tasks.

4.3.2. NLP Transformers

- **BERT** [7]: Employs bidirectional pretraining for contextual language understanding.
- **GPT** [8]: Focuses on autoregressive language modeling for text generation.
- **T5** [36]: Provides a unified framework for various NLP tasks through text-to-text paradigm.
- **Transformer-XL** [5]: Extends context length for improved language modeling.

4.3.3. Graph Attention

- **Hypergraph Attention** [27]: Extends attention to hypergraphs, capturing complex relationships in non-Euclidean data.
- **GAAN** [28]: Focuses on enhancing attention mechanisms for graph-structured data.

4.3.4. Multimodal Transformers

- **MulT** [22]: Handles unaligned multimodal inputs (text, audio, vision) without explicit alignment.
- **IMRAM** [31]: Enhances cross-modal image-text retrieval through iterative matching with recurrent attention.

Table 2 compares various transformer mechanisms by focusing on their strengths and limitations. Table 3 summarizes our classification of attention mechanisms based on their strengths and limitations.

Table 2. Classifying Attention Mechanisms.

Metric	Mechanism	Strengths	Limitations
Efficiency	Scaled Dot-Product Attention [1], Linear Attention[9], Linformer [3], Sparse Attention [34], Performer [29], Big Bird [10], Switch transformers [11], Efficient Attention [9, 34]	Improves computational efficiency, reduces memory overhead, suitable for large-scale data processing, and faster inference times.	Trade-offs in accuracy or generalizability for some tasks, complexity in implementation, and limited adoption in diverse domains.
Accuracy	Multi-Head Attention[1], Relative Position Representations [4], Axial Attention [12], Dual Attention [13]	Achieves high precision for tasks requiring fine-grained understanding, improved contextual representations, and captures complex dependencies effectively.	Increased computational and memory costs, potential overfitting, and diminishing returns for larger models.
Scalability	Longformer [35], Swin Transformer [15], Linformer [3], Reformer [30], Infini-attention [23]	Handles long sequences effectively, supports training on massive datasets, and reduces quadratic complexity in attention.	May lose granularity in local attention, requires tuning for specific domains, and complex architectures can hinder ease of use.
Adaptability	Generalized Attention [26], IMRAM [31], Roformer [25], Set Transformer [24], Lightweight Attention [37]	Flexible across multiple modalities, robust to varying input structures, and transferable across tasks.	May require extensive domain-specific fine-tuning, can introduce interpretability challenges, and performance may degrade on non-standard data.
Vision Transformers	ViT [14], Multiscale vision transformers [20], Swin transformer [15], Conformer [18], SegFormer [17], Adavit [19], Memvit [38], Mvitv2 [39]	Revolutionizes vision tasks with state-of-the-art performance, handles global context efficiently, and supports multiscale inputs.	Requires large-scale pretraining, computationally expensive for high-resolution images, and may lack inductive biases inherent in convolutional architectures.
NLP Transformers	Transformer-xl [6], Longformer [35], T5 [36], Big bird [10], Synthesizer [40], BERT [7], GPT[8]	Excels at a wide range of NLP tasks, efficient pretraining for transfer learning, and models contextualized embeddings effectively.	High resource requirements for training, susceptibility to bias in training data, and can lack domain-specific knowledge without additional fine-tuning.
Graph Attention	Hypergraph Attention [27], GAAN [28]	Captures relationships in graph-structured data, enhances graph learning tasks like node classification and link prediction.	Scalability challenges for very large graphs, may struggle with dynamic graphs, and limited benchmarks compared to other transformer-based approaches.
Multimodal Transformers	Multimodal learning [41], MulT [22]	Integrates information from diverse modalities, enables improved performance in multimodal tasks, and supports applications like vision-language learning.	Complex to train and optimize, requires significant amounts of labeled multimodal data, and potential misalignment between different modalities.

Table 3. Summarizes Classifying Attention Mechanisms.

Metric	Strengths	Limitations
Efficiency	Reduced computation, faster inference, lower memory usage	Potential accuracy trade-offs, implementation complexity
Accuracy	High precision, improved context modeling, fine-grained understanding	Increased computational costs, potential overfitting
Scalability	Handles longer sequences, supports larger datasets, reduces complexity	May sacrifice local attention granularity, requires domain tuning
Domain-Specific	Optimized for particular tasks, state-of-the-art performance	Limited generalization, requires specialized knowledge

The summarized table provides a concise comparison of key aspects of attention mechanisms across various domains. It highlights four main metrics: Efficiency, Accuracy, Scalability, and Domain-Specific Applications. Each category presents notable strengths for instance, reduced computational cost, enhanced contextual understanding, the ability to process long sequences, and optimized performance for specialized tasks. However, these advantages are accompanied by certain limitations such as potential accuracy trade-offs, increased resource demands, reduced generalization, and the need for domain-specific tuning. Overall, the table offers a clear overview of the trade-offs involved in the design and application of attention mechanisms.

5. Future Directions

In recent decades, attention mechanisms have developed quickly and have already changed various areas such as natural language processing (NLP), computer vision (CV), and much more. While attention mechanisms have revolutionized deep learning, several challenges and opportunities remain for future research:

5.1. Enhancing Efficiency for Large-Scale Apps

The increase in model size and complexity has presented a bottleneck problem in terms of the computational and memory requirements of attention mechanisms. Scalability remains a critical challenge for attention mechanisms, particularly in processing long sequences. Models like Linformer [3], Performer [29], and Big Bird [10] have introduced methods to reduce computational overhead. However, further advancements are necessary to address the quadratic complexity of traditional attention. Innovations such as landmark attention [32] and lean attention [33] point toward achieving hardware-aware and linear complexity, offering avenues for real-time processing in constrained environments. These innovations change the way for real-time applications in constrained environments, such as edge computing and low-power systems.

The computational and memory requirements of attention mechanisms present challenges as model size increases:

- **Hardware-Aware Designs**: Developing attention mechanisms optimized for specific hardware architectures (GPUs, TPUs, edge devices).
- **Sub-Linear Complexity**: Pursuing attention variants that achieve sub-linear scaling with respect to sequence length.
- **Conditional Computation**: Implementing dynamic attention that adapts its computation based on input complexity.

5.2 Expanding Contextual Understanding

Advanced contextual embeddings, utilize rich data, which aids in attaining a greater comprehension of intricate datasets. Relative positional embeddings and rotary position embeddings in Roformer [25] techniques have improved the model ability to incorporate context by refining sensitivity to positional relationships within input sequences. Approaches like Infini-attention [23] aim to capture context over infinitely long sequences. This capability has potential to revolutionize fields requiring extensive contextual reasoning, such as storytelling or analyzing genomic data. Multimodal transformers could enhance understanding in complex tasks, such as visual-language reasoning or decision-making in autonomous systems.

Improving the ability to model and utilize context remains an important research direction:

- **Long-Range Dependencies**: Enhancing mechanisms like Infini-attention [18] to capture dependencies over extremely long contexts.
- **Hierarchical Understanding**: Developing attention that operates at multiple levels of abstraction simultaneously.
- **Contextual Integration**: Combining local and global context more effectively in unified attention frameworks.

5.3. Broader Applications and Adaptability

Attention mechanisms are increasingly applied in dynamic systems and multimodal tasks. Dynamic attention mechanisms, such as lightweight attention [37], can handle time-variant and event-driven data in areas like robotics and real-time analytics. In graph-based domains, techniques like GAAN [28] and Hyper-SAGNN [27] can capture complex relational structures. New approaches may extend attention to tensor-structured data, opening opportunities in chemistry, biology, and environmental modeling. However, hypernetwork attention models and meta-learning frameworks offer adaptability for multi-task learning and domain-specific applications.

Extending attention mechanisms to new domains and enhancing their adaptability:

- **Cross-Domain Generalization**: Creating attention mechanisms that transfer effectively between different data types and domains.
- **Dynamic Systems**: Adapting attention for time-varying and event-driven data in robotics and real-time analytics.
- **Multi-Task Learning**: Developing attention mechanisms that can simultaneously serve multiple objectives without performance degradation.

5.4 Interpretability and Robustness

Addressing concerns about the transparency and reliability of attention-based models:

- **Explainable Attention**: Developing visualization techniques and architectures that provide insight into attention decisions.
- **Robust Attention**: Creating mechanisms resistant to adversarial attacks and input perturbations.
- **Calibrated Confidence**: Ensuring attention weights accurately reflect confidence in model predictions.

Improving attention mechanisms is key for reliable use in sensitive fields like healthcare and finance. Techniques such as visualization and anomaly control help manage biased attention patterns. Using probabilistic attention and context control can further boost robustness. Comparing human and AI attention may also lead to more transparent and explainable models.

The future of attention mechanisms lies in addressing these challenges while expanding applications to new domains. As research progresses, we anticipate attention mechanisms will continue to evolve, becoming more efficient, interpretable, and adaptable across diverse AI tasks.

6. Conclusion

In this survey, we have explored the evolution, key developments, and challenges of attention mechanisms in transformer architectures. The emergence of attention mechanisms has revolutionized machine learning, enabling breakthrough advances in natural language processing, computer vision, and multimodal applications. The self-attention concept introduced in the seminal "Attention Is All You Need" paper [24] transformed modelling approaches by enabling selective focus on input elements rather than processing them sequentially.

Attention mechanisms have significantly improved model efficiency, adaptability, and scalability. However, challenges persist in computational complexity, contextual understanding, interpretability, and robustness. Our classification framework provides a systematic approach to understanding different attention variants based on efficiency, accuracy, scalability, and domain-specific optimization.

Future research directions include developing more efficient attention mechanisms for large-scale applications, enhancing contextual understanding across longer sequences, expanding to new domains, and improving interpretability and robustness. As attention mechanisms continue to evolve, they will undoubtedly play a central role in advancing deep learning and artificial intelligence, enabling more efficient, interpretable, and capable systems across diverse applications.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *"Attention Is All You Need,"* Advances in Neural Information Processing Systems (NeurIPS), vol. 30, Long Beach, CA, USA, 2017.

[2] D. Bahdanau, *"Neural machine translation by* jointly *learning to align and translate,"* in 3rd International Conference on Learning Representation (ICLR 2015), 2015.

[3] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, *"Linformer: Self-attention with linear complexity,"* arXiv: 2006.04768, 2020.

[4] P. Shaw, J. Uszkoreit, and A. Vaswani, "*Self-attention with relative position representations,"* in NAACL, 2018.

[5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, *Transformer-xl: "Attentive language models beyond a fixed-length context,"* in 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). ACL, 2019, pp. 2978–2988.

[6] R. Child, S. Gray, A. Radford, and I. Sutskever, *"Generating long sequences with sparse transformers,"* arXiv:1904.10509, 2019.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "*Bert: Pre-training of deep bidirectional transformers for language understanding,"* arXiv: 1810.04805, 2018.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *"Language models are* unsupervised *multitask learners,"* OpenAI blog, 2019. 1(8): p. 9.

[9] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, *"Transformers are rnns: Fast autoregressive transformers with linear attention,"* in *International conference on machine learning*. 2020. PMLR.

[10] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, *"Big bird: Transformers for longer sequences,"* Advances in neural information processing systems, 2020. 33: p. 17283-17297.

[11] W. Fedus, B. Zoph, and N. Shazeer, *"Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,"* Journal of Machine Learning Research, 2022. 23(120): p. 1-39.

[12] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, *"Axial attention in multidimensional transformers,"* arXiv: 1912.12180, 2019.

[13] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, *"Dual attention network for scene segmentation,"* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *"An image is worth 16x16 words: Transformers for image recognition at scale,"* arXiv: 2010.11929, 2020.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, *"Swin transformer: Hierarchical vision transformer using shifted windows,"* in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[16] H. Wu, B. Xiao, N. Codella, and M. Liu, *"Cvt: Introducing convolutions to vision transformers,"* in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, *"SegFormer: Simple and efficient design for semantic segmentation with transformers,"* Advances in neural information processing systems, 2021. 34: p. 12077-12090.

[18] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, *"Conformer: Local features coupling global representations for visual recognition,"* in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[19] L. Meng, H. Li, B. Chen, S. Lan, Z. Wu, Y. Jiang, and S. Lim, *"Adavit: Adaptive vision transformers for efficient image recognition,"* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[20] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "*Multiscale vision transformers,"* in *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[21] H. Lin, X. Cheng, X. Wu, F. Yang, D. Shen, Z. Wang, Q. Song, and W. Yuan, *"Cat: Cross attention in vision transformer,"* in *2022 IEEE international conference on multimedia and expo (ICME)*. 2022. IEEE.

[22] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov, *"Multimodal transformer for unaligned multimodal language sequences,"* in *Proceedings of the conference. Association for computational linguistics. Meeting*. 2019. NIH Public Access.

[23] T. Munkhdalai, M. Faruqui, and S. Gopal, "*Leave no context behind: Efficient infinite context transformers with infini-attention,"* arXiv: 2404.07143, 2024.

[24] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. The, "*Set transformer: A framework for attention-based permutation-invariant neural networks,"* in *International conference on machine learning*. 2019. PMLR.

[25] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen, and Y. Liu, *"Roformer: Enhanced transformer with rotary position embedding,"* Neurocomputing, 2024. 568: p. 127063.

[26] D. Heo, and H. Choi, *"Generalized Probabilistic Attention Mechanism in Transformers,"* arXiv: 2410.15578, 2024.

[27] R. Zhang, Y. Zou, and J. Ma, *"Hyper-SAGNN: a self-attention based graph neural network for hypergraphs,"* arXiv: 1911.02613, 2019.

[28] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D. Yeung, *"Gaan: Gated attention networks for learning on large and spatiotemporal graphs,"* arXiv: 1803.07294, 2018.

[29] K. Choromanski, V. Likhosherstov, D. Dohan, and X. Song, *"Rethinking attention with performers,"* arXiv: 2009.14794, 2020.

[30] N. Kitaev, Ł. Kaiser, and A. Levskaya, *"Reformer: The efficient transformer,"* in 8th International Conference on Learning Representations (ICLR 2020), 2020.

[31] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, *"Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,"* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

[32] A. Mohtashami, and M. Jaggi, *"Landmark attention: Random-access infinite context length for transformers,"* arXiv: 2305.16300, 2023.

[33] R. Sanovar, S. Bharadwaj, R. S. Amant, V. Rühle, and S. Rajmohan, *"Lean Attention: Hardware-Aware Scalable Attention Mechanism for the Decode-Phase of Transformers,"* arXiv: 2405.10480, 2024.

[34] A. Roy, M. Saffar, A. Vaswani, and D. Grangier, *"Efficient content-based sparse attention with routing transformers,"* Transactions of the Association for Computational Linguistics, 2021. 9: p. 53-68.

[35] I. Beltagy, M. E. Peters, and A. Cohan, *"Longformer: The long-document transformer,"* arXiv: 2004.05150, 2020.

[36] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *"Exploring the limits of transfer learning with a unified text-to-text transformer,"* Journal of machine learning research, 2020. 21(140): p. 1-67.

[37] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, *"Payless attention with lightweight and dynamic convolutions,"* arXiv: 1901.10430, 2019.

[38] C. Wu, Y. Li, K. Mangalam, H. Fan, B. Xiong, J. Malik, and C. Feichtenhofer, *"Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition,"* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[39] Y. Li, C. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, *"Mvitv2: Improved*

multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[40] Y. Tay, D. Bahri, D. Metzler, D. Juan, Z. Zhao, and C. Zheng, *"Synthesizer: Rethinking self-attention for transformer models,"* in *International conference on machine learning*. 2021. PMLR.

[41] P. Xu, X. Zhu, and D. A. Clifton, *"Multimodal learning with transformers: A survey,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023. 45(10): p. 12113-12132.

مروری کلی بر سازوکارهای توجه در مدل‌های ترنسفورمر

رسول حسین‌زاده* و مهدی صادق‌زاده

گروه مهندسی کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

ارسال ۱۴۰۳/۱۰/۲۹؛ بازنگری ۱۴۰۳/۱۱/۱۰؛ پذیرش ۱۴۰۴/۰۳/۰۶

چکیده:

سازوکارهای توجه پیشرفت چشمگیری در حوزه یادگیری ماشین و یادگیری عمیق در زمینه‌های مختلفی مانند پردازش زبان طبیعی، بینایی ماشین و سیستم‌های چندوجهی ایجاد کرده‌اند. این مقاله یک مرور جامع از سازوکارهای توجه در معماری‌های ترنسفورمر ارائه می‌دهد و بر سیر تحول، روش‌های طراحی و کاربردهای اختصاصی آن‌ها در حوزه‌های پردازش زبان طبیعی، بینایی ماشین و یادگیری چندوجهی تمرکز دارد. ما انواع سازوکارهای توجه را بر اساس اهدافی مانند کارایی، مقیاس‌پذیری و قابلیت تفسیرپذیری دسته‌بندی کرده و تحلیلی تطبیقی از نقاط قوت، محدودیت‌ها و موارد کاربرد مناسب آن‌ها ارائه می‌کنیم. این مقاله به خلأ موجود در زمینه شهود ب صری نیز می‌پردازد و با ارائه یک طبقه‌بندی شفاف‌تر و برر سی رویکردهای ترکیبی، همچون ترکیب‌های پراکنده- سل سله‌مراتبی، درک جامع‌تری از سازوکارهای توجه فراهم می‌سازد. علاوه بر بررسی سازوکارهای بنیادین، مقاله به رویکردهای ترکیبی، مبانی نظری و ملاحظات عملی نیز توجه مناسبی دارد. در ادامه، چالش‌های فعلی همچون هزینه محاسباتی، پایداری عملکرد و شفافیت شناسایی و تحلیل شده و طبقه‌بندی منظمی همراه با پیشنهادهایی برای مسیرهای آینده ارائه شده است. با مقایسه جدیدترین تکنیک‌های روز، این مقاله تلاش دارد تا راهنمایی برای پژوهشگران در انتخاب و طراحی سازوکارهای توجه متناسب با کاربردهای خاص هوش مصنوعی ارائه دهد و در نهایت موجب توسعه مدل‌های ترنسفورمر کارآمدتر، قابل تفسیرتر و سازگارتر شود.

کلمات کلیدی: سازوکار توجه، ترنسفورمر، یادگیری عمیق.