Journal of Electronic Research and Application, 2025, Volume 9, Issue 3
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

# The Latest Research Progress of Attention Mechanism in Deep Learning

**Xu Jiang[1*], Xiaoling Bai[1], Lifeng Yin[2]**

[1]School of Software, Harbin Institute of Information Technology, Harbin 150431, Heilongjiang, China
[2]School of Intelligent Railway Engineering, Dalian Jiaotong University, Dalian 116028, Liaoning, *China*

**Author to whom correspondence should be addressed.*



**Abstract:** With the development of artificial intelligence and deep learning, the attention mechanism has become a key technology for enhancing the performance of complex tasks. This paper reviews the evolution of attention mechanisms, including soft attention, hard attention, and recent innovations such as multi-head latent attention and cross-attention. It focuses on the latest research outcomes, such as lightning attention, the PADRe polynomial attention replacement algorithm, the context anchor attention module, and improvements in attention mechanisms for large models. These advancements improve the efficiency and accuracy of models, expanding the application potential of attention mechanisms in fields such as computer vision, natural language processing, and remote sensing object detection, aiming to provide readers with a comprehensive understanding and stimulate innovative thinking.

**Keywords:** Natural language processing; Computer vision; Attention mechanism; Large models

**Online publication:** 29 May, 2025

## 1. Introduction

In the field of artificial intelligence, particularly with the rapid development of deep learning, the attention mechanism has emerged as an indispensable core technology [1–3]. This mechanism simulates the human ability to "focus," allowing models to actively filter out the most critical information when faced with massive amounts of data. Through this approach, models can not only enhance their ability to capture important information but also effectively reduce unnecessary computational burdens, thereby improving overall performance and efficiency. The introduction of the attention mechanism has enabled deep learning models to perform more intelligently in complex tasks, better understanding and analyzing complex structures and relationships within data. In recent years, significant progress has been made in the study of attention mechanisms, driving innovative applications in various domains. Especially in the two primary directions of computer vision and natural language processing [4–7], researchers have continuously explored and optimized various attention mechanisms to enhance model performance and adaptability. In computer vision, attention mechanisms assist models in more precisely

focusing on key features within images, leading to remarkable results in tasks such as object detection and image segmentation. In natural language processing, attention mechanisms have facilitated breakthroughs in machine translation, text generation, sentiment analysis, and other tasks, enabling models to better understand contextual information and semantic relationships. This article delves into the latest research advancements of attention mechanisms from these two perspectives, providing detailed introductions to their working principles and application examples in computer vision and natural language processing. It demonstrates how attention mechanisms are pushing the boundaries of artificial intelligence technology. By analyzing these recent research achievements, the article offers a more comprehensive understanding and recognition of attention mechanisms, thereby inspiring further innovative thinking and applications.

## 2. Overview of the development of attention mechanisms

In 2014, Bahdanau *et al*. introduced the soft attention mechanism or additive attention [6], a deterministic approach to attention allocation. This mechanism assigns weights to each part of the input data, which are calculated based on similarity with a query vector. The model then performs a weighted sum of the input data using these weights, generating a weighted representation for subsequent processing. Its main advantage is that the computation process is differentiable, facilitating backpropagation training. However, a drawback is that it may assign weights to some unimportant information, affecting the model's ability to focus.

In 2015, Luong *et al*. proposed the hard attention mechanism, also known as stochastic attention [7]. This mechanism samples among different parts of the input data, focusing on a specific part at a time instead of assigning weights to all parts. This randomness allows hard attention to explore data more effectively in certain tasks, but also poses training challenges due to the non-differentiable nature of the random sampling process, requiring special optimization techniques. In the same year, Sukhbaatar *et al*. introduced a memory-based attention mechanism that incorporates a memory module [8], considering both input data and memory information when calculating attention weights. Memory can store previously processed data or prior knowledge, enabling the model to reference these memories in current tasks and make more accurate decisions. For example, in dialogue systems, memory-based attention helps the model better understand dialogue history, improving the coherence and accuracy of responses. Additionally, Yao *et al*. proposed a temporal attention mechanism in that year [9], focusing on time series features in data. This mechanism assigns attention weights based on temporal order, enabling the model to better understand temporal dependencies and dynamic changes when processing time series data such as speech recognition and video analysis. By focusing on data points at different time steps, the model can capture key information in the time series, enhancing modeling and prediction capabilities.

In 2016, Martins and Astudillo presented the sparse attention mechanism, which introduces sparsity constraints to reduce redundant information, resulting in a sparser distribution of attention weights with only a few elements having larger weights [10]. This mechanism improves the model's computational efficiency and scalability, effectively reducing computational resource consumption when handling large-scale data and complex tasks such as long texts and high-resolution images. Meanwhile, Yang *et al*. introduced a hierarchical attention mechanism that divides attention into multiple levels, with lower levels focusing on word features and higher levels focusing on sentence or paragraph features [11]. This hierarchical approach enhances the model's understanding of hierarchical structures and semantic information, particularly excelling in natural language processing tasks such as text generation and question answering systems.

In 2017, Vaswani *et al*. introduced the self-attention mechanism and multi-head self-attention mechanism [1]. Self-attention focuses on relationships between internal parts of the input data, assigning attention weights by calculating similarities between elements without requiring additional context information. It effectively captures long-distance dependencies and enhances the understanding of semantic relationships between words. The multi-head attention mechanism extends self-attention by dividing the input data into multiple "heads," independently calculating attention weights, and then concatenating or performing a weighted sum. This allows the model to attend to data from multiple perspectives and capture more features. This mechanism is the core of the Transformer architecture and has been widely applied in natural language processing tasks, achieving significant results. The adaptive attention mechanism, proposed by Lu *et al*., aimed to enable the model to automatically adjust attention allocation based on task requirements and input data [12]. By learning a set of parameters, this mechanism improves the model's ability to select the best attention strategy in different scenarios, enhancing flexibility and generalization capabilities to adapt to complex and varying task environments, such as automatically adjusting the level of focus on different regions in image recognition.

In 2021, Gheini *et al*. introduced the cross-attention mechanism [13], primarily applied in encoder-decoder architectures. It allows the decoder to calculate attention weights based on the current decoding state and encoder output, selectively focusing on information relevant to the current decoding task. This mechanism is particularly important in tasks like machine translation and text summarization, helping the model better establish correspondences between the source and target languages.

## 3. Latest research achievements in attention mechanisms

The attention mechanism originates from observations of the human visual system, where humans automatically focus on interesting parts of a scene while ignoring other irrelevant information. Inspired by this, researchers have introduced attention mechanisms into deep learning models to improve their performance in handling complex tasks. In recent years, attention mechanisms have achieved significant results in computer vision, natural language processing, speech recognition, and other fields.

### 3.1. Next-generation attention mechanism – lightning attention

Qin *et al*. proposed a new attention mechanism called lightning attention [14]. This mechanism demonstrates significant improvements in computational efficiency and effectiveness, potentially becoming an important alternative to future Transformer architectures. The network architecture diagram is shown in **Figure 1**.

The framework of lightning attention is shown in **Figure 1**, which illustrates its algorithmic diagram. In the i-th iteration, the tiled blocks of matrices $Q_i$, $K_i$, and $V_i$ are transferred from High-Bandwidth Memory (HBM) to Static Random-Access Memory (SRAM). In SRAM, the outputs Ointra and Ointer are independently calculated, followed by an update to the KV matrix. Finally, the output $Q_i$ (the sum of Ointra and Ointer) is written back from SRAM to HBM.

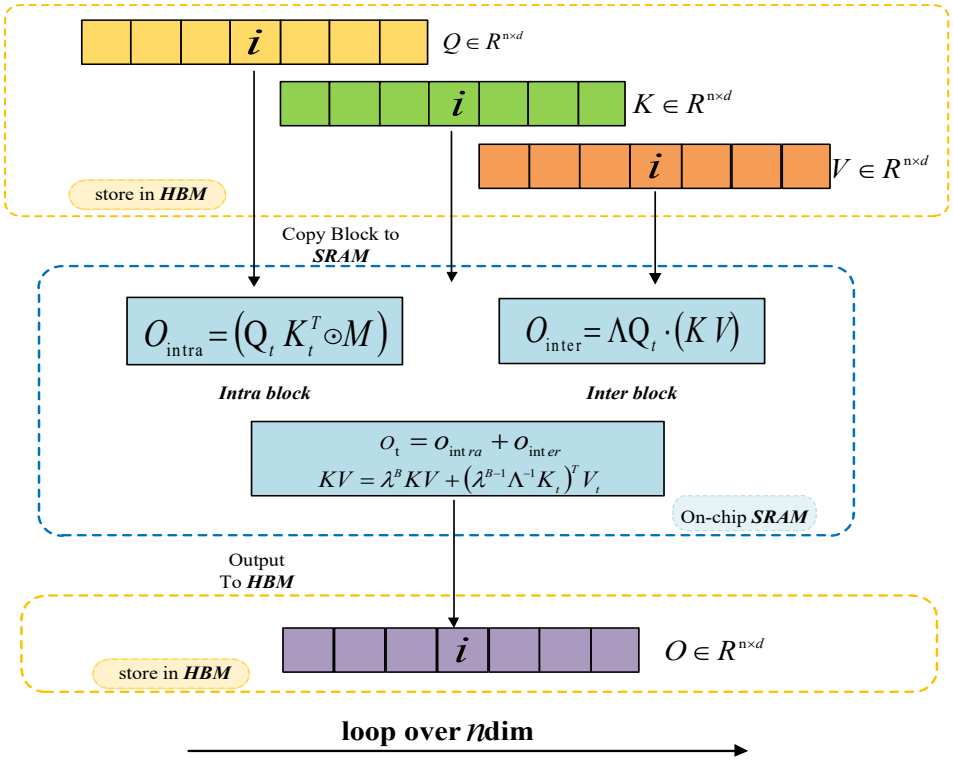


**Figure 1.** Structural framework of lightning attention

## 3.2. Multi-head latent attention mechanism

Multi-head latent attention introduces latent variables and based on multi-head attention [15], and its structural framework is shown in **Figure 2**. It aims to better capture the complex structures and uncertainties in data by learning the latent attention distribution. Compared to traditional multi-head attention, it can more flexibly model changes in attention weights, thus demonstrating better adaptability and robustness when processing noisy or ambiguous data, effectively extracting useful information.

## 3.3. Sparse attention in large models

Native Sparse Attention (NSA) is an efficient solution for modeling long contexts, combining algorithmic innovations with hardware-aligned optimizations [16]. NSA employs a dynamic hierarchical sparsity strategy, reducing computation through coarse-grained token compression while maintaining global contextual awareness and local precision through fine-grained token selection. The model's two key innovations lie in achieving arithmetically intensity-balanced algorithmic design, significantly accelerating computations, and optimizing for modern hardware. Additionally, NSA supports end-to-end training, reducing pre-training computations without sacrificing performance. Its network architecture is shown in **Figure 3**.

**Figure 3** (Left) shows the framework processes the input sequence through three parallel attention branches. For a given query, previous keys and values are transformed into compressed attention to identify coarse-grained patterns; important token blocks are focused on through selective attention; and local contexts are captured through sliding attention. The right shows a visualization of different attention patterns generated by each branch. The green areas represent parts where attention scores need to be calculated, while the white areas indicate parts that can be skipped.

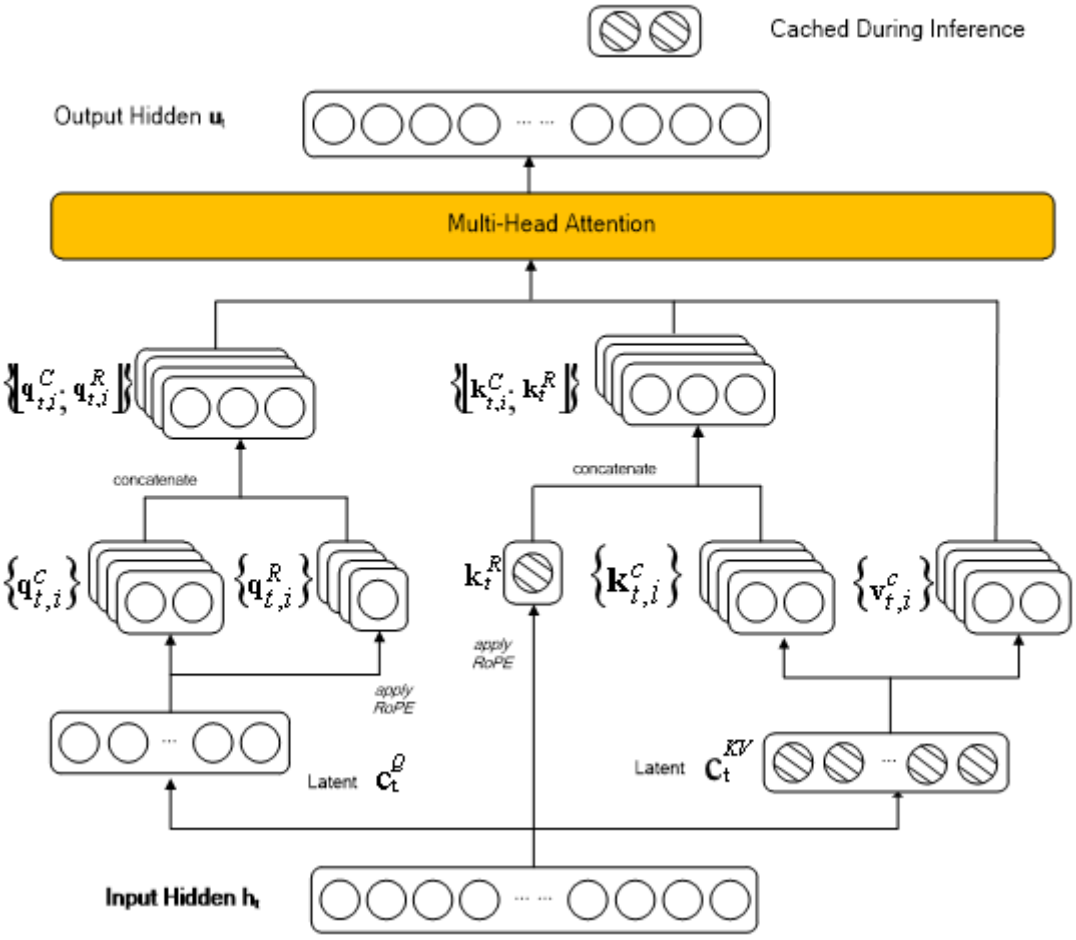


**Figure 2.** Framework diagram of multi-head latent attention

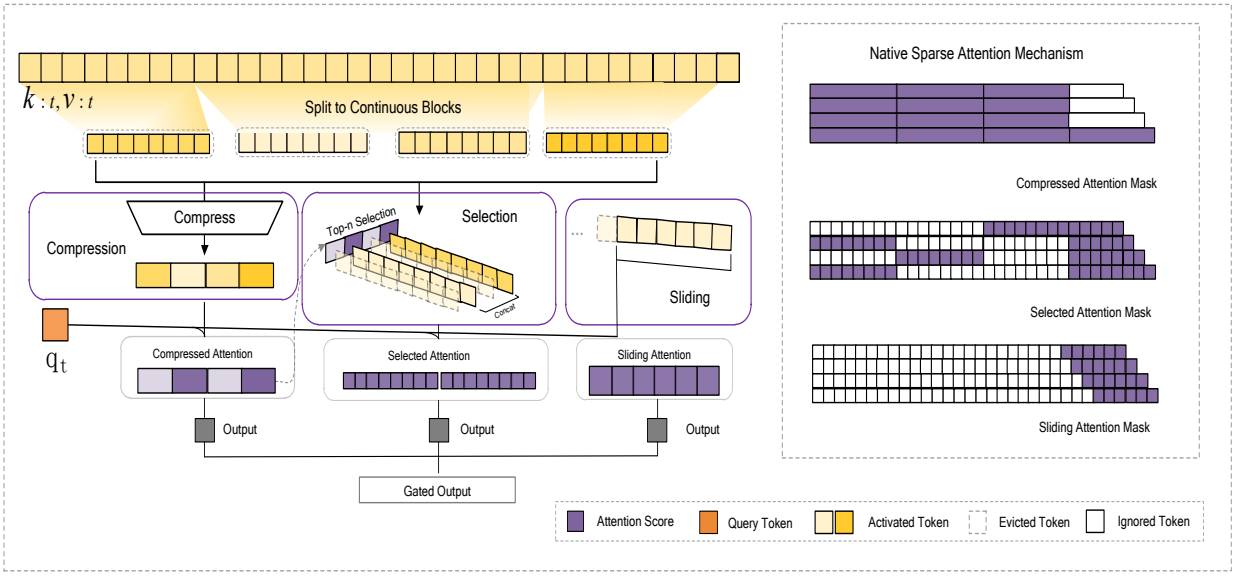


**Figure 3**. Sparse attention network architecture diagram

### 3.4. Attention mechanism in remote sensing target detection

In the YOLOv11 network, researchers have added a new attention mechanism [17], namely the Context Anchor Attention (CAA) module, which captures long-distance contextual information in parallel for remote sensing target detection.

### 3.5. Polynomial attention alternative algorithm for efficient visual transformers

PADRe is a novel polynomial attention alternative algorithm designed to improve the computational efficiency of visual Transformers [18]. This algorithm provides a unified framework that can effectively replace traditional self-attention mechanisms, thereby enhancing overall performance. PADRe utilizes polynomial functions and draws on relevant achievements in approximation theory to enhance computational efficiency without sacrificing accuracy. Its key components include multiplicative nonlinearity, employing intuitive and hardware-friendly operations such as the Hadamard product, resulting in only linear computational and memory overhead. Additionally, PADRe avoids using complex functions such as Softmax while achieving comparable or superior accuracy to traditional self-attention mechanisms.

### 3.6. Scaled dot-product attention

According to **Figure 4**, scaled dot-product attention is a commonly used attention computation method. It calculates attention scores by computing the dot product of query vectors and key vectors, then scales the scores to prevent numerical instability. Finally, it uses the Softmax function for normalization to obtain attention weights. This mechanism is simple and efficient, widely used in many Transformer-based models, giving the models powerful representation capabilities.

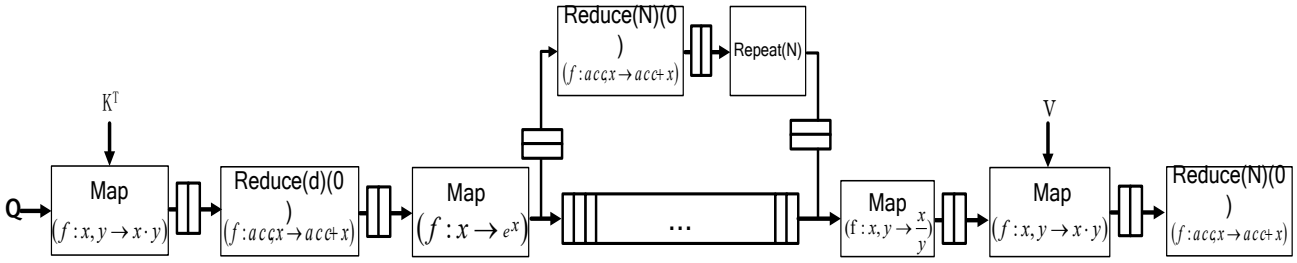


**Figure 4.** Scaled dot-product attention

## 4. Conclusion

This article systematically reviews and analyzes the latest research progress of attention mechanisms, demonstrating their wide range of applications in multiple fields and the trend of continuous evolution. New attention mechanisms, such as lightning attention and multi-head latent attention, exhibit superior performance and flexibility in handling complex tasks. Meanwhile, with the continuous development of technology, improvements in attention mechanisms for large models and multimodal tasks also point to future research directions. By deeply understanding these mechanisms and their applications, researchers and engineers can better utilize attention mechanisms to enhance the intelligence level of artificial intelligence models, promoting innovation and development in various fields.

## Funding

Dalian Social Science Project (2024dlsky095)

## Disclosure statement

The authors declare no conflict of interest.

## References

[1] Vaswani A, Shazeer N, Parmar N, et al., 2017, Attention is All You Need. Advances in Neural Information Processing Systems, 30: 5998–6008.

[2] Hu J, Shen L, Sun G, 2018, Squeeze and Excitation Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 7132–7141.

[3] Devlin J, Chang MW, Lee K, et al., 2019, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers), Minneapolis, Minnesota, 4171–4186.

[4] Wang X, Girshick R, Gupta A, et al., 2018, Non-local Neural Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 7794–7803.

[5] Anderson P, He X, Buehler C, et al., 2018, Bottom-up and top-down Attention for Image Captioning and Visual Question Answering, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6077–6086.

[6] Bahdanau D, Cho K, Bengio Y, 2014, Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arXiv.1409.0473

[7] Luong MT, Pham H, Manning CD, 2015, Effective Approaches to Attention-based Neural Machine Translation. https://doi.org/10.48550/arXiv.1508.04025

[8] Sukhbaatar S, Weston J, Fergus R, 2015, End-to-end Memory Networks. Advances in Neural Information Processing Systems, 28.

[9] Yao L, Torabi A, Cho K, et al., 2015, Describing Videos by Exploiting Temporal Structure, Proceedings of the IEEE International Conference on Computer Vision, 4507–4515.

[10] Martins A, Astudillo R, 2016, From Softmax to Sparsemax: A Sparse Model of Attention and Multi-label Classification, International Conference on Machine Learning. PMLR, 1614–1623.

[11] Yang Z, Yang D, Dyer C, et al., 2016, Hierarchical Attention Networks for Document Classification. Association for Computational Linguistics, 2016: 1480–1489.

[12] Lu J, Xiong C, Parikh D, et al., 2017, Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 375–383.

[13] Gheini M, Ren X, May J, 2021, Cross-attention is All You Need: Adapting Pretrained Transformers for Machine Translation. https://arxiv.org/abs/2104.08771

[14] Qin Z, Sun W, Li D, et al., 2024, Lightning Attention-2: A Free Lunch for Handling Unlimited Sequence Lengths in Large Language Models. https://arxiv.org/abs/2401.04658

[15] Liu A, Feng B, Wang B, et al., 2024, Deepseek-v2: A Strong, Economical, and Efficient Mixture-of-experts Language Model. https://arxiv.org/abs/2405.04434

[16] Yuan J, Gao H, Dai D, et al., 2025, Native Sparse Attention: Hardware-aligned and Natively Trainable Sparse Attention. https://arxiv.org/abs/2502.11089

[17] Cai X, Lai Q, Wang Y, et al., 2024, Poly Kernel Inception Network for Remote Sensing Detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 27706–27716.

[18] Letourneau PD, Singh MK, Cheng HP, et al., 2024, Padre: A Unifying Polynomial Attention Drop-in Replacement for Efficient Vision Transformer. https://arxiv.org/abs/2407.11306