Immersive Neural Network Exploration: A VR Approach to Human-Centered AI Understanding

Medet Inkarbekov
medet.inkarbekov@mu.ie
Department of Computer Science,
Maynooth University
Maynooth, Ireland

Barak A. Pearlmutter
Department of Computer Science,
Hamilton Institute, Maynooth
University
Maynooth, Ireland

Rosemary Monahan
Department of Computer Science,
Hamilton Institute, Maynooth
University
Maynooth, Ireland

ABSTRACT

In today's rapidly evolving artificial intelligence (AI) landscape, the complexity of neural networks, especially in deep learning, presents significant challenges for intuitive human understanding. Compelling visualization methods have become crucial with AI systems integrating into everyday experiences and environments. Virtual Reality (VR), as an immersive and interactive technology, offers a novel approach to visualizing intricate AI processes. This work introduces transformative updates to the DeepVisionVR platform, a pioneering tool for 3D visualization of Convolutional Neural Networks (CNNs) in VR. A cornerstone of our enhancements is the Sensitivity Analysis module, which offers real-time interactivity, allowing users to adjust pixels within the VR space, shedding light on the intricacies of model outcomes. In addition, advanced model interpretation methodologies have been integrated, including Integrated Gradients, GradientShap, and Occlusion, enriching the depth of insight into model rationale. Our Adversarial Analysis, utilizing the Fast Gradient Sign Method (FGSM), unveils potential weak points in models, emphasizing their vulnerability to minor input alterations. The new features aim to provide a deeper understanding of how neural networks interpret and react to various inputs, thereby bridging the gap between complex machine learning models and human interpretability.

ACM Reference Format:
Medet Inkarbekov, Barak A. Pearlmutter, and Rosemary Monahan. 2023. Immersive Neural Network Exploration: A VR Approach to Human-Centered AI Understanding . In *Human Centered AI Education and Practice Conference 2023 (HCAIep '23), December 14–15, 2023, Dublin, Ireland.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3633083.3633221

CONTENT

The poster aims to highlight how immersive VR technology can help in better understanding and interpreting complex neural networks. Key sections of the poster include:

1. Visualization of AI Systems in VR: Overview. This section provides an introduction to visualizing AI systems using VR, emphasizing the necessity for better interpretation and understanding of AI models. It discusses various visualization methods like heat maps, saliency maps, Layer-wise Relevance Propagation, and Local Interpretable Model-Agnostic Explanations, which aid in interpreting complex AI system behaviors. Following this, the exploration of VR is highlighted for its substantial benefits in enhancing user engagement, spatial awareness, and intuitive model exploration through an immersive 3D experience, transcending the traditional 2D visualization limitations. Additionally, the section references a survey that identifies a significant gap in the current VR in AI visualization landscape [1], particularly the absence of integrated analysis and interpretability tools in existing VR platforms for AI. This insight steers the focus towards enhancing the analytical capabilities of VR platforms to enrich the AI visualization experience.

2. DeepVisionVR. Here, the DeepVisionVR platform developed by Linse et al. [3] is introduced. It is a VR platform for engaging with large-scale CNNs, linked real-time with PyTorch. The platform allows an in-depth examination of network outputs and activated features, facilitating better understanding of the networks. It's open-source, adaptable and optimized for large-scale models, promoting collaborative exploration of CNNs.

3. Current Innovations and Future Horizons. This part presents the ongoing enhancements and future plans for the project. It includes the development of a user-centric interface, a Sensitivity Analysis Module, integration with Captum for model interpretation [2], and an Adversarial Analysis Feature for exploring model vulnerabilities. Future works highlight further integration with Captum, enhanced CNN visualization through various techniques, and expanded adversarial techniques to provide more insights into model vulnerabilities.

The poster's content delves into the pressing need for rendering intricate AI models more transparent and interpretable, echoing the core principles of human-centered AI education and practice. By offering a glimpse into pioneering strategies, notably the incorporation of VR in AI visualization, the poster is primed to introduce a compelling and pertinent discussion topic to the conference attendees. The integration of VR to elucidate the sophisticated workings of CNNs marks a notable stride towards unraveling the mystique surrounding AI, aligning well with the objectives of the Human-Centred AI community to bolster human understanding and interpretability of complex machine learning models.

REFERENCES

[1] Medet Inkarbekov, Rosemary Monahan, and Barak A. Pearlmutter. 2023. Visualization of AI Systems in Virtual Reality: A Comprehensive Review. *International Journal of Advanced Computer Science and Applications* 14, 8 (2023). https://doi.org/10.14569/IJACSA.2023.0140805

[2] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

[3] Christoph Linse, Hammam Alshazly, and Thomas Martinetz. 2022. A walk in the black-box: 3D visualization of large neural networks in virtual reality. *Neural Computing and Applications* 34, 23 (2022), 21237–21252. https://doi.org/10.1007/s00521-022-07608-4

HCAIep '23, December 14–15, 2023, Dublin, Ireland

ACM ISBN 979-8-4007-1646-1/23/12.
https://doi.org/10.1145/3633083.3633221