Explaining probabilistic Artificial Intelligence (AI) models by discretizing Deep Neural Networks

Explaining probabilistic Artificial Intelligence (AI) models by discretizing Deep Neural Networks

Rabia Saleem[†], Bo Yuan[1], Fatih Kurugollu[2]
School of Electronics, Computing, and Mathematics
University of Derby
Derby, United Kingdom
[†]r.saleem3@unimail.derby.ac.uk
[1]B.Yuan@derby.ac.uk [2]F.Kurugollu@derby.ac.uk

Ashiq Anjum[3]
College of Science and Engineering
University of Leicester,
Leicester, United Kingdom
[3]a.anjum@leicester.ac.uk

***Abstract*—Artificial Intelligence (AI) models can learn from data and make decisions without any human intervention. However, the deployment of such models is challenging and risky because we do not know how the internal decision-making is happening in these models. Especially, the high-risk decisions such as medical diagnosis or automated navigation demand explainability and verification of the decision making process in AI algorithms. This research paper aims to explain Artificial Intelligence (AI) models by discretizing the black-box process model of deep neural networks using partial differential equations. The PDEs based deterministic models would minimize the time and computational cost of the decision-making process and reduce the chances of uncertainty that make the prediction more trustworthy.**

Keywords: Artificial Intelligence, Deep Neural Networks, Partial differential equations, Discretization.

I. INTRODUCTION

Deep learning is a subset of machine learning that learns from big data and uses more hidden layers than the classical architecture of the neural networks. With the aid of hundreds to thousands of iterations, deep learning classifiers can give an accurate decision, albeit without knowing the learning path, which is basically an audit trail of the decision-making process. We cannot trust the outcomes of these models if we do not know the process they follow to make predictions. We need to understand, analyse, and interpret the deep learning models to improve trust and confidence in them so that we can reliably exploit them for extracting and processing complex information [1].

There are three approaches to explain or interpret the black-box AI models, namely, Model-Explanation (global-interpretation), Outcomes-Explanation (local interpretation), and Model-Inspection (visual-interpretation). Simple and modest size models can be explained by decision trees and decision rules. However, it is hard to explain the deep neural networks by using these approaches [2]. By highlighting the salient parts of input images, the authors in [3] explain the outcomes of the black-box model. The visual explanation can be generated with the aid of an attention mechanism and this explanation provides direction for textual justification [4].

The continuous interpretation of deep neural networks has been recently discussed in [5,6,7], where the researchers connect the different architectures (PolyNet, FractalNet, RevNet) with the approximation methods (backward Euler, Runge-Kutta, forward Euler) of ordinary differential equations. Due to the limited availability of memory, the authors in [8,9] offer stable and reversible networks. The reversible architectures in neural networks allow us to reconstruct activations from the input layer to the output layer and vice versa [10] also, to train a deep architecture with minimal memory storage.

As a powerful modelling tool, partial differential equations (PDEs) interpret many image processing tasks such as denoising, image segmentation. With the help of a strong and well- established theoretical approach of PDEs, [11] proposed three architectures to interpret the deep architecture of one of the variants of convolutional neural networks (CNNs) known as the residual neural network (ResNet). Our aim is to establish a bridge between the residual neural network (ResNet) and discretized PDE that explains the ResNet more accurately with the aid of more stable approximate methods.

II. RESEARCH PROBLEM

Existing approaches emphasize the local and visual explanation of the black-box model consisting of deep neural networks. However, due to the complex architectures, these networks are still facing the challenge of complete or global explanation. We aim to globally explain the black-box in the AI models by using deterministic models such as PDEs. Consequently, whenever an error occurs, instead of unfolding all the neural network, this deterministic model will identify the issue by observing the parameters in a PDE, improve it with less trial and error, and quickly fix the issues leading to reliable and trustworthy decisions.

In the recent past, partial differential equations (PDEs) have been widely used for the modelling of many complex physical phenomena, thus as a powerful mathematical tool, PDEs can provide a transparent model that mimics the black-box model and explains it at the global level. To accomplish the real-life and scientific problems, PDEs approximate their solution by choosing a suitable approximation method, such as the finite difference method, finite element method.

III. Proposed and Planned Work

To explain the black-box model containing deep neural architectures, a novel approach has been proposed, where we will discretize the deep neural network (DNN) using the mathematical model such as Partial differential equations (PDEs). A combination of a suitable approximation method and the well-established theory of PDEs with the coefficient's values and initial constraints would help us to explain the deep architectures of the neural network and make AI models trustworthy. PDEs can solve the existing gaps as follows:

- The partial derivatives can keep one variable constant to check the impact of other variables on the system. This would help to identify the more significant neurons of deep neural networks that contribute to predictions.
- The discretization techniques transfer the continuous domain of PDEs into discrete parts. As a result, the complex non-linear PDE appears as a simple set of linear algebraic equations or ordinary differential equations.
- Lastly, we can choose the appropriate approximation methods such as the finite difference method and finite element method for the resultant equations for the ultimate explanation and interpretation of the deep architecture of neural networks.

Hence, PDEs based deterministic models reduce the chances of uncertainty and make the prediction more trustworthy as these models give the same results for specific data sets, no matter how many times we compute it. To address the black-box model explanation problem we will

- Discretize the deep neural networks (DNN) using mathematical models such as Partial differential equations (PDEs).
- Tune the parameters such as weights, learning rate and, activation function during the training of DNN that would enable us to design a set of partial differential equations (PDEs) that can accurately represent the functionality of the original DNN.
- Apply the proposed system of PDEs on DNNs to optimize their functionality and to explain the behavior of the deep neural network.
- Validate the accuracy and transparency of the proposed explainable AI model by comparing its functionality against the original DNN.

IV. Preliminary Work

At the initial stage of this research, we exploit a simple ResNet architecture as an initial value problem (IVP) with a set of hyperparameters [11][12]. The assumptions on the convolutional operators ($K = K_o = -K_i{}^T$), normalization layer ($\aleph(U) = U$) and activation function ($\sigma(x) = x$) connect the deep ResNet architecture to a parabolic PDE known as heat equation. Also, we replace the convolutional operator with the operator ∇ to get the Laplacian operator to get the following heat equation.

$$U'(t,\theta) = -KK^T U = -\nabla\nabla^T U = \nabla^2 U \quad (1)$$

The architecture of the deep ResNet has already been interpreted by constructing a stable network such as the Hamiltonian network and leapfrog network [7] by using the explicit approximate method with good accuracy, however, we want to improve the accuracy level with unconditional stability while explaining the deep architecture of ResNet.

We propose the architectures by connecting the deep ResNet with the discretized heat equation through the implicit and Crank-Nicolson approximation methods as these approximation methods are unconditionally stable with higher accuracy as compared to the explicit method.

The implicit method uses the backward difference at the time t_{k+1} and central difference for space variable at x_i. On the other hand, the Crank-Nicolson method is an average of the explicit method and implicit method [13]. Although, the Crank- Nicolson method has become more useful after the approximating Black Scholes equation, but originally, this method was first applied to heat equation to get an approximate solution by approximating derivatives in time t and space x. As can be seen below, the discretized heat equation by the Implicit (equation 2) and Crank-Nicolson (equation 3) methods resembles the simple ResNet architecture.

$$U_{k+1} = E(U_k + F_{k+1}) \quad (2)$$

$$U_{k+1} = CU_k + S^{-1}F_k \quad (3)$$

TABLE I: Comparison of Finite-Difference Methods

Explicit method	**Implicit method**	**Crank-Nicolson scheme**
Easy implementation	Complex implementation	Complex implementation
Order of accuracy: $\mathcal{O}(\Delta t, \Delta x^2)$	Order of accuracy: $\mathcal{O}(\Delta t, \Delta x^2)$	Order of accuracy: $\mathcal{O}(\Delta t^2, \Delta x^2)$
Conditional stability	Unconditional stability	Unconditional stability

From Table I, one can say that an explicit method is easy to implement but this approach is conditionally stable with low accuracy. On the other side, the Crank-Nicolson and implicit methods are unconditionally stable and also provide a more accurate numerical solution as compared to the explicit method.

V. Conclusions and Future Directions

As a conclusion, the aims of this work are

- to connect the deep neural network with the partial differential equation such as heat equation.
- to develop the unconditional stable architectures via numerical discretization techniques that offer the best and more accurate way to explain the deep neural architectures.

- to provide a better and strong interpretation of deep neural networks through the deterministic models.

To validate the proposed approach, we would evaluate our proposed architectures on three classical benchmarks, SLT-10, CIFAR10, and CIFAR-100, and compare against the results mentioned in Ruthotto's paper. Our aim with this numerical comparison would be achieving better accuracy along with the stability advantage. In addition to improved accuracy, we would provide a trustworthy and more reliable explanation of the deep ResNet architecture.

In future, we will like to implement the proposed system using the data that is stored in federated and distributed architectures so that we can learn and inference in almost real-time. This will lead to a high-performance analytics system and will build upon our previous work that has been reported in [14][15][16]. In addition, we will run the proposed algorithms as a set of workflows [17] that can make use of the cloud and distributed infrastructures leading to auditable and verifiable systems similar to the one reported in [18].

ACKNOWLEDGMENT

This work was carried out in the Data Science Research Centre (DSRC) at the University of Derby.

REFERENCES

[1] Qingchen Zhangab, Laurence T.Yangab, ZhikuiChenc, PengLic, "A survey on deep learning for big data.," Inf. Fusion 42, vol. 42, pp. 146-157, 2018.

[2] Sina Mohseni, Niloofar Zarei, Eric D. Ragan , " A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," ACM Trans. Interact. Intell. Syst., 2020.

[3] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, Bart Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models.," Decis. Support Syst. , pp. 141-154, 2011.

[4] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, Marcus Rohrbach, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018.

[5] Lars Ruthotto, Eldad Haber. Stable architectures for deep neural networks. Inverse Problems, 34, 014004, 2018.

[6] Eldad Haber, Lars Ruthotto, Elliot Holtham, Seong-Hwan Jun. Learning across scales-a multiscale method for convolution neural networks. in The Thirty-Second AAAI Conference.2017.

[7] Yiping Lu, Aoxiao Zhong, Quanzheng Li, Bin Dong, " Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations," in International Conference on Machine Learning (ICML), 2017.

[8] Sergey Zagoruyko, Nikos Komodakis, "DiracNets: Training Very Deep Neural Networks Without Skip-Connections," Clinical Orthopaedics and Related Research (CORR), vol. abs/1706.00388, 2017.

[9] Aidan N. Gomez, Mengye Ren, Raquel Urtasun, Roger B. Grosse, "The Reversible Residual Network: Backpropagation Without Storing Activations," in Conference on Neural Information Processing Systems, 2019.

[10] Bo Chang, Lili Meng, Eldad Haber, Lars Ruthotto, David Begert, Elliot Holtham, "Reversible architectures for arbitrarily deep residual neural networks," in In AAAI Conference on AI, 2018.

[11] Lars Ruthotto, Eldad Haber.Deep Neural Networks Motivated by Partial Differential Equations. Journal of Mathematical Imaging and Vision, vol. 62, p. 352–364, 2020.

[12] Muhammad Usman Yaseen, Ashiq Anjum, Omer F. Rana, Nikolaos Antonopoulos , "Deep Learning Hyper-Parameter Optimization for Video Analytics in Clouds.," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, pp. 253-264, 2019.

[13] Fadugba, Sunday Emmanuel and C. R. Nwozo., "On the Comparative Study of Some Numerical Methods for Vanilla Option Valuation," Communications in Applied Sciences, vol. 2(1), no. 2201-7372, pp. 65-84, 2014.

[14] Saad Liaquat Kiani, Ashiq Anjum, Michael Knappmeyer, Nik Bessis, Nikolaos Antonopoulos, "Federated broker system for pervasive context provisioning," Journal of Systems and Software, vol. 86, no. 4, pp. 1107-1123, 2013.

[15] Conrad Steenberg, Julian Bunn, Iosif Legrand, Harvey Newman, Michael Thomas, Frank van Lingen, Ashiq Anjum, Tahir Azim."The clarens grid-enabled web services framework: Services and implementation," in Proceedings of Computing in High Energy Physics and Nuclear Physics, Interlaken, Switzerland, 2004.

[16] F. Van Lingen, P. Avery, M. Thomas, J. In, D. Bourilkov, J. Bunn, R. Cavanaugh, L. Chitnis, M. Kulkarni, I. Legrand, H. Newman, C. Steenberg, A. Anjum, T. Azim. "Grid Enabled Analysis : Architecture, prototype and status," in Proceedings of Computing in High Energy Physics and Nuclear Physics, 2005.

[17] Khawar Hasham, Antonio Delgado Peris, Ashiq Anjum, Dave Evans, Dirk Hufnagel, Eduardo Huedo, Jose M. Hernandez, Richard McClatchey, Stephen Gowdy, Simon Metson, "CMS Workflow Execution Using Intelligent Job Scheduling and Data Access Strategies," IEEE Transactions on Nuclear Science, vol. 58, no. 3, pp. 1221-1232, 2011.

[18] Richard McClatchey, Andrew Branson, Ashiq Anjum, Peter Bloodsworth, Irfan Habib, Kamran Munir, Jetendr Shamdasani, Kamran Soomro, The neuGRID Consortium3, "Providing traceability for neuroimaging analyses," International journal of medical informatics, vol. 82, no. 9, pp. 882-894, 2013.