Towards Reliable Explanations of Deep Neural Networks

Sédrick STASSIN

Abstract

Today, we are fortunate to live in an unprecedented era. Day by day, humanity develops a technology that surpasses itself, revolutionizes, and advances at an extraordinary pace: artificial intelligence (AI), challenging our ability to regulate and comprehend it. This rapid progress has been largely fueled by deep neural networks (DNNs), whose increasingly complex architectures have significantly improved performance and precision. Additionally, the rise of transformers and their various adaptations, such as Vision Transformers (ViTs), multimodal models like Vision-Language Transformers (VL), and large language models (LLMs), represent a new frontier of complexity. However, these advances have come at the cost of reduced interpretability, raising critical concerns about transparency and trust. Amidst this rapid evolution, ensuring the transparency and reliability of AI systems has become a pressing priority, particularly in high-stakes applications. This thesis addresses this challenge through the lens of explainable AI (XAI), a field dedicated to elucidating the decision-making processes of AI systems to ensure that they are interpretable and trustworthy.

The central aim of this work is to develop and guide towards more *reliable* explanations for neural networks, with a particular focus on XAI for vision tasks. Through this approach, this research seeks to lay the foundation for building more robust and trustworthy AI systems in the future. This thesis is structured around three major contributions:

1. **Bias Detection and Mitigation**: The thesis investigates how biases in convolutional neural network (CNN) models can be identified using XAI methods, and then explores strategies to mitigate these biases. A case study on two datasets (X-ray lung; biased colored digits) illustrates how dataset bias impacts model understanding and offers pathways for improvement.
2. **Evaluation of Explainability Methods**: A comprehensive framework is proposed for selecting and evaluating XAI methods using explainability metrics. This framework is applied to convolutional and transformer-based neural networks, providing insights into their interpretability and correlation, and exposing limitations in current evaluation practices.
3. **Advancing Explainability for Novel Architectures**: A novel explainability method is introduced for Vision Transformers, more reliable than its perturbation-based counterparts, along with an adaptable extension for multimodal models handling diverse data types such as images and text. Our results stand out consistently by securing top rankings across various state-of-the-art metrics.

Together, these contributions advance the field of AI explainability, offering practical tools and insights for developing transparent and reliable AI systems. By addressing critical challenges in bias mitigation, method evaluation, and architectural adaptability, this work aims to ensure that AI technologies can be safely and confidently integrated into society, fostering trust in their potential.