A Scale-Invariant Diagnostic Approach Towards Understanding Dynamics of Deep Neural Networks

Ambarish Moharil*
Jhernonimus Academy of Data-Science,
Tilburg University
a.s.moharil@tilburguniversity.edu

Damian Tamburri
Eindhoven University of Technology
d.a.tamburri@tue.nl

Indika Kumara
Willem-Jan Van Den Heuvel
Tilburg University
i.p.k.weerasingha.dewage@tue.nl
w.j.a.m.v.d.heuvel@jads.nl

ABSTRACT

This paper introduces a scale-invariant methodology employing *Fractal Geometry* to analyze and explain the nonlinear dynamics of complex connectionist systems. By leveraging architectural self-similarity in Deep Neural Networks (DNNs), we quantify fractal dimensions and *roughness* to deeply understand their dynamics and enhance the quality of *intrinsic* explanations. Our approach integrates principles from Chaos Theory to improve visualizations of fractal evolution and utilizes a Graph-Based Neural Network for reconstructing network topology. This strategy aims at advancing the *intrinsic* explainability of connectionist Artificial Intelligence (AI) systems.

1 INTRODUCTION

Explainable Artificial Intelligence (XAI) seeks to demystify decision-making in complex Machine Learning and Deep Learning systems [1]. While there is no universal definition of explainability, Liao et al. [2] describe it simply as *"an answer to a question"*. Adopting *connectionism* has greatly enhanced modeling capabilities regarding physical and informational complexities through complex non-linear dynamical systems of independently communicating units [3, 4]. These systems, often referred to as *black-box* and partially *chaotic*, show a sensitive dependence on initial conditions, complicating predictions about their long-term behavior [5]. Moreover, the inherent non-linearity across the network architecture suggests a connectionist self-symmetry, invariant across different scales of observation [6].

This understanding is crucial for comprehending *non-linearities* and *emergence* in such networks. Traditional *surrogate* methods like LIME, which offer post-hoc explanations via sparse linear feature representations, fail to capture the network's non-linear interactions and dynamic behaviors during optimization [1, 7]. However, recent advancements in the use of fractal features for network analysis and the computation of fractal dimensions in complex systems suggest a new segmentation approach for creating a *non-linear* connectionist representation across multiple scales [6, 8]. Inspired by Mandelbrot's work in fractal geometry and discoveries of non-linear attractor behaviors [9–11], our research leverages fractal analysis to delve deeply into connectionist network architectures. By evaluating fractal dimensions and roughness, we gain insights into network connectivity and emergent phenomena, thereby enhancing our understanding of system dynamics and aiding in the identification of cyclic *attractors*, as demonstrated in Kauffman's studies on Random Boolean Networks (RBNs) [12]. Our approach aims to augment the *intrinsic* explainability of connectionist networks and emergent phenomena by addressing two fundamental questions: **RQ1.** *"How can we create a non-linear connectionist representation across multiple scales for DNNs?"* and **RQ2.** *"To what extent does such a representation enhance the explainability of nonlinear interactions in DNNs?"*

KDD '24, August 25–29, 2024, Barcelona, ESP
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PROPOSED METHODOLOGY

Our methodology segments the network at specific scales during and after optimization, generating a fractal representation of network connections using a graph-based surrogate. Segmenting the layers of a network (DNN), in turn, implies partitioning the associated parameter matrices, as it is the parameters that *form* and *deform* the network. Formally, consider a connectionist network f with L_p layers. For any given layer L_i, connected to subsequent layer L_j where $i \neq j$, the parameter matrix $W_{n\times m}$ represents the connections, where n and m denote the number of nodes in layers L_i and L_j respectively. The fractal dimension FD of this matrix is calculated using the box-counting method [8], defined as: $FD = \frac{\ln(N)}{\ln(1/r)} \in \mathbb{R}$ where N is the number of $r \times r$ boxes needed to cover the matrix. The segment size r, crucial for fractal analysis, must satisfy: $r > 1$, to avoid granularity at the level of individual matrix elements, and: $r < \min(n, m)$, to prevent oversimplification by covering the matrix with a single box. The specific ranges for r are defined as follows:

$$r \in \begin{cases} [2, \lfloor \frac{n+1}{2} \rfloor], & \text{if } n = m \text{ and } n \mod 2 \neq 0 \\ [2, \frac{n}{2}], & \text{if } n = m \text{ and } n \mod 2 = 0 \\ [2, \lfloor \frac{\min(n,m)+1}{2} \rfloor], & \text{if } n \neq m \text{ and } \min(n,m) \mod 2 \neq 0 \\ [2, \frac{\min(n,m)}{2}], & \text{if } n \neq m \text{ and } \min(n,m) \mod 2 = 0 \end{cases}$$

The segmentation extracts sub-matrices based on r, iterating over

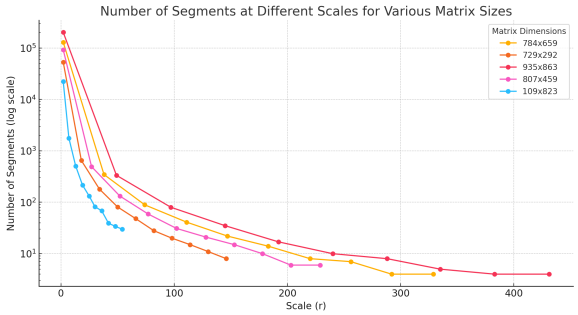

Figure 1: Visualizing the #segments (log scale) at various *valid* scales r for different matrix dimensions $n \times m$.

dimensions n and m where $n, m > 2$ as (Fig 1):

$$S_{ij} = W_{nm}[i \cdot r : (i+1) \cdot r, j \cdot r : (j+1) \cdot r] \tag{1}$$

for $i = 0$ to $\lceil \frac{n}{r} \rceil - 1$ and $j = 0$ to $\lceil \frac{m}{r} \rceil - 1$. [1] This adaptive approach to segmentation ensures optimal coverage and granularity for analytical purposes, accounting for structural variations in the matrix dimensions. We furthermore use the activation map $A^i_{g,h}, \ni g, h \in \mathbb{R}$, of layer L_i to study the non-linear interactions between fractal segments of subsequent layers, computed at scale r_q. Given two fractal segments S_ω and S_λ, we apply the exponential kernel taking inspiration from [7, 13], computing the edge values between segments across layers $L_i : L_p$ accounting for the local influence of the parameter segments, as follows :

$$e_{\omega,\lambda} = \gamma \cdot exp(||\alpha_\omega - \alpha_\lambda||) \tag{2}$$

Where γ is the spread of the kernel, α_ω and α_λ and are segment specific features at scale r_q, computed as :

$$\alpha_{S_{i,j}} = A^{'i}_{g,h} \cdot FD(S_{i,j}, r_q) \cdot H(S_{i,j}) \tag{3}$$

, Where $A^{'i}_{g,h}$ is the flattened feature map of $A^i_{g,h}$ and $H(S_{ij})$ is the entropy of the segment, calculated as $H(S_{ij}) = -\sum p(x) \log p(x)$ with $p(x)$ representing the probability distribution within S_{ij}. Following the extraction of non-linear relationships among fractal segments, represented as an adjacency matrix $\mathcal{A}$, we utilize a Graph-Based Neural Network to learn the network representation at scale r_q, inspired by Wang et al. [13]. This approach allows for the exploration and learning of structural space ($\mathcal{A}$), culminating in the aggregation of learned graphs into a fractal hypergraph and achieving comprehensive insights into the underlying informational dynamics as:

$$\Omega(A^{(i,k)}_{g,h}, \mathcal{A}) = \sigma\left(\hat{D}^{-\frac{1}{2}} \hat{\mathcal{A}} \hat{D}^{\frac{1}{2}} A^{(i,k)}_{g,h} S^{(k)}_{i,j}\right) \tag{4}$$

where σ is the activation function, $A^{(i,k)}_{g,h}$ the feature map of k^{th} segment of i^{th} layer along with the parameter segment $S^k_{i,j}$ and $\hat{D}$ is the diagonal degree of $\mathcal{A}$.

3 PRELIMINARY RESULTS

In our study, we examine multiclass classification on the MNIST dataset [14] using a CNN with two convolutional layers of 32 and 64 neurons and two fully connected layers, each convolutional layer utilizing a kernel size of 3 with padding and stride set to 1. We train the model over 50 epochs with a learning rate of $6e-4$ using the Adam Optimizer, and analyze gradients and loss post each epoch through a segmentation of model weights, as described in equation 1. The weights of the first convolutional layer are a $4-D$ tensor $[32, 3, 3, 3]$ and the second layer $[64, 32, 3, 3]$, segmented into four overlapping segments per neuron, scaled by $r = 2$. We evaluate segment features to determine their influence on inputs as detailed in equation 3, with Figure 2 (left) illustrating the captured segment features, and Figure 2 (right) displaying the exponential kernel interactions between segments, indicating local influence across layers. We analyze the phase flow graph $\forall^M_{j=1} \forall^Q_{i=1} \frac{\partial \mathcal{L}}{\partial (W^j_{s_i})}$ Vs $\mathcal{L}$ for a network with M layers and Q segments per layer, where $\mathcal{L}$ is the cross-entropy loss. Figure 3 (left) depicts the learning trajectory of a segment, highlighting an initial acceleration followed by a gradual deceleration, and Figure 3 (right) shows a plot of $\frac{\partial}{\partial t}$ vs $\frac{\partial^2}{\partial t^2}$, evidencing a positive trend. The epochs, color-coded, reveal a decreasing trend in gradient norms, suggesting model stabilization. Data convergence in later epochs toward the center indicates the formation of an attractor, enhancing predictability and training stability.

[1] Segmentation overlap depends on the parity of matrix dimensions n and m in W_{nm}. For square matrices ($n = m$), an odd n means a stride of 1, causing edge overlaps, while an even n allows perfect tiling without overlaps. For non-square matrices, odd $\min(n, m)$ results in overlaps for coverage, whereas even $\min(n, m)$ ensures perfect alignment without overlaps.

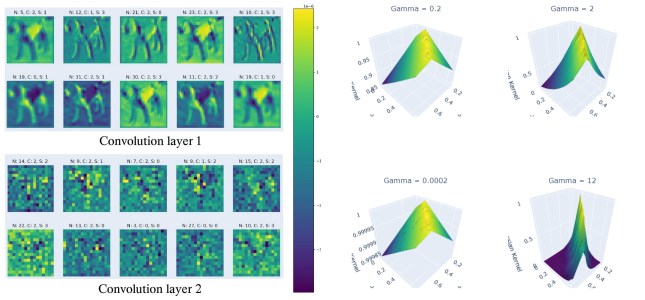

Figure 2: Left image shows features learned by neuron-channels across two convolution layers. Right image visualizes the exponential kernel interactions between fractal segment features for various γ values.

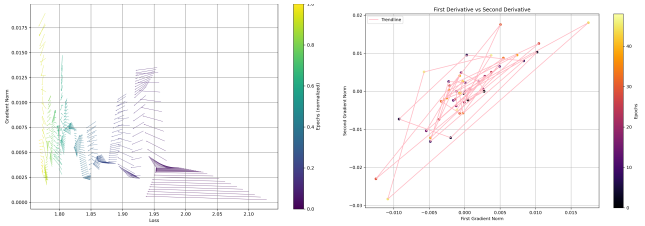

Figure 3: Phase-Flow Diagrams

4 RESEARCH ROADMAP

Our future research in Explainable Artificial Intelligence (XAI) encompasses three interconnected strategic categories, aiming to deepen the integration and sophistication of graphical models and analytical techniques. In *Enhancements in Graph-Based Surrogates*, we focus on advancing fractal feature learning using graph-based surrogates, improving model capabilities through node embeddings and neural message passing [15], and extending our approaches to incorporate semantic feature analysis across various DNN architectures using saliency models [16]. The *Advanced Visualization and Theoretical Approaches* involve applying renormalization theory for efficient feature distillation [17] and exploring attractor behaviors (fixed, periodic, quasi, aperiodic) [9, 18] within network segments along with identifying *answerable* questions regarding system dynamics [2]. Finally, *Comprehensive System Analysis* leverages Automated Machine Learning (AutoML) for hyperparameter sampling to facilitate a robust analysis of system behaviors as either *dissipative* or *conservative* [18], enhancing the intrinsic explainability of AI models [19, 20]. Together, these efforts aim to advance the transparency and interpretability of AI systems within the field of XAI, using the concepts from non-linear system dynamics.

REFERENCES

[1] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey, 2021.

[2] Q. Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. ACM, April 2020.

[3] Ivan M. Havel. Artificial intelligence and connectionism: Some philosophical implications. In Vladimír Mařík, Olga Štěpánková, and Rorbert Trappl, editors, *Advanced Topics in Artificial Intelligence*, pages 25–41, Berlin, Heidelberg, 1992. Springer Berlin Heidelberg.

[4] Jieshu Wang. Symbol vs. connectionism: A closing gap in artificial intelligence. http://wangjieshu.com/2017/12/23/symbol-vs-connectionism-a-closing-gap-in-artificial-intelligence/, 2017.

[5] Edward N Lorenz. Predictability: Does the flap of a butterfly's wings in brazil set off a tornado in texas? *Journal of the Atmospheric Sciences*, 20(2):130–141, 1973.

[6] Tao Wen and Kang Hao Cheong. The fractal dimension of complex networks: A review. *Information Fusion*, 73:87–102, 2021.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[8] Julia El Zini, Bassel Musharrafieh, and Mariette Awad. On the potential of the fractal geometry and the cnns ability to encode it, 2024.

[9] A.-M. Leventi-Peetz, T. Östreich, W. Lennartz, and K. Weber. Scope and sense of explainability for ai-systems. In Kohei Arai, editor, *Intelligent Systems and Applications*, pages 291–308, Cham, 2022. Springer International Publishing.

[10] Benoit B. Mandelbrot. How long is the coast of britain? statistical self-similarity and fractional dimension. *Science*, 156(3775):636–638, 1967.

[11] M. J. Feigenbaum. Universal behavior in nonlinear systems. *Los Alamos Science*, 1:4–27, 1983.

[12] Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, 1993.

[13] Tao Wang, Xiangwei Zheng, Lifeng Zhang, Zhen Cui, and Chunyan Xu. A graph-based interpretability method for deep neural networks. *Neurocomputing*, 555:126651, 2023.

[14] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[16] Elizabeth M. Hou and Gregory Castanon. Decoding layer saliency in language transformers, 2023.

[17] Harold Erbin, Vincent Lahoche, and Dine Ousmane Samary. Renormalization in the neural network-quantum field theory correspondence, 2022.

[18] Sangit Chatterjee and Mustafa R. Yilmaz. Chaos, fractals and statistics. *Statistical Science*, 7(1):49–68, 1992.

[19] Marc-André Zöller and Marco F. Huber. Benchmark and survey of automated machine learning frameworks, 2021.

[20] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017.