Thesis Synthesis

Artificial intelligence (AI) and deep learning (DL) have significantly advanced in computer vision (CV), particularly in image classification and object recognition. However, ensuring the trustworthiness of AI systems, especially in safety-critical applications remains a major challenge. Enhancing model robustness and interpretability is crucial for making AI systems reliable and explainable, which in turn improves user trust and facilitates widespread adoption. In this context, the primary goal of the research during the Ph.D. program was to address the challenges of improving the robustness and interpretability of AI systems. Specifically, innovative techniques were leveraged to ensure data-quality and provide new insights into the opaque and complex nature of DL models.

On one hand, the recent discipline of data-centric AI emphasized the importance of high-quality data for building effective AI systems, moving the focus towards improving data treatment steps such as data augmentation and monitoring after deployment to ensure fairness and robustness. On the other hand, continuous advances are being made in the field of Explainable AI (XAI), where techniques such as attention mechanisms and saliency maps were developed to provide insights into the decision-making process of these models. Expanding upon these domains, my research endeavors to address several unresolved challenges pertaining to the areas of robustness and interpretability. These include the exigency of studying novel systematic approaches to data preprocessing in data-centric AI, which arises from the need to structure datasets in a way that avoids or reduces biases in data, leading to more reliable and trustworthy AI systems. The lack of publicly available visual collections of XAI methods is another challenge that hinders a widespread adoption of interpretability in AI. Additionally, the need for more research in efficient unsupervised drift detection and for a systematic approach to create challenging real-world benchmark datasets for OOD detection arises due to the complexity and diversity of real-world data. Addressing these challenges is crucial for the development of more reliable and robust data-centric AI models that can be trusted to make decisions in real-world scenarios.

The research described in this thesis aims to provide innovative solutions to the challenges mentioned earlier. By delving into each topic, the ultimate goal of this document is to propose solutions to address some of the most pressing issues in trustworthy AI development and provide directions for future research in related domains.