Short Communication
Volume 16:4, 2023

Journal of Computer Science & Systems Biology

ISSN: 0974-7230

Open Access

Interpretable Neural Networks: Bridging the Gap between AI and Human Understanding

Luca Borghesi*

Department of Business Information Systems, Humboldt University of Berlin, Berlin, Germany

Introduction

As artificial intelligence continues to advance and integrate into various facets of our lives, the demand for understanding AI decisions and predictions is growing exponentially. Interpretable neural networks, designed to make AI systems more transparent and comprehensible to humans, have emerged as a crucial research frontier. This research article explores the significance of interpretability in neural networks, discusses current approaches and challenges, and highlights the role of interpretable AI in building trust, ethics, and responsible AI deployment.

Artificial neural networks, often regarded as black boxes due to their complex architectures, have delivered impressive results in diverse domains [1-3]. However, this complexity comes at a cost - a lack of transparency in decision-making. As AI systems become more integral to society, ensuring that they are interpretable and accountable is paramount. This article delves into the realm of interpretable neural networks and their role in bridging the gap between AI and human understanding. Artificial intelligence has made remarkable strides in recent years, permeating various aspects of our lives.

However, as AI systems grow more intricate, so does the challenge of understanding their decisions and actions. Interpretable neural networks, a cutting-edge area of AI research, have emerged as a vital bridge between the complexity of AI models and human comprehension. In this research article, we explore the crucial role of interpretability in neural networks, examine current approaches, and address the challenges and prospects of creating AI systems that are not only powerful but also transparent and interpretable. By doing so, we aim to illuminate the path toward building trust, ethics, and responsible AI deployment in our increasingly AI-driven world.

Description

Interpretable AI systems instill trust and confidence in users and stakeholders. Understanding how and why an AI system makes a decision is essential for its acceptance in critical applications like healthcare, finance, and autonomous vehicles. Interpretable neural networks empower us to uncover and rectify biases, unfair decisions, and unethical behavior in AI models. By exposing the decision-making process, we can mitigate these issues and promote fairness and accountability. Feature attribution methods like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) provide insights into the importance of input features in neural network predictions. These methods generate explanations that humans can understand, aiding in decision interpretation.

Saliency maps highlight the regions in input data that significantly influence neural network outputs. They are particularly valuable in computer vision and are used for visualizing what parts of an image are crucial for a particular classification. Incorporating rule-based models like decision trees or symbolic reasoning into neural networks can make them more interpretable. Hybrid models combine the strengths of both rule-based and deep learning approaches. There exists a trade-off between model accuracy and interpretability. Striking the right balance is a significant challenge, especially in domains where precision is critical [4,5].

Many interpretable techniques are computationally expensive and may not scale well to larger neural networks or real-time applications. Developing efficient methods is crucial for practical deployment. The field of interpretability lacks standardized evaluation metrics, making it difficult to compare and benchmark different techniques. Establishing common evaluation criteria is essential for driving progress.

Conclusion

Interpretable neural networks are pivotal in ensuring that AI aligns with human values, ethics, and expectations. By providing transparent and understandable decision-making processes, these models enhance trust, accountability, and the responsible deployment of AI systems across various domains. As we continue to develop and refine interpretability techniques, the bridge between AI and human understanding will strengthen, fostering a harmonious integration of AI technology into our lives. In an era where AI's influence continues to expand, interpretable neural networks represent a beacon of transparency and accountability, facilitating a synergy between artificial and human intelligence.

Acknowledgement

None.

Conflict of Interest

Authors declare no conflict of interest.

References

1. Guo, Yulan, Mohammed Bennamoun, Ferdous Sohel and Min Lu, et al. "3D object recognition in cluttered scenes with local surface features: A survey." *IEEE Trans Pattern Anal Mach Intell* 36 (2014): 2270-2287.
2. Schmidhuber, Jürgen and Sepp Hochreiter. "Long short-term memory." Neural Comput 9 (1997): 1735-1780.
3. Al Badawi, Ahmad, Louie Hoang, Chan Fook Mun and Kim Laine, et al. "Privft: Private and fast text classification with homomorphic encryption." *IEEE Access* 8 (2020): 226544-226556.
4. Jung, Wonkyung, Sangpyo Kim, Jung Ho Ahn and Jung Hee Cheon, et al. "Over 100x faster bootstrapping in fully homomorphic encryption through memory-centric optimization with gpus." *IACR Trans Cryptogr Hardw Embed Syst* (2021): 114-148.
5. Eltayieb, Nabeil, Rashad Elhabob, Alzubair Hassan and Fagen Li. "A blockchain-based attribute-based signcryption scheme to secure data sharing in the cloud." *J Syst Archit* 102 (2020): 101653.

How to cite this article: Borghesi, Luca. "Interpretable Neural Networks: Bridging the Gap between AI and Human Understanding." *J Comput Sci Syst Biol* 16 (2023): 476.

***Address for Correspondence:** Luca Borghesi, Department of Business Information Systems, Humboldt University of Berlin, Berlin, Germany, E-mail: lucaborghesi31@gmail.com*

Received: 01 July, 2023, Manuscript No. jcsb-23-113749; **Editor Assigned:** 03 July, 2023, Pre QC No. P-113749; **Reviewed:** 17 July, 2023, QC No. Q-113749; **Revised:** 22 July, 2023, Manuscript No. R-113749; **Published:** 31 July, 2023, DOI: 10.37421/0974-7230.2023.16.476