BAI Ph.D. Proposal

Exploring Explainable AI for Transformer and Diffusion Models

Faculty: R. Venkatesh Babu (CDS) and Chandra Sekhar Seelamantula (EE)

Introduction:

Transformers and diffusion models have become popular models for many machine learning tasks such as natural language processing and image processing. However, these models are often considered black boxes, and it is challenging to understand how these models make their predictions. The field of explainable AI aims to address this challenge by developing methods for understanding and interpreting the decision-making process of AI models. In this thesis proposal, we propose to investigate how explainable AI techniques can be applied to models involving Transformers and diffusion models to enhance their interpretability and reliability.

Research Problem:

The research problem is to develop novel explainable AI techniques for models involving Transformers and diffusion models that can provide interpretable insights into their decision-making process. The proposed techniques will aim to overcome the challenges associated with the opacity of these models and enable users to understand how they arrive at their predictions. Specifically, the research problem will involve the following steps:

Conduct a comprehensive review of the literature on explainable AI techniques for models involving Transformers and diffusion models.

Develop a framework for applying these techniques to models involving Transformers and diffusion models.

Implement and evaluate the proposed framework on various machine learning tasks, such as natural language processing, image processing, and video processing.

Investigate the impact of the framework on the interpretability of the models, including identifying the most important features or attention mechanisms that contribute to their decisions.

Explore the potential applications of the framework in various domains, such as healthcare, finance, and legal domains, where interpretable AI models are critical.

Expected Outcomes:

The expected outcomes of this research include the development of novel explainable AI techniques for models involving Transformers and diffusion models that can provide interpretable insights into their decision-making process. The proposed techniques will enable users to understand how these models arrive at their predictions, which can enhance their reliability and trustworthiness. The research will also contribute to the development of interpretable AI models that can be deployed in critical applications where transparency and accountability are essential.

Conclusion:

In conclusion, the proposed research aims to address the challenge of interpretability in models involving Transformers and diffusion models by developing novel explainable AI techniques. The expected outcomes of this research include the development of interpretable models that can be deployed in critical applications, which can enhance their reliability and trustworthiness.