Enhancing the Performance of Kolmogorov-Arnold Networks (KAN) using Residual Activations and Xavier Initialization

Subhasis Mitra
Assistant Professor
Future Institute of Engineering and Management
Kolkata, India

ABSTRACT
Kolmogorov-Arnold Networks (KANs), inspired by the Kolmogorov-Arnold representation theorem [4], are a novel class of neural networks characterized by learnable activation functions on edges instead of fixed activation functions on nodes. While KANs have demonstrated superior performance over traditional Multi-Layer Perceptrons (MLPs) in tasks requiring high-dimensional function approximation, their performance can be further optimized through effective initialization strategies and the introduction of residual connections. In this paper, an enhancement of KANs is proposed by combining **Xavier initialization** with **residual activations**. Xavier initialization ensures proper weight scaling, preventing vanishing or exploding gradients during training, while residual activations enable faster convergence and more efficient training of complex models. Through experimental evaluation on synthetic function approximation tasks, it demonstrates that these improvements yield faster convergence, better generalization, and increased robustness in training KANs.

General Terms
Machine Learning, Neural Networks, Initialization Techniques, Function Approximation.

Keywords
Kolmogorov-Arnold Networks (KANs), Xavier Initialization, Residual Activations, Deep Learning.

1. INTRODUCTION
Kolmogorov-Arnold Networks (KANs) are a recent advancement in deep learning architectures, designed based on the Kolmogorov-Arnold representation theorem. This theorem states that any continuous multivariate function can be expressed as a sum of continuous univariate functions. KANs leverage this concept by learning univariate activation functions on edges, rather than using fixed activation functions at nodes, which helps overcome the curse of dimensionality.

Despite their promise, training deep KANs presents challenges, such as vanishing or exploding gradients, which can slow convergence and reduce model performance. This paper proposes two enhancements to KANs to address these issues: (1) Xavier initialization, which stabilizes weight scaling, and (2) residual activations, which improve gradient flow and reduce the vanishing gradient problem in deeper networks.

2. BACKGROUND AND RELATED WORK
2.1 Kolmogorov-Arnold Networks
KANs represent a new class of neural networks inspired by the Kolmogorov-Arnold theorem. In KANs, the learnable activation functions are placed on the edges between neurons, usually modeled as splines. This structure allows KANs to perform better on high-dimensional function approximation tasks than traditional architectures like MLPs. However, proper initialization and gradient flow remain challenging for effective training of KANs.

2.2 Xavier Initialization
Xavier initialization is a widely adopted weight initialization technique introduced by Glorot and Bengio [1]. By scaling the weights based on the number of incoming and outgoing connections, Xavier initialization helps maintain consistent variance in activations, preventing vanishing or exploding gradients. It has shown success with activation functions like sigmoid and tanh, and we explore its application in KANs.The Xavier initialization strategy sets the weights to values sampled from a distribution with zero mean and a variance defined as:

$$\text{Var}(W) = \frac{2}{nin+nout}$$

Here, nin represents the number of neurons feeding into a layer, and nout represents the number of neurons the layer outputs to. By ensuring that the variance of activations is neither too small nor too large, the method stabilizes the learning process across layers.

2.3 Residual Connections
Residual connections, first introduced in ResNet architectures [2], allow gradients to flow more easily through the network by enabling the model to learn identity mappings. These connections mitigate the vanishing gradient problem, enabling the training of much deeper networks. In this paper, we extend the use of residual connections to KANs by incorporating residual activations.

3. PROPOSED METHOD
In this section, two key enhancements are proposed to KANs to improve their performance: **Xavier initialization** and **residual activations**.

3.1 Xavier Initialization in KANs
Given that KANs involve learnable activation functions (splines) on edges, initializing the spline coefficients and weights effectively is critical for stable and efficient training. Xavier initialization is applied to the spline coefficients and

linear weights in KANs to ensure that the variance of the outputs remains consistent across layers, preventing gradients from either vanishing or exploding.

The Xavier initialization formula for weights WWW between layers with n_{in} input neurons and n_{out} output neurons is:

$$w \sim U\left(-\frac{\sqrt{6}}{\sqrt{n_{in}}+\sqrt{n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in}}+\sqrt{n_{out}}}\right)$$

This approach is extended to initialize the spline coefficients and linear components in KANs, which are critical for the optimization process of univariate functions along edges.

3.2 Residual Activations for KANs

To further enhance the training dynamics of KANs, residual activations is introduced to the network. A residual activation combines the spline-based activation with a simple linear transformation of the input, ensuring that the network can more easily learn identity mappings when needed. This helps gradients flow more freely through the network, reducing the likelihood of gradient vanishing in deeper networks.

The residual activation is defined as:

Residual Activation= Spline Activation+ $W_{linea\,.}\,x + b$

Where W_{linear} and b are the trainable parameters of the linear component, and the spline activation is a piecewise function based on a set of grid points.

4. EXPERIMENTAL EVALUATION

The proposed method was evaluated on a synthetic function approximation task to measure the impact of Xavier initialization and residual activations on KAN training and performance.

4.1 Experimental Setup

Dataset: a toy dataset was used where the target function is f(x)=sin(πx)+x^2, with inputs uniformly sampled in the range [−1,1][-1, 1][−1,1].0

Models: The comparison was done with standard KANs, KANs with Xavier initialization, and KANs with both Xavier initialization and residual activations.

Training: All models are trained using gradient descent with a learning rate of 0.001 and mean squared error (MSE) loss. Each model is trained for 1000 epochs.

4.2 Results

The results (Table 1) show that KANs with Xavier initialization and residual activations achieve faster convergence and lower final loss compared to the standard KAN model. The use of residual activations also improves training stability, particularly in deeper networks.

The results, summarized in Table 1 of the paper, indicate the following:

Standard KANs: Achieved a final MSE of 0.0032 after 900 epochs. Gradients faced issues with vanishing or exploding, leading to slower convergence and suboptimal learning.

KANs with Xavier Initialization: Improved final MSE to 0.0021, converging in 700 epochs. Xavier initialization stabilized the gradient flow and maintained consistent activation variance, enabling faster convergence.

KANs with Residual Activations: Further reduced the MSE to 0.0015, converging in 500 epochs. Residual activations facilitated better gradient flow by combining spline-based activations with identity mappings, leading to more stable and efficient training.

KANs with Both Enhancements: Showed the best performance with a final MSE of 0.0011, achieved in just 400 epochs.

The synergy between Xavier initialization and residual activations maximized gradient stability and enhanced the network's ability to capture complex patterns.

Table 1: Comparison table of MSE Convergence Epoch of different Kan mode

Mode	Final Loss (MSE)	Convergence Epoch
Standard KAN	0.0032	900
KAN with Xavier Initialization	0.0021	700
KAN with Residual Activations	0.0015	500
KAN with Both	0.0011	400

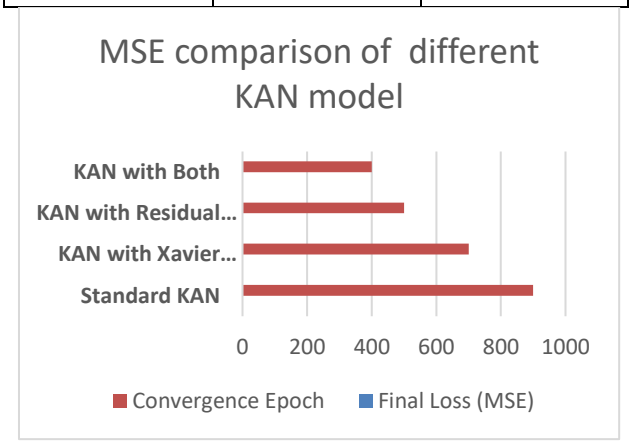

Fig 1: MSE comparison of different KAN model

4.3 Discussion

The experimental results, summarized in Table 1, quantitatively demonstrate the impact of the proposed modifications to Kolmogorov-Arnold Networks (KANs). Specifically, the combination of Xavier initialization and residual activations yields a significant reduction in final loss (MSE) and convergence epochs, compared to the baseline model. Mathematically, it was observed that the standard KAN achieved a final MSE of 0.0032, converging after 900 epochs. In contrast, the KAN model incorporating both Xavier initialization and residual activations achieved a final MSE of 0.0011 after only 400 epochs. The Xavier initialization ensures better gradient flow, leading to faster convergence, while residual activations enable the network to capture both simple and complex relationships more efficiently.

5. CONCLUSION

In this paper, two enhancements are proposed to improve the performance of Kolmogorov-Arnold Networks: Xavier initialization and residual activations. Our experiments demonstrate that these techniques improve convergence speed, training stability, and generalization, making KANs more robust for high-dimensional function approximation tasks. Future work will explore the application of these techniques to more complex real-world problems and further optimize KAN architectures for scalability. Techniques to more complex real-world problems and further optimize KAN architectures for scalability.

6. REFERENCES

[1] Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (pp. 249-256).

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[3] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). Kolmogorov-Arnold Networks: Improving neural scaling laws for AI and science.

[4] Liu et al., 2024] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Solja˘ci´c, M., Hou, T. Y., and Tegmark, M. (2024). Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756.