The 14th IEEE International Conference on Dependable Systems, Services and Technologies, DESSERT'2024
11-13 October, 2024, Athens, Greece

Integrating Mixture of Experts into Transformers Architecture to Control UAV Swarms

Vadym Slyusar
The Central Research Institute of Armaments and Military Equipment of the Armed Forces of Ukraine
Kyiv, Ukraine
swadim@ukr.net

Nataliia Bihun
The Central Research Institute of Armaments and Military Equipment of the Armed Forces of Ukraine
Kyiv, Ukraine
bigun0714@ukr.net

***Abstract* — This article explores the application of Mixture of Experts technology to improve the efficiency of controlling autonomous swarms of unmanned aerial vehicles (UAVs). The authors analyze and develop methods for integrating Mixture of Experts into neural transformers used for data processing in U unmanned aerial vehicle swarms. The authors propose the use of specialized submodels ("experts") to solve specific tasks, such as visual data processing, navigation, meteorological data analysis, and communication between UAVs. The study demonstrates that integrating Mixture of Experts with transformers significantly improves the performance of unmanned aerial vehicle swarms through more efficient task allocation and adaptation to changing environmental conditions. This is achieved by ensuring that each "expert" processes data relevant to its specialization, and the system dynamically selects "experts" depending on the situation. In addition, the authors consider the problem of limited communication channel bandwidth and propose a solution based on quadrature amplitude modulation to ensure reliable data transmission. The scientific novelty of the study is the development of an optimized Mixture of Experts architectures that consider the computational and resource limitations of unmanned aerial vehicles. The practical significance is to create more efficient autonomous systems for reconnaissance and rescue missions. The research opens up new avenues for developing algorithms to determine the relevance of "experts" to specific tasks and study the impact of different Mixture of Experts architectures on the overall system performance.**

Keywords — unmanned aerial vehicle swarm, Mixture of Experts, neural network, transformers, autonomous systems, data processing

I. Introduction

The improvement of neural networks is of particular relevance as the field of artificial intelligence is rapidly evolving and the need for automated and optimized data processing is growing. A frequent challenge in solving this problem is to ensure high data processing accuracy while minimizing computational costs and speeding up system response. This issue is also relevant in managing swarms of unmanned aerial vehicles (UAVs), as the efficiency of data processing directly affects mission success and flight safety.

Due to scalability limitations, traditional neural network architectures [1], [2] are not always able to efficiently handle complex data processing tasks, making them ineffective for specialized tasks requiring high accuracy.

The concept of the Mixture of Experts (MoE) offers an alternative approach [3], [4], it allows the distribution of tasks among specialized subnets ("experts"), increasing the adaptability and performance of models [5], [6]. With MoE, each UAV or group of UAVs in a swarm can focus on specific tasks during missions. However, the integration of MoE into UAV swarm management requires careful analysis and development of optimized approaches that will ensure high system performance without compromising computational efficiency.

The research problem is to develop and analyse effective methods for integrating MoE into a neural network architecture to optimize the performance of UAV swarms. This includes finding ways to increase the adaptability and specialization of models to handle various tasks in dynamic environments and developing strategies for optimal allocation of computing resources.

The rest of the paper is structured as follows.

In the following section, an overview of current research on the integration of MoE architectures into transformers and their application in managing swarms of unmanned aerial vehicles (UAVs) is presented. Special attention is given to the challenges of resource optimization and communication efficiency.

Section III contains a detailed analysis of MoE architectures, focusing on models and methods for expert selection and task specialization in the context of UAV swarms. Additionally, an example of practical implementation is presented, demonstrating the performance improvement of UAV swarms using attention mechanism models and task allocation optimization.

Finally, in Section IV, the key conclusions are summarized and potential directions for future research are outlined.

II. Analysis of the Latest Research and Publications

Modern artificial intelligence research is increasingly turning to MoE architectures, noting their ability to efficiently process large amounts of data. In particular, recent works demonstrate the potential of MoE to solve the problems faced by UAV swarms, especially when integrating this technology into the architecture of transformers.

Gormley and Frühwirth-Schnatter [7] laid the fundamental theoretical foundations for understanding MoE, emphasizing its versatility and suitability for various tasks. Subsequent work, building on this foundation, has focused on optimizing MoEs, developing new architectures, and expanding their applications. For example, Krishnamurthy, Watkins, and Gaertner [8] proposed a new gateway architecture and regularisation techniques to improve the performance and specialization of "experts", while Jawahar et al. [9] demonstrated the effectiveness of MoE in neural machine translation.

Further research, such as the development of the "Mixture of Tokens" model [10] and the implementation of MoE in

979-8-3315-1093-0/24/$31.00 ©2024 IEEE

transformer-based architectures for pattern recognition [11], [12], demonstrates the versatility and promise of MoE. Puigcerver, Riquelme, Mustafa, and Houlsby [13] contributed significantly to solving the problems of scalability and instability of MoE learning by introducing the concept of Soft MoE.

Recent studies have further explored the integration of Mixture of Experts (MoE) architectures into Unmanned Aerial Vehicle (UAV) swarm control systems. For instance, Verdoucq et al. (2022) proposed a bio-inspired control model for collective UAV swarm motion, demonstrating that reactive algorithms derived from natural phenomena can enhance coordination and task efficiency in dynamic environments [14]. Similarly, Wu et al. (2022) developed a UAV swarm formation transformation algorithm that optimizes formation control and enhances scalability in UAV operations—a key requirement for large-scale MoE implementations [15].

Furthermore, Karampelia et al. (2023) presented task allocation models specifically designed for UAV swarms, which optimize energy consumption and task distribution by considering the individual capabilities of the drones [16]. This research addresses a critical challenge of MoE systems in dynamic environments. Finally, Sharma et al. (2023) developed a cloud-based control system for UAV swarms, leveraging real-time localization and communication to manage UAV formations efficiently [17]. This study underscores the potential of MoE architectures to enhance real-time decision-making in resource-constrained environments.

Despite significant progress, the use of MoE in UAV swarm control, especially with the use of neural transformers, is still an under-researched area. Existing problems and unresolved issues make the development of optimized MoE frameworks for UAV swarms an urgent research task.

This research aims to address these issues by studying in detail the integration of MoE into transformer architecture to improve the functioning of UAV swarms. The work aims to analyze and develop methods that will expand the capabilities of UAV swarms by improving neural transformers using Mixture of Experts technology.

III. Summary of the Main Research Material

Using the advanced capabilities of neural transformers, this study considers the integration of Mixture of Experts models into the UAV swarm management system. The research methodology is based on a comprehensive approach that includes the selection of optimal MoE models, analysis of the features of neural transformers, and consideration of the dynamics of UAV swarms. A UAV swarm functions as a network of interconnected drones equipped with sensors and communication modules, which allows them to exchange data and make decisions in real-time (Fig. 1).

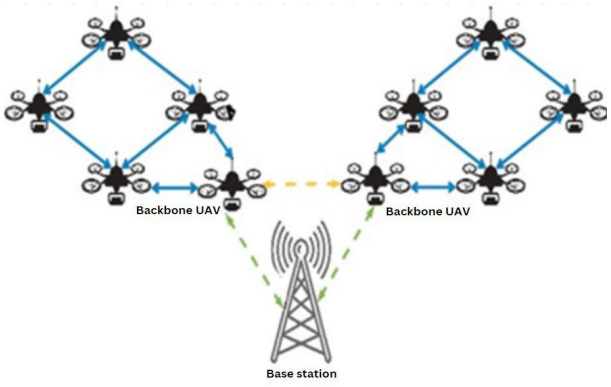

Fig. 1. UAV swarm network option.

This system is designed for autonomous missions, and the MoE models built based on neural transformers provide an intelligent distribution of tasks among the UAV swarm. The MoE technology in machine learning is based on combining several "expert" models to achieve better performance [3] (Fig. 2).

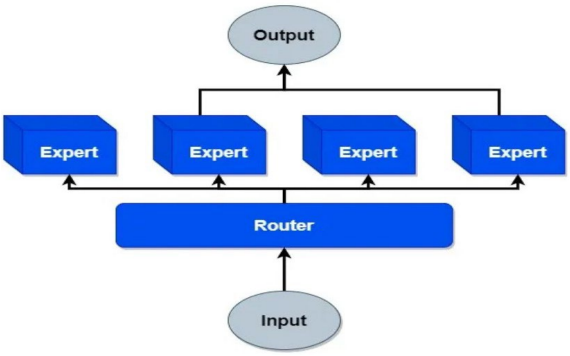

Fig. 2. An example of the Mixture of Experts model.

Each "expert" specializes in a particular task: visual data processing (image and video analysis to detect objects and monitor changes, e.g., the use of modified U-Net and PSP models [18] by "experts" will allow UAV swarms to autonomously detect and classify objects over large areas with high accuracy [1], navigation data processing (optimization of flight routes, ensuring effective obstacle avoidance and adaptation to changing conditions), meteorological data analysis (real-time adjustment of flight plans to improve flight safety), communication optimization (synchronization of actions and automated distribution of tasks between UAVs in a swarm). Like a team of specialists, MoE combines the knowledge and skills of individual neural networks to process complex data.

Given the basic MoE methodology, it is important to consider the specific implementation of the MoE structure (Fig. 3) [3]. The MoE structure is a complex architectural component that enhances the computational capabilities of neural networks by dynamically selecting "experts". Each "expert" in this layer specializes in analyzing different parameters or features of the input data, which allows for a more detailed and comprehensive analysis than is possible with a single fully monolith model.

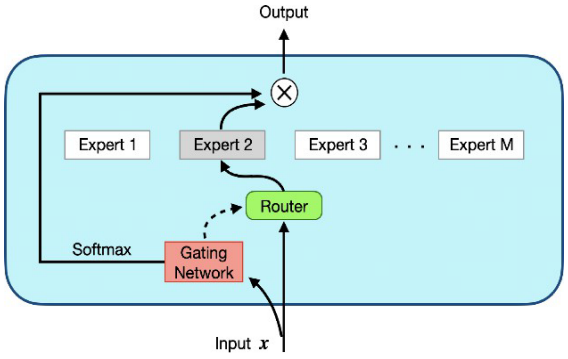

Fig. 3. Mixture of Experts model with routing procedure [4].

The mechanism for selecting experts in that structure is implemented through a gating network or routing procedure that evaluates the input data and determines which expert is most suitable for processing a particular data set. This approach allows the model to use the specialized knowledge of each "expert", which significantly increases the efficiency and accuracy of complex tasks.

The routing procedure analyses the input data and determines how well each "expert" is suited to process a particular data set. It assigns a weight to each expert's output, determining its contribution to the final output of the MoE framework. This process allows the model to utilize the specialized knowledge of each expert, resulting in improved performance on complex tasks. Integrating the MoE layer into transformers increases their ability to adapt to different environmental conditions and ensures more efficient use of computing resources.

Different machine learning models can represent each "expert" in the MoE structure (Fig. 3). "Experts" can be traditional neural networks as well as modern models such as large language models (LLMs), object detection systems (e.g. Yolo), and other specialized models such as ResNet for image processing, GPT-3 for text generation, VGG for object recognition, etc. As a result, each type of expert can be configured to perform specific tasks using their unique capabilities and strengths.

Fig. 4 illustrates the practical implementation and potential benefits of MoE technology [19] using a neural network model as an example. Fig. 4 demonstrates how MoE layers can be integrated into the architecture of more complex neural networks, such as transformers, to efficiently process data sequences. This approach reduces unnecessary load on the system and increases its efficiency. In addition, MoE makes the system more transparent, as each "expert" is responsible for a specific aspect of data processing.

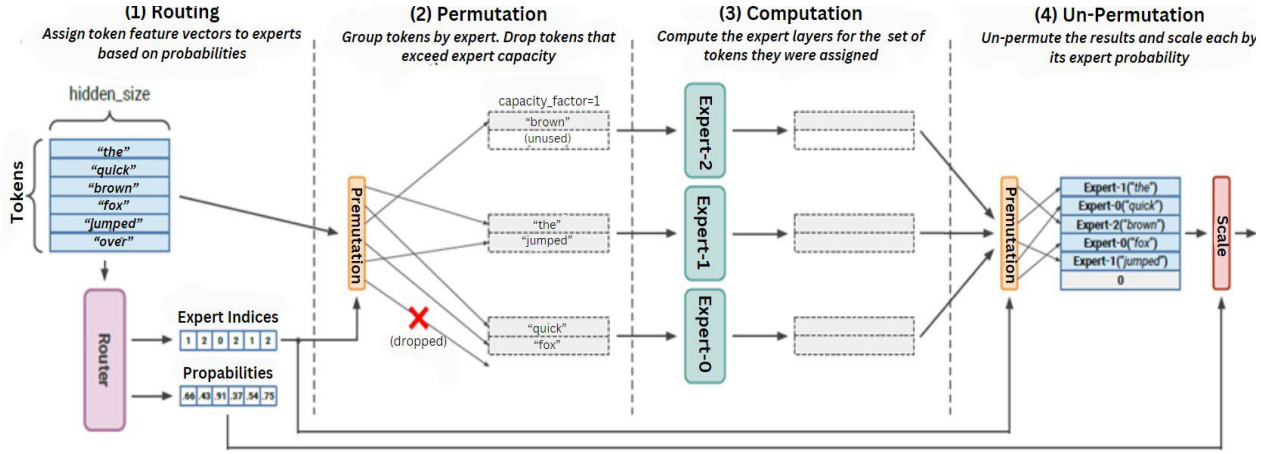

Fig. 4. Mixture of Experts model layer [15].

MoE technology is particularly effective when used based on neural transformers, where input data is pre-processed. The effectiveness of MoE depends heavily on pre-training, which determines which "experts" are best suited to process different types of data. This allows the gateway network to dynamically distribute tasks among the experts, adapting the load and resource distribution to the needs of the system. During the training process, the "experts" focus on certain aspects of the data, developing specialized knowledge and skills, which ultimately improve the overall accuracy and efficiency of the MoE system.

Neural transformers, which serve as the computational backbone of the system, use "self-attention" mechanisms to efficiently process sequential data with high accuracy and minimal latency. Transformers play an important role in interpreting and responding to the large amounts of data that swarms of UAVs encounter during their flights. Their architecture allows for the efficient processing of sequential data, analyzing the importance of different data segments using a "self-attention" mechanism. The transformers consist of an encoder and a decoder, each of which contains layers of self-focus, feedforward, normalization, and residual connections. This structure provides deep learning with parallel data processing, which is especially important for working with large amounts of information.

Neural transformers (Fig. 5) form the basis of the computing system and use "self-focus" mechanisms to process sequential data with high accuracy and speed [12], [20]. The ability of transformers to interpret and analyze complex data streams encountered by UAV swarms allows them to respond effectively to changes in real time. The transformer architecture is distinguished by its ability to efficiently process sequential data, analyzing the importance of different segments of information using a "self-attention" mechanism, without the need for looping or convolutional structures. The transformer consists of an encoder and a decoder, each with self-awareness, direct coupling, and normalization layers, as well as residual connections. This structure enables deep learning with parallel data processing, making it an effective tool for analyzing large data sets.

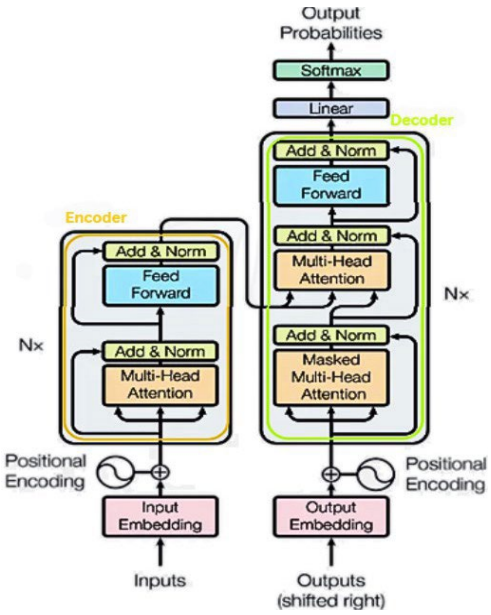

Fig. 5. Transformer architecture [12].

The transformer architecture, thanks to its attention mechanism, can efficiently analyze contextual relationships in sequential data and provide parallel processing. This feature makes transformers an ideal platform for integrating Mixture

of Experts (MoE) technology, as it allows for flexible distribution of computational tasks among "experts" depending on the nature of the input data.

As can be seen from Fig. 6, MoE integration into the transformer architecture is realized by replacing some of the fully connected layers in the encoder and decoder with specialized MoE layers. An important advantage of this approach is the ability to effectively scale the model to many devices. In distributed use, the MoE layer is distributed among devices, while other layers are duplicated, which ensures optimal use of computing resources. Each device processes part of the data based on its own set of "experts", which increases the speed and efficiency of the entire system [21].

Such flexibility and scalability open up new prospects for the development of high-performance data processing systems for UAV swarms capable of performing complex tasks in a high-quality and resource-efficient manner [22]. The dynamic distribution of computations among the "experts" in MoE not only optimizes resource usage and improves system performance, but also makes the system more understandable and easier to analyze and configure. Focusing each "expert" on a specific type of data makes it easier to interpret the model's results. The implementation of MoE for managing UAV swarms creates new opportunities for the development of adaptive and efficient autonomous systems capable of functioning in diverse environments.

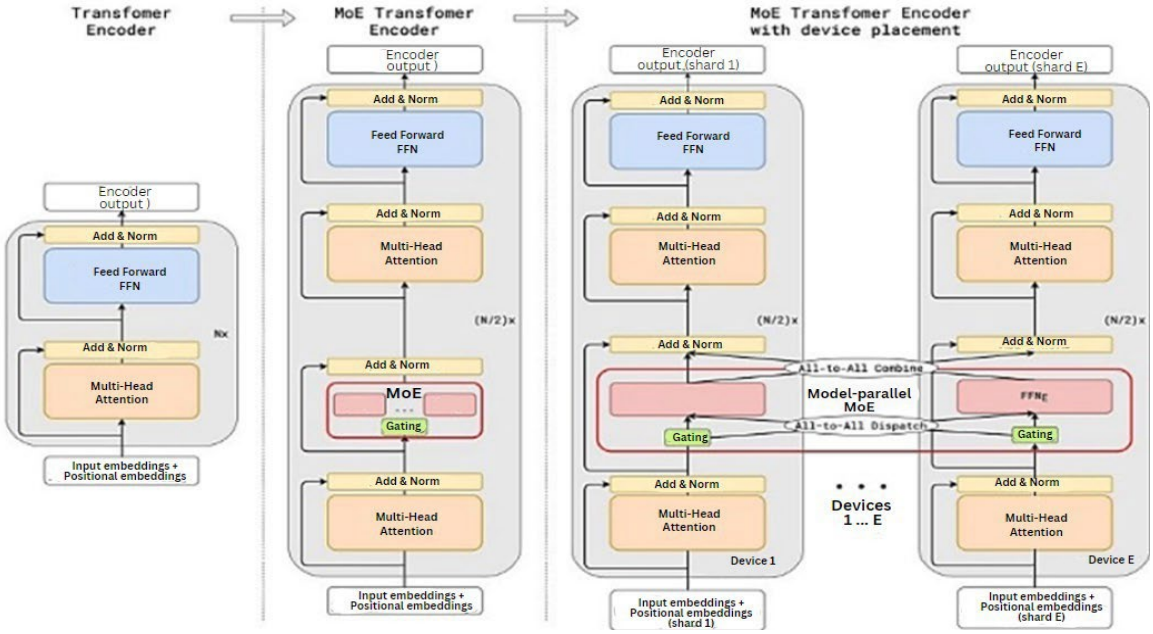

Fig. 6. Transformer encoder scaling by integrating MoE layers [21].

Fig. 7 shows the proposed advanced architecture of UAV swarm management systems using MoE technology. This architecture allows "experts" on UAV1 to interact not only with similar "experts" on UAV2 but also with "experts" on other MoE layers, even when radio communication between different layers of several UAVs is performed at various frequencies.

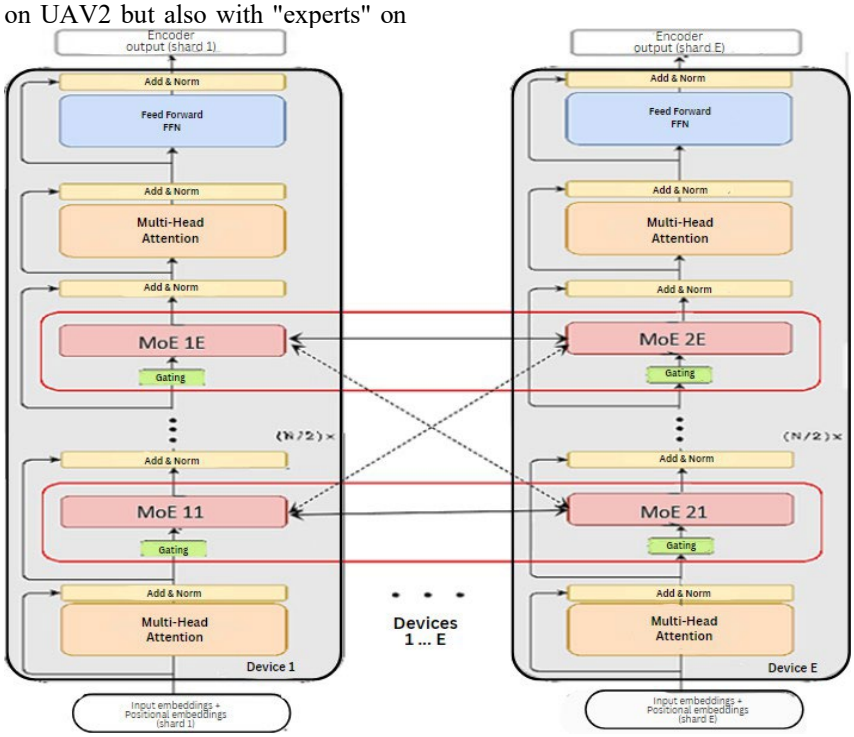

Fig. 7. The architecture of UAV swarm control systems with MoE technology.

A gating network determines the relevance of each "expert" to the processing of certain data, which allows for more rational use of computing resources, activating only those MoE layers that are necessary for a particular task. For example, during nighttime missions, you can disable the MoE layers responsible for processing visual data and activate "experts" specialized in processing infrared images or other sensor data that work better in the dark. Likewise, during daytime missions, experts processing high-resolution visual data can be activated, while layers specializing in nighttime conditions can be deactivated. Additionally, the ability to interact between experts of different levels and devices (UAVs) contributes to a more efficient allocation of tasks and resources, which is critical for the success of missions.

An important challenge in managing UAV swarms is the limited bandwidth of communication channels, which can lead to data loss. The limited computing resources of drones and the need for resilience to external influences make this problem particularly urgent. An efficient data transmission system should ensure reliable communication between UAVs in a swarm and with ground stations, making the most of the available radio frequency resources.

To solve this problem, it is proposed to use an approach similar to quadrature amplitude modulation (QAM) [23]. The encoded data is represented in the form of two components - cosine and sinusoidal, formed based on 2T data points in the latent space of the encoder and decoder. This separation, due to the orthogonality of the cosine and sinusoidal functions, reduces the impact of noise during data transmission and recovery, thereby increasing the speed and reliability of transmission, and providing a compact data representation.

It is worth noting that there are alternatives to the traditional multilayer perceptrons (MLPs) on which transformers are based. For example, Kolmogorov-Arnold networks (KANs) offer a more flexible approach to training, allowing for the adaptation of activation functions on the edges of the network (Fig. 8) [24], [25]. This can provide higher accuracy with fewer parameters compared to MLP.

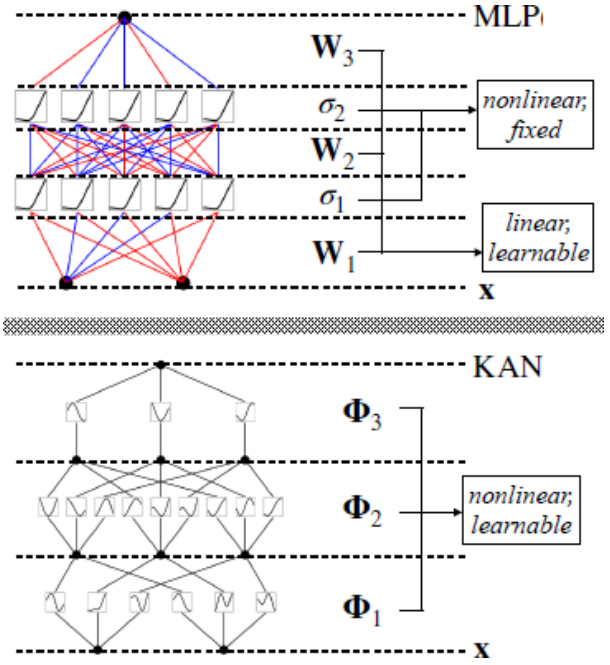

Fig. 8. Comparison of structural schemes MLP and KAN [23].

KANs are also highly interpretable, their structure is easy to visualize, and each parameter has a clear meaning in the context of a functional approximation, which makes KANs promising for scientific research, where it is important to understand the decision-making process of a network and use the knowledge gained for further analysis.

KAN networks can be combined with the self-attention mechanisms inherent in neural transformers. This combination allows using the advantages of both approaches: the flexibility and adaptability of KANs in setting up activation functions and the ability of transformers to efficiently process sequential data and identify dependencies in it [26].

Self-attention mechanisms with which KAN can be integrated include Scaled Dot-Product Attention, which uses key, query, and value matrices to calculate the weights of each element in a sequence relative to other elements, providing contextual dependency. Multi-Head Attention includes several parallel layers of attention, which allows the model to learn different representations of dependencies in the data and increases the model's ability to handle complex input data structures. Self-Attention Mechanism allows each element of the sequence to interact with all other elements, facilitating the capture of long-term dependencies in the data [27]. Relative Position Encoding is used to preserve information about the positional relationships between sequence elements, which is important for processing text data and other sequences. Layer Normalisation normalizes the output of each attention layer, contributing to training stability and accelerating model convergence.

The combination of KAN with these self-attention mechanisms opens up new possibilities for developing advanced machine learning models that can adapt to a wide range of tasks and conditions.

IV. CONCLUSION AND FUTURE WORK

This study demonstrates the prospects of integrating a Mixture of Experts technology into the architecture of transformers to create efficient and adaptive UAV swarm management systems. The use of MoE not only increases the efficiency of data processing and ensures the scalability of the system, but also makes it more understandable and easier to analyze. Transformers, with their "self-attention" mechanism, are an effective tool for analyzing the complex data streams faced by UAV swarms, allowing them to quickly adapt to changing conditions.

An important issue for further research is ensuring reliable data transmission in limited communication channel bandwidth conditions. The proposed approach, based on QAM principles, can increase the speed and reliability of transmission but requires further study and optimization. In addition, a promising direction is to explore the possibilities of integrating alternative neural network architectures, such as Kolmogorov-Arnold networks, which can potentially provide higher accuracy, better interpretability, and learning efficiency on smaller data sets. Further research could lead to more advanced and robust UAV swarm management systems capable of efficiently performing complex tasks in real-world environments.

REFERENCES

[1] V. Slyusar *et al.*, “Improving the model of object detection on aerial photographs and video in unmanned aerial systems”, *EEJET*, vol. 1, no. 9(115), pp. 24–34, Feb. 2022, doi: 10.15587/1729-4061.2022.252876.

[2] V. Slyusar *et al.*, “Improving a neural network model for semantic segmentation of images of monitored objects in aerial photographs,” *EEJET*, vol. 2, no. 6 (114) pp. 86–95, Dec. 2021, doi: 10.15587/1729-4061.2021.248390.

[3] Z.-H. Zhou, *Ensemble Methods*. Chapman Hall/CRC, 2012, doi: 10.1201/b12207.

[4] Z. Chen, Y. Deng, Y. Wu, Q. Gu, and Y.-F. Li, "Towards understanding mixture of experts in deep learning," *ArXiv*, abs/2208.02813, 2022. [Online]. Available at: https://api.semanticscholar.org/CorpusID:251320183.

[5] V. Slyusar, Y. Kondratenko, A. Shevchenko, and T. Yeroshenko, "Some aspects of artificial intelligence development strategy for mobile technologies," *J. Mobile Multimedia,* pp. 525–554, May 2024, doi: 10.13052/jmm1550- 4646.2031.

[6] W. Zhou, J. Li, and Q. Zhang, "Joint communication and action learning in multi-target tracking of UAV swarms with deep reinforcement learning," *Drones*, vol. 6, № 11, 2022. [Online]. Available at: https://doi.org/10.3390/drones6110339.

[7] I. C. Gormley and S. Fruhwirth-Schnatter, "Mixtures of experts models". 2018. [Online]. Available at: https://api.semanticscholar.org/CorpusID:196182122.

[8] Y. Krishnamurthy, C. J. Watkins, and T. Gaertner, "Improving expert specialization in mixture of experts," *ArXiv*, abs/2302.14703, 2023. [Online]. Available at: https://api.semanticscholar.org/CorpusID:257232540.

[9] G. Jawahar *et al*., "AutoMoE: Heterogeneous mixture-of-experts with adaptive computation for efficient neural machine translation," *Findings Assoc. Comput. Linguistics: ACL 2023*, Toronto, Canada. Stroudsburg, PA, USA: Assoc. Comput. Linguistics, 2023. [Online]. Available at: https://doi.org/10.18653/v1/2023.findings-acl.580.

[10] S. Antoniak *et al*., "Mixture of Tokens: Efficient LLMs through Cross-Example Aggregation," *ArXiv*, abs/2310.15961, 2023. [Online]. Available at: https://api.semanticscholar.org/CorpusID:264439523.

[11] L. Ajay, "An intuitive introduction to Transformers". *Medium*. [Online]. Available at: https://lakshmi1212.medium.com/an- intuitive-introduction-to-transformers-6f574c8e7df6.

[12] E. B. Thomas, "A clear explanation of transformer neural networks". *Medium*. [Online]. Available at: https://medium.com/@ebinbabuthomas_21082/decoding-the- enigma-a-deep-dive-into-transformer-model-architecture- 749b49883628.

[13] J. Puigcerver, C. Riquelme, B. Mustafa, and N. Houlsby, "From sparse to soft mixtures of experts," *ArXiv*, abs/2308.00951, 2023. [Online]. Available at: https://api.semanticscholar.org/CorpusID:260378993.

[14] M. Verdoucq, G. Theraulaz, R. Escobedo, C. Sire, and G. Hattenberger, "Bio-inspired control for collective motion in swarms of drones," *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, Dubrovnik, Croatia, 2022, pp. 1626–1631, doi: 10.1109/ICUAS54217.2022.9836112.

[15] P. Wu, Z. Wang, Z. Zhang, G. Hu, W. Dou, and J. Zheng, "UAV swarm regular geometric formation transformation and interactive control algorithm," *2022 5th International Conference on Data Science and Information Technology (DSIT)*, Shanghai, China, 2022, pp. 01–09. doi: 10.1109/DSIT55514.2022.9943960.

[16] I. Karampelia, T. Kyriakidis, and M. Louta, "UAV swarms & Task allocation: the way ahead in precision agriculture," *2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA),* Volos, Greece, 2023, pp. 1-8, doi: 10.1109/IISA59645.2023.10345854.

[17] D. Sharma, Annu, N. Praveen Babu Mannam, and P. Rajlakshmi, "Cloud-Based Control of Drone Swarm with Localization via Ultra-Wideband," *2023 International Conference on Unmanned Aircraft Systems (ICUAS)*, Warsaw, Poland, 2023, pp. 738-744. doi: 10.1109/ICUAS57906.2023.10156277.

[18] M. Verdoucq, G. Theraulaz, R. Escobedo, C. Sire, and G. Hattenberger, "Bio-inspired control for collective motion in swarms of drones," *2022 Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Dubrovnik, Croatia, 2022. [Online]. Available at: https://doi.org/10.1109/icuas54217.2022.9836112.

[19] T. Gale, D. Narayanan, C. Young, and M. Zaharia, *MegaBlocks: Efficient Sparse Training With Mixture-of-Experts*. 2022. doi: 10.48550/arXiv.2211.15841.

[20] M. Vakulenko and V. Slyusar, "Automatic smart subword segmentation for the reverse Ukrainian physical dictionary task," in *Proc. Modern Data Science Technologies Workshop*, Lviv, Ukraine, May 31–June 1, 2024, pp. 59-73. [Online]. Available at: https://www.researchgate.net/publication/382828784_Automatic_smart_subword_segmentation_for_the_reverse_Ukrainian_physical_dictionary_task.

[21] D. Lepikhin *et al*., "GShard: Scaling giant models with conditional computation and automatic sharding," *ArXiv*, abs/2006.16668, 2020. [Online]. Available at: https://api.semanticscholar.org/ CorpusID:220265858.

[22] L. Shazeer *et al*., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," *ArXiv*, abs/1701.06538, 2017. [Online]. Available at: https://arxiv.org/abs/1701.06538.

[23] V. Slyusar and N. Bihun, "The method of increasing the immunity of data transmission in communication channels," in *2022 IEEE 9th Int. Conf. Problems Infocommun., Sci. Technol. (PIC S&T)*, Kharkiv, Ukraine, 10–12 Oct. 2022. IEEE, 2022. doi: 10.1109/picst57299.2022.10238546.

[24] Z. Liu *et al*., "KAN: Kolmogorov-Arnold Networks," *ArXiv*, abs/2404.19756, 2024. [Online]. Available at: https://api.semanticscholar.org/CorpusID:269457619.

[25] R. Shergill, "Kolmogorov-Arnold networks with Gaussian functions". *Medium*. [Online]. Available at: https://riteshshergill.medium.com/kolmogorov-arnold-networks-with-gaussian-functions-84dc463f480a.

[26] R. Genet and H. Inzirillo, "A temporal Kolmogorov-Arnold transformer for time series forecasting," *ArXiv*, abs/2406.02486, 2024. [Online]. Available at: https://api.semanticscholar.org/CorpusID:270226121.

[27] M. Sanchis-Agudo, Y. Wang, K. Duraisamy, and R. Vinuesa, "Easy attention: A simple self-attention mechanism for Transformers," *ArXiv*, Aug. 2023. [Online]. Available at: https://doi.org/10.48550/arXiv.2308.12874.