A U S H A N G

FREIE UNIVERSITÄT BERLIN

Fachbereich Mathematik und Informatik

Promotionsbüro, Arnimallee 14, 14195 Berlin

D I S P U T A T I O N

Freitag, 16. Februar 2024, 12:00 Uhr

Ort: Seminarraum 115

(Fachbereich Mathematik und Informatik, Arnimallee 3, 14195 Berlin)

Disputation über die Doktorarbeit von

Leon Nikolaus Sixt

Thema der Dissertation:

Enhancing And Evaluating Interpretability InMachine Learning Through Theory And Practice

Thema der Disputation:

Explainable AI: Past Challenges and Future Opportunities

Die Arbeit wurde unter der Betreuung von **Prof. Dr. T. Landgraf** durchgeführt.

Abstract: In my first talk, I present an information theoretic perspective on the Variational Autoencoder (VAE). The VAE is a generative model that compresses the input into learned latent representation and then reconstructs the input. Usually, VAEs are parameterized using deep neural networks. Introduced in 2013, VAEs have become a popular tool in machine learning, with applications in image generation, data compression, and representation learning. In my talk, I focus on the variational bottleneck and how it can be seen as a noisy communication channel. This viewpoint helps in understanding how VAEs encode and reconstruct data.
In my second talk, I present the main ideas of my dissertation in the field of Explainable Artificial Intelligence (XAI). The field of XAI develops methods to increase the transparency of neural networks and other models. The lack of transparency remains a major obstacle to the adoption of AI systems in critical applications. As a first topic, I discuss the application of the information bottleneck concept to saliency estimation in images, i.e. finding out which image areas were most relevant. Furthermore, I address the limitations of current XAI methods, particularly the Deep-Taylor Decomposition. The talk concludes with insights from a human subject study on the effectiveness of AI explanations. Overall, I present both theoretical and practical insights into the field of XAI and its applications.

Die Disputation besteht aus dem o. g. Vortrag, danach der Vorstellung der Dissertation einschließlich jeweils anschließenden Aussprachen.

Interessierte werden hiermit herzlich eingeladen

Der Vorsitzende der Promotionskommission
Prof. Dr. T. Landgraf