Enhancing Social Motion Prediction Using Attention Mechanisms and Hierarchical Structures

Botao Dong[1][a], Yumo Ji[2][b,*] and Yongze Miao[3][c]

[1]*School of Mechanical and Electrical and vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China*

[2]*Software and Systems Engineering, LUT University, Lahti, Finland*

[3]*School of Science Technology, Hong Kong Metropolitan University, Hong Kong, China*

Keywords: Multi-Agent Motion Prediction, Kolmogorov-Arnold Networks, Behavioural Cloning, Generative Adversarial Imitation Learning, Transformer Model.

Abstract: With the continuous deepening of artificial intelligence applications in multi-agent systems, the accuracy and efficiency of motion prediction have become key research challenges. This paper proposes a novel neural network model based on Kolmogorov-Arnold Networks (KANs) aimed at enhancing the generalization ability and prediction accuracy of models in multi-agent motion prediction tasks. The study first analyses the limitations of existing behavioural cloning methods and Generative Adversarial Imitation Learning (GAIL) in handling complex dynamic interactions and nonlinear feature data. To address these issues, this paper introduces KANs, a model that replaces the weight parameters in traditional multi-layer perceptron with learnable univariate spline functions, thereby enhancing the model's nonlinear feature extraction capability and adaptability. In the experiments, this paper adopts the Wusi dataset proposed by Zhu et al. in 2024, which contains historical motion sequences of multiple participants. The model designed in this study combines Transformer encoders and decoders, along with KANs, to process local and global features and generate motion predictions for all participants in future time periods. Through feature fusion nodes and multi-level strategy networks, the model can generate more natural and accurate motion sequences. The experimental results show that compared with traditional Transformer-based models, the model in this paper has significantly improved prediction accuracy and training efficiency. Moreover, the model demonstrates better generalization ability on unseen complex patterns, providing new perspectives and methods for the practical application of multi-agent systems.

1 INTRODUCTION

Social motion prediction is an active and challenging research topic in the field of artificial intelligence, involving the understanding and prediction of human behaviour patterns in social environments. This research is not only crucial for applications such as autonomous driving and team sports but also has profound implications for improving machine interaction capabilities and intelligence levels. The anticipatory ability demonstrated by humans in social activities enables them to make rapid and accurate responses in complex environments. Current research mainly focuses on motion prediction in simple interaction scenarios, with limited capability to capture complex and fine-grained human behaviours. To overcome these limitations, the main research method of this paper includes the adoption of an innovative deep learning framework to improve the accuracy and generalization ability of multi-agent motion prediction. The core of the research method is the introduction of Kolmogorov-Arnold Networks (KANs), a novel neural network model that enhances the model's flexibility and adaptability by replacing traditional weight parameters with univariate spline functions. The learnable activation functions of

[a] https://orcid.org/0009-0001-4734-9977
[b] https://orcid.org/0009-0007-7805-1640
[c] https://orcid.org/0009-0006-2200-2287

Dong, B., Ji, Y. and Miao, Y.
Enhancing Social Motion Prediction Using Attention Mechanisms and Hierarchical Structures.
DOI: 10.5220/0013253300004558

In *Proceedings of the 1st International Conference on Modern Logistics and Supply Chain Management (MLSCM 2024)*, pages 185-189
ISBN: 978-989-758-738-2

KANs can adjust adaptively according to the training data, thereby better capturing nonlinear relationships in the data (Helbing, & Molnar, 1995).

The study also employs the Transformer model, which consists of encoders and decoders, to process local and global features. Each Transformer contains three layers and eight attention heads, and the strategy network parameters are shared. KANs are introduced into the Transformer encoders and decoders to enhance the model's ability to capture complex dynamic interactions. In addition, the study also adjusts the model's dimensions to adapt to a specific dataset—the Wusi dataset, proposed by Zhu et al. in 2024, which includes historical motion sequences of multiple participants.

By combining KANs and Transformers, the study aims to generate more natural and accurate future motion predictions while improving the model's adaptability to unseen complex patterns (Alahi, Goel, Ramanathan, et al, 2016). The proposal of this new algorithm aims to address the limitations of existing algorithms in handling highly nonlinear and periodic data, as well as the problem of overfitting to training data (Guo, Bennewitz, 2019).

2 METHOD

2.1 Dataset

This study cites the first large-scale multiplayer 3D sports dataset, Wusi (Wusi Basketball Dataset), proposed in 2024 by Zhu et al (Zhu, Qin, Lou, et al, 2024). The Wusi dataset shows advantages when faced with the task of multiplayer sports prediction, outperforming other datasets in terms of size (duration and number of people) and intensity of interactions.

The input data is composed of historical movement sequences from multiple participants. Given P participants, the movement history of each participant p may be represented as a time series of length T, where each time step t records the body posture of the participant in 3D space: $\{\mathcal{X}_{\tau}^{\rho}\} 1 \leq t \leq \mathrm{T}, 1 \leq p \leq \mathrm{P}$. The output data consists of motion predictions for all participants in the future time period, and the goal of the output is to predict the sequence of postures from time T to $\mathrm{T} + \mathrm{T}'$. For each participant p, the sequence of predicted future poses is represented as: $\{\mathcal{X}_{t}^{p}\} \mathrm{T} \leq t \leq \mathrm{T} + \mathrm{T}', 1 \leq p \leq \mathrm{P}$. $\mathcal{X}_{t}^{p}$ represents the 3D pose at time step t.

2.2 Existing Algorithm

The motion prediction model in the study was modelled using Markov Decision Process (MDP). Behavioural cloning uses expert demonstration data to train the model by supervised learning, which minimizes the discrepancy between the model-generated actions and the expert's behaviours. Behavioural cloning methods excel in terms of computational and sample efficiency (Caude, Behavioural, 2010), but there are some problems; the strategies tend to overfit the presentation of the expert in the region of the state space, limiting the ability to generalize (Borui, Ehsan, Hsu, 2019). To address these problems, Generative Adversarial Imitation Learning (GAIL) (Jonathan and Stefano, 2016) was introduced. The policy network is regularized by adversarial training to match its distribution of state-action pairs with that of the policy of the expert, while a specific cognitive hierarchy model is used to express the recursive reasoning process (Colin, Ho, and Chong, 2004).

2.3 Limitations of Baseline

During the algorithmic implementation of the baseline, multiple performance bottlenecks limit the accuracy and training efficiency of the model.

In multi-agent motion prediction tasks with input data having complex dynamic interactions and potentially nonlinear features, the nonlinear feature extraction capability of the model is crucial for prediction accuracy. Transformer models are known for their expertise in identifying semantic correlations (Alharthi, & Mahmood, 2024), but Transformer-based deep learning model performs obvious limitations when dealing with highly nonlinear and periodic data (Nie, et al, 2022; Zeng, et al, 2023). This study argues that when the input data contains complex dynamic interactions, the prediction accuracy of the model decreases significantly due to its inability to effectively capture these nonlinear relationships. Meanwhile, due to the inability of the model to adequately capture the complex interaction characteristics between the participants, the generated motion sequences do not behave naturally enough in certain scenarios. This phenomenon not only affects the generative ability of the model, but also reduces its credibility in practical applications.

As the model underperforms under complex models, Transformer-based deep learning models may perform overfitting on the data, which means that the model performs well on the training data but generalizes poorly on the validation or test dataset.

This phenomenon reveals that the model is poorly adapted to unseen complex patterns, limiting its potential for application in different scenarios.

2.4 Kolmogorov-Arnold Networks

Kolmogorov-Arnold Networks (KANs) (Liu, Wang, Vaidya, 2024) is an innovative neural network model that challenges the Multilayer Perceptron (MLP). The MLP occupies almost all the non-embedded parameters in the Transformer (Ashish, 2017) model and is often hard to interpret in the lack of post-analysis tools as compared to the Attention Layer (Hoagy, Aidan, Logan, 2023). KANs replace each weight parameter with a univariate spline function. The learnable activation function will adaptively adjust and learn based on the training data, showing higher flexibility and adaptability compared to fixed activation functions.

The study implements training and testing on the Wusi dataset using the proposed framework. This study uses Transformer encoder for local and global state encoding, while Transformer decoder is used for policy network. Each Transformer consists of three layers and eight attention heads, while sharing the policy network parameters $\emptyset_{(1)} \cdots \emptyset_{(K)}$. The study introduces KANs in the pair of Transformer encoder and decoder, while the dimensionality of the model is adjusted.

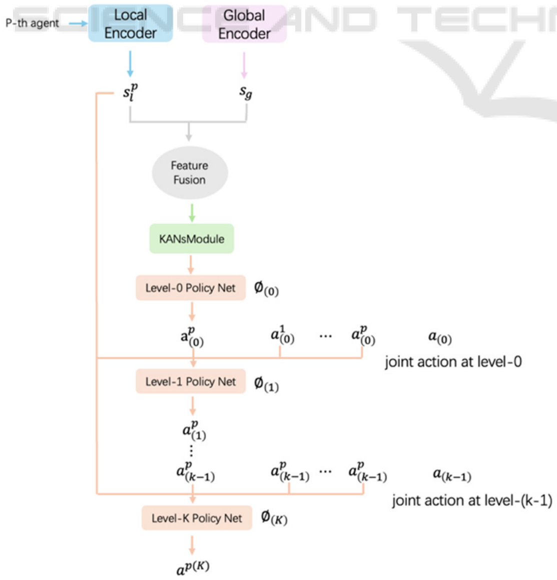

Figure 1: Framework overview(Photo/Picture credit: Original).

Figure 1 shows an overview of the framework. For the p-th agent, two state encoders are responsible for processing local and global state features s_l^p and s_g, after integration by feature fusion nodes, the fused features are passed to the KANs. The processed enhanced features are passed to the Level-0 policy network to generate the initial action $a_{(0)}^p$. The Level-K policy network is based on s_l^p and the joint actions of the previous level $a_{(k-1)}$ to produce action $a_{(K)}^p$.

3 EXPERIMENTS

3.1 Setup

Evaluation metrics: This study calculates the Mean Per Joint Position Error (MPJPE) between the prediction of future motion and the ground truth. The trajectory prediction and position prediction results are distinguished by counting the mean root position error and the mean local position error (MPJPE after root alignment). Results are reported in millimetres.

3.2 Evaluation and Comparison

In this study, the proposed model is tested to evaluate the performance of the existing model against the optimized model in this study. Table 1 reveals that the optimized model in this study achieved competitive results with the SOTA method. It is worth noting that the baseline model is very similar to the optimized model in this study in terms of long-term motion prediction accuracy. However, the baseline model is limited in short-term and root trajectory motion prediction. In contrast, the optimized model in this study significantly outperforms the baseline model in short-term and root-trajectory motion prediction, which implies that the optimized model in this study demonstrates an advantage in capturing motion patterns and trends in the short-term, while it is able to demonstrate a high quality of motion prediction when dealing with complex motion scenarios. The copyright form is located on the authors' reserved area.

3.3 Ablation Study

Table 2 shows the ablation study of model size and architectural design. The dimensionality of the model is adjusted in this study. By comparing the models with different dimensions, model dimensions that are too large lead to overfitting of the model, which cannot be generalized well to new data, while the

Table 1: Performance comparison with the baseline method.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
Baseline	54.6	**86.2**	**119.3**	152.5	43.7	60.8	**74.6**	**86.6**	41.7	66.9	94.8	124.0
Ours	**54.5**	86.4	119.4	**152.4**	**42.8**	**60.7**	75.7	88.3	**40.0**	**64.5**	**91.6**	**119.9**

Table 2: Ablation study of model dimensions and architectural designs.

Milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
(a) D_ model = 128	61.3	92.5	124.8	157.6	48.4	65.2	79.1	90.6	43.9	68.6	96.0	124.9
(b) D_ model = 256	55.0	86.8	120.0	153.7	43.3	**60.6**	**75.1**	**87.3**	40.8	65.6	93.3	122.7
(c) D_ model = 1024	60.9	92.9	126.0	158.9	47.3	64.2	78.3	90.1	44.7	70.1	98.0	126.9
(d) D_ model = 512	**54.5**	**86.4**	**119.4**	**152.4**	**42.8**	60.7	75.7	88.3	**40.0**	**64.5**	**91.6**	**119.9**

computational cost of the model increases significantly. Model dimensions that are too small lead to higher errors, which cannot effectively capture complex features in the data and have limitations on generalization to new data. In (d), the model performance is the best in terms of error performance, excelling in both short-term and long-term predictions, showing good generalization to unseen data, especially outperforming the other dimensions in terms of root position error.

4 CONCLUSIONS

The method proposed in this study outperforms the baseline model in terms of average root position error. The model in this study shows an advantage in the prediction accuracy of the core parts of human movement, and the accurate prediction of root position effectively reduces the accumulation of errors caused by unstable postures. In comparison with the baseline model, the performance of the model is similar to that of the baseline model at different time steps, but the method proposed in this study has improved short-term prediction ability. The model proposed in this study is able to respond quickly and accurately capture sudden changes and short-term dynamics in motion, providing accurate short-term prediction results.

Although the model performs well in short-term prediction, the performance at long time steps still needs to be improved. It needs to be ensured that the model is not excessively complex for further enhancement of the expressive power of the model. The current model's main application is in motion prediction, showing limitations in multi-domain applications. Future research could try to extend to more domains to explore the optimization directions and challenges of the model architecture.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Leung, T., & Fei-Fei, L. (2016). Social-LSTM: Human trajectory prediction in crowded spaces. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3419-3428. https://doi.org/10.1109/CVPR.2016.332

Alharthi, M., & Mahmood, A. (2024). Enhanced Linear and Vision Transformer-Based Architectures for Time Series Forecasting. Big Data and Cognitive Computing, 8(5), 48.

Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., Łukasz, K., and Illia, P. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Borui, W., Ehsan, A., Hsu, C., Huang, D. A., and Niebles, J. C. 2019. Imitation learning for human pose prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7124–7133.

Caude, S., Behavioral, C. 2010. Encyclopedia of Machine Learning, pages 93–97.

Colin, F., Ho, T. H., and Chong, J. K. 2004. A cognitive hierarchy model of games. The Quarterly Journal of Economics, 119(3):861–89.

Guo, Y., & Bennewitz, M. (2019). Predicting pedestrian trajectories with social-aware deep reinforcement learning. IEEE Robotics and Automation Letters, 4(2), 3443-3450. https://doi.org/10.1109/LRA.2019.2894916

Helbing, D., & Molnar, P. (1995). Social force model for pedestrian dynamics. Nature, 376(6517), 47-50. https://doi.org/10.1038/376047a0

Hoagy, C., Aidan, E., Logan, R., Robert, H., and Lee, S. 2023. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600.

Jonathan, H., and Stefano, E. 2016. Generative adversarial imitation learning. In Advances in neural information processing systems, pages 4565–4573.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., ... & Tegmark, M. (2024). Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756.

Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J. A. 2022. Time Series is worth 64 words: Long-term forecasting with Transformers. arXiv 2022, arXiv:2211.14730.

Zeng, A., Chen, M., Zhang, L., Xu, Q. 2023. Are Transformers effective for Time Series Forecasting? Proc. AAAI Conf. Artif. Intell. 2023, 37, 11121–11128.

Zhu, W., Qin, J., Lou, Y., Ye, H., Ma, X., Ci, H., & Wang, Y. (2024). Social motion prediction with cognitive hierarchies. Advances in Neural Information Processing Systems, 36.