Karlsruher Institut für Technologie

Institut für Angewandte Informatik
und Formale Beschreibungsverfahren

Graduiertenkolloquium Angewandte Informatik

Methods for Strengthening Explainable AI in Industrial Applications

Jacqueline Höllig (FZI Forschungszentrum Informatik, Germany)

With the growing adoption of Artificial Intelligence (AI) and the successful application of deep learning methods in various domains, AI, and particularly deep learning, increasingly influences people's lives. However, depending on the use case, wrong decisions can be costly and dangerous (e.g., an AI medical diagnosis system misclassifies patients' diseases). The emerging topic of Explainable Artificial Intelligence (xAI) offers approaches and algorithms that introduce transparency into black-box models by producing explanations of AI Systems' inner workings and decisions. Specifically, in industrial use cases, where complex problems and decision-making processes are widespread, enabling transparent automation and decision support is crucial. However, while research in xAI is trending, applying xAI in industrial use cases is challenging. For many data types (e.g., images or tabular data), xAI methods are well studied. Nevertheless, support for time series, which are ubiquitous in industrial settings, is missing. Further, to use any xAI method in deployment, understanding the explainers' quality, strengths, and weaknesses is of utmost importance to prevent ambiguous and incorrect explanations. Well-performing xAI methods can help users to understand the reasons behind a deep learners' prediction and enable the recognition of spurious correlations learned by a deep learner or missing information in the collected data. Especially in industrial settings, where only a limited amount of (often) noisy data is available, reverting incorrect model decisions and explanations provides the opportunity to include domain knowledge of users. Providing human feedback on the explanation enables a deep learner to infer the missing context and close this gap. To address these application obstacles of xAI in industrial settings, we introduce methods for xAI on time series, the evaluation of xAI, and xAI-based model revisions.

Termin: **Freitag, 19. April 2024, 14:00 Uhr**
Ort: Kaiserstr. 89, 76133 Karlsruhe
Kollegiengebäude am Kronenplatz (Geb. 05.20), 1. OG, Raum 1C-04
(Hinweise für Besucher: www.aifb.kit.edu/web/Kontakt)

Veranstalter: Institut AIFB, Forschungsgruppe Web Science

Zu diesem Vortrag lädt das Institut für Angewandte Informatik und Formale Beschreibungsverfahren alle Interessierten herzlich ein.

M. Färber, S. Lazarova-Molnar, A. Oberweis, H. Sack, A. Sunyaev, Y. Sure-Vetter (Org.),
A. Vinel, M. Volkammer, J. M. Zöllner